Lecture 9: Mathematics for Biomolecular Data

Guowei Wei

Mathematics

Michigan State University

https://users.math.msu.edu/users/wei/

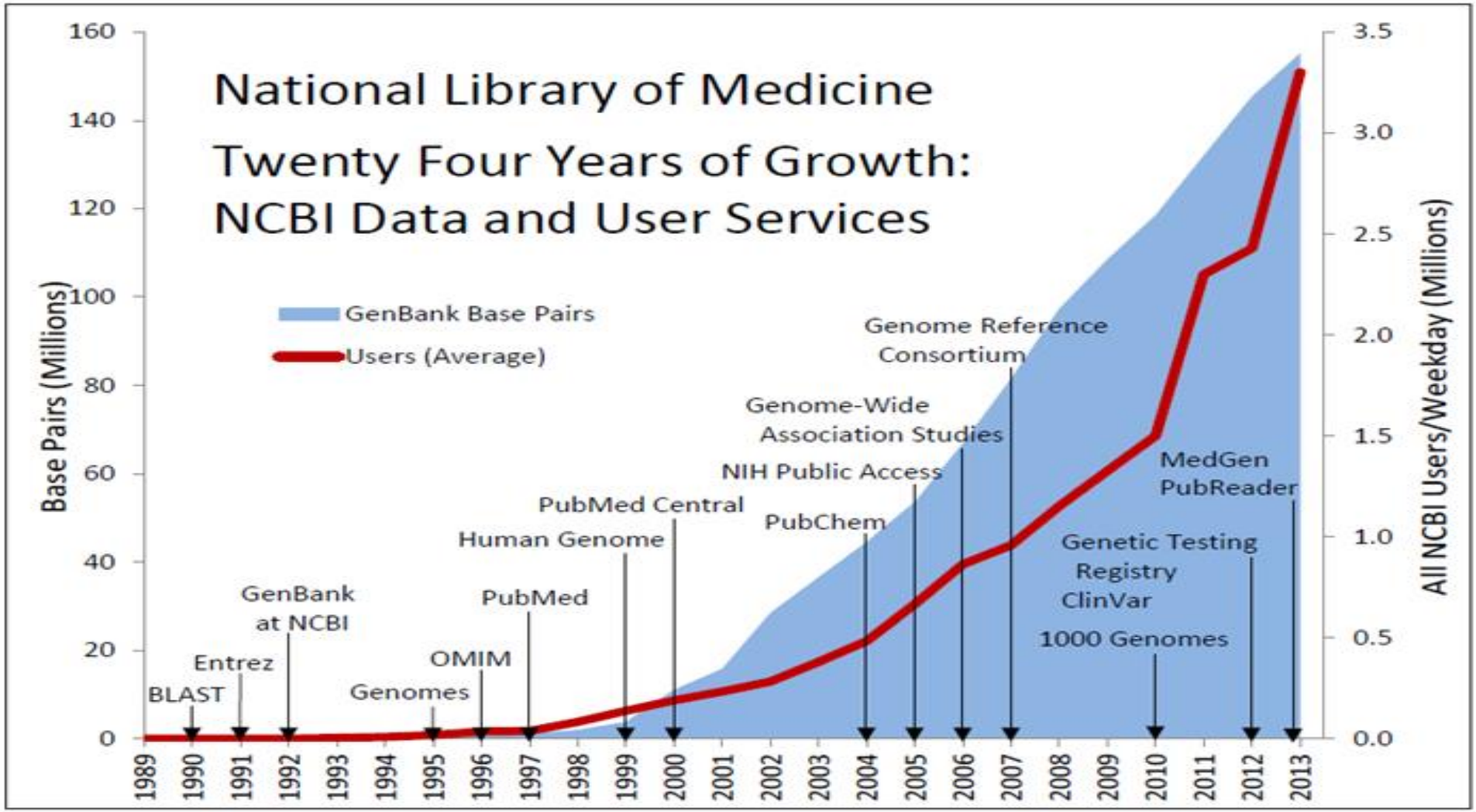
NSF-CBMS Conference on Mathematical Molecular Bioscience and Biophysics

University of Alabama

Tuscaloosa, May, 13-17, 2019

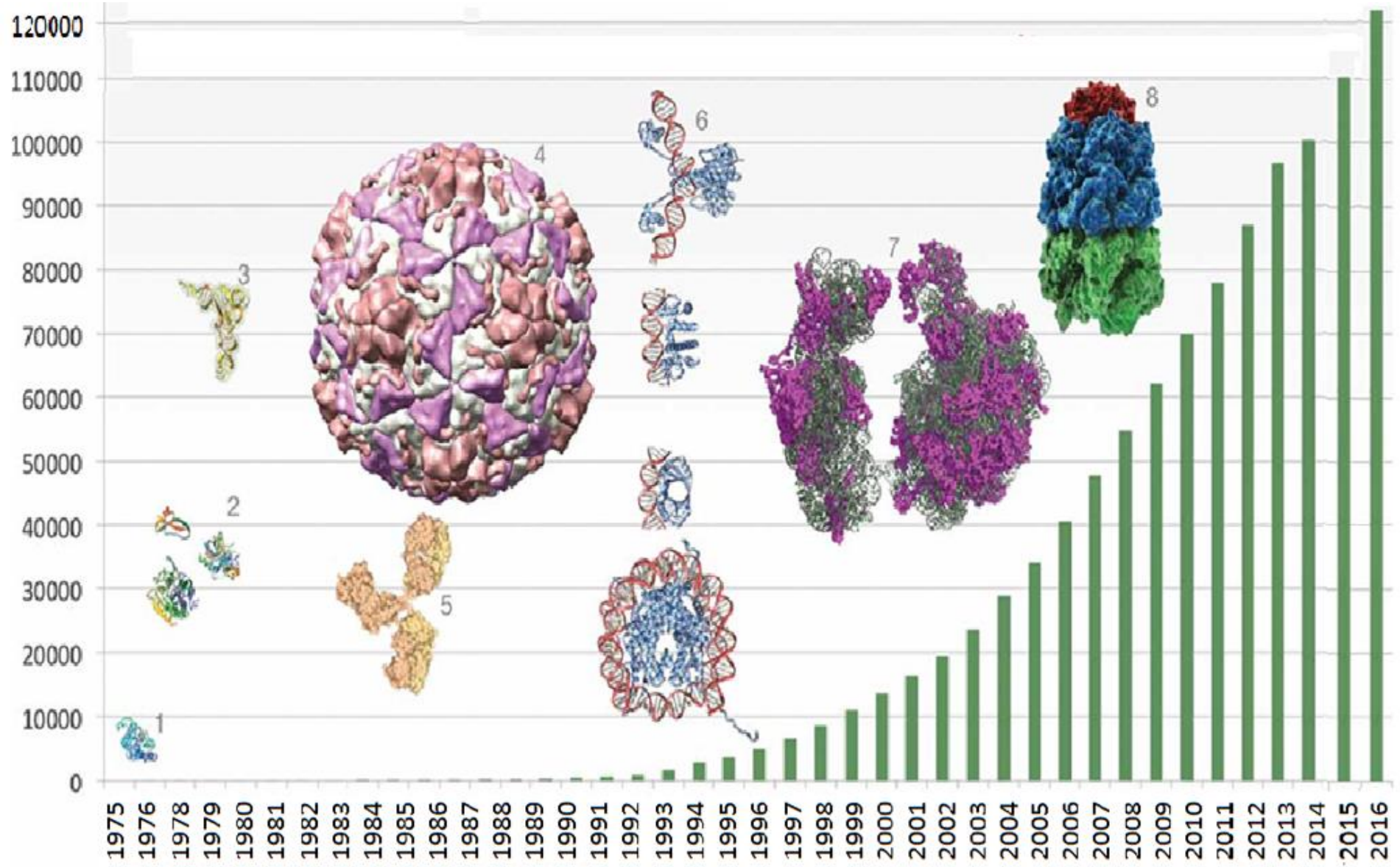
Grant support: NSF, NIH, MSU, BMS, and Pfizer

GenBank (GB)



		GenBank		WGS	
Release	Date	Bases	Sequences	Bases	Sequences
231	Apr 2019	321680566570	212775414	4421986382065	993732214

Protein Data Bank (PDB)



PDB: 151,579 Biological macromolecular structures as 04/30/2019

Protein-ligand (-protein) binding

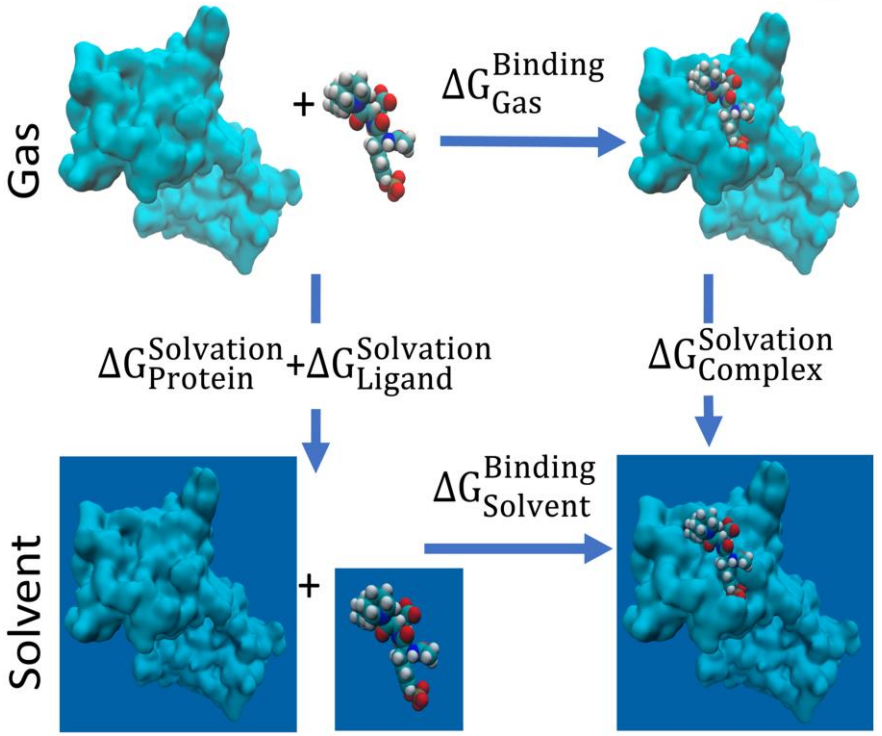
Protein (P) and ligand (L) form a protein-ligand complex (PL):

$$\mathrm{P} + \mathrm{L} \leftrightharpoons \mathrm{PL}$$

The association and dissociation constants are

$$K_a = \frac{[\mathrm{PL}]}{[\mathrm{P}][\mathrm{L}]}, \qquad K_d = \frac{[\mathrm{P}][\mathrm{L}]}{[\mathrm{PL}]}.$$

Binding affinity: $\Delta G = RT \ln K_d$



Database:

ChEMBL (as 02/28/2019):

Targets: 12,091

Compound records: 2,275,906

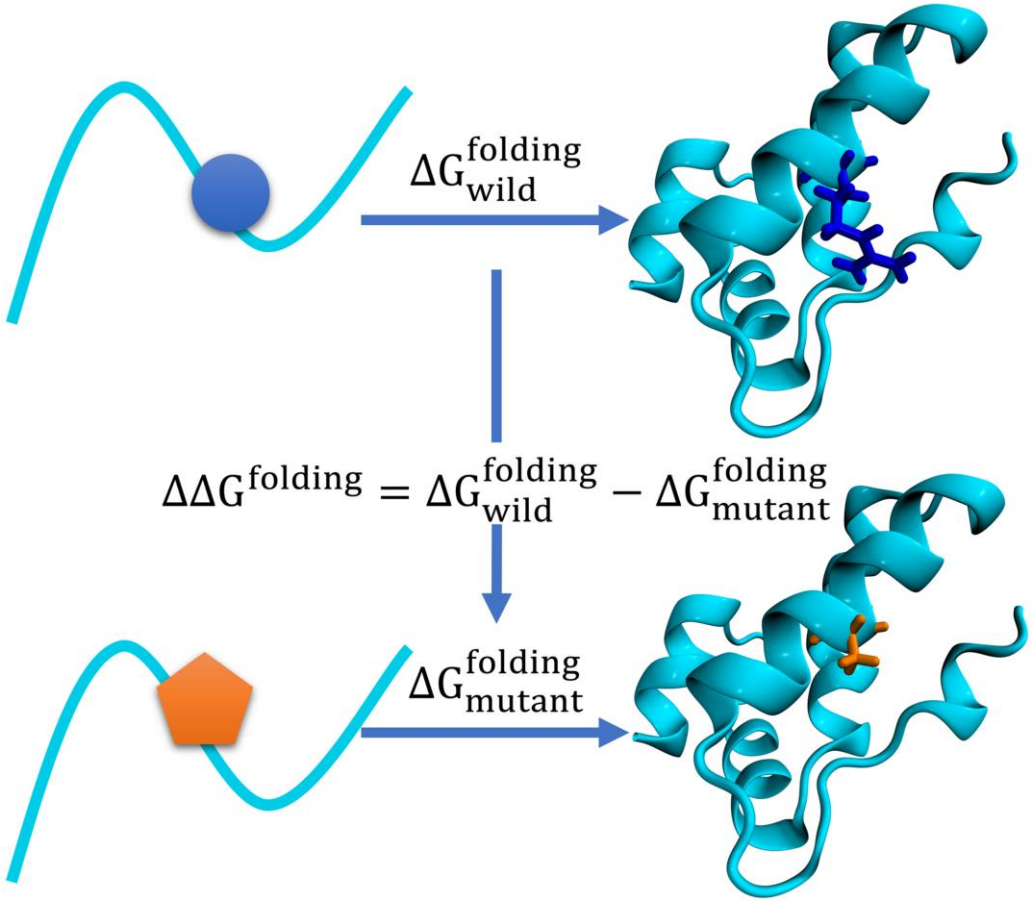
Distinct compounds: 1,828,820

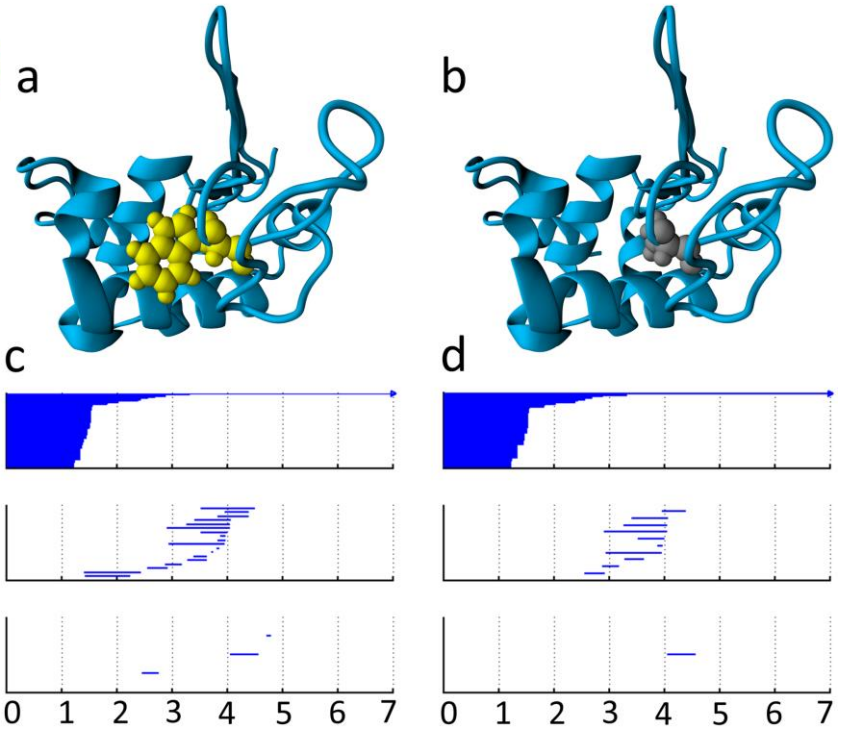
Activities: 15,207,914

Publications: 69,861

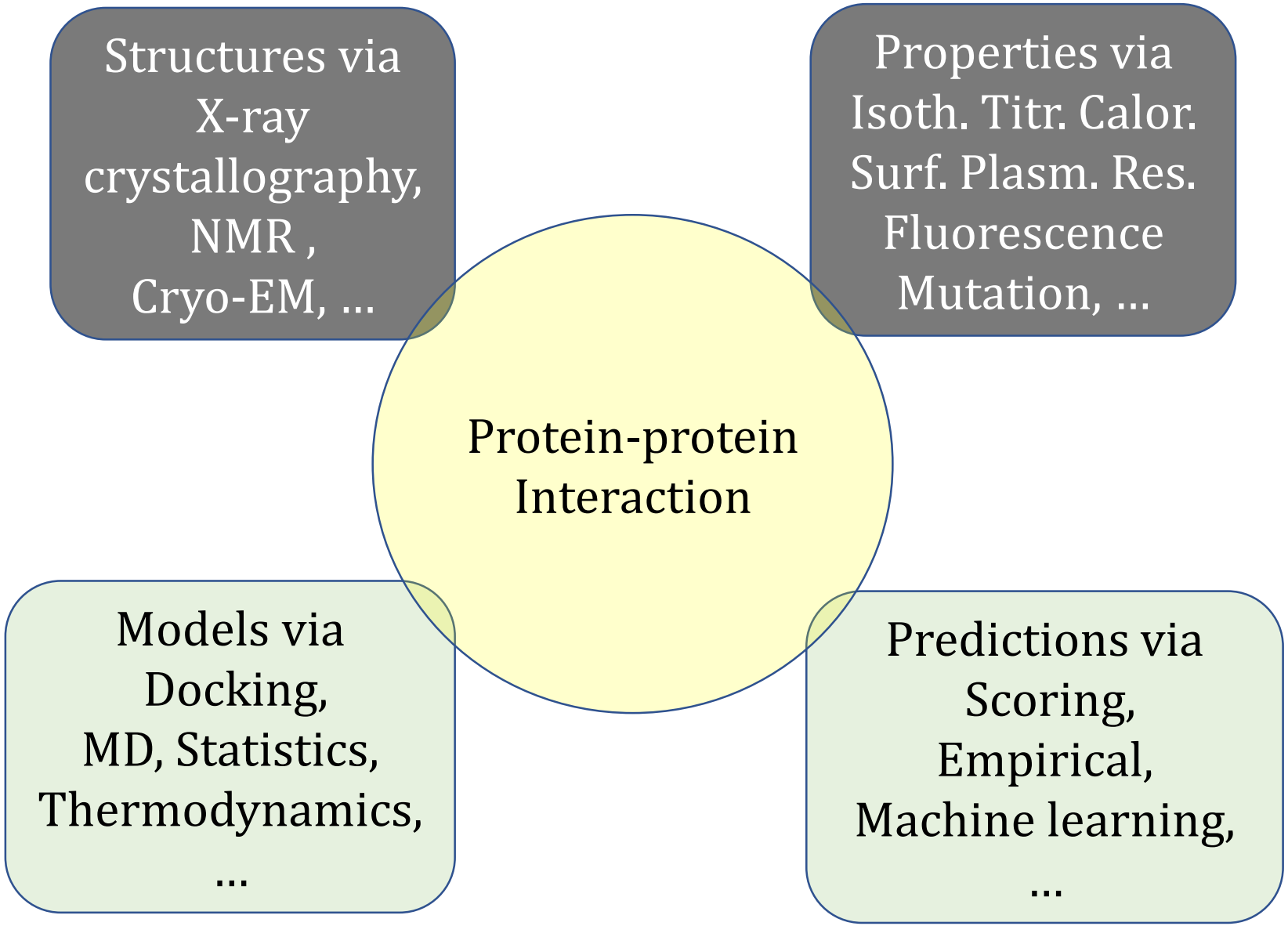
Binding DB, PubChem, PDBbind, K_i Database.

Mutation induced free energy change

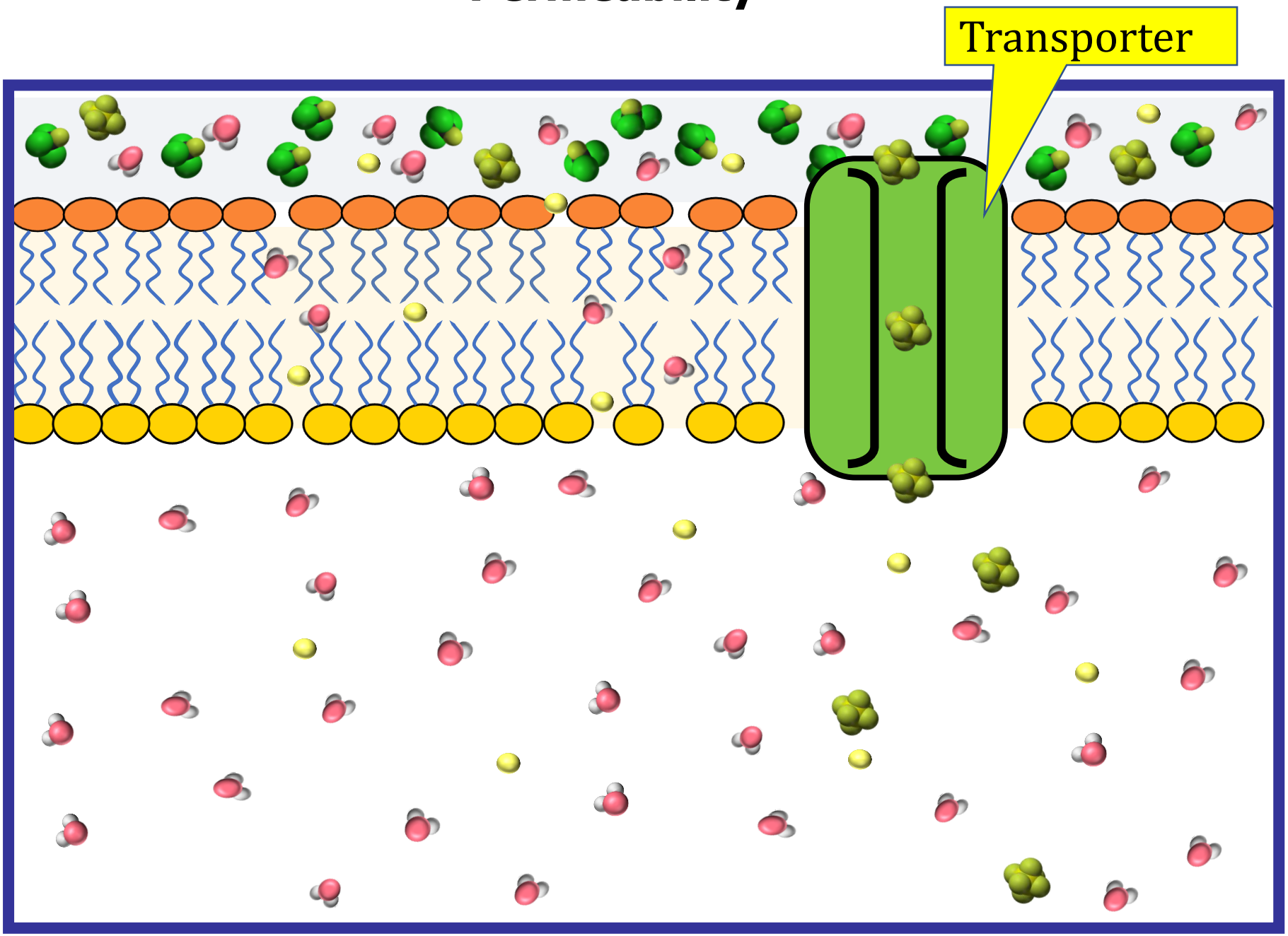




Protein-Protein interaction



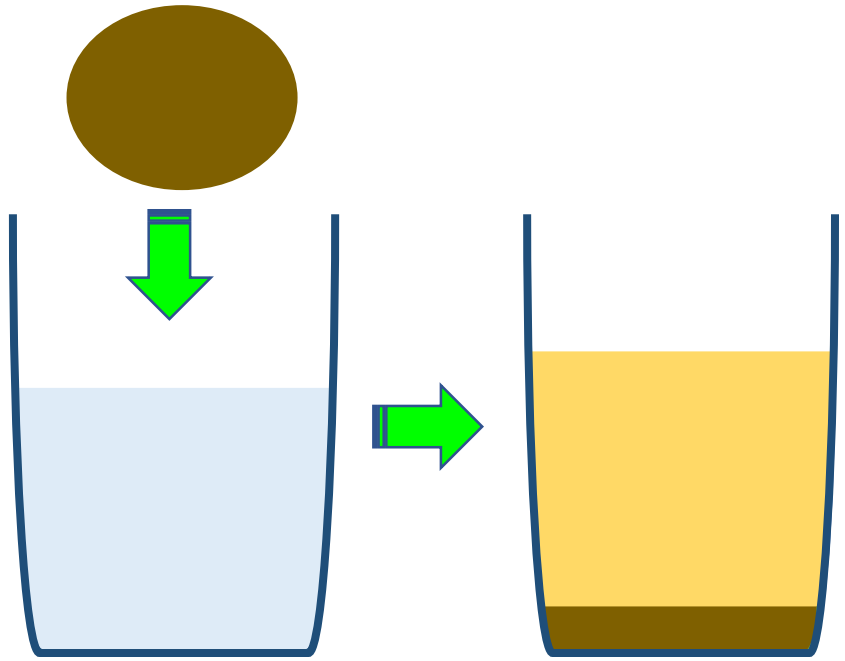
Permeability



Solubility and partition coefficient

Solubility is commonly expressed as a concentration; for example, as *g* of solute per *kg* of solvent.

Partition coefficient is defined as a particular ratio of the concentrations of a solute between the two solvents



Toxicity

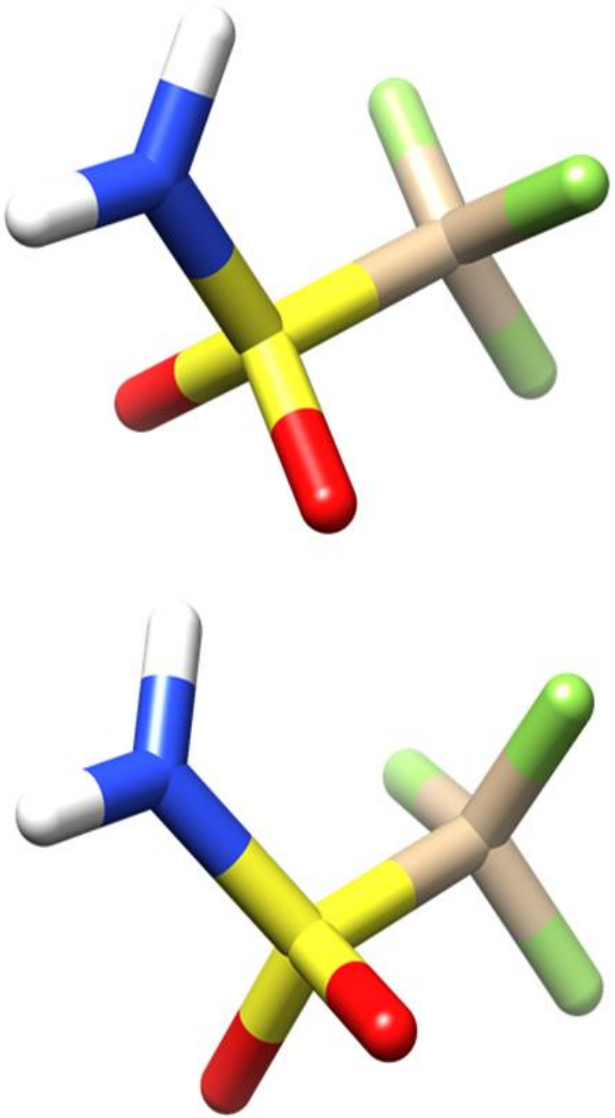
Toxicity: The degree to which a substance (a toxin or poison) can harm humans or animals. **Drug toxicity** occurs when a person has accumulated too much of a drug in his bloodstream, leading to adverse effects on the body.

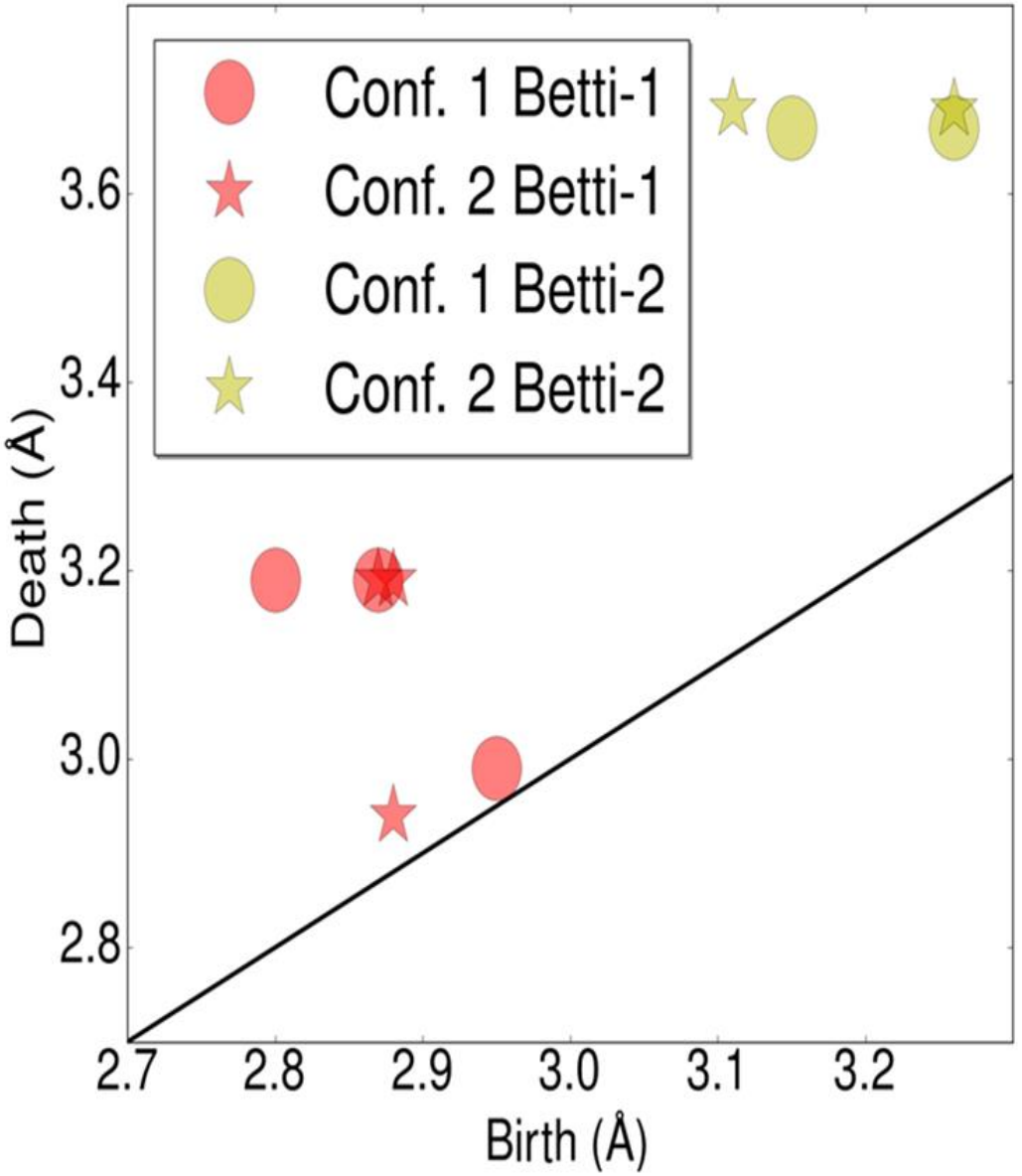
Bioassays:

LD50 is defined as the lethal dose at which 50% of the population if killed in a given period of time.

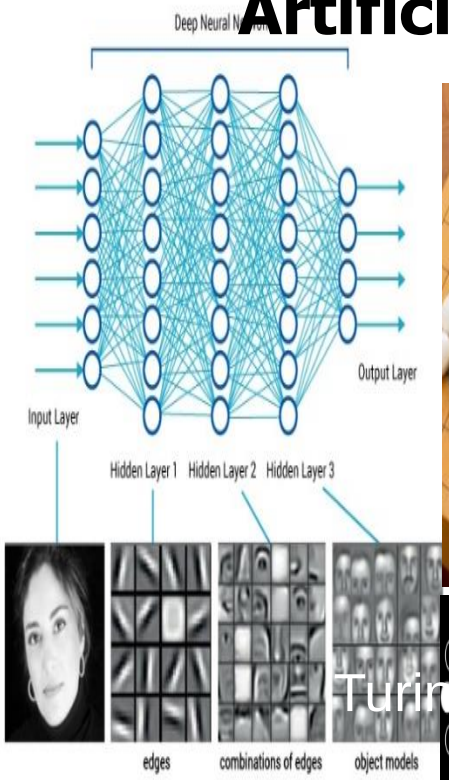
LC50 is the lethal concentration required to kill 50% of the population. The LC50 is a measure, *e.g.* in mg/l, of the concentration of the toxin whereas a dose is a more general term.

Topological fingerprints of stereoisomers (rotamers)





Artificial intelligence and machine learning

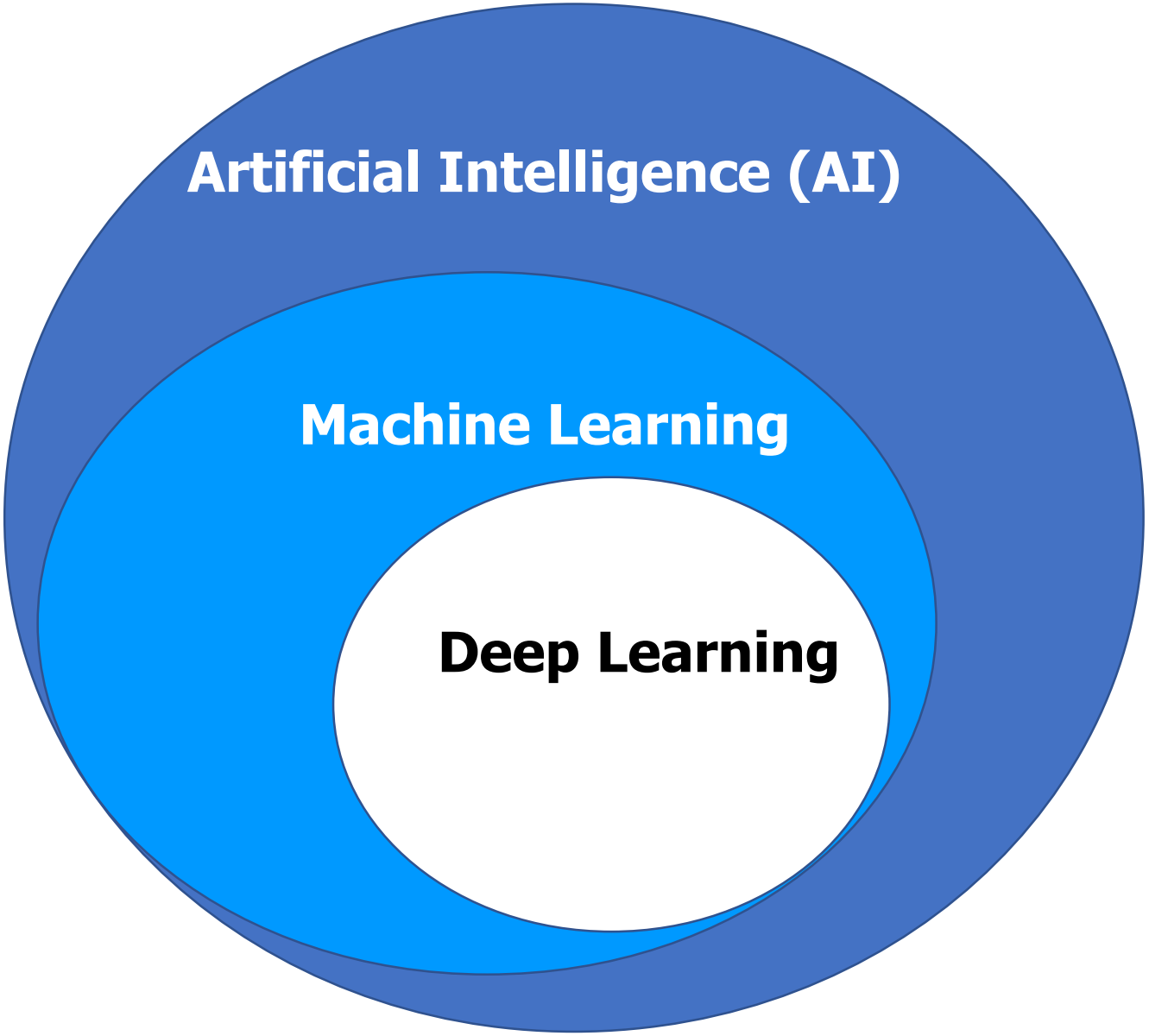


Turing Award 2019

Turing Award 2019

Geoffrey Hinton, Yann LeCun, and Yoshua Bengio

AI and Machine Learning



Basic math needed in Machine Learning

- Calculus
- Linear algebra
- Probability
- Statistics

Advanced math involved in Machine Learning

- Topology
- Geometry
- Analysis
- Algebra
- Graph theory
- Dynamical system
- ODE and PDE
- Combinatorics

Machine Learning Methods and Data

- Unsupervised learning (clustering):

$$\text{Data: } \mathcal{D} = \left\{ \mathbf{x}^{(i)} \middle| \mathbf{x}^{(i)} \in \mathbb{R}^n \right\}_{i=1}^{M}$$

- Supervised learning (classification):

$$\mathcal{D} = \left\{ \left(\mathbf{x}^{(i)}, y^{(i)} \right) \middle| \mathbf{x}^{(i)} \in \mathbb{R}^n, y^{(i)} \in \{C_1, C_2, \dots, C_K\} \right\}_{i=1}^{M}$$

- Supervised learning (regression):

$$\mathcal{D} = \left\{ \left(\mathbf{x}^{(i)}, y^{(i)} \right) \middle| \mathbf{x}^{(i)} \in \mathbb{R}^n, y^{(i)} \in \mathbb{R} \right\}_{i=1}^{M}$$

- Semi-supervised learning (regression):

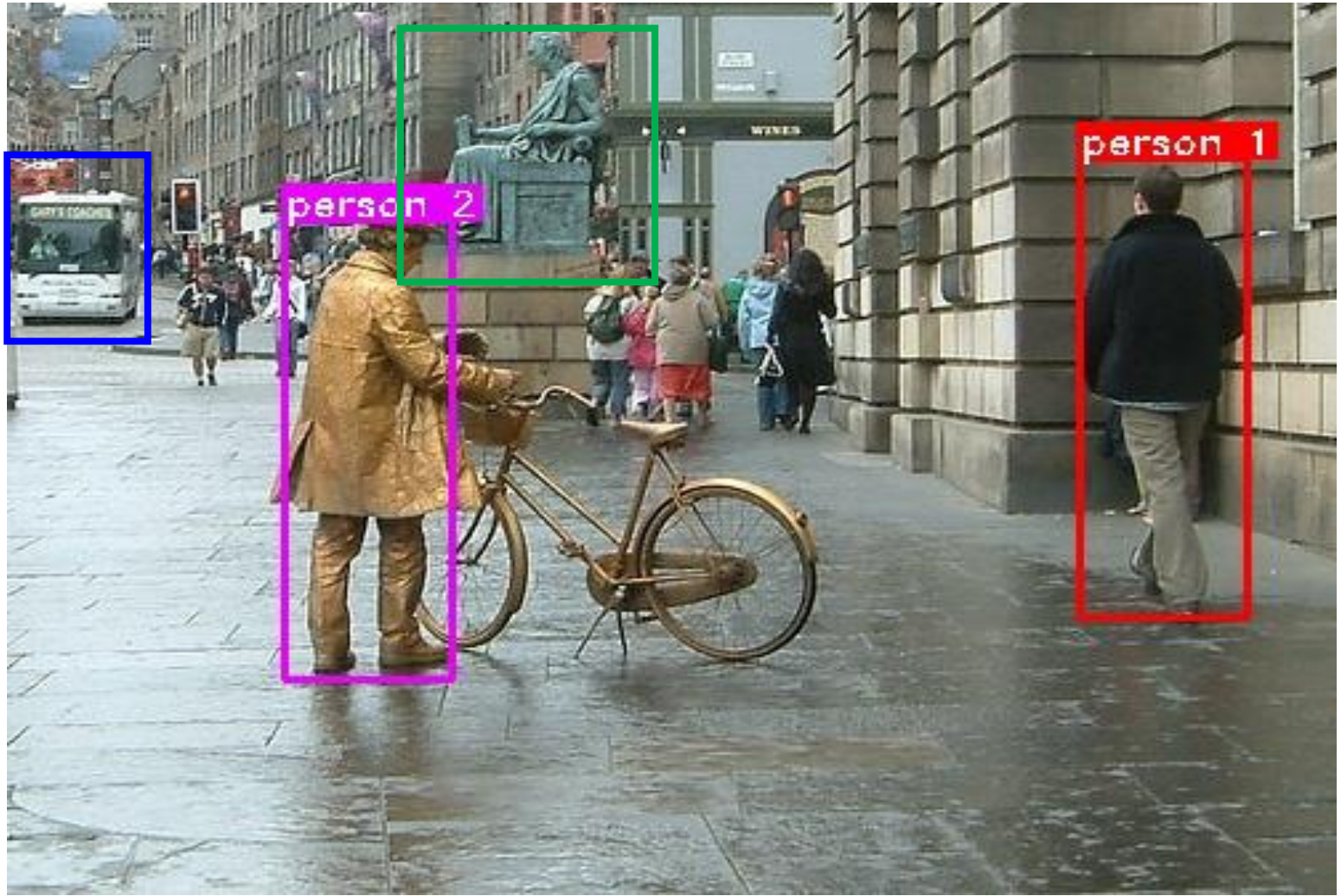
$$\mathcal{D} = \left\{ \left(\mathbf{x}^{(i)}, y^{(i)} \right) | \mathbf{x}^{(i)} \in \mathbb{R}^n, i = 1,2, \cdots M; y^{(i)} \in \mathbb{R}, i = 1,2, \cdots M' < M \right\}$$

- Semi-supervised learning (Classification):

$$\mathcal{D} = \left\{ \left(\mathbf{x}^{(i)}, y^{(i)} \right) \middle| \begin{array}{c} \mathbf{x}^{(i)} \in \mathbb{R}^n, i = 1,2, \cdots M; \\ y^{(i)} \in \{C_1, C_2, \dots, C_K\}, i = 1,2, \cdots M' < M \end{array} \right\}$$

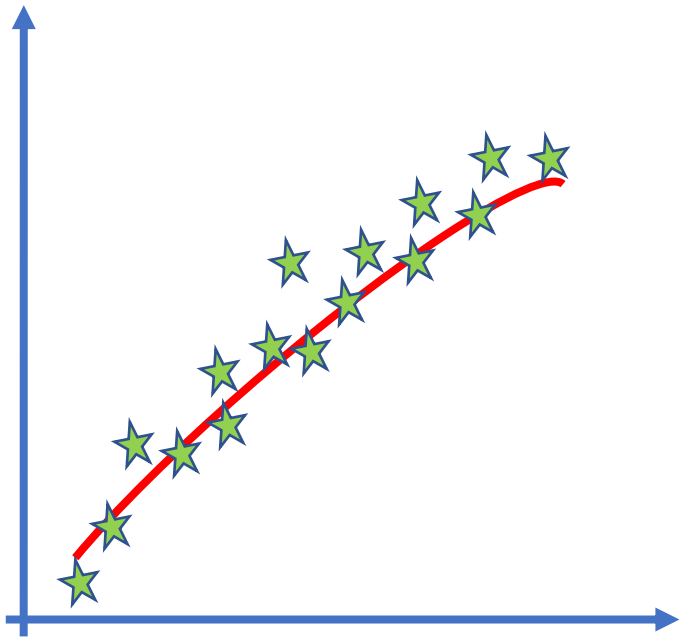
Unsupervised Learning

- K-means
- Generative Adversarial Networks
- Reinforced learning, ….

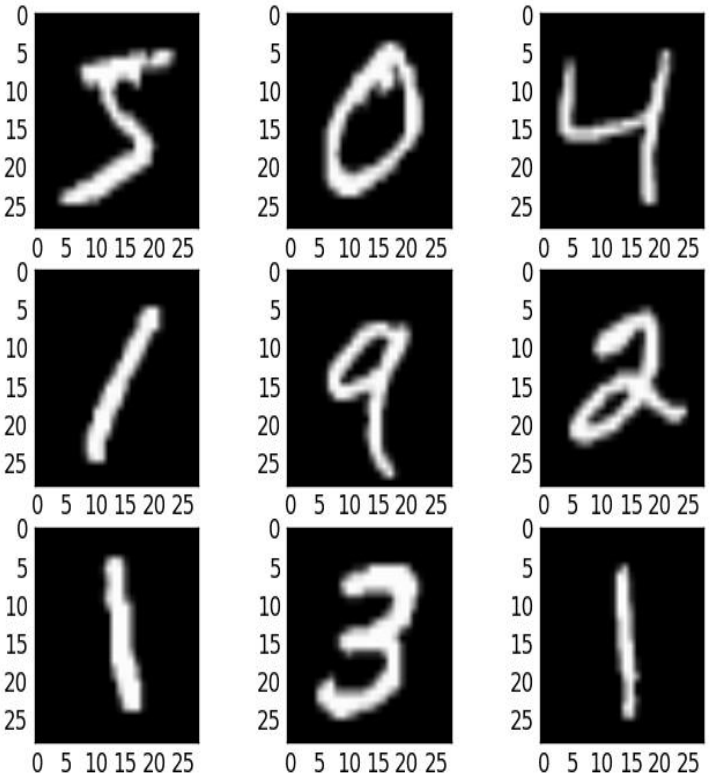


Supervised Learning

- **Regression**: Linear regression, Logistic regression, SVM, Random Forest, GBDT, CNN, RNN, etc.



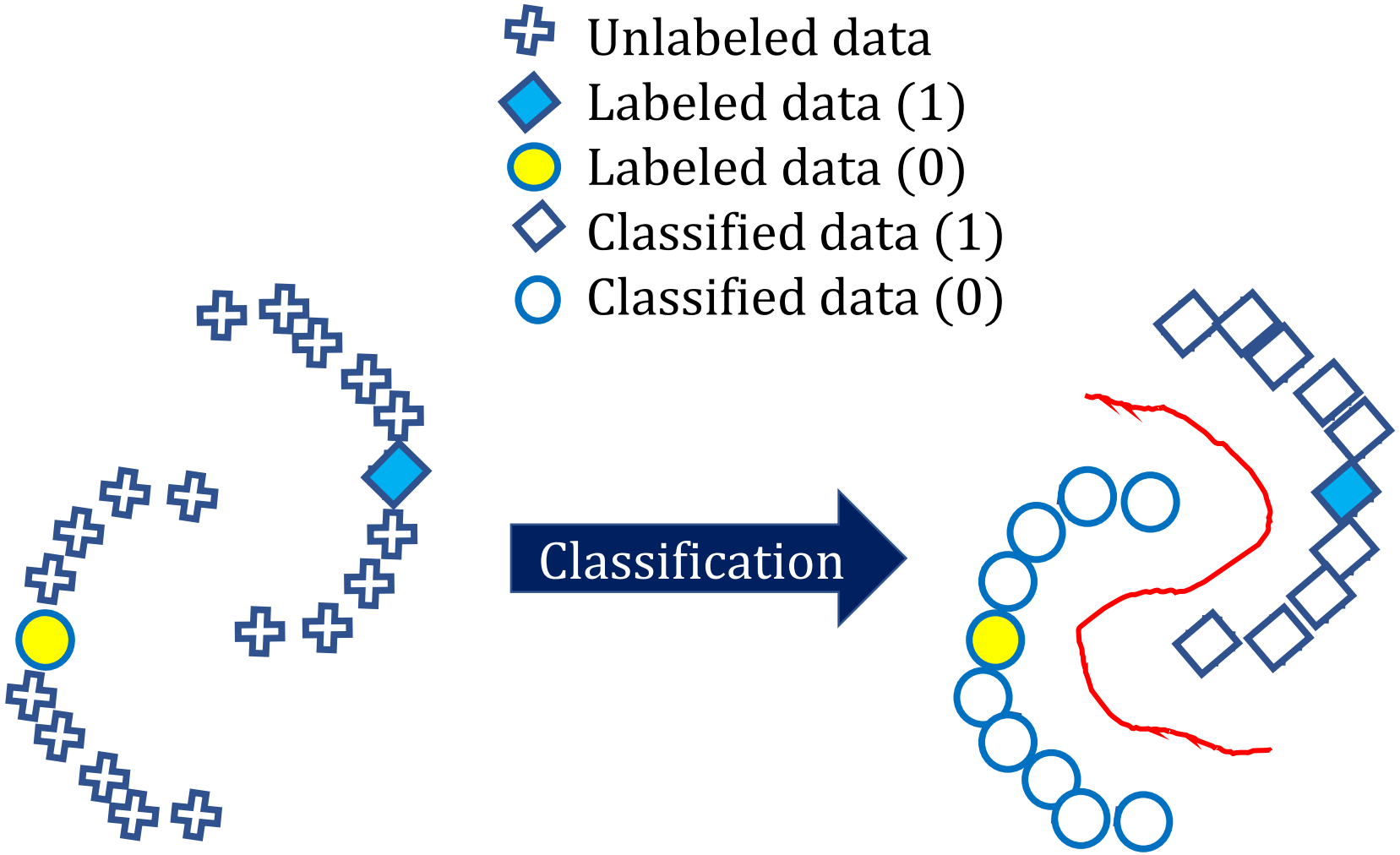
- **Classification**: K-NN, SVM, Random Forest, CNN, RNN, etc.



Semi-supervised Learning (classification)

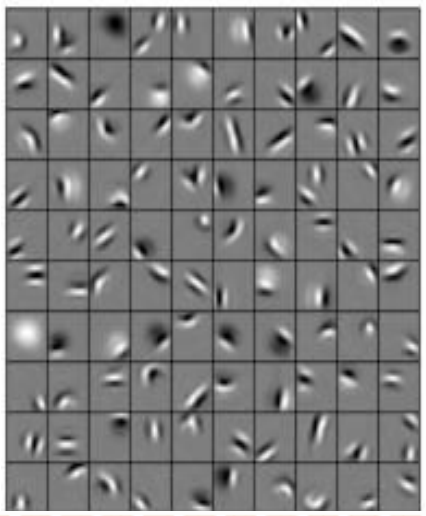
Predict unlabeled data from labeled information

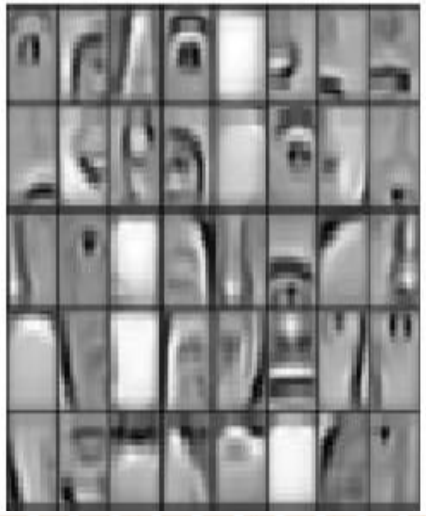
- Manifold learning
- Kernel SVM

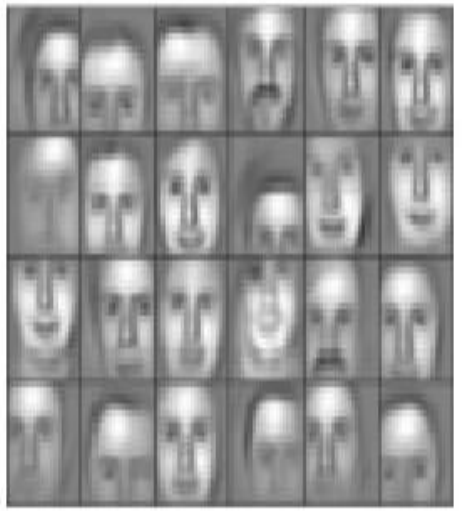


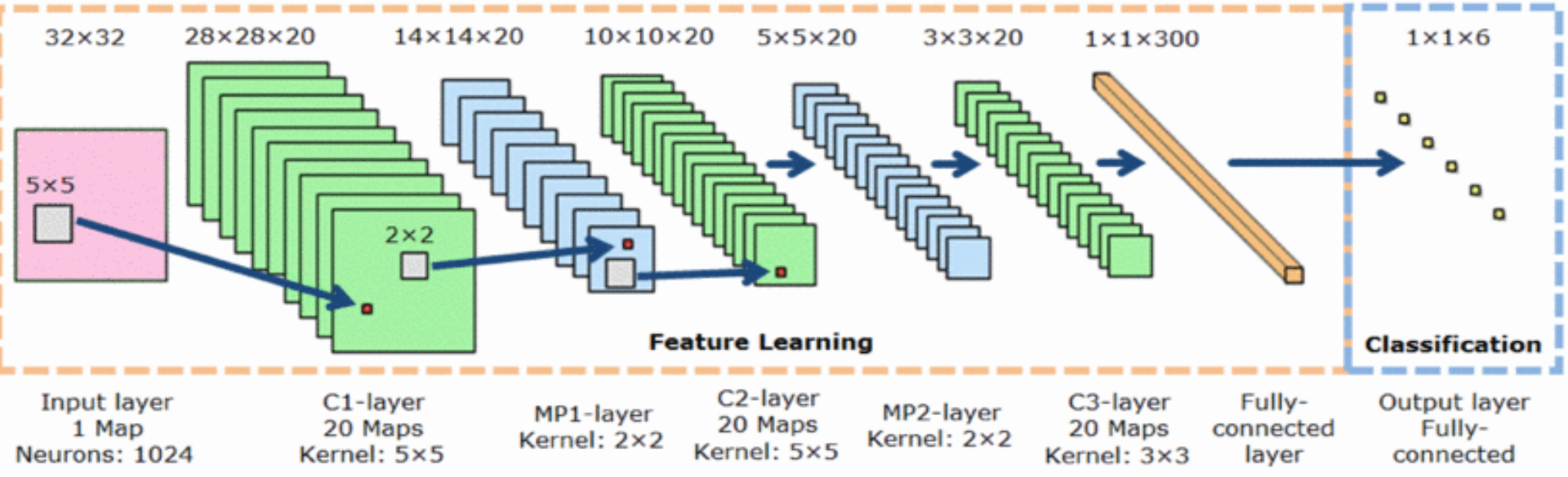
Artificial Intelligence & Deep learning

Bryson and Ho (Backpropagation 1969); Fukushima (Neo-Cognitron 1980); LeCun (CNN 1998); Hopfield (RNN 1982); Hochreiter and Schmidhuber (LSTM 1997); Goodfellow et al (GAN 2014); Autoencoder; Image translation, …

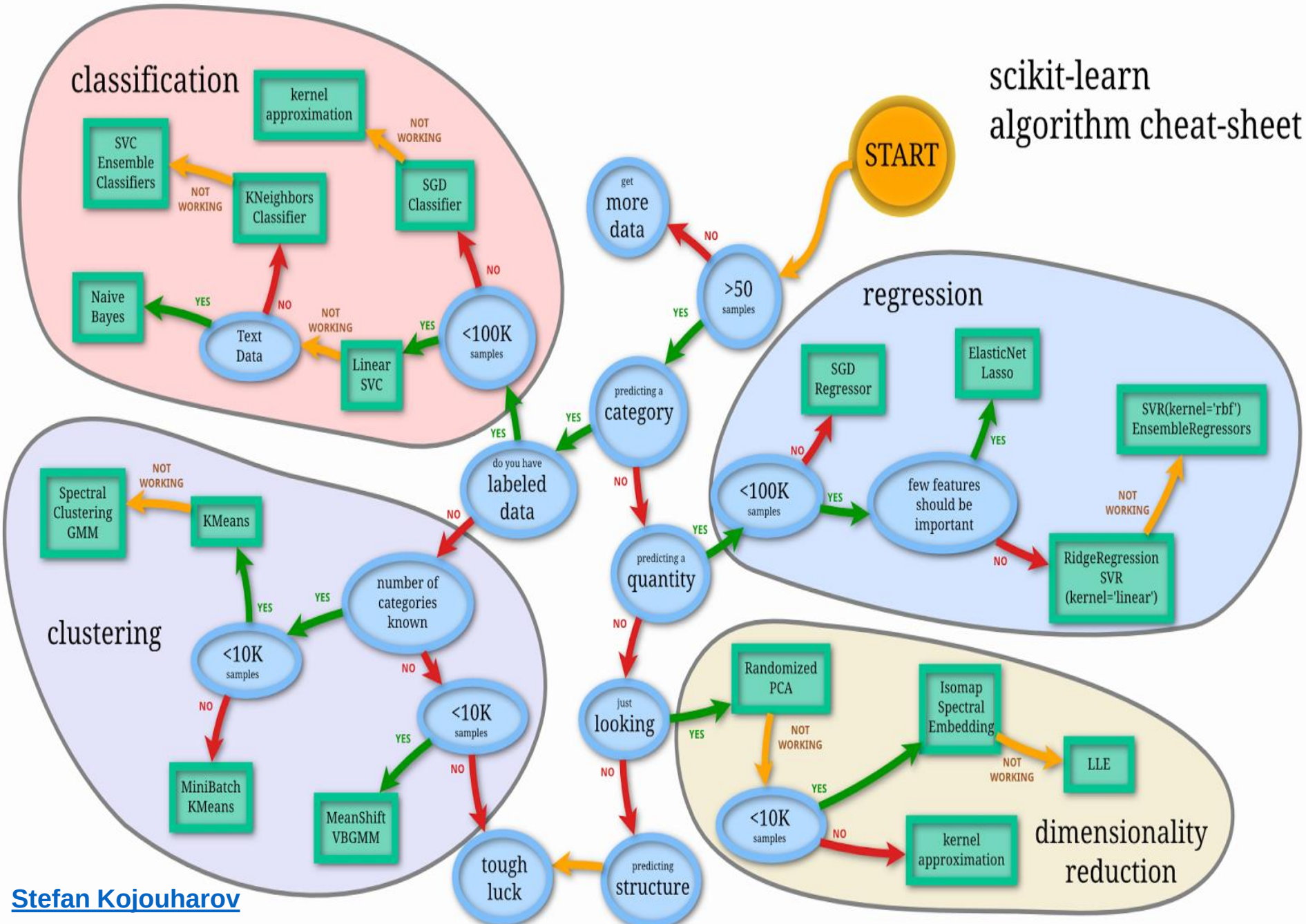




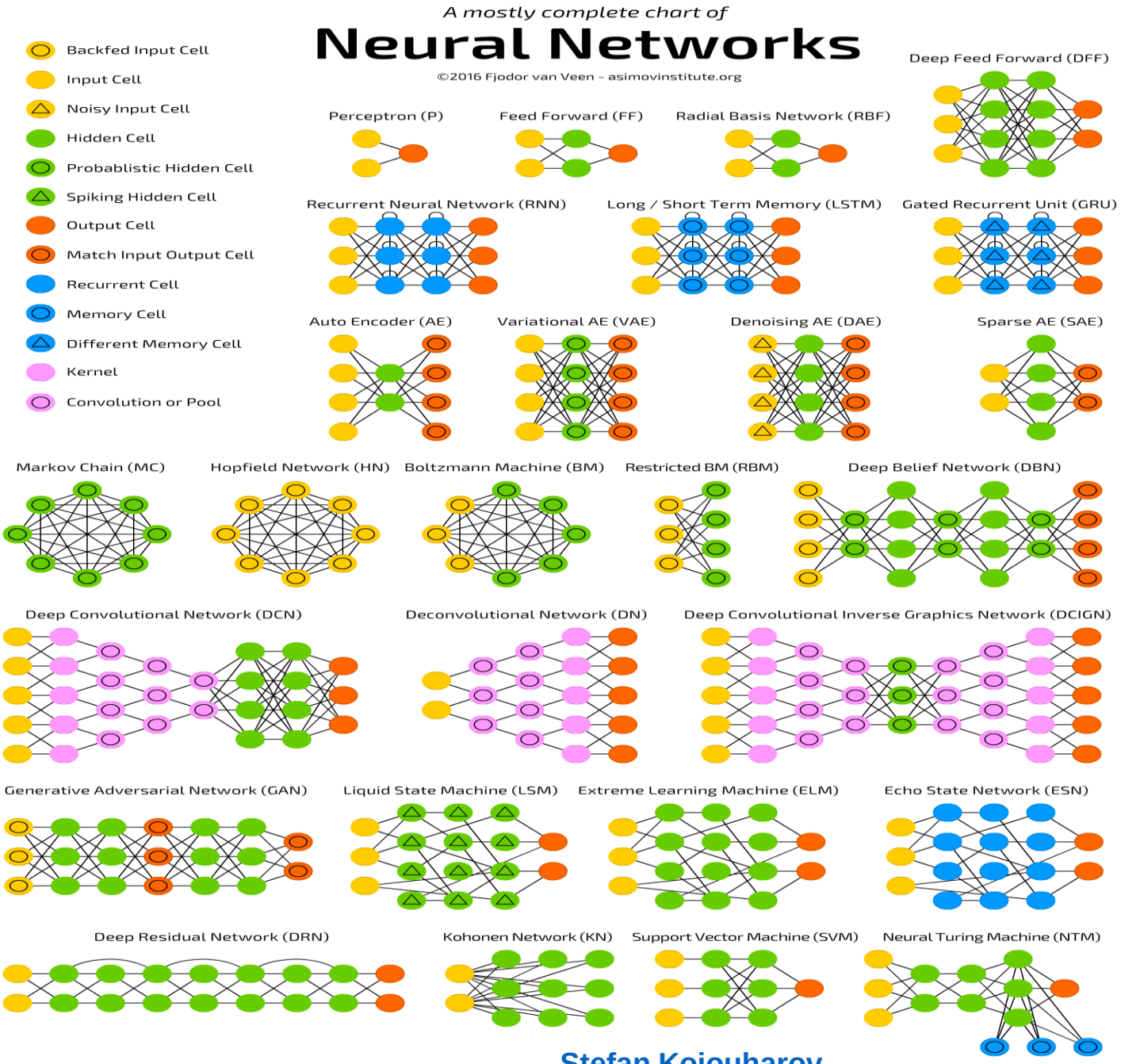




Machine Learning Summary



Stefan Kojouharov



Potential mathematics:

Persistent homology
Topological complexity
Poset
Directed graph
Spectral graph
Hypergraph
Algebraic geometry
K-theory?
Exterior algebra?
Homotopy
Matroid
Differential geometry

Computational issues:

Linear algebra
Matrix analysis
Sophistical analysis
Gradient descend
…

Stefan Kojouharov

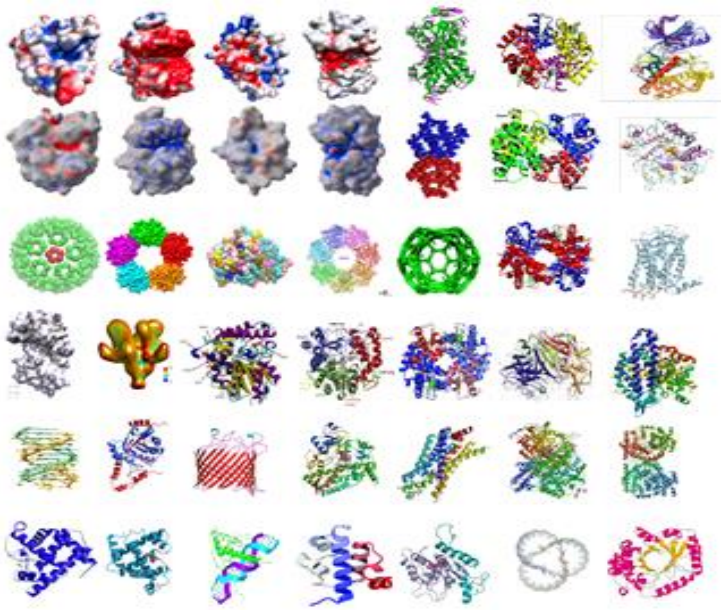
How to do deep learning for 3D biomolecular data?

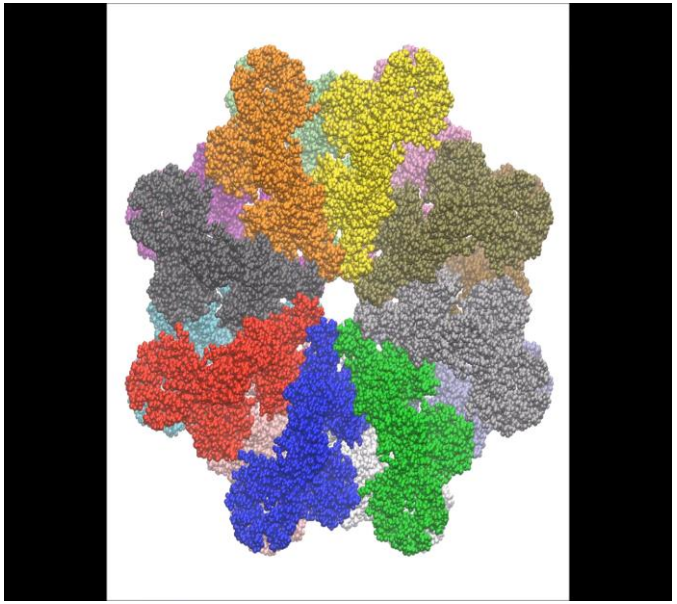
Obstacles for deep learning of 3D biomolecules:

- **Geometric dimensionality: $\mathbb{R}^{3N}$, where *N* ~5000 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: intermolecular & intramolecular interactions**

Solution:

- **Geometric simplification, dimension reduction & scale unification**





Two schools of thinking

Given a protein with ***N*** atom and an average of ***n*** electrons in each atom

Fundamentalism

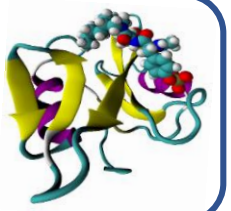
Quantum Mechanics $\mathbb{R}^{3Nn+3N}$

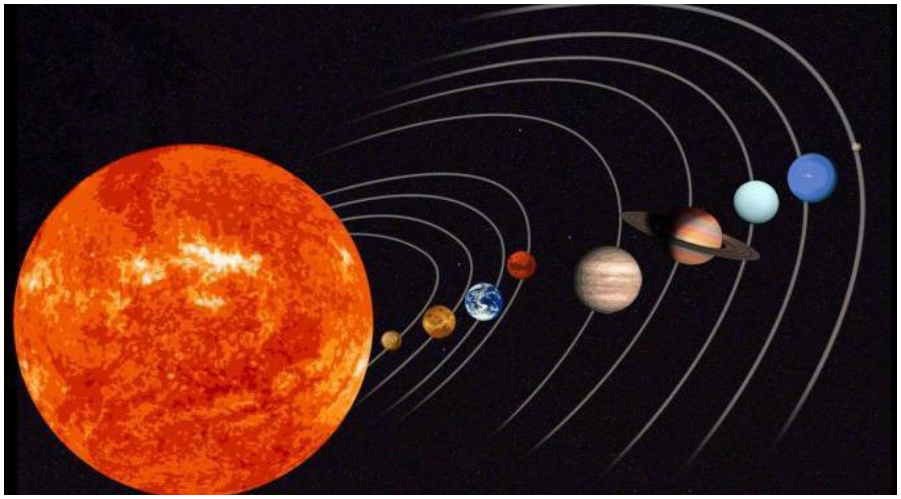
QM/MM $\mathbb{R}^{K}$ $3N< K <3N(n+1)$

Molecular Mechanics $\mathbb{R}^{3N}$

Multiscale Coarse-grain $\mathbb{R}^{M}$ $(3<M<3N)$

Poisson-Boltzmann, PNP, etc. $\mathbb{R}^{3}$





Differentiable Manifold $\mathbb{R}^{2}$

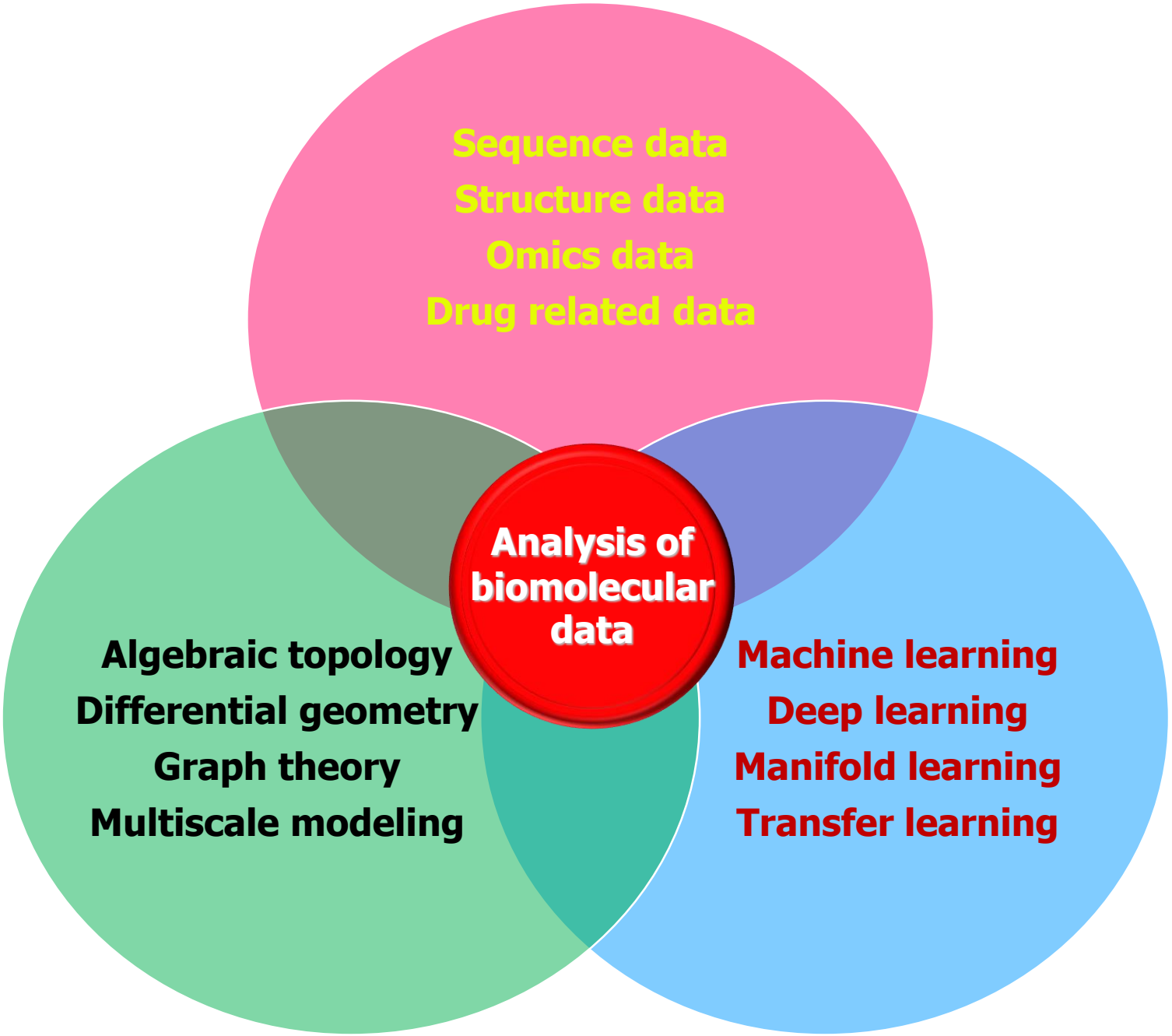
Algebraic Topology $\mathbb{R}^{1}$

Graph Theory $\mathbb{R}^{0}$

Geo-Top Indices $\mathbb{R}^{0}$

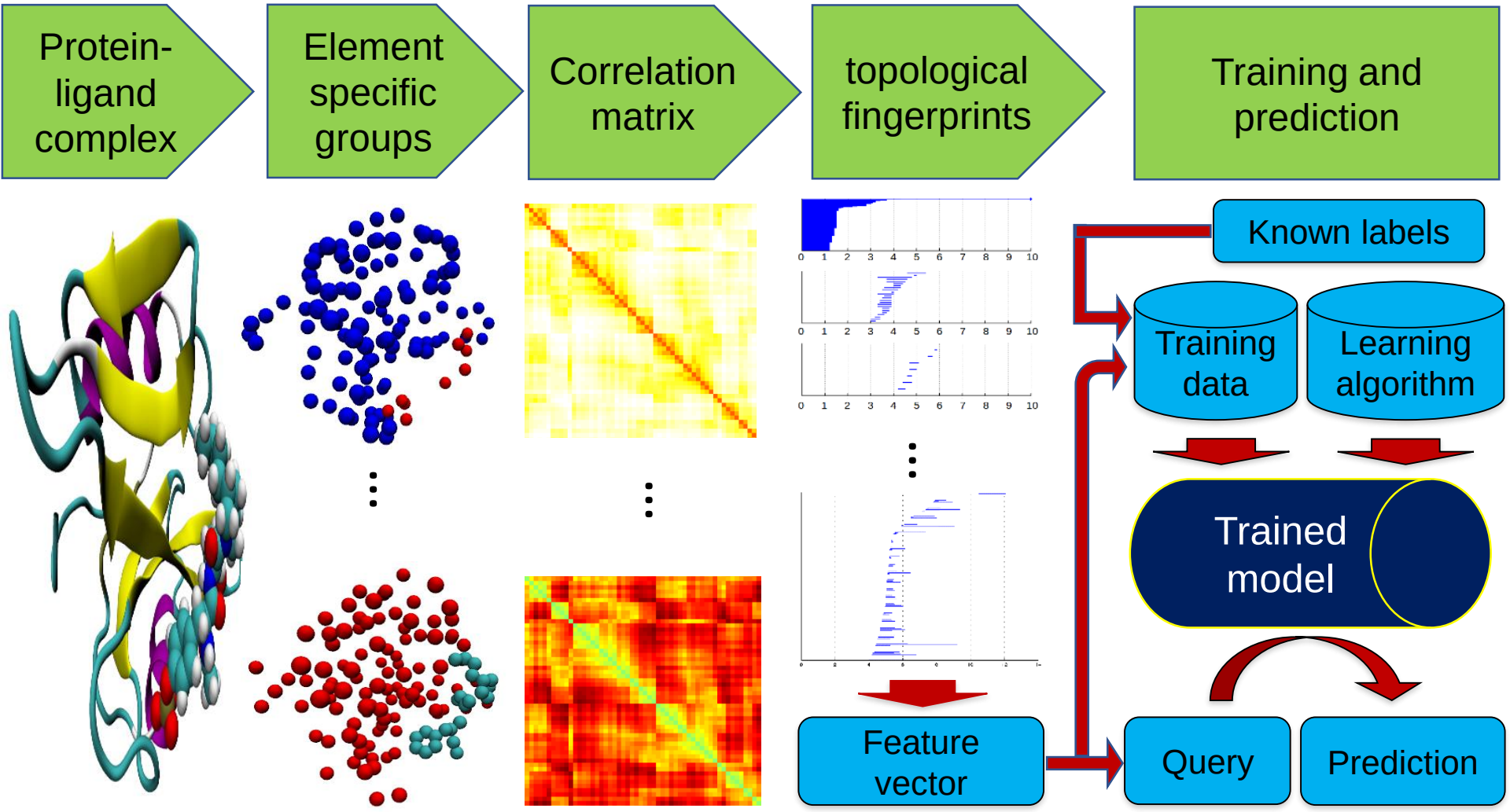
Reductionism

Basic hypothesis:
Intrinsic physics lies on low-dimensional manifolds in a high dimensional space

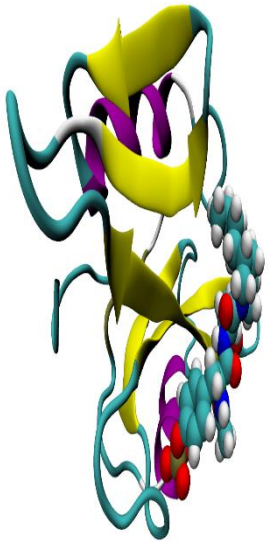

Sequence data
Structure data
Omics data
Drug related data
Analysis of biomolecular data
Algebraic topology
Differential geometry
Graph theory
Multiscale modeling
Machine learning
Deep learning
Manifold learning
Transfer learning

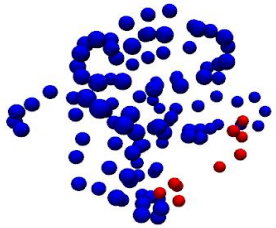
Topology based learning architecture

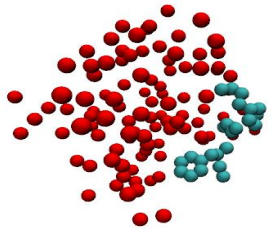
(Cang & Wei, IJNMBE, 2017)

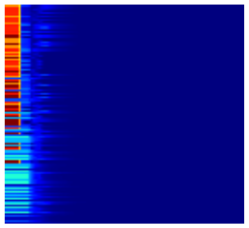


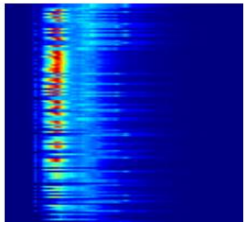
Topological convolutional deep learning architecture









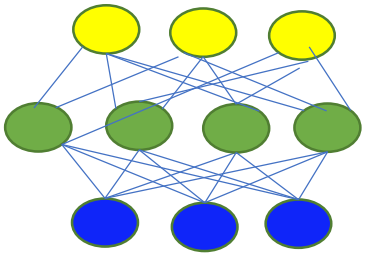


Multichannel images (54x200)

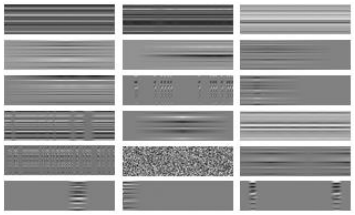
Convolution (128x200)

Pooling (128x100)

Flattening (1xN)



Prediction



Original protein-ligand Complex

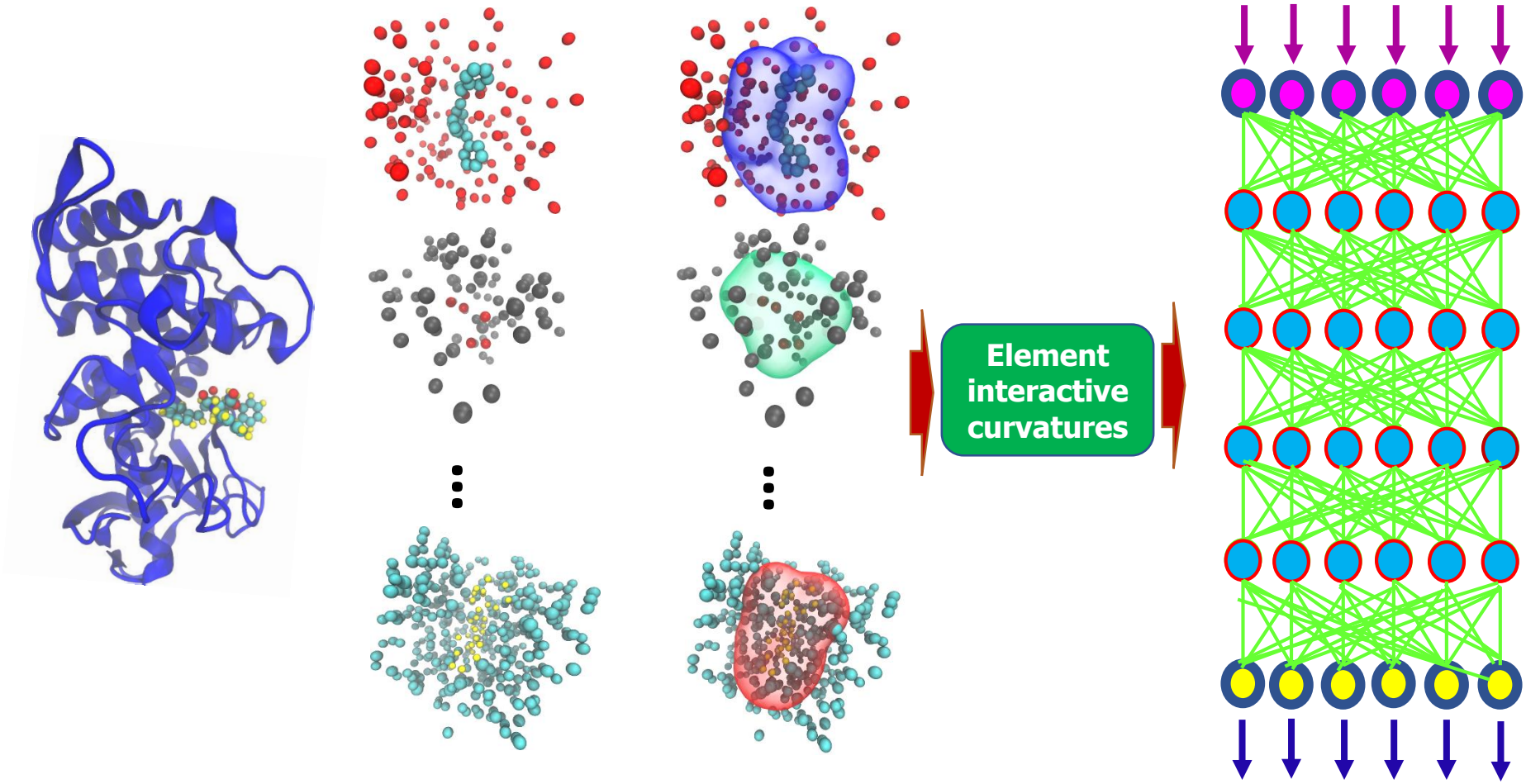
Classify atoms into element specific groups

Generate topological, DG & graph fingerprints

Convolutional deep learning neural network

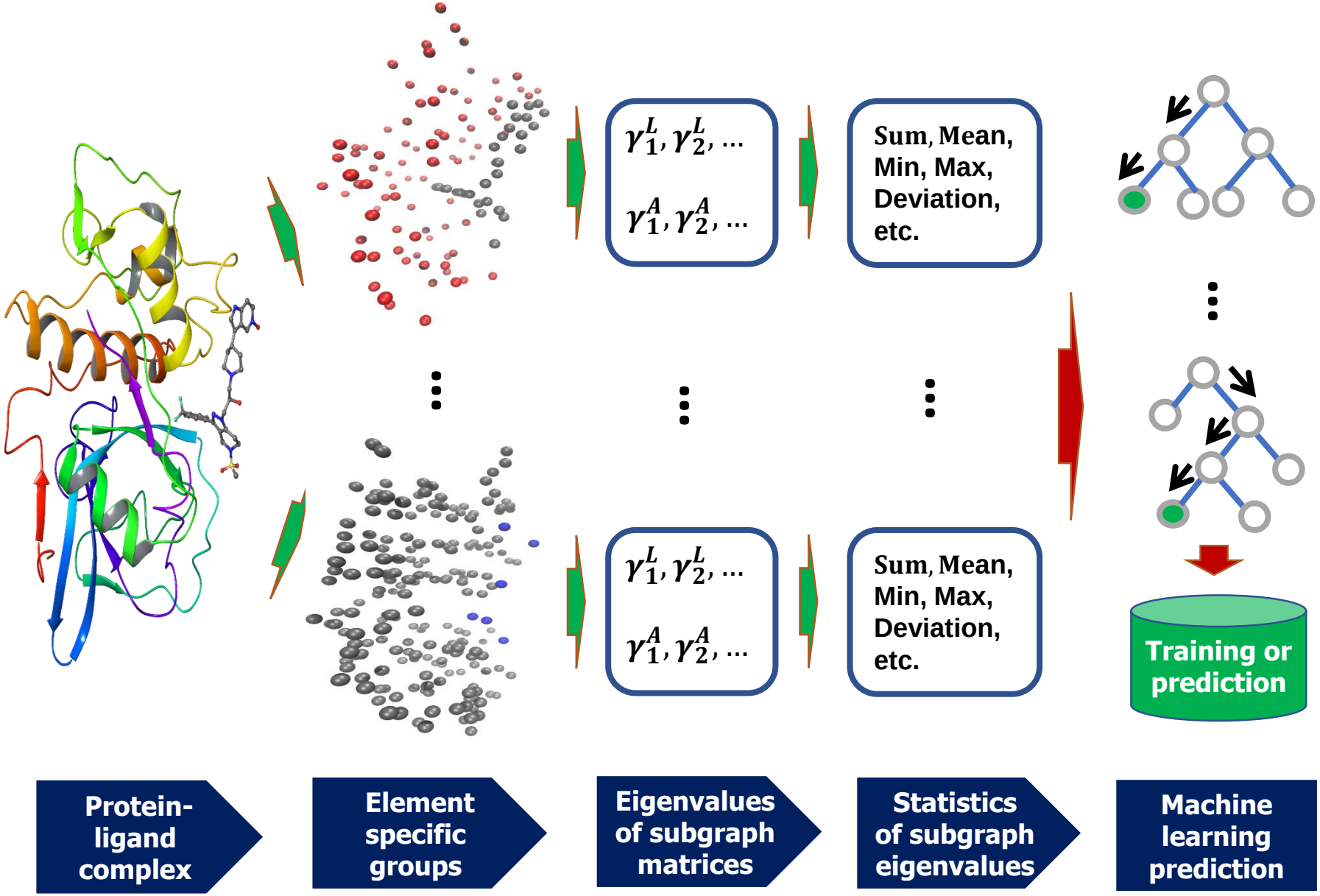
(Cang & Wei, PLOS CB, 2017)

Differential geometry based deep learning



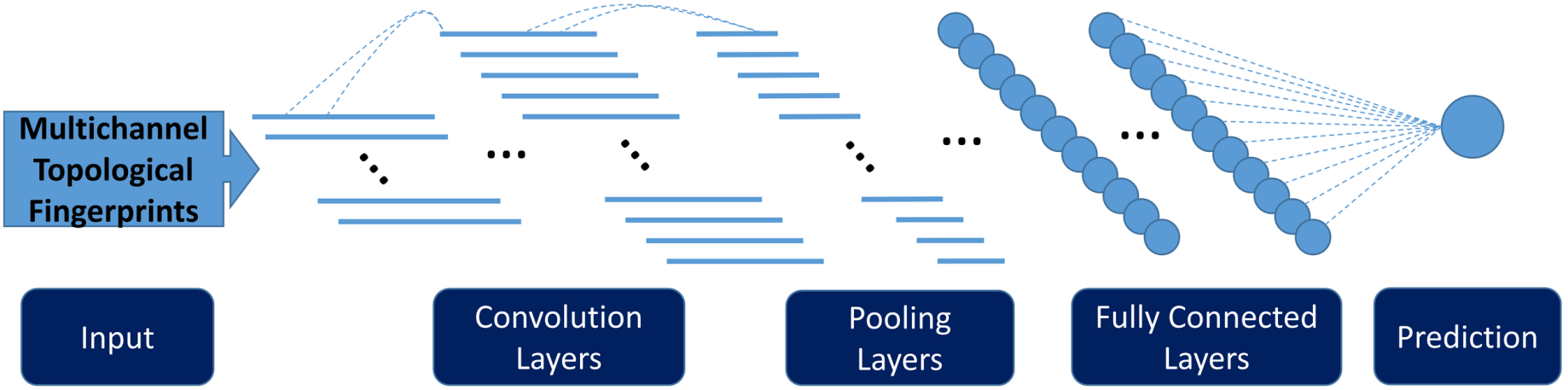
(Nguyen & Wei, 2018)

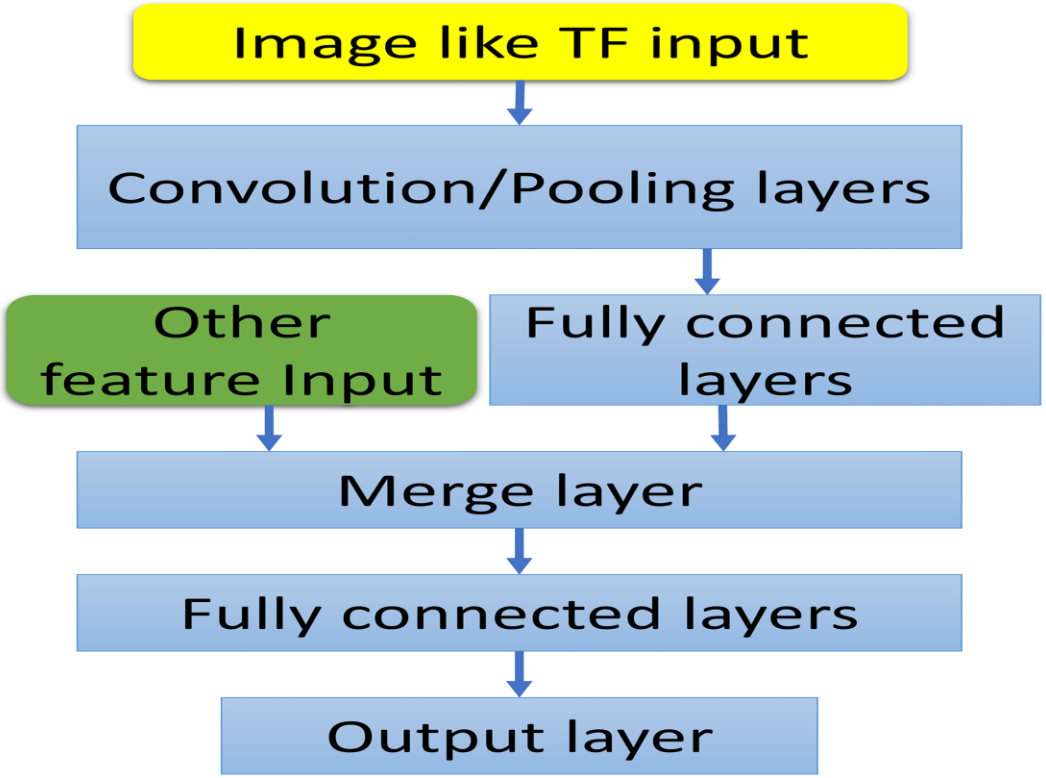
Algebraic graph theory based learning



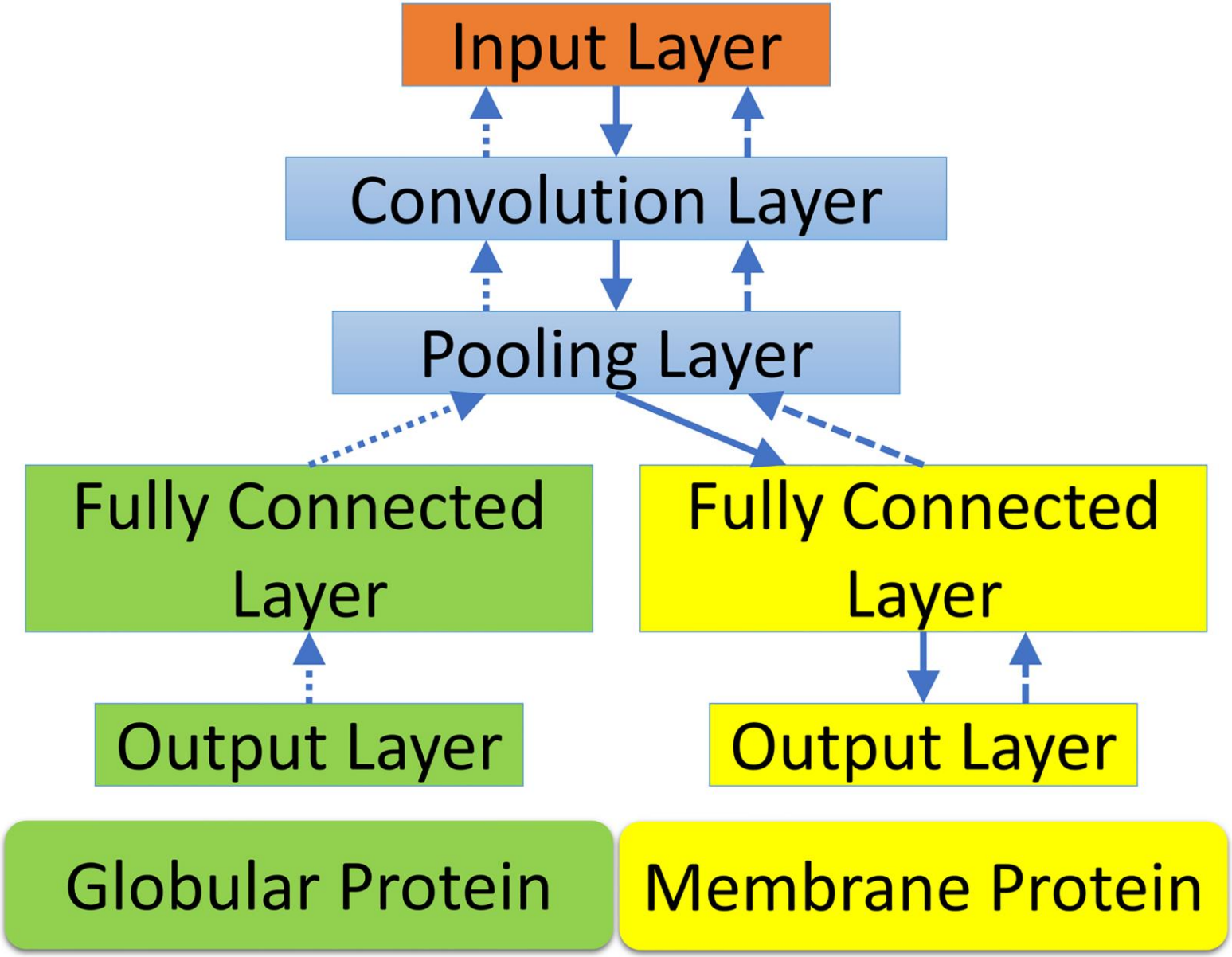
(Nguyen & Wei, 2018)

1D convolutional neural networks for mixed features

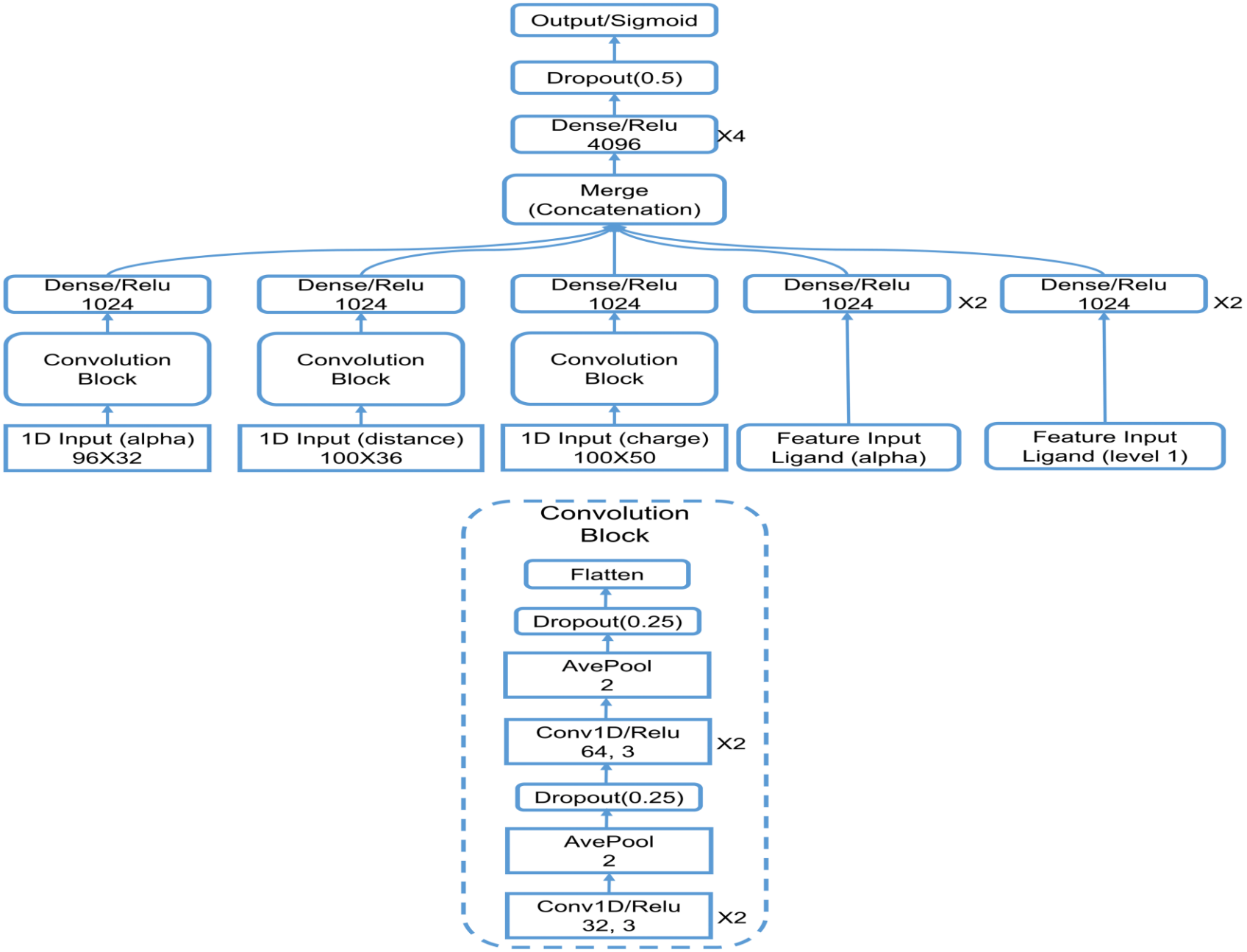




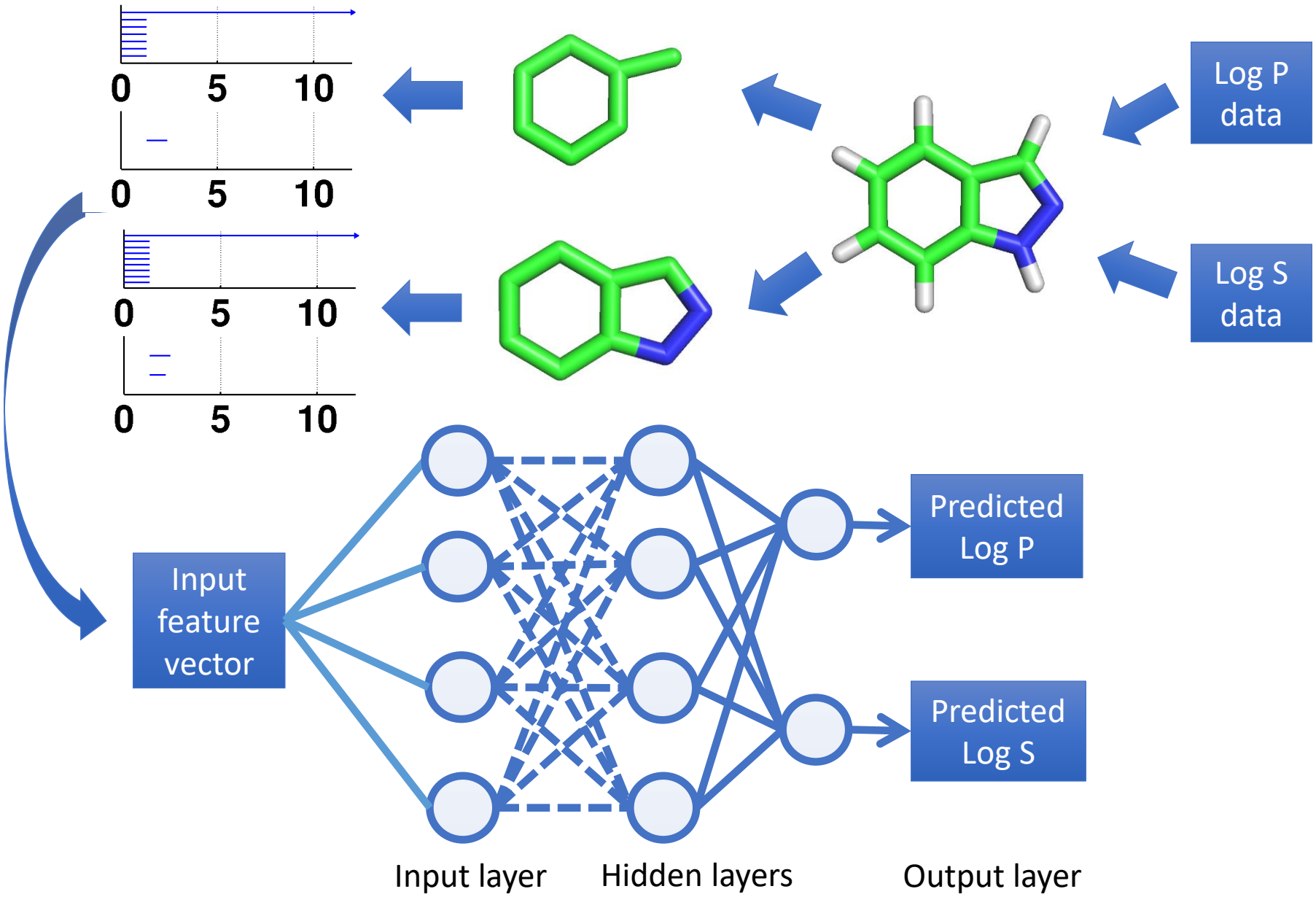
Multitask convolutional neural networks



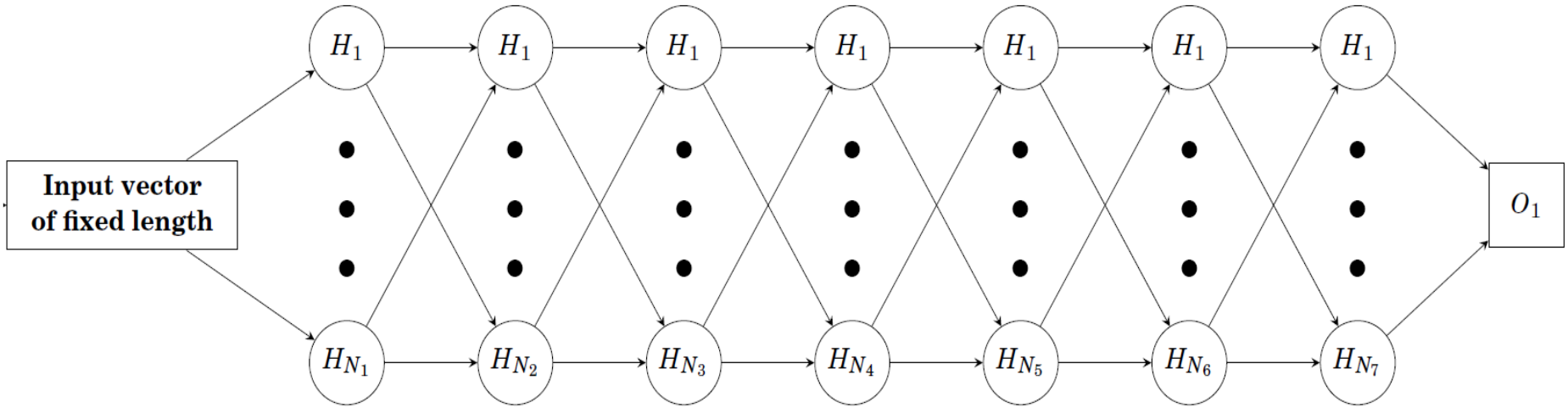
Convolutional neural networks for mixed features

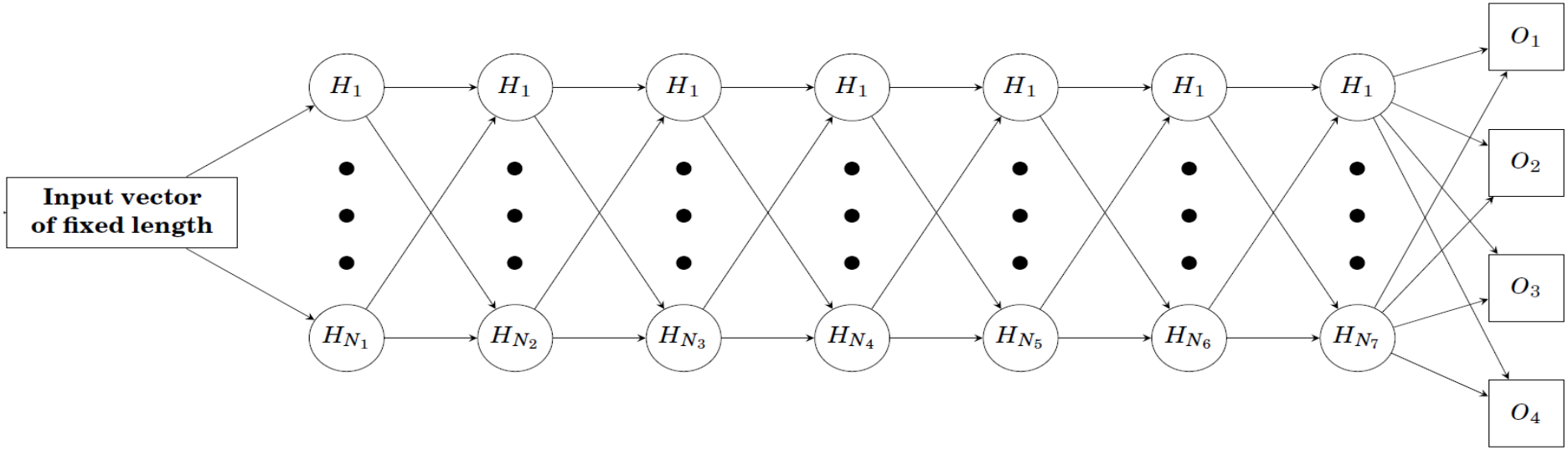


Artificial neural networks for multi-task solubility and partition coefficient predictions

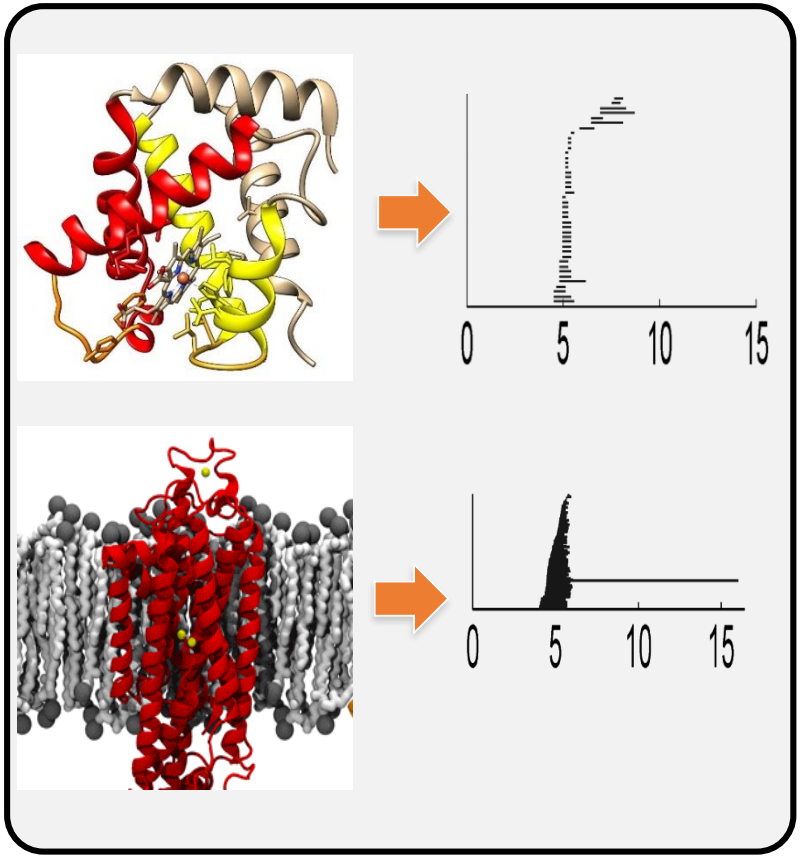


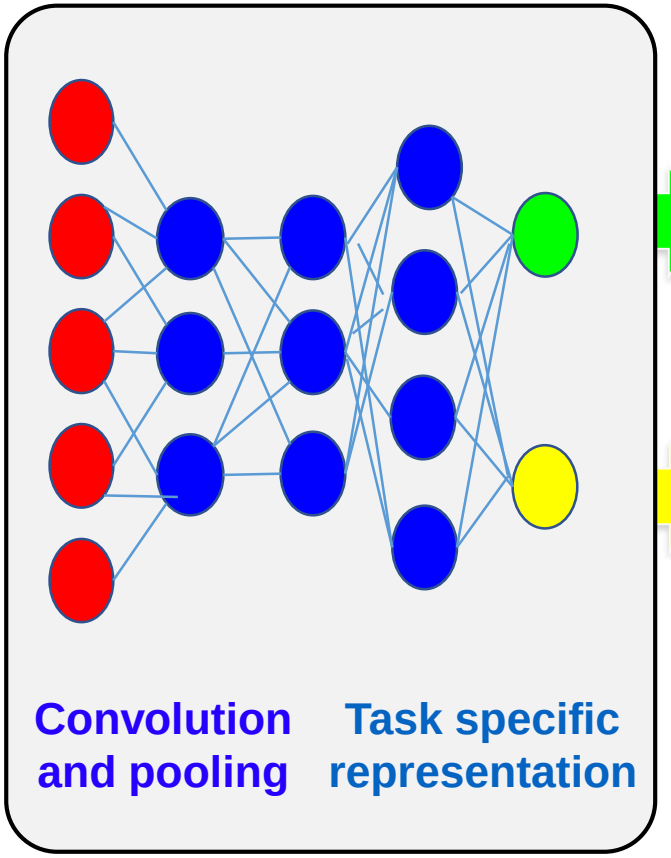
Artificial neural networks for single-task and multi-task toxicity predictions





Topological Multi-Task Deep Learning





Globular protein mutation impacts

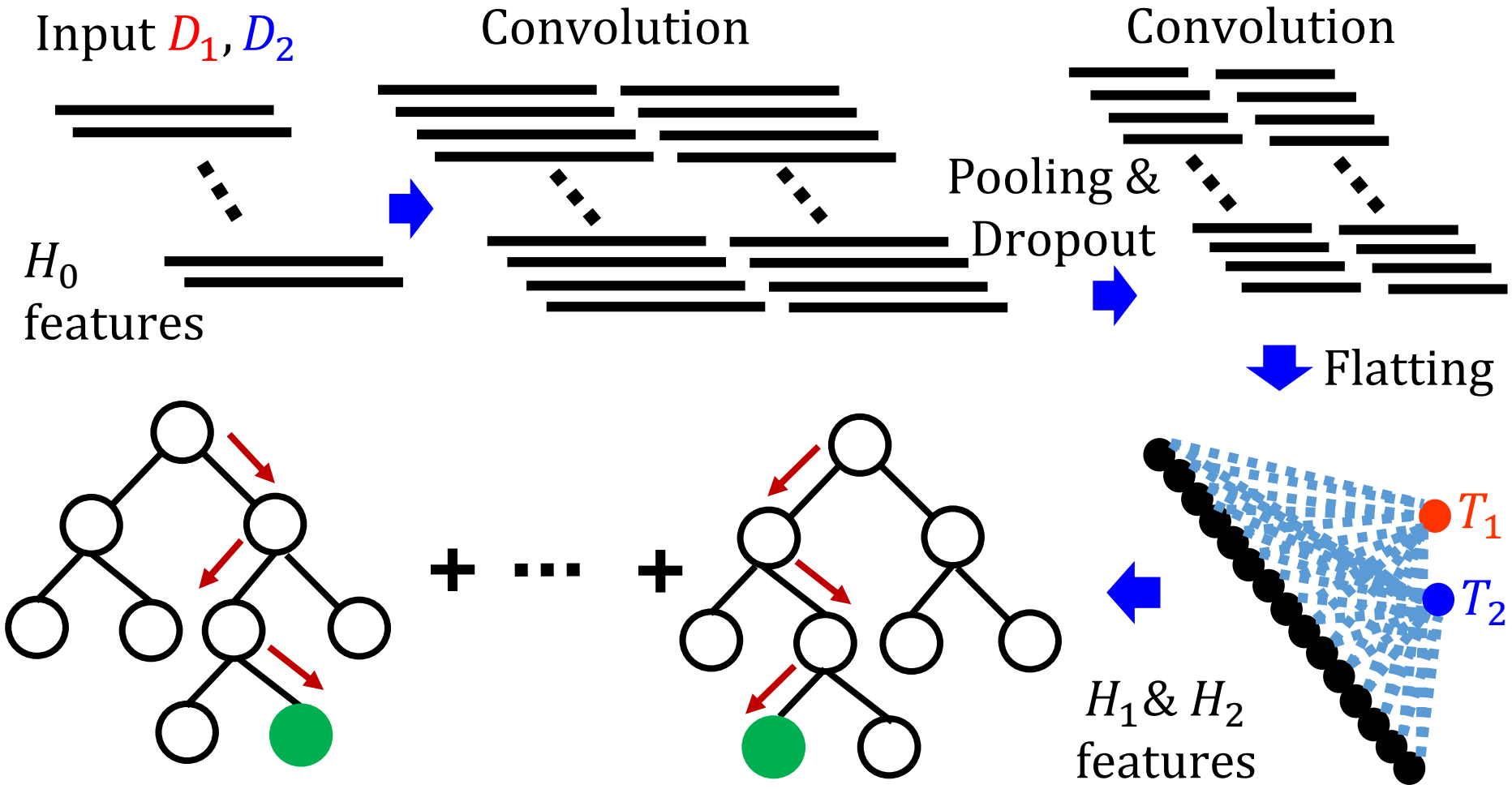
Membrane protein mutation impacts

Topological feature extraction

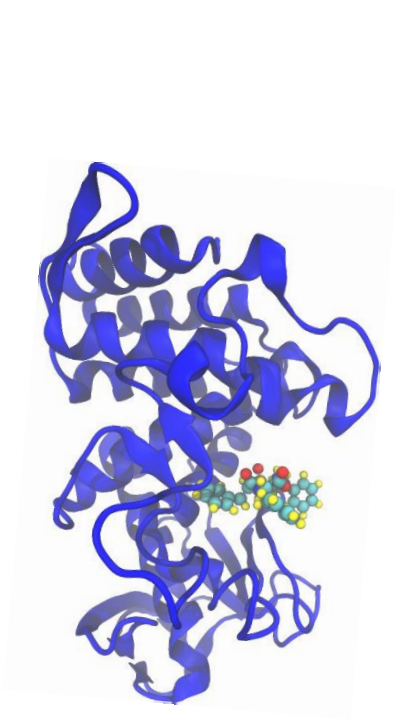
Multi-task topological deep learning

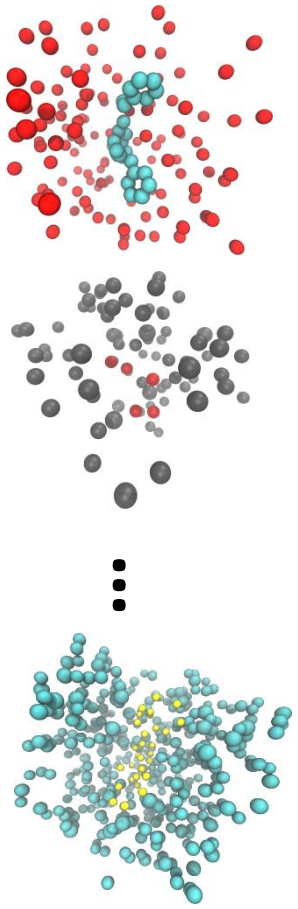
(Cang & Wei, PLOS CB, 2017)

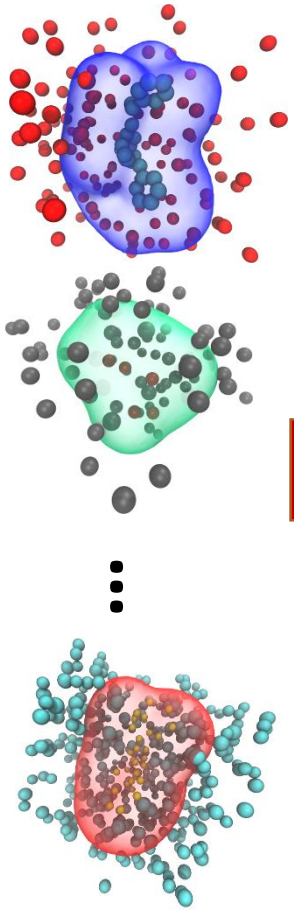
Convolutional neural network assisted multitask gradient boosting trees



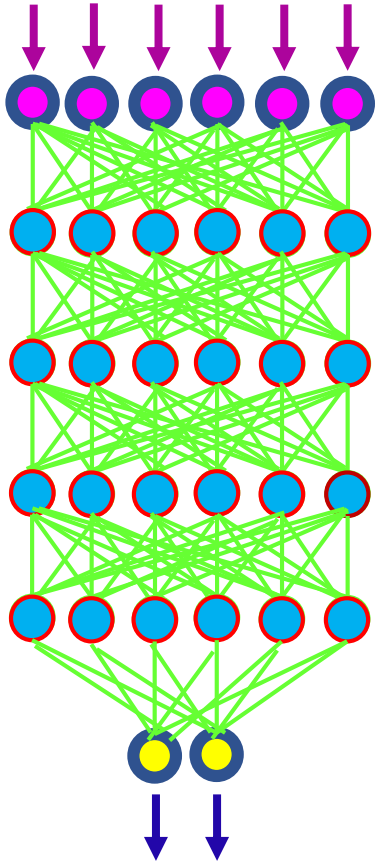
Mathematical deep learning







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



Protein-ligand complex

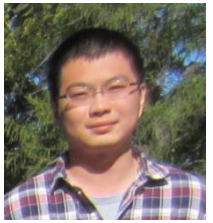
Element specific groups

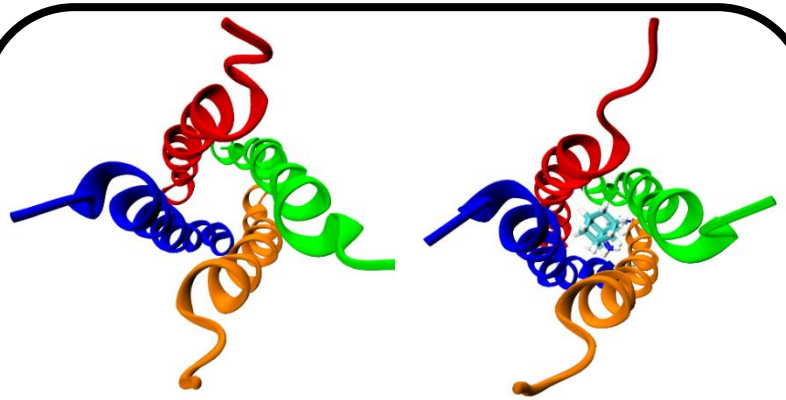
Element interactive manifolds

Various Mathematical features

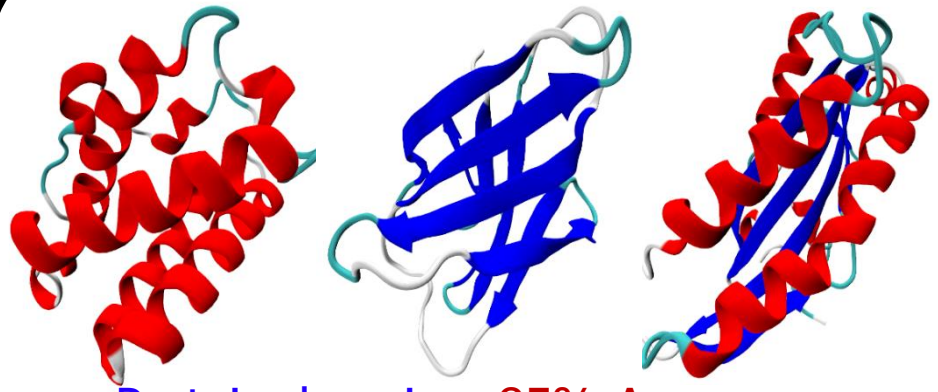
Machine learning prediction

Topological fingerprint based machine learning (SVM) for the classification of 2400 proteins



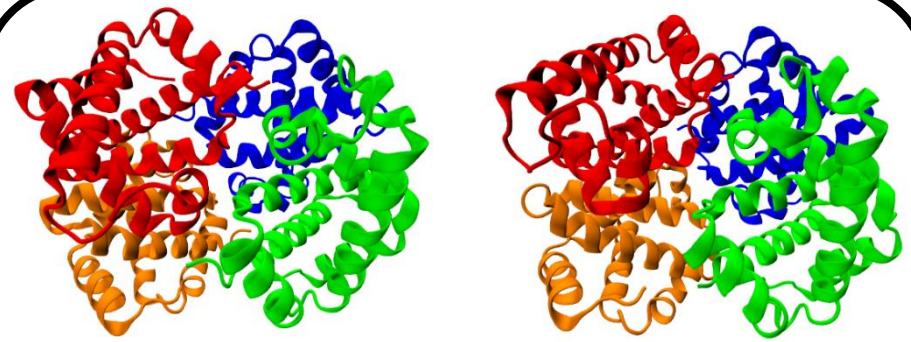


Influenza A virus drug inhibition: 96% Accuracy



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

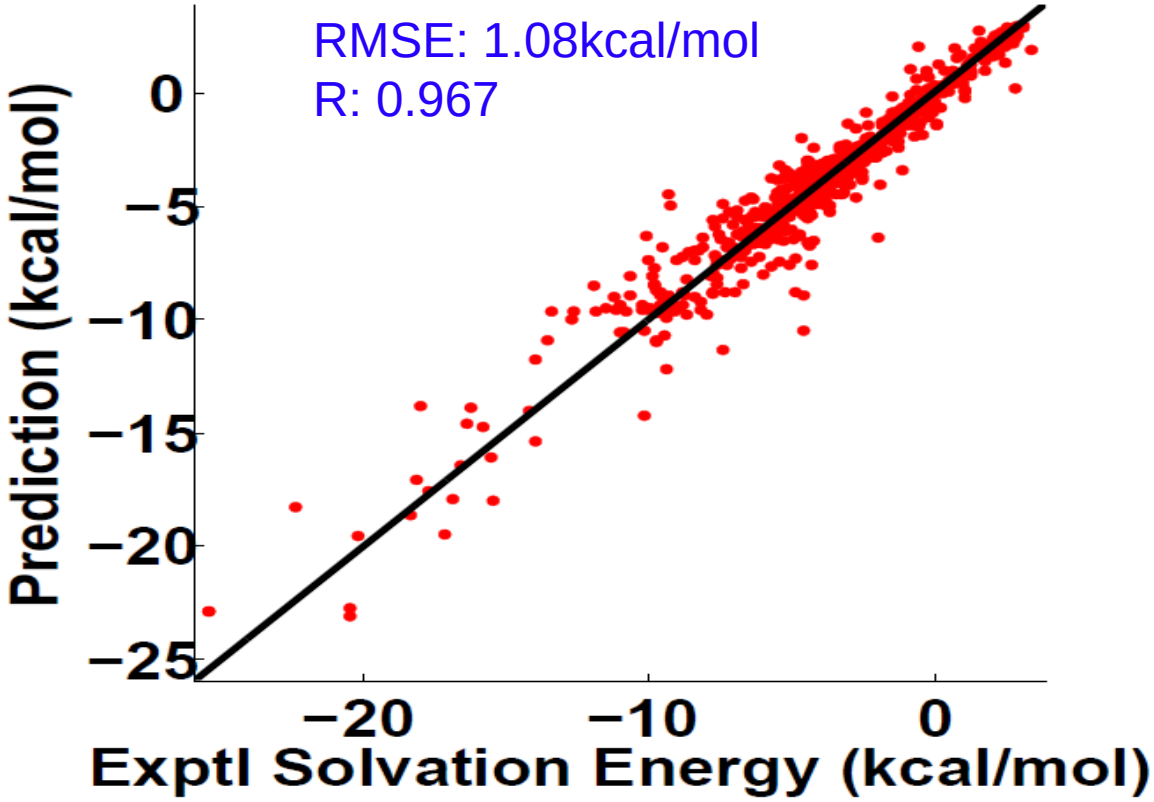
(Cang et al, MBMB, 2015)

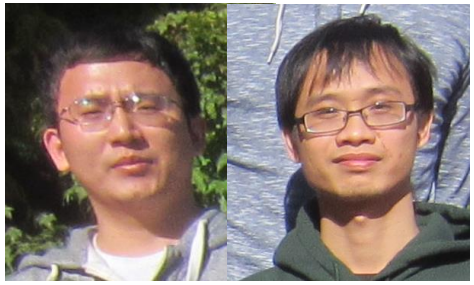


Hemoglobins in their relaxed and taut forms: 80% accuracy

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

Leave-one-out prediction for solvation free energies of 668 molecules

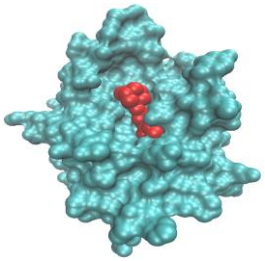


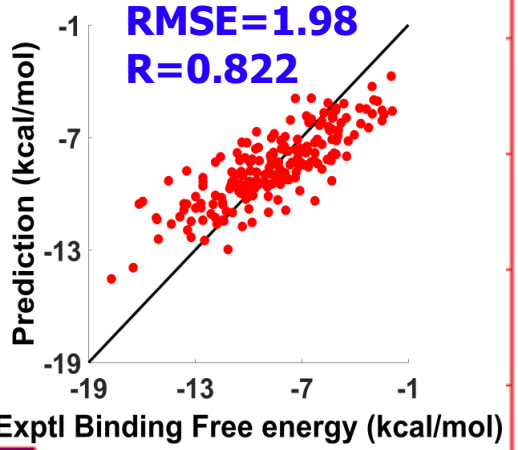


Wang, Zhao, Wei, JCP 2016

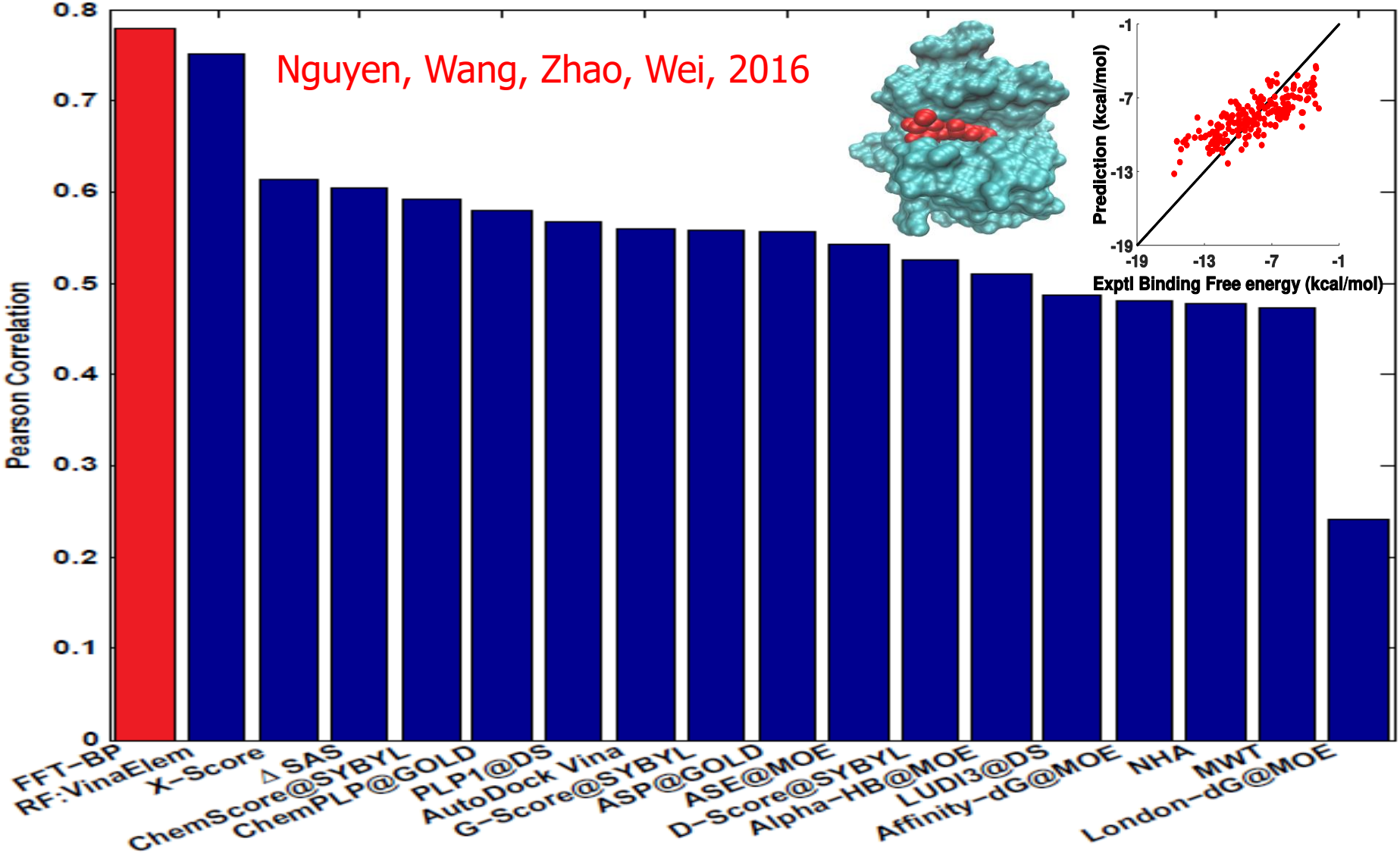
Blind binding affinity prediction of PDBBind v2007 core set of 195 protein-ligand complexes

Nguyen, Wang, Xiao, Wei, 2016



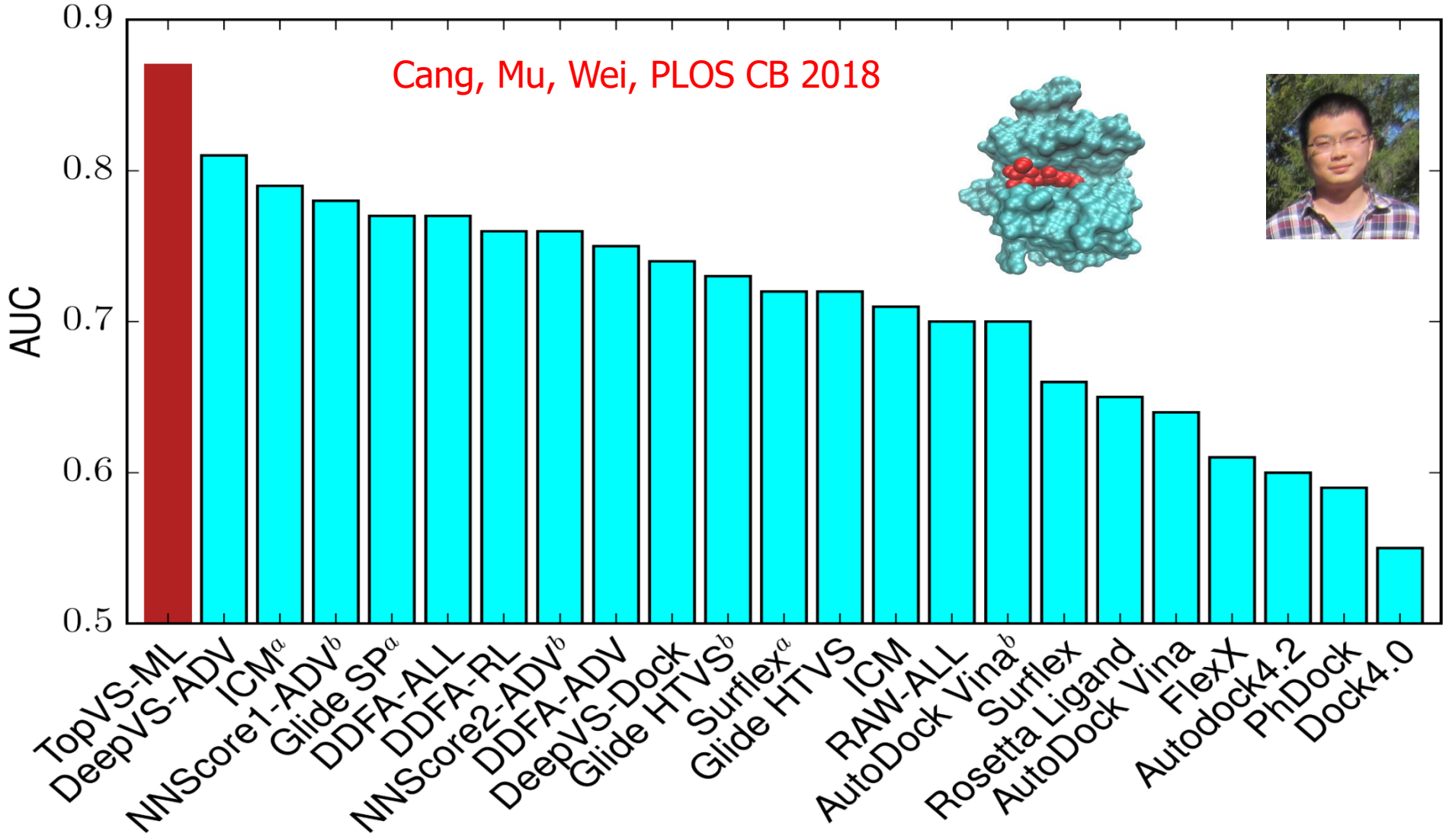


Blind binding affinity prediction of PDBBind v2013 core set of 195 protein-ligand complexes



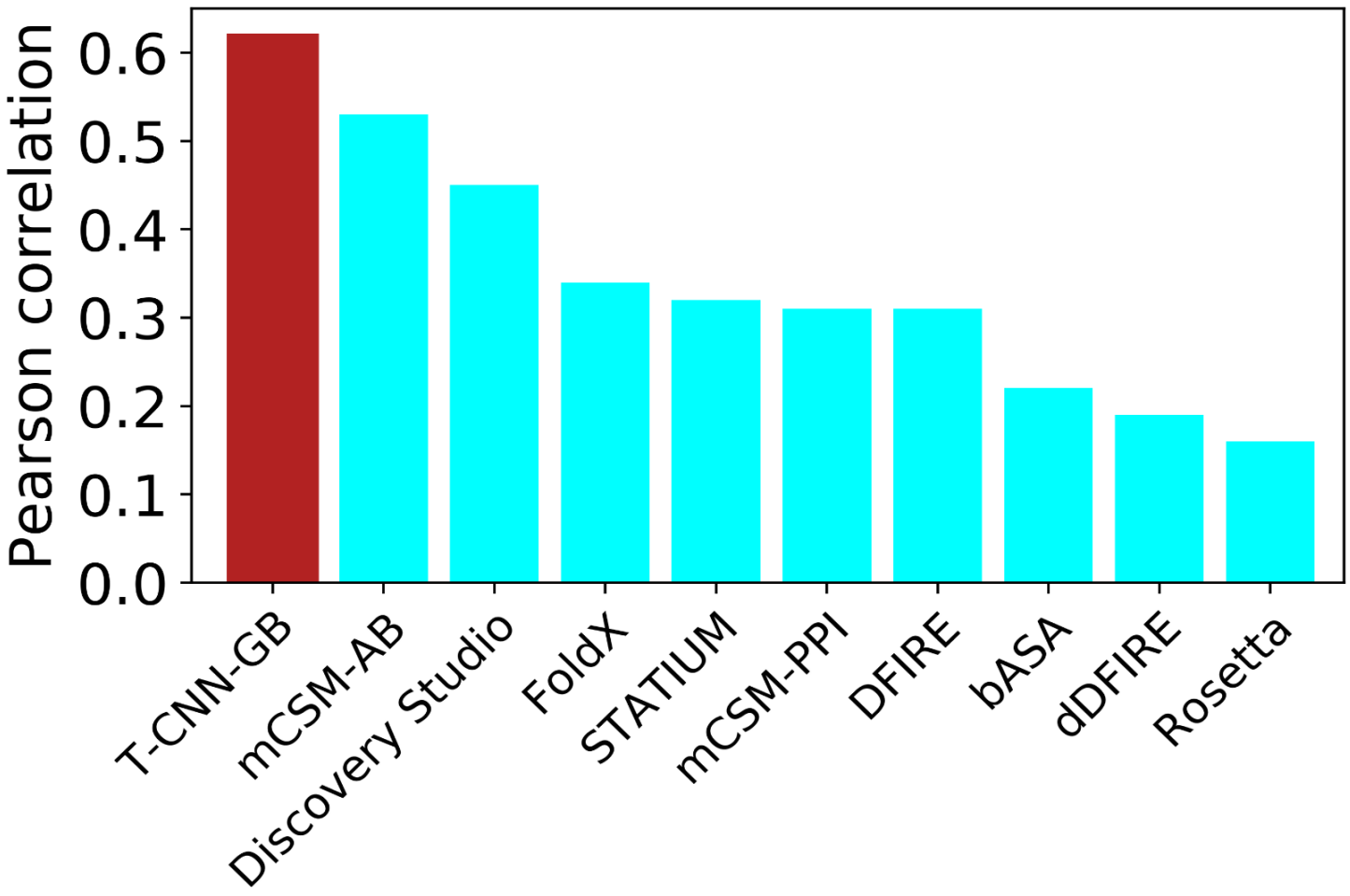
Directory of Useful Decoy (DUD)

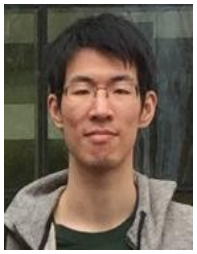
Classification of 98266 compounds containing 95316 decoys and 2950 active ligands binding to 40 targets from six families



Antibody-antigen binding free energy changes upon mutation (AB-Bind dataset)

Wang, Cang, Wei, 2019





Menglun Wang

Prediction of partition coefficients: Star Set (223 molecules)

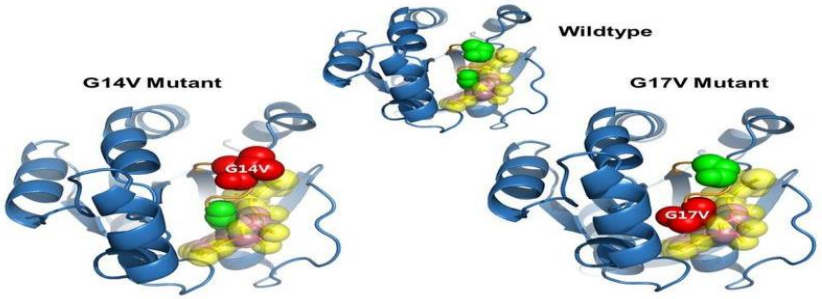
Wu, Wang, Zhao, Wang, Wei, 2016

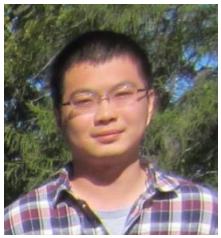
RMSE (log units)

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

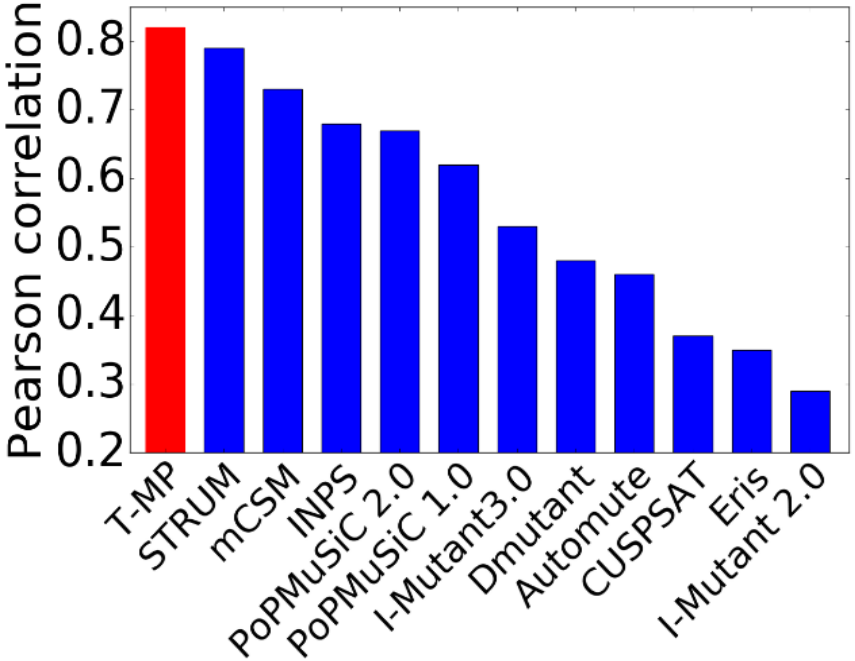
Blind prediction of mutation energies



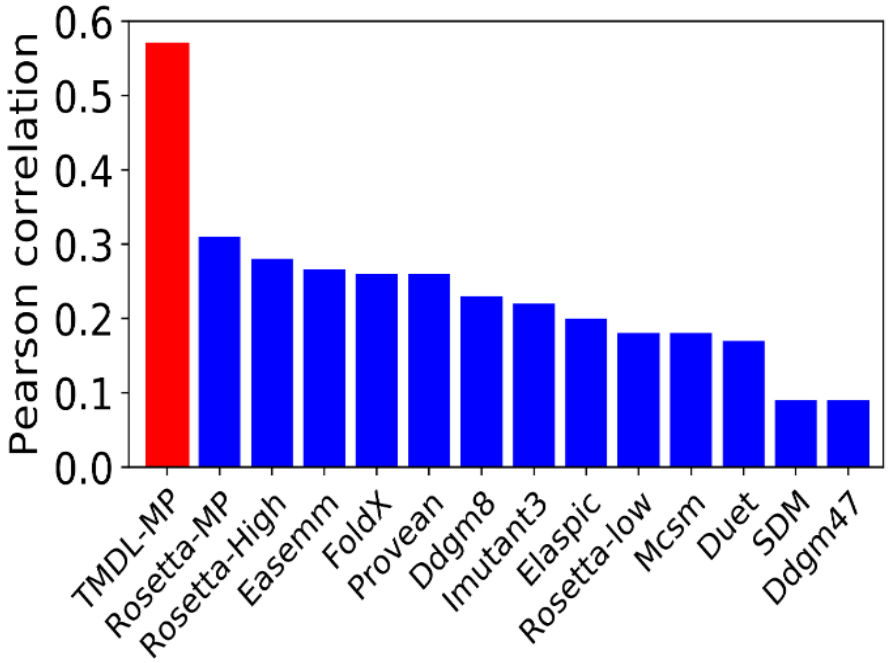


Cang, Wei, Bioinformatics 2017

Predicting mutations on 2648 globular proteins



Predicting mutations on 223 membrane proteins



Further topics and future directions

- Omics data analysis: proteomics, metabolomics, metagenomics, lipidomics, glycomics, transcriptomics, epigenomics, etc.
- Integration of molecular data and genomic data.
- Data fusion from molecule, subcellular organelle, cell, tissue, organism, to cross species.
- Using mathematics (geometry, topology, algebra, analysis, PDE, numerical analysis, etc.) to analyze the foundation of machine learning and AI.
- Creating new mathematical tools to reduce the dimensionality of massive, diverse and complex biological data.
- Developing and application of new machine learning algorithms for biomolecular data analysis and prediction.
- Gromov-Hausdorff and Gromov-Wasserstein distances for biomolecular data.
- Symplectic geometry based gradient descent.
- Predictability analysis.
- Biology inspired new mathematics to better understand the rule of life.


thank you