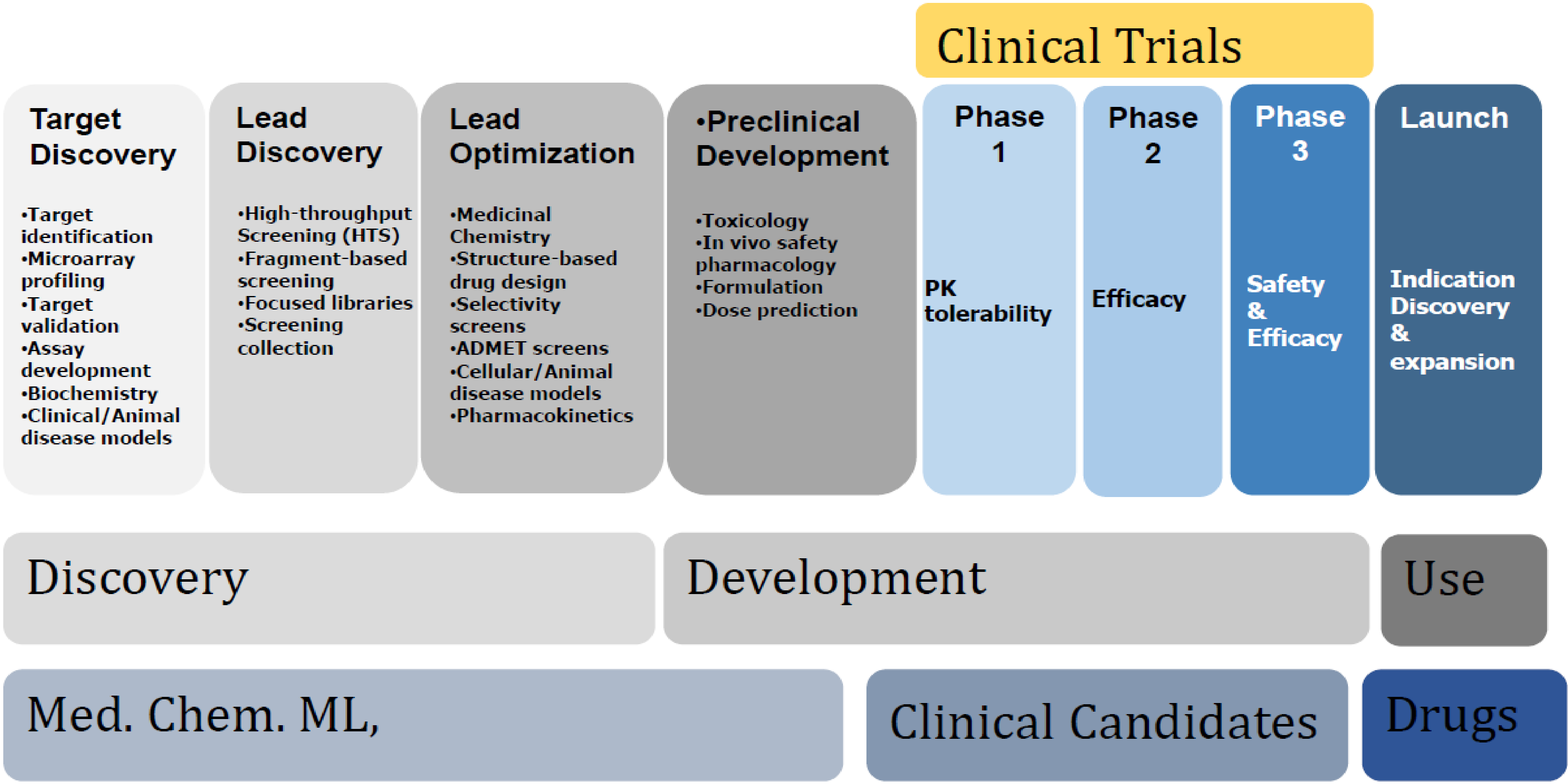
Persistent function based machine learning for drug design

Xiang Liu

Nankai University

December 8, 2023

Drug Discovery Process (Simplified)



Clinical Trials

Target Discovery

- Target identification
- Microarray profiling
- Target validation
- Assay development
- Biochemistry
- Clinical/Animal disease models

Lead Discovery

- High-throughput Screening (HTS)
- Fragment-based screening
- Focused libraries
- Screening collection

Lead Optimization

- Medicinal Chemistry
- Structure-based drug design
- Selectivity screens
- ADMET screens
- Cellular/Animal disease models
- Pharmacokinetics

•Preclinical Development

- Toxicology
- In vivo safety pharmacology
- Formulation
- Dose prediction

Phase 1

PK tolerability

Phase 2

Efficacy

Phase 3

Safety & Efficacy

Launch

Indication Discovery & expansion

Discovery

Development

Use

Med. Chem. ML,

Clinical Candidates

Drugs

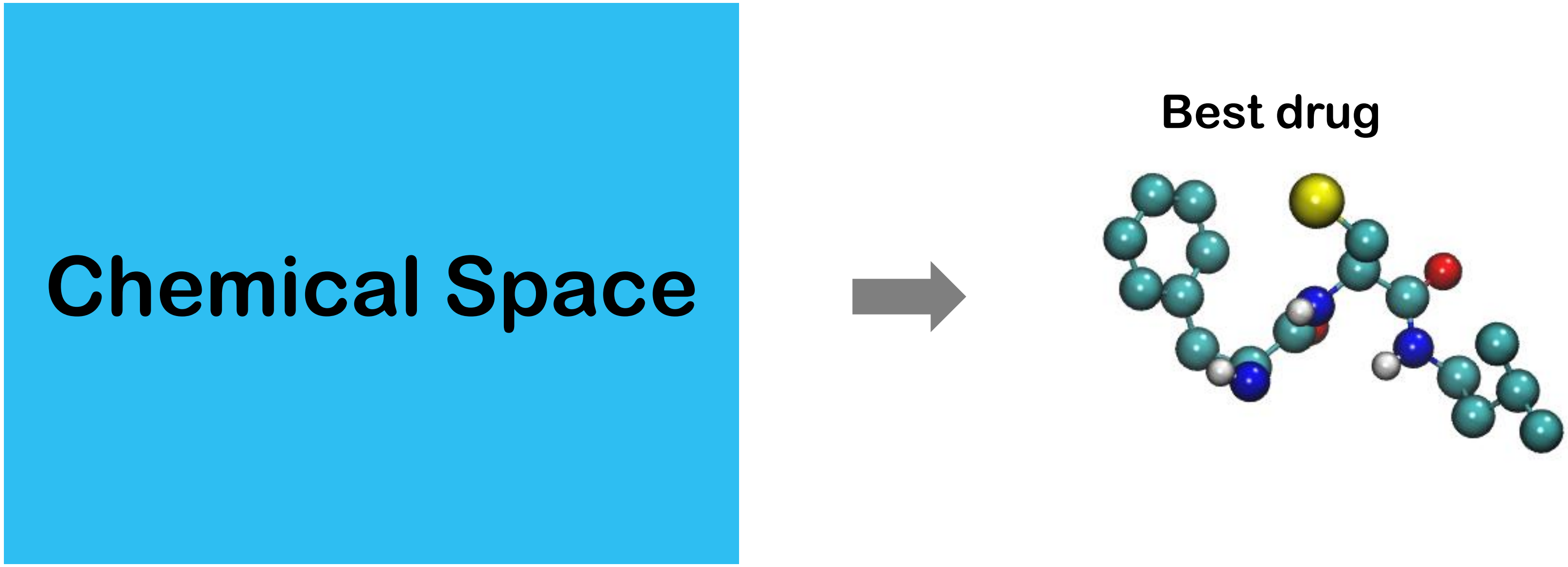
>450,000 distinct compounds
~25,000 distinct lead series

~12,000 candidates

~1,200 drugs

Time: > 10 years **Cost: > 2.6 billion$** **High failure rate**

Drug discovery is a challenging search problem



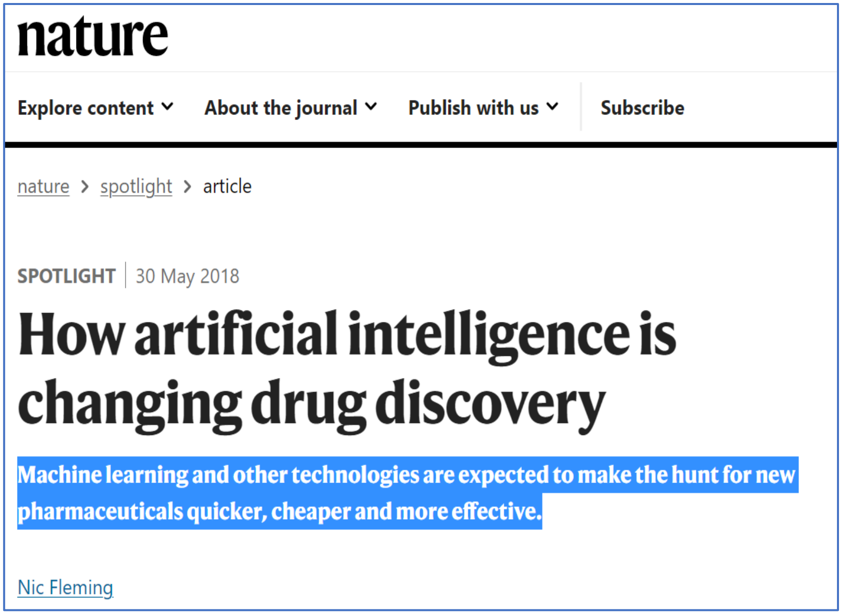
Chemical Space

Best drug

Number of possible drug-like molecules $\approx 10^{60}$ **obeying Lipinski's rule-of-five for oral bioavailability**

Kirkpatrick, P., Ellis, C. Nature (2004); Acc. Chem. Res. 2015, 48, 3, 722–730

AI in drug design and discovery

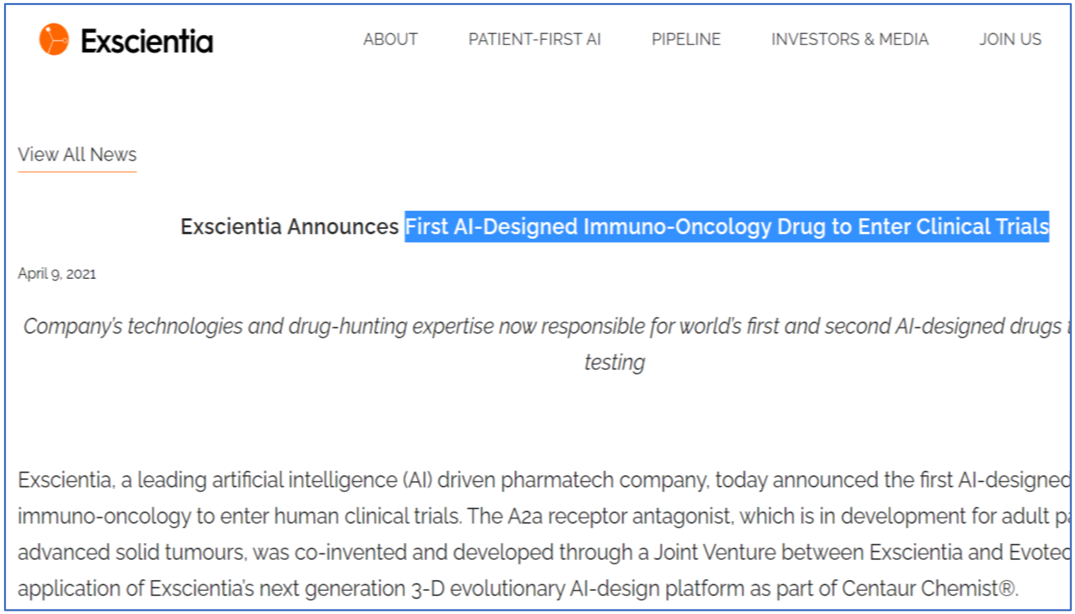


SPOTLIGHT | 30 May 2018

How artificial intelligence is changing drug discovery

Machine learning and other technologies are expected to make the hunt for new pharmaceuticals quicker, cheaper and more effective.

Nic Fleming

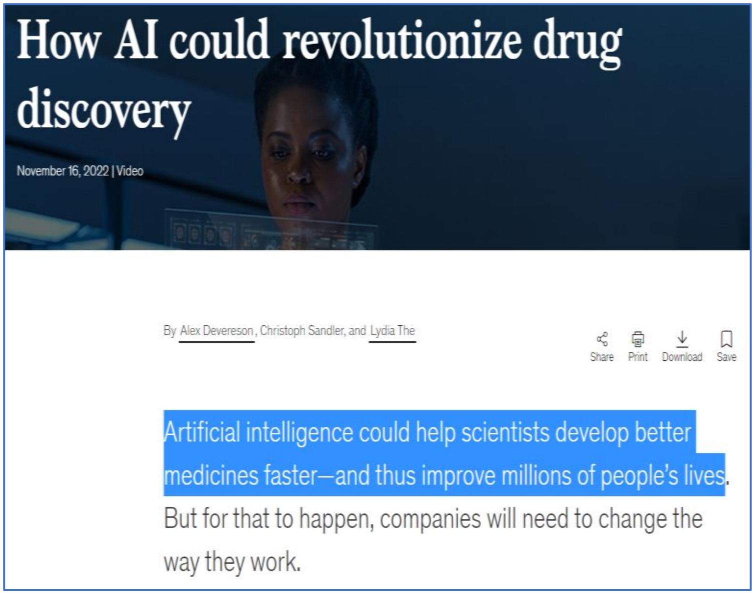


Exscientia Announces First AI-Designed Immuno-Oncology Drug to Enter Clinical Trials

April 9, 2021

Company's technologies and drug-hunting expertise now responsible for world's first and second AI-designed drugs i testing

Exscientia, a leading artificial intelligence (AI) driven pharmatech company, today announced the first AI-designed immuno-oncology to enter human clinical trials. The A2a receptor antagonist, which is in development for adult p advanced solid tumours, was co-invented and developed through a Joint Venture between Exscientia and Evotec application of Exscientia's next generation 3-D evolutionary AI-design platform as part of Centaur Chemist®.

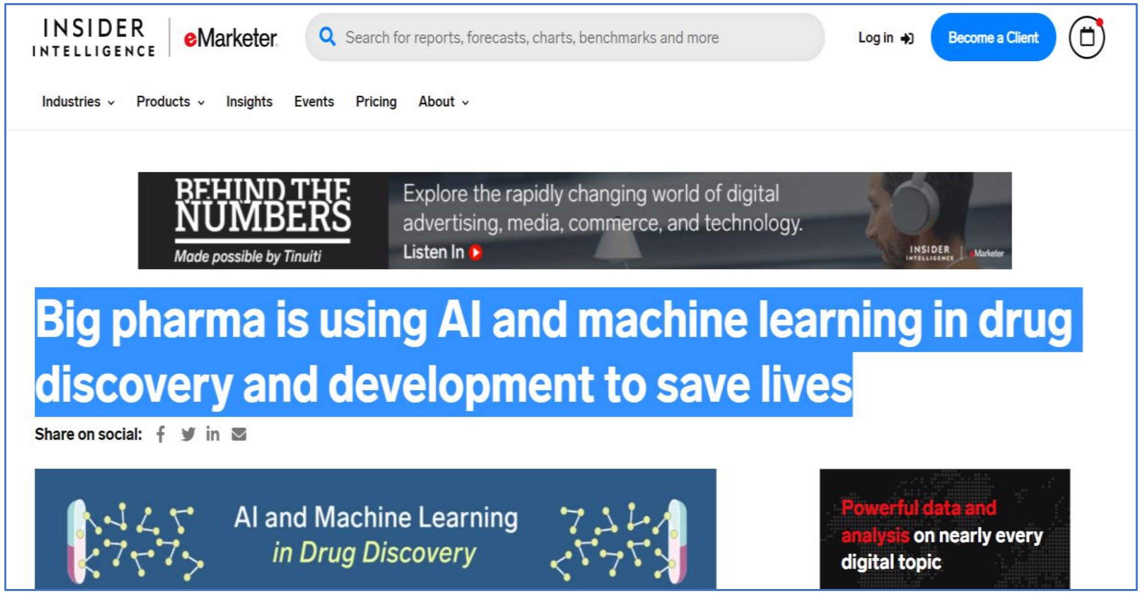


How AI could revolutionize drug discovery

November 16, 2022 | Video

By Alex Devereson, Christoph Sandler, and Lydia The

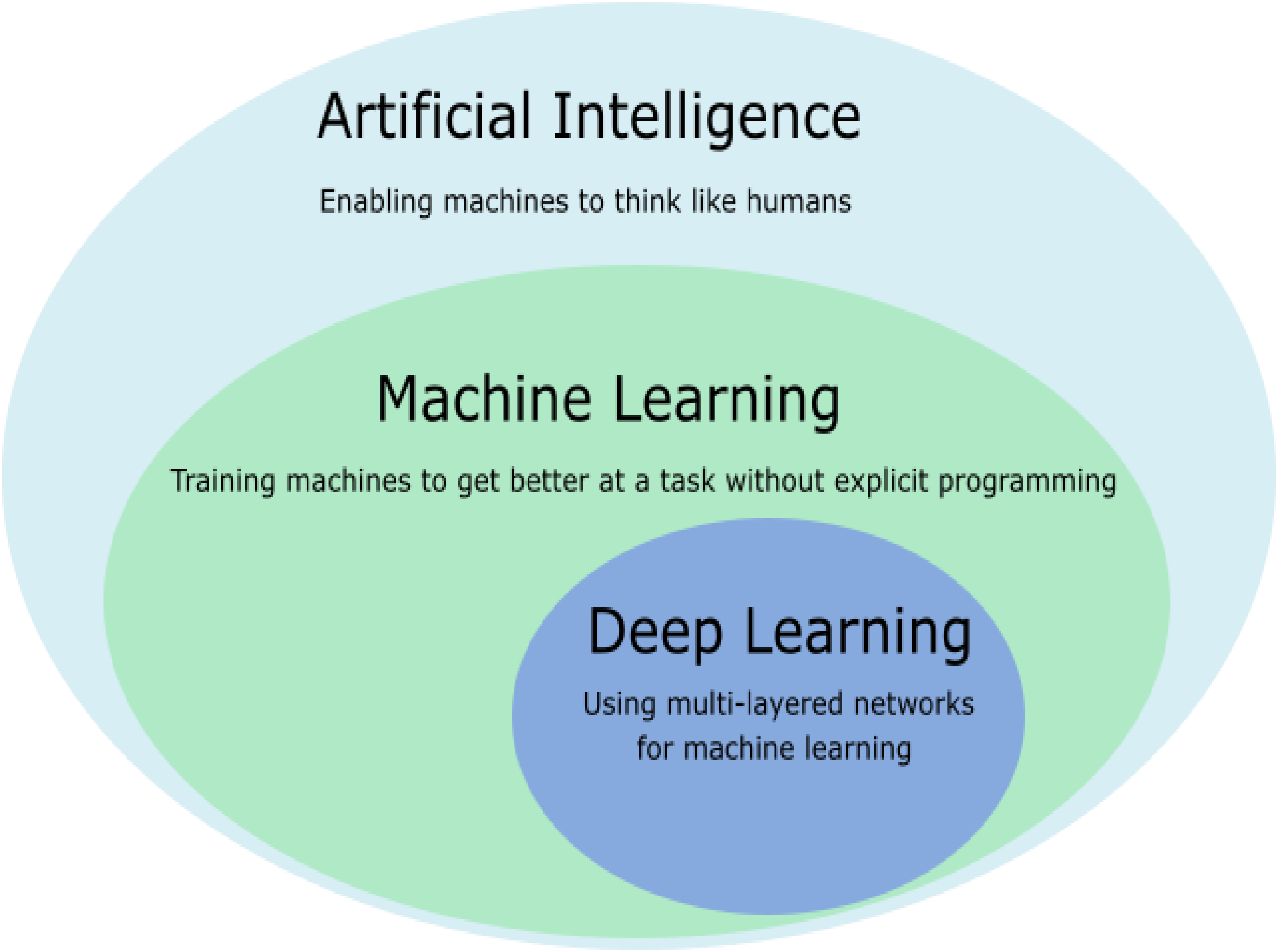
Artificial intelligence could help scientists develop better medicines faster—and thus improve millions of people's lives. But for that to happen, companies will need to change the way they work.


INSIDER INTELLIGENCE
eMarketer
Search for reports, forecasts, charts, benchmarks and more
Log in
Become a Client
Industries ⌄
Products ⌄
Insights
Events
Pricing
About ⌄
BEHIND THE NUMBERS
Made possible by Tinuiti
Explore the rapidly changing world of digital advertising, media, commerce, and technology.
Listen In

Big pharma is using AI and machine learning in drug discovery and development to save lives

Share on social:

AI and Machine Learning in Drug Discovery
Powerful data and analysis on nearly every digital topic


Artificial Intelligence
Enabling machines to think like humans
Machine Learning
Training machines to get better at a task without explicit programming
Deep Learning
Using multi-layered networks for machine learning

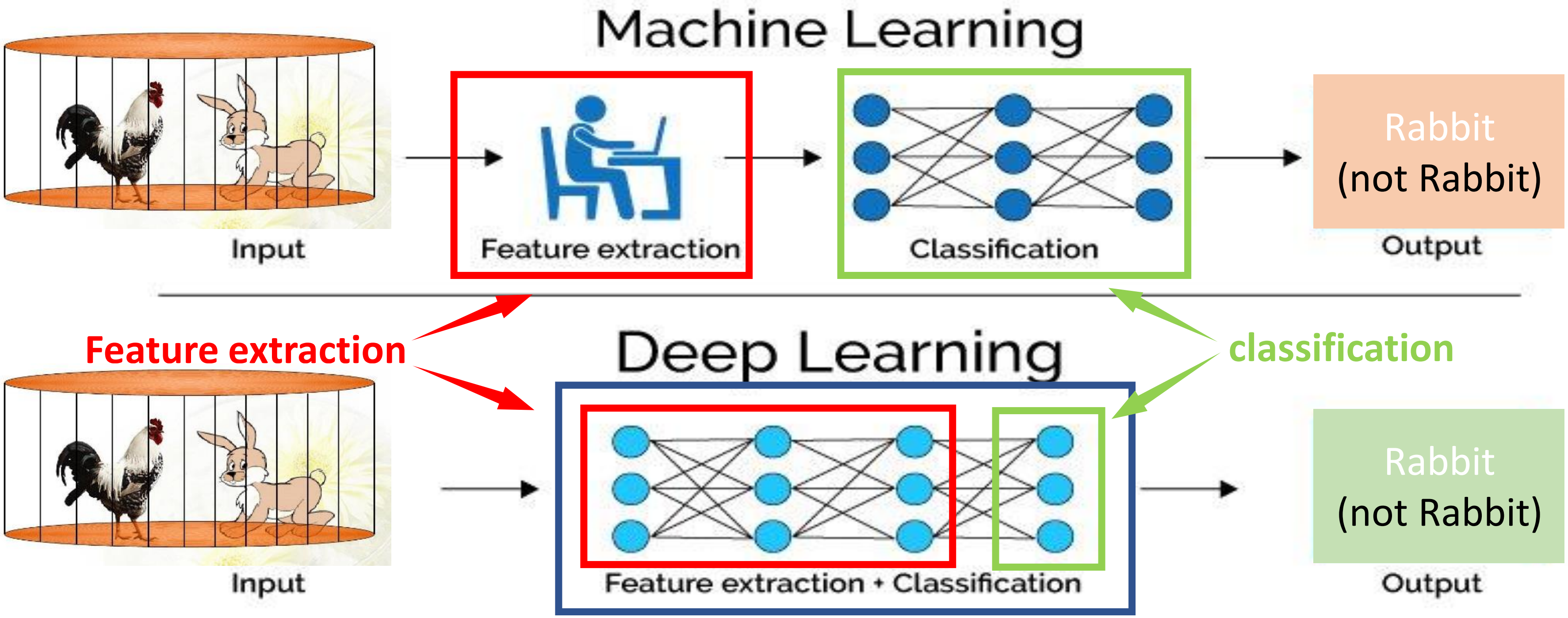
Feature extraction and feature learning

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

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

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

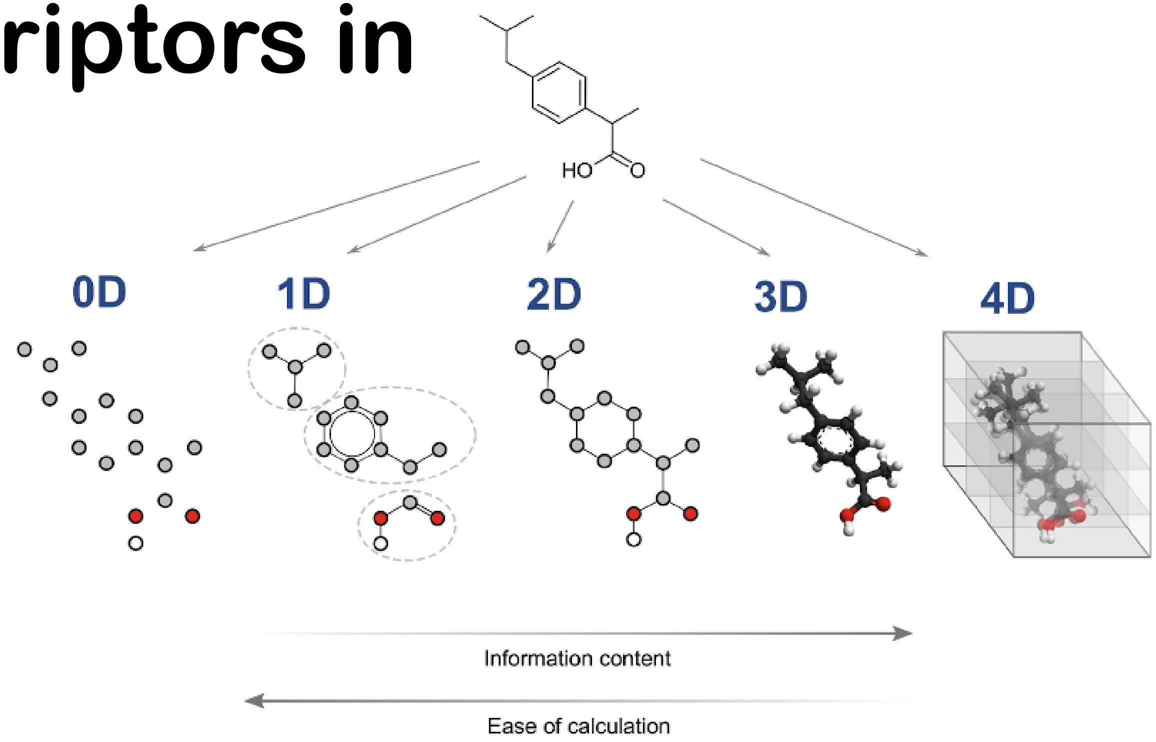
Y. Bengio, "Deep Learning of Representations: Looking Forward

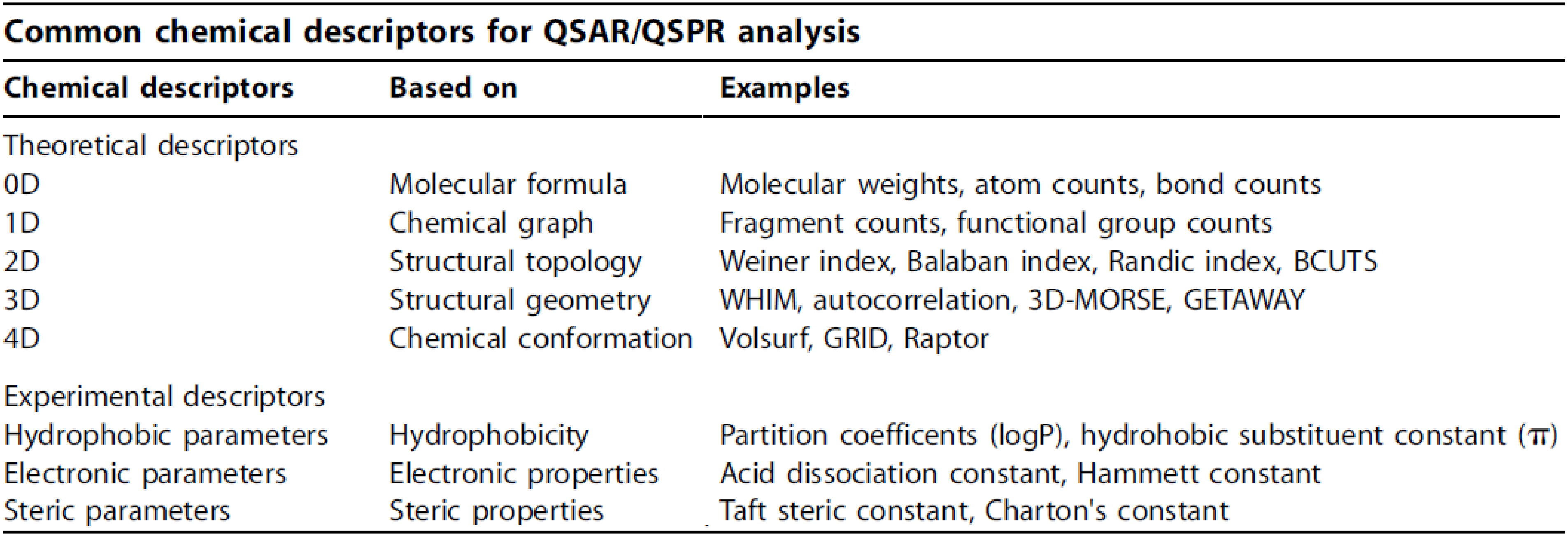


Molecular Descriptors in QSAR models

More than 5000 Molecular descriptors in Quantitative Structure Activity relationship (QSAR) models.

Grisoni F, Ballabio D, Todeschini R, et al. Molecular descriptors for structure–activity applications: a hands-on approach[M]// Computational Toxicology. Humana Press, New York, NY, 2018: 3-53.



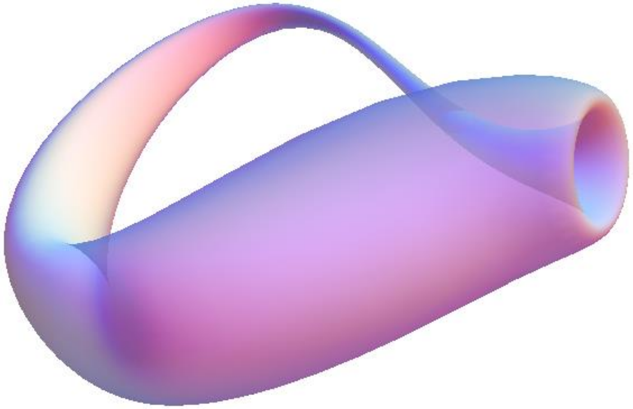


Common chemical descriptors for QSAR/QSPR analysis

Chemical descriptors	Based on	Examples
Theoretical descriptors		
0D	Molecular formula	Molecular weights, atom counts, bond counts
1D	Chemical graph	Fragment counts, functional group counts
2D	Structural topology	Weiner index, Balaban index, Randic index, BCUTS
3D	Structural geometry	WHIM, autocorrelation, 3D-MORSE, GETAWAY
4D	Chemical conformation	Volsurf, GRID, Raptor
Experimental descriptors		
Hydrophobic parameters	Hydrophobicity	Partition coefficents (logP), hydrohobic substituent constant (π)
Electronic parameters	Electronic properties	Acid dissociation constant, Hammett constant
Steric parameters	Steric properties	Taft steric constant, Charton's constant

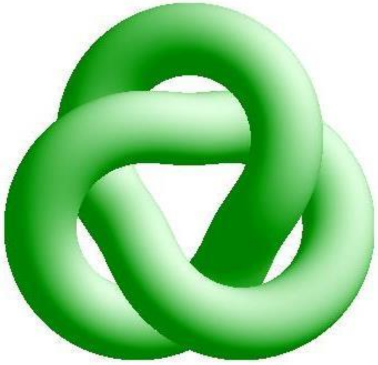
Topological Data Analysis (TDA)

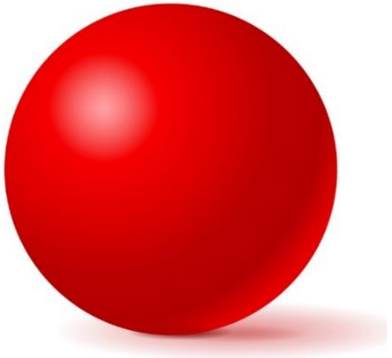
Topological invariant:
Homology Group
Homotopy Group
Cohomology Ring
Steenrod Module
……


Klein bottle

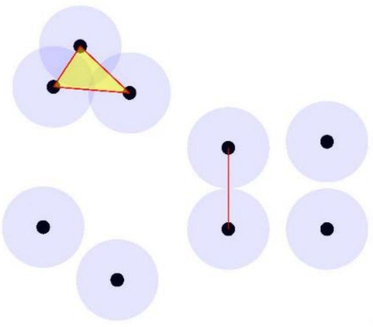
Torus

Double Torus


Knot

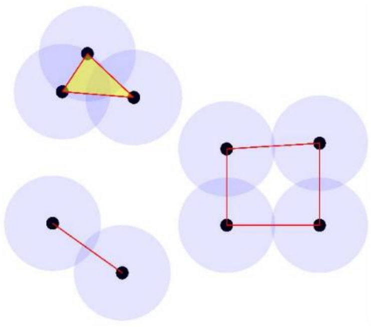

Sphere

Topological Data Analysis---- Persistent Homology



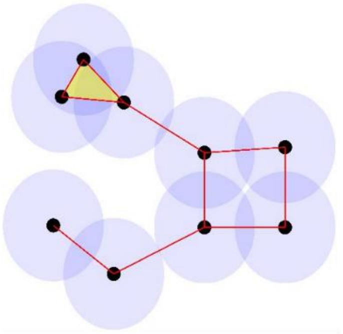
$f_1=0.4$

$\beta_0: 6 \quad \beta_1: 0$



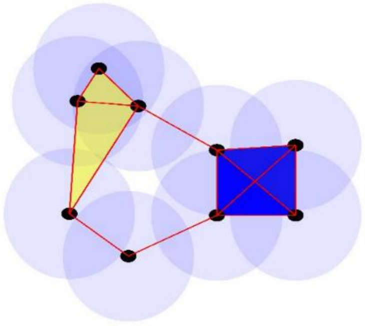
$f_2=0.5$

$\beta_0: 3 \quad \beta_1: 1$



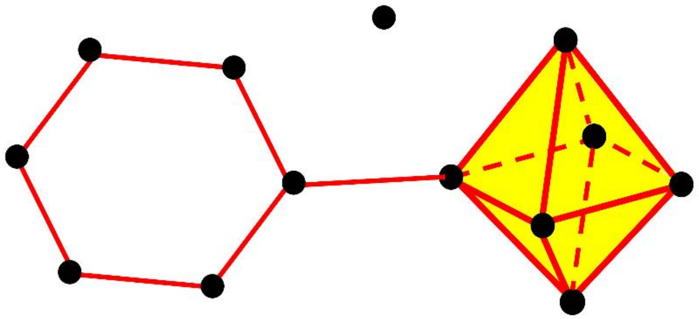
$f_3=0.6$

$\beta_0: 1 \quad \beta_1: 1$

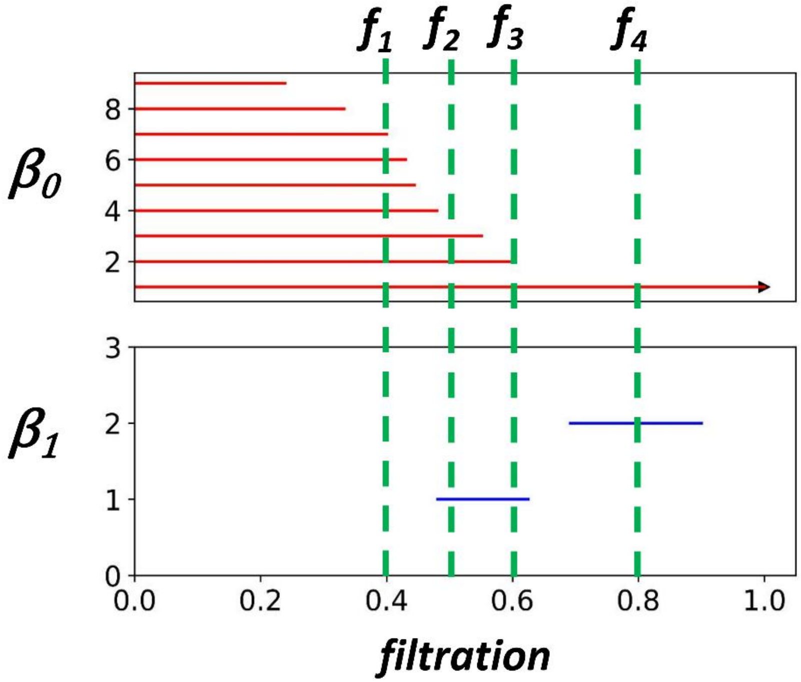


$f_4=0.8$

$\beta_0: 1 \quad \beta_1: 1$

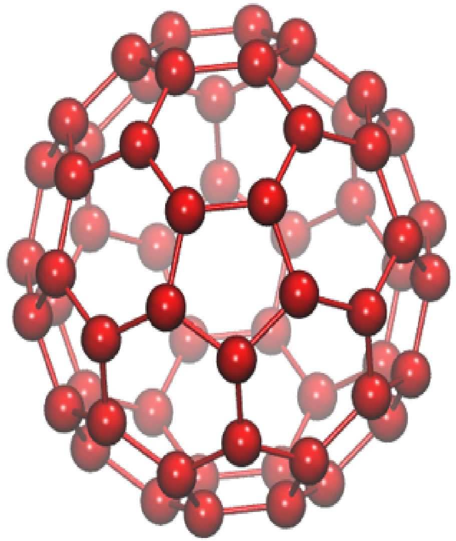


$\beta_0 = 2 \quad \beta_1 = 1 \quad \beta_2 = 1$

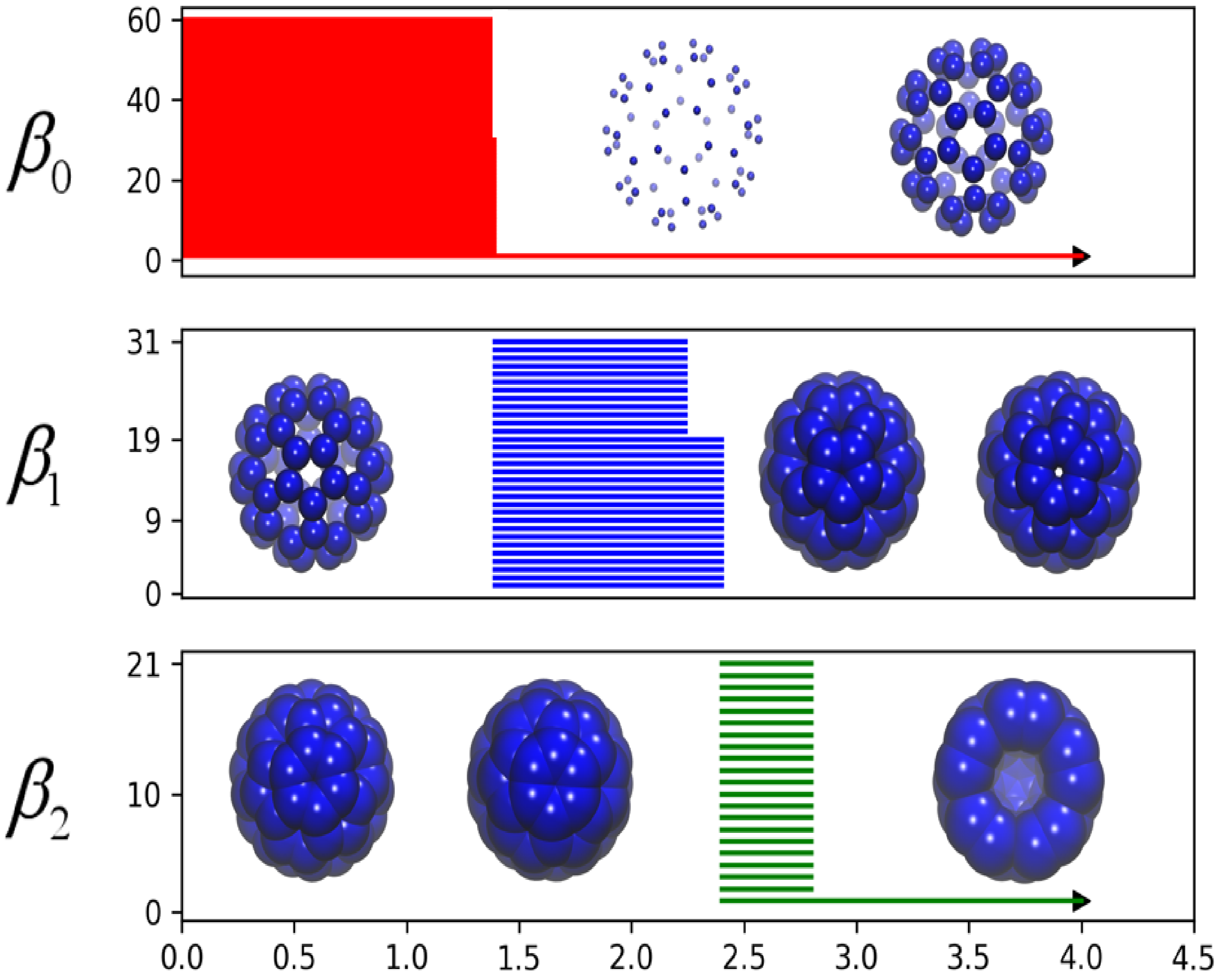


Persistent Homology Analysis of Carbon-60

(Xia, Feng, Tong & Wei, JCC, 2015)



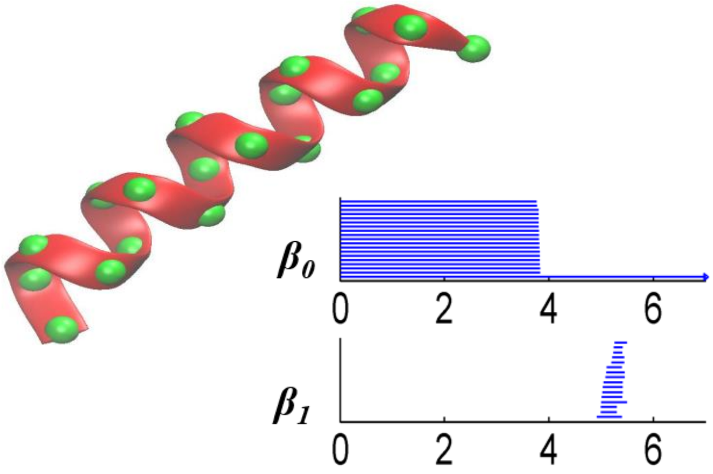
C_{60}



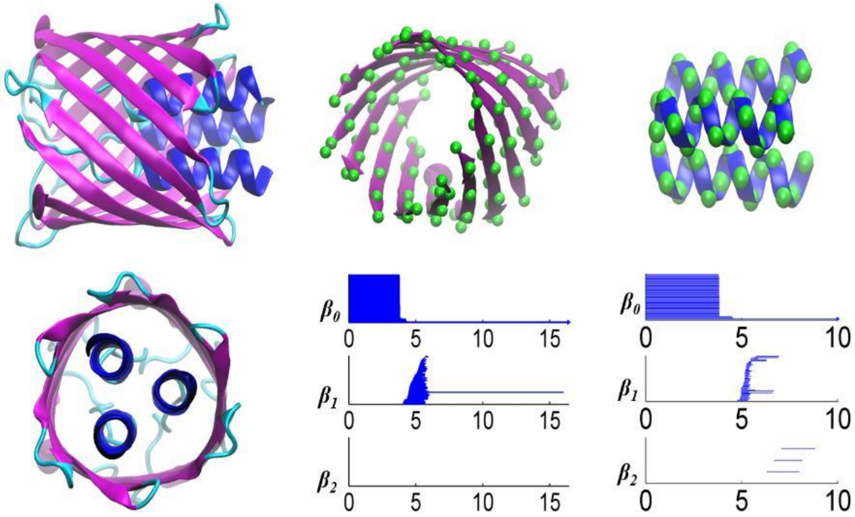
Biomolecular Topological Fingerprints

(Xia & Wei, IJNMBE, 2014)

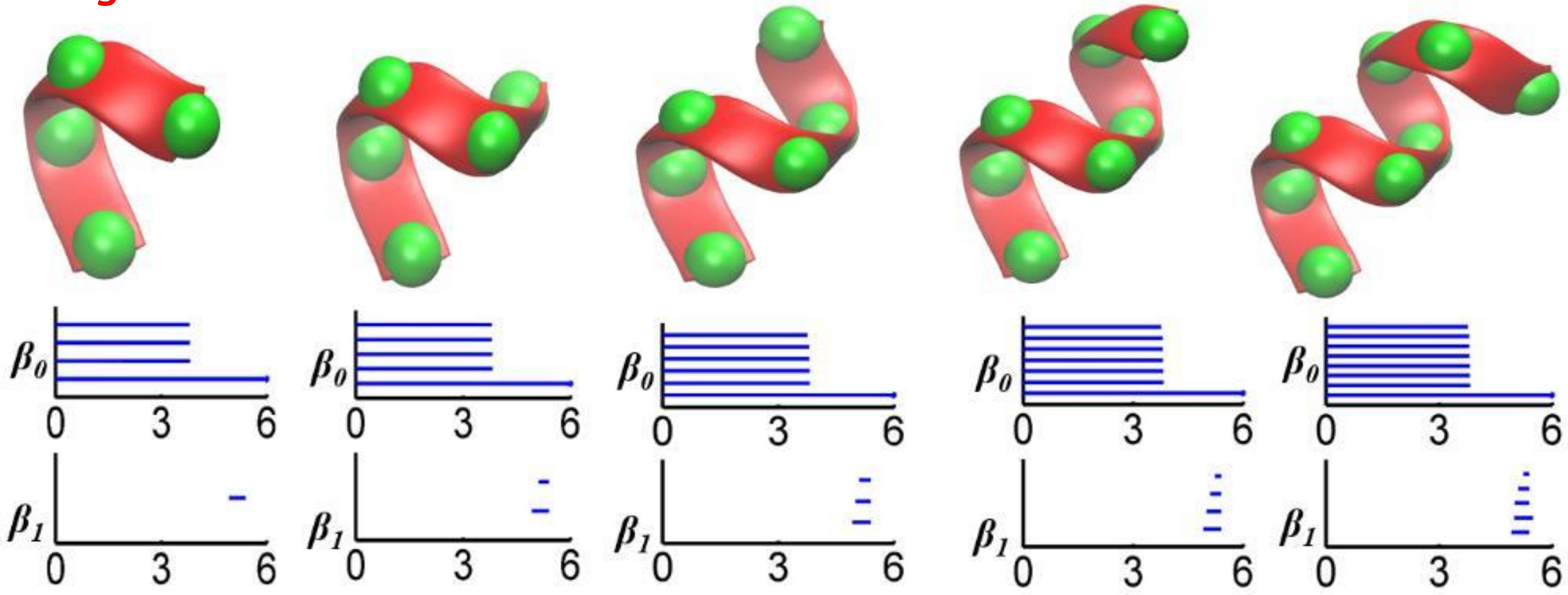
TF for alpha helix



TF for beta barrel

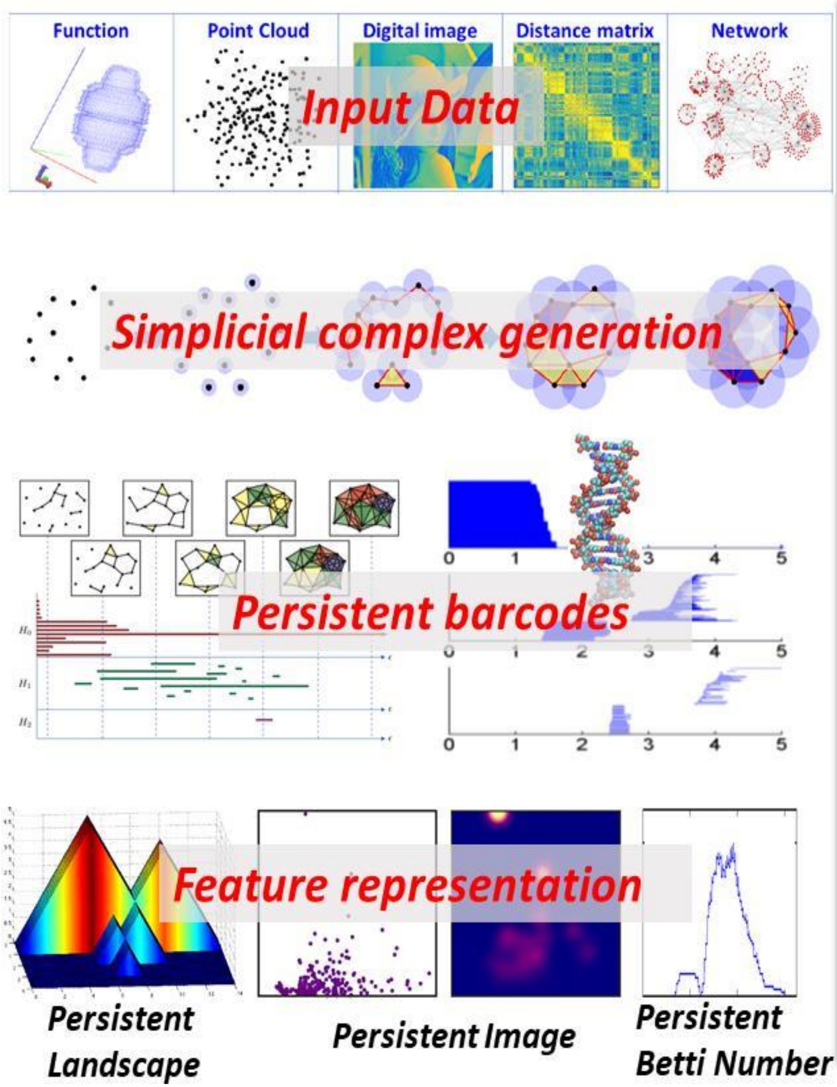


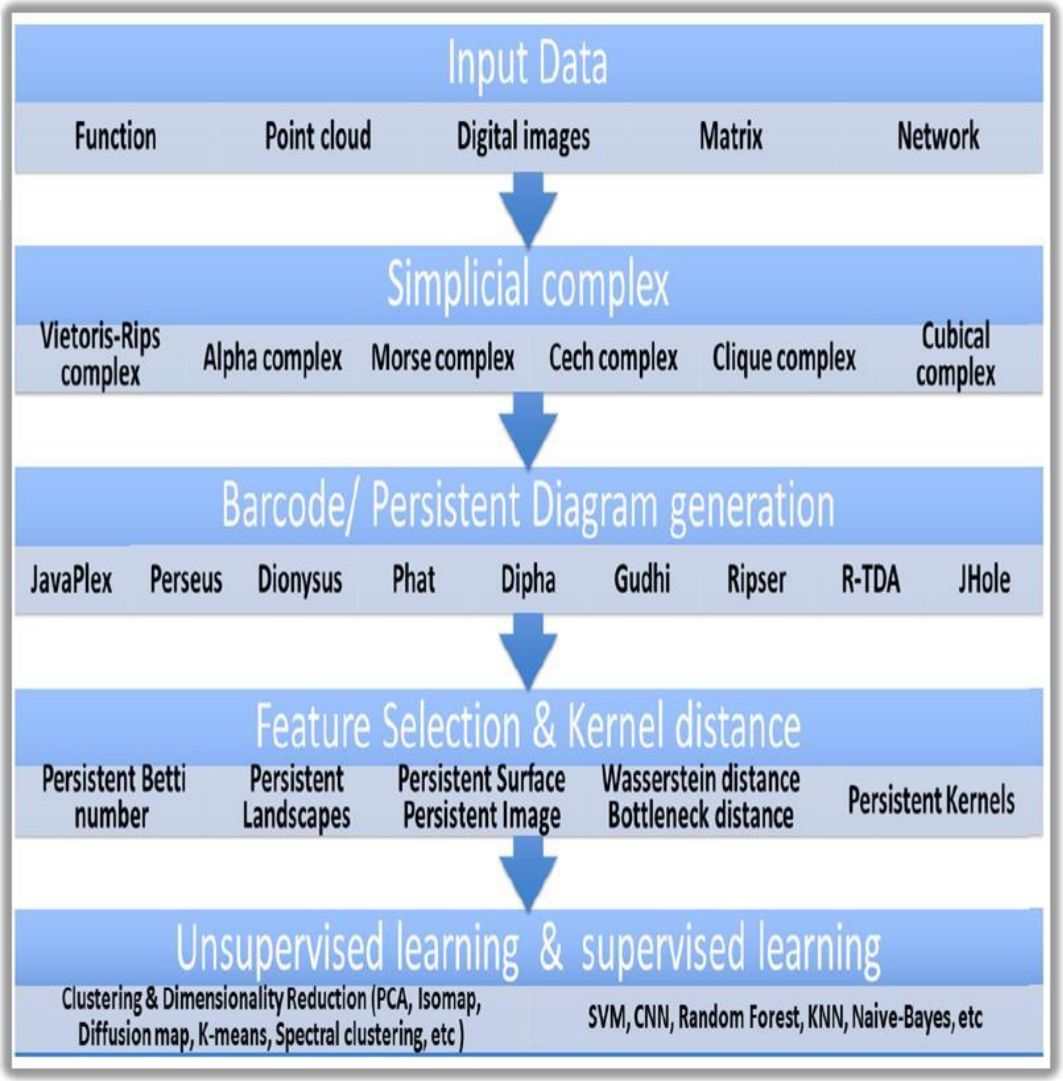
Slicing method



TDA based machine learning models

(Pun, Lee and Xia, AIR, 2021)





Input Data

Function | Point cloud | Digital images | Matrix | Network

Simplicial complex

Vietoris-Rips complex | Alpha complex | Morse complex | Cech complex | Clique complex | Cubical complex

Barcode/ Persistent Diagram generation

JavaPlex | Perseus | Dionysus | Phat | Dipha | Gudhi | Ripser | R-TDA | JHole

Feature Selection & Kernel distance

Persistent Betti number | Persistent Landscapes | Persistent Surface Persistent Image | Wasserstein distance Bottleneck distance | Persistent Kernels

Unsupervised learning & supervised learning

Clustering & Dimensionality Reduction (PCA, Isomap, Diffusion map, K-means, Spectral clustering, etc)

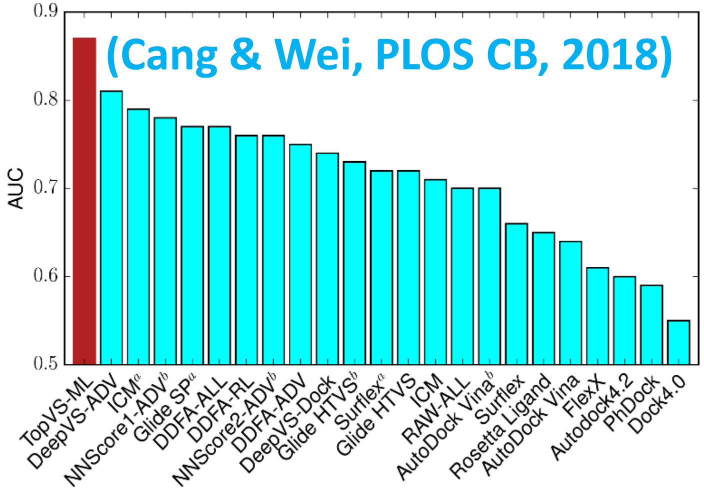
SVM, CNN, Random Forest, KNN, Naive-Bayes, etc

Recent progress of TDA based drug design

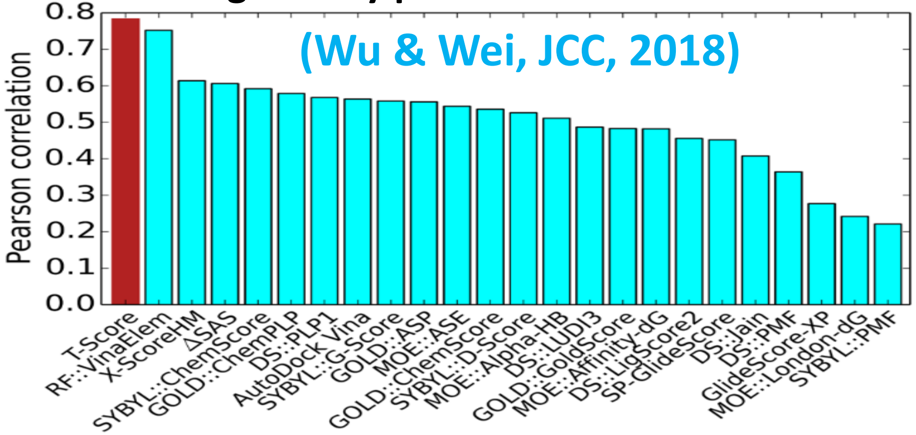
Guowei Wei

MSU Foundation professor

DUD database 128374 protein-ligand/decoy pairs

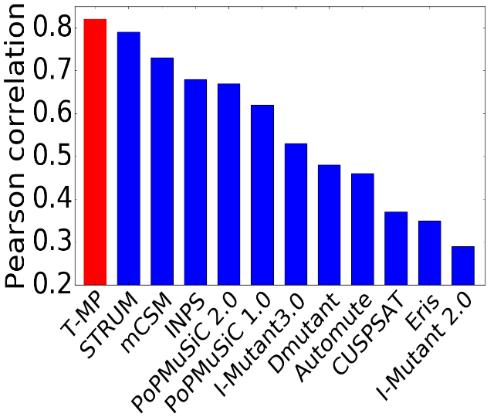

(Cang & Wei, PLOS CB, 2018)

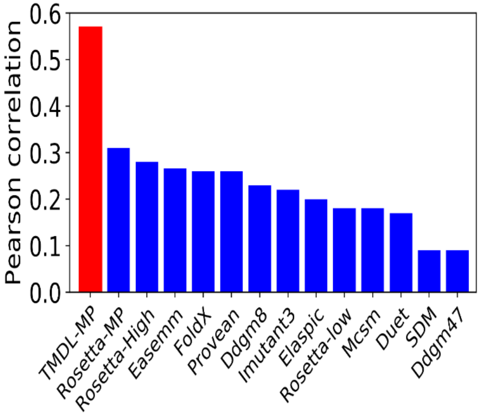
Binding affinity prediction of PDBbind v2013


(Wu & Wei, JCC, 2018)

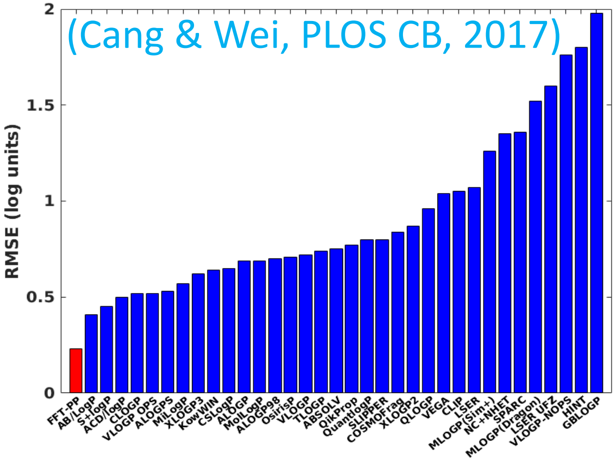
Prediction correlations for 2648 mutations on globular proteins

(Cang & Wei, PLOS CS, 2017)





Prediction RMSD of logP(star set)


(Cang & Wei, PLOS CB, 2017)

Recent progress of TDA based drug design

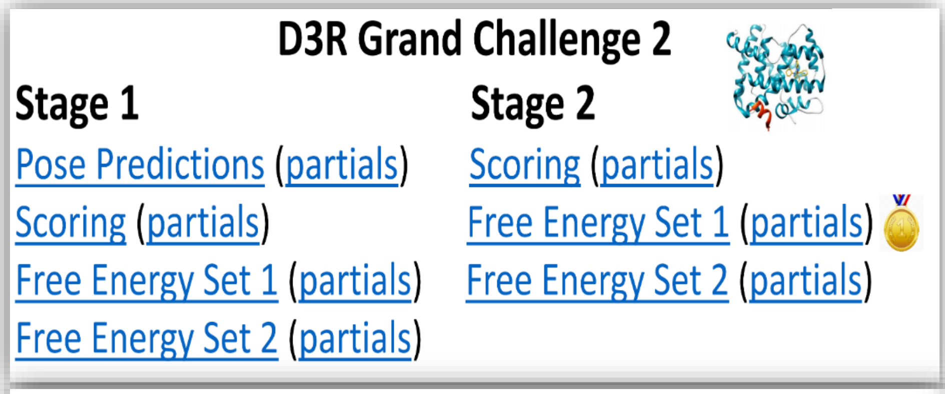
Drug Design Data Resource (D3R) Grand Challenges

Grand Challenge 2： win 14%

Grand Challenge 3: win 38% while the second winner had a rate of 19%

Grand Challenge 4: win 50%

Wei Team's performance at D3R Grand Challenge



D3R Grand Challenge 2

Stage 1

Pose Predictions (partials)

Scoring (partials)

Free Energy Set 1 (partials)

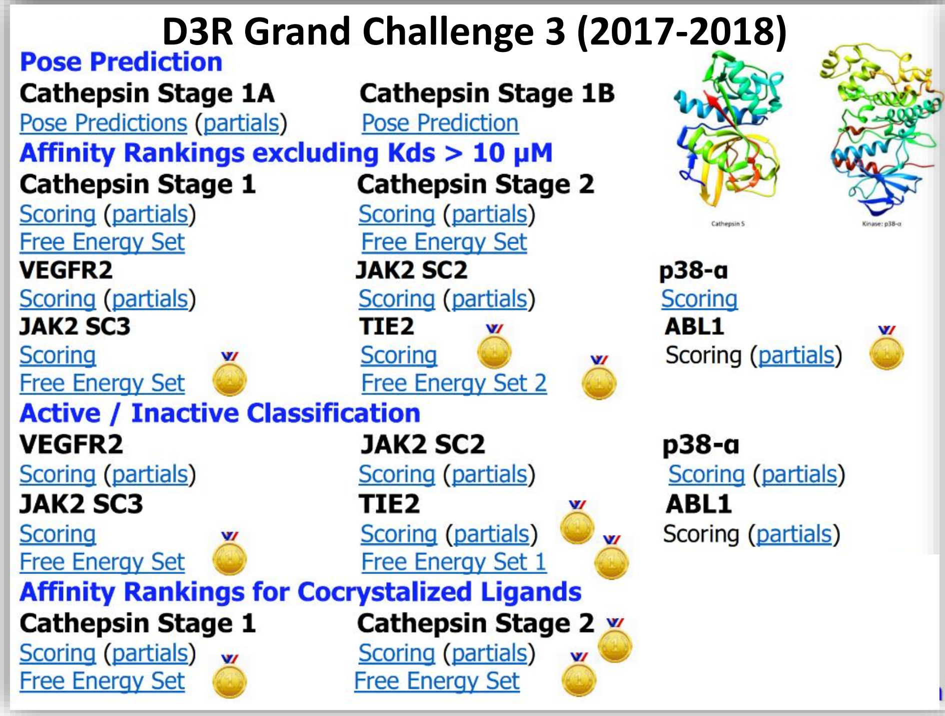
Free Energy Set 2 (partials)

Stage 2

Scoring (partials)

Free Energy Set 1 (partials) 🥇

Free Energy Set 2 (partials)



D3R Grand Challenge 3 (2017-2018)

Pose Prediction

Cathepsin Stage 1A
Pose Predictions (partials)

Cathepsin Stage 1B
Pose Prediction

Affinity Rankings excluding Kds > 10 μM

Cathepsin Stage 1
Scoring (partials)
Free Energy Set

Cathepsin Stage 2
Scoring (partials)
Free Energy Set

VEGFR2
Scoring (partials)

JAK2 SC2
Scoring (partials)

p38-α
Scoring

JAK2 SC3
Scoring
Free Energy Set 🥇

TIE2
Scoring 🥇
Free Energy Set 2 🥇

ABL1
Scoring (partials) 🥇

Active / Inactive Classification

VEGFR2
Scoring (partials)

JAK2 SC2
Scoring (partials)

p38-α
Scoring (partials)

JAK2 SC3
Scoring
Free Energy Set 🥇

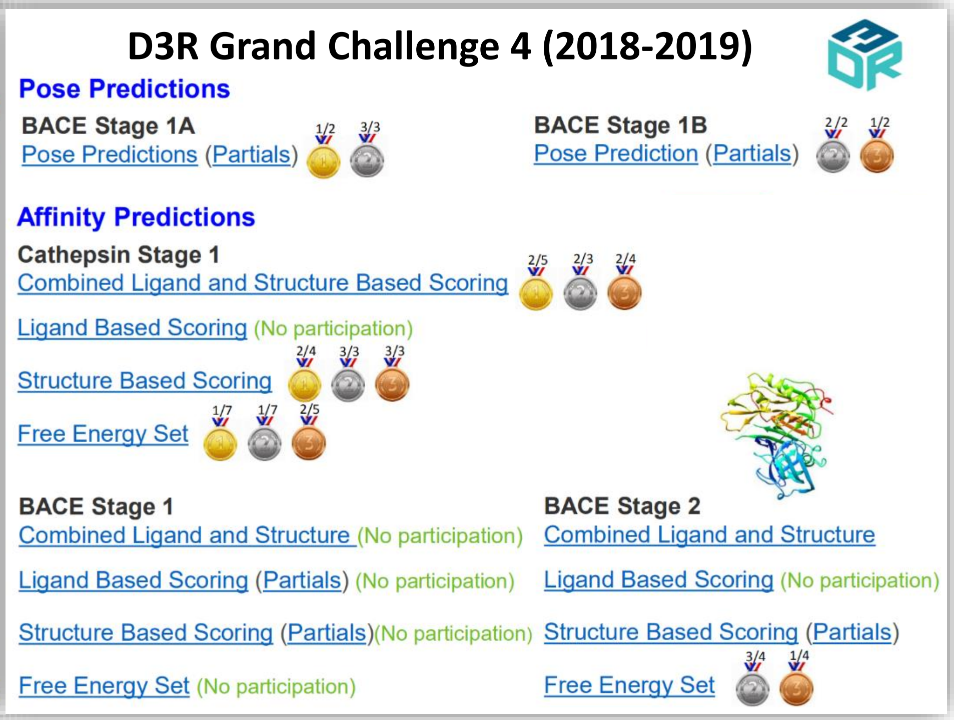
TIE2
Scoring (partials) 🥇
Free Energy Set 1 🥇

ABL1
Scoring (partials)

Affinity Rankings for Cocrystalized Ligands

Cathepsin Stage 1
Scoring (partials)
Free Energy Set 🥇

Cathepsin Stage 2 🥇
Scoring (partials) 🥇
Free Energy Set



D3R Grand Challenge 4 (2018-2019)

Pose Predictions

BACE Stage 1A
Pose Predictions (Partials) 1/2 🥇 3/3 🥈

BACE Stage 1B
Pose Prediction (Partials) 2/2 🥈 1/2 🥉

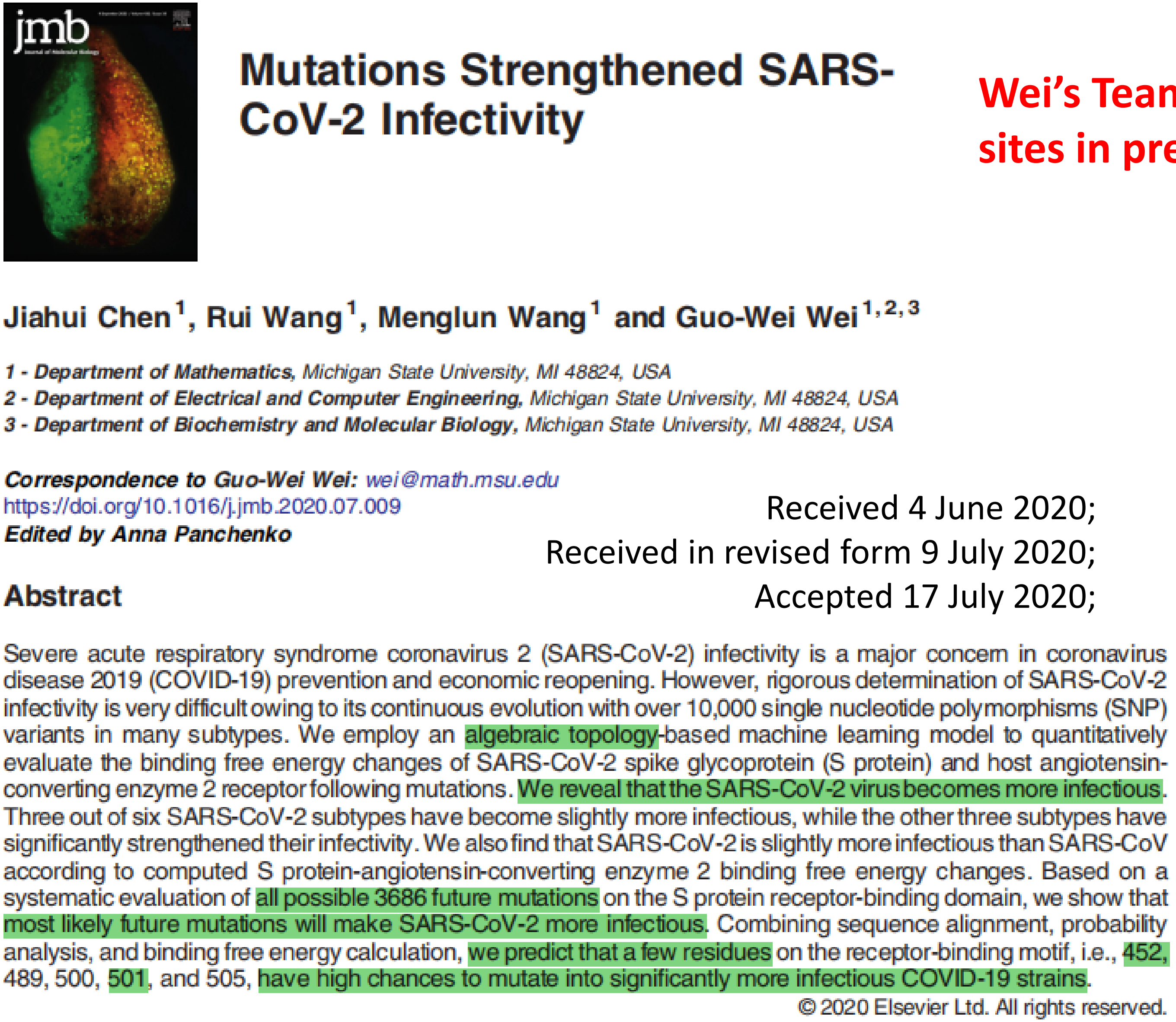
Affinity Predictions

Cathepsin Stage 1
Combined Ligand and Structure Based Scoring 2/5 🥇 2/3 🥈 2/4 🥉
Ligand Based Scoring (No participation)
Structure Based Scoring 2/4 🥇 3/3 🥈 3/3 🥉
Free Energy Set 1/7 🥇 1/7 🥈 2/5 🥉

BACE Stage 1
Combined Ligand and Structure (No participation)
Ligand Based Scoring (Partials) (No participation)
Structure Based Scoring (Partials)(No participation)
Free Energy Set (No participation)

BACE Stage 2
Combined Ligand and Structure
Ligand Based Scoring (No participation)
Structure Based Scoring (Partials)
Free Energy Set 3/4 🥈 1/4 🥉

TDA-based learning models in SARS-Cov-2



Mutations Strengthened SARS-CoV-2 Infectivity

Jiahui Chen[1], Rui Wang[1], Menglun Wang[1] and Guo-Wei Wei[1,2,3]

1 - Department of Mathematics, Michigan State University, MI 48824, USA
2 - Department of Electrical and Computer Engineering, Michigan State University, MI 48824, USA
3 - Department of Biochemistry and Molecular Biology, Michigan State University, MI 48824, USA

Correspondence to Guo-Wei Wei: *wei@math.msu.edu*

https://doi.org/10.1016/j.jmb.2020.07.009
Edited by Anna Panchenko

Received 4 June 2020;
Received in revised form 9 July 2020;
Accepted 17 July 2020;

Abstract

Severe acute respiratory syndrome coronavirus 2 (SARS-CoV-2) infectivity is a major concern in coronavirus disease 2019 (COVID-19) prevention and economic reopening. However, rigorous determination of SARS-CoV-2 infectivity is very difficult owing to its continuous evolution with over 10,000 single nucleotide polymorphisms (SNP) variants in many subtypes. We employ an algebraic topology-based machine learning model to quantitatively evaluate the binding free energy changes of SARS-CoV-2 spike glycoprotein (S protein) and host angiotensin-converting enzyme 2 receptor following mutations. We reveal that the SARS-CoV-2 virus becomes more infectious. Three out of six SARS-CoV-2 subtypes have become slightly more infectious, while the other three subtypes have significantly strengthened their infectivity. We also find that SARS-CoV-2 is slightly more infectious than SARS-CoV according to computed S protein-angiotensin-converting enzyme 2 binding free energy changes. Based on a systematic evaluation of all possible 3686 future mutations on the S protein receptor-binding domain, we show that most likely future mutations will make SARS-CoV-2 more infectious. Combining sequence alignment, probability analysis, and binding free energy calculation, we predict that a few residues on the receptor-binding motif, i.e., 452, 489, 500, 501, and 505, have high chances to mutate into significantly more infectious COVID-19 strains.

Wei's Team predicts key mutation sites in prevailing variants

Mutations at 501 and 452 in prevailing SARS-Cov-2 variants

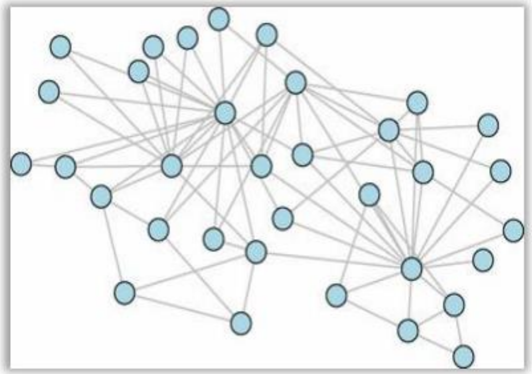
Alpha: N501Y
Beta: K417N, E484K, N501Y
Gamma: K417T, E484K, N501Y
Delta: L452R, T478K
Epsilon: L452R
Kappa: L452R, E484Q
Omicron: N501,…

They discovered the mechanism of viral transmission and evolution: more infectious

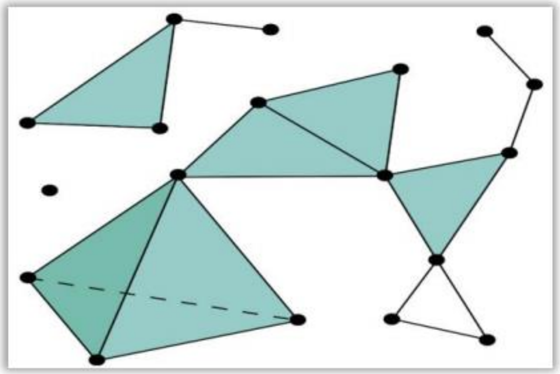
Why is TDA so powerful ?

Representation

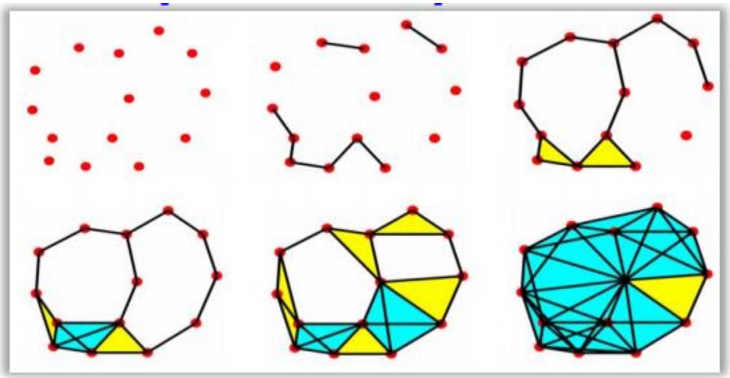
Graph



Simplicial complex



Filtered simplicial complex

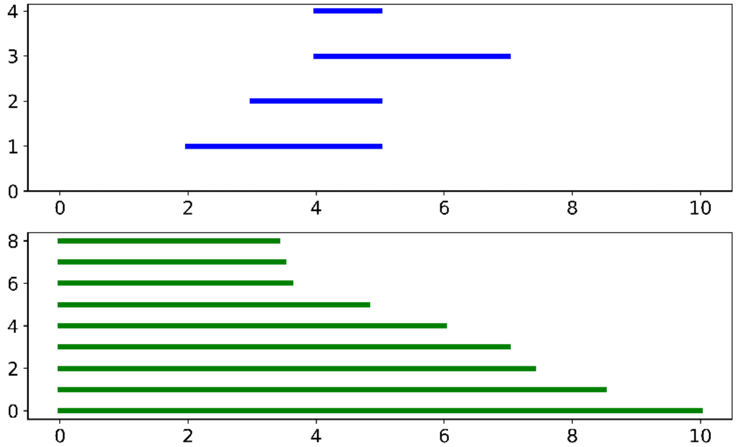


Featurization

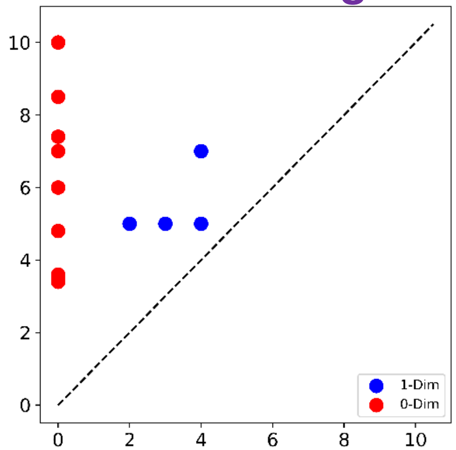
Topological invariants

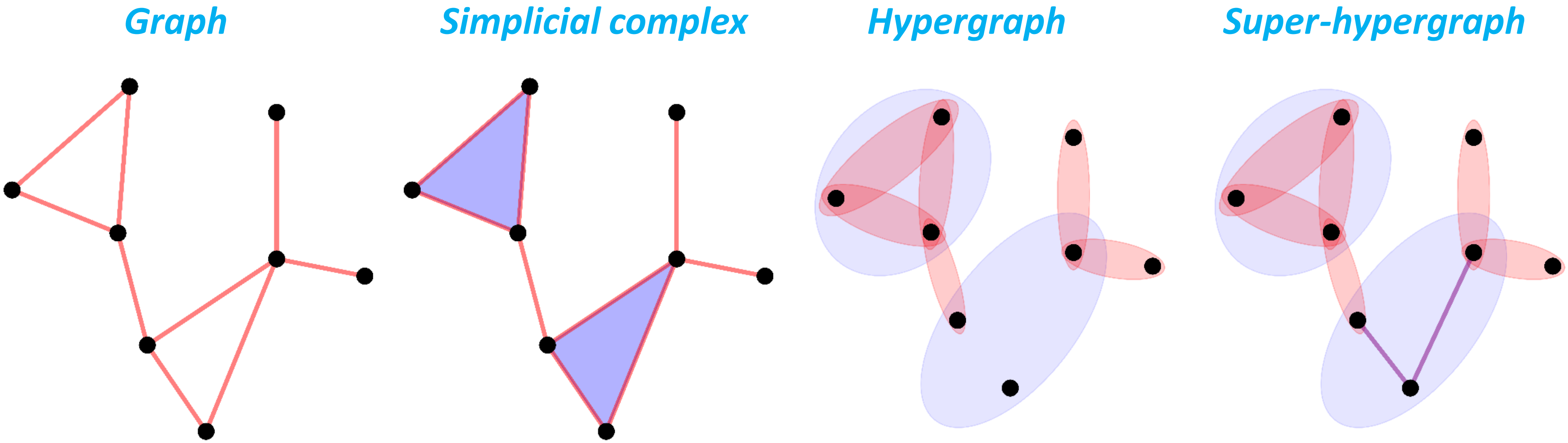
Homology Group
Homotopy Group
Cohomology Ring
Steenrod Module
……

Persistent barcode



Persistent diagram





Jie Wu,
BIMSA

Hypergraph based data representation

Grbic J, Wu J, Xia K, Wei GW. Aspects of topological approaches for data science[J]. Foundations of Data Science, 2022.

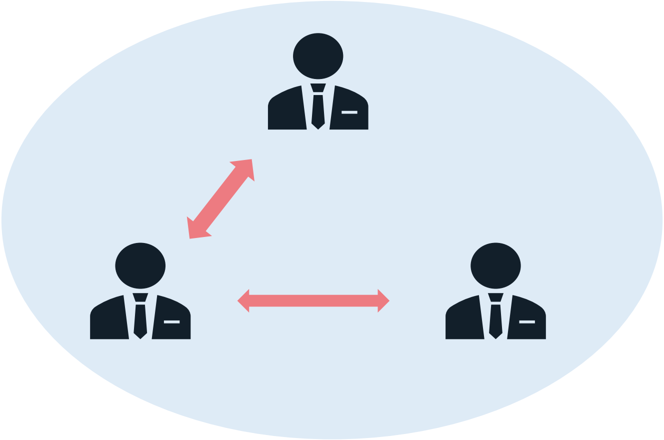
Bressan, Li, Ren, Wu. The embedded homology of hypergraphs and applications , 2016

Ren, Shiquan, et al. "Computing the Homology of Hypergraphs." *arXiv preprint arXiv:1705.00151* (2017).

Ren, Shiquan, Chengyuan Wu, and Jie Wu. "Operators on random hypergraphs and random simplicial complexes." *arXiv preprint arXiv:1712.02045* (2017).

Ren, Shiquan, and Jie Wu. "Stability of persistent homology for hypergraphs." *arXiv preprint arXiv:2002.02237* (2020).

Ren, Shiquan, et al. "A Discrete Morse Theory for Hypergraphs." *arXiv preprint arXiv:1804.07132* (2018).



Embedded homology of hypergraph

Definition (infimum chain complex)

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

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

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

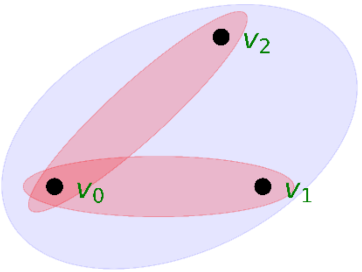
Definition (supremum chain complex)

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

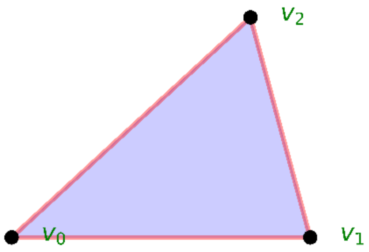
$$Sup_n(\mathcal{H}, R) = \bigcap\{C_n | \ C_\star \text{ is a subchain complex of } R((K_{\mathcal{H}})_\star) \text{ and } R(\mathcal{H}_n) \subset C_n\}$$

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

Bressan, Li, Ren, Wu. AJM, 2019



Hypergraph *H*



Associated simplicial complex K_H

Proposition

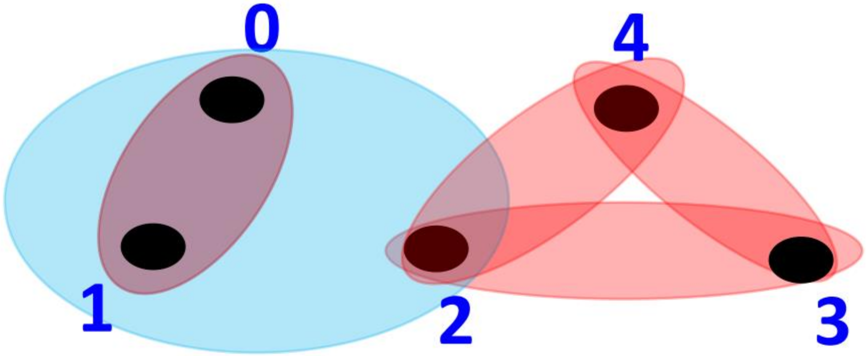
Given a hypergraph $\mathcal{H}$, the homology of the infimum chain complex of and supremum chain complex of $\mathcal{H}$ with coefficient R are isomorphic.

Definition (Hypergraph embedded homology)

Given a hypergraph $\mathcal{H}$, the n-th embedded homology of $\mathcal{H}$ with coefficient R is defined as

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

$$C_0 = Z\{\{0\},\{1\},\{2\},\{3\},\{4\}\}$$
$$C_1 = Z\{\{0,1\},\{2,3\},\{2,4\},\{3,4\}\}$$
$$C_2 = Z\{\{0,1,2\}\}$$
$$A_0 = Z\{\{0\},\{1\},\{2\},\{3\},\{4\}\}$$
$$A_1 = Z\{\{0,1\},\{0,2\},\{1,2\},\{2,3\},\{2,4\},\{3,4\}\}$$
$$A_2 = Z\{\{0,1,2\}\}$$
$$\to A_3 \xrightarrow{\partial_3} A_2 \xrightarrow{\partial_2} A_1 \xrightarrow{\partial_1} A_0$$
$$S_n = C_n + \partial_{n+1}(C_{n+1}), I_n = C_n \cap \partial_n^{-1}(C_{n-1})$$



$$I_0 = Z\{\{0\},\{1\},\{2\},\{3\},\{4\}\}$$
$$I_1 = Z\{\{0,1\},\{2,3\},\{2,4\},\{3,4\}\}$$
$$I_2 = 0$$
$$S_0 = Z\{\{0\},\{1\},\{2\},\{3\},\{4\}\}$$
$$S_1 = Z\{\{0,1\},\{2,3\},\{2,4\},\{3,4\},\partial\{0,1,2\}\}$$
$$S_2 = Z\{\{0,1,2\}\}$$

$$H_0^s = Ker(\partial_0^s) / \mathrm{Im}(\partial_1^s)$$
$$= S_0 / \mathrm{Im}(\partial_1^s)$$
$$= Z\{\{0\},\{1\},\{2\},\{3\},\{4\}\} / Z\{\{1\}-\{0\},\{3\}-\{2\},\{4\}-\{2\},\{4\}-\{3\}\}$$
$$= I_0 / \mathrm{Im}(\partial_1^i)$$
$$= Ker(\partial_0^i) / \mathrm{Im}(\partial_1^i)$$
$$= H_0^i$$

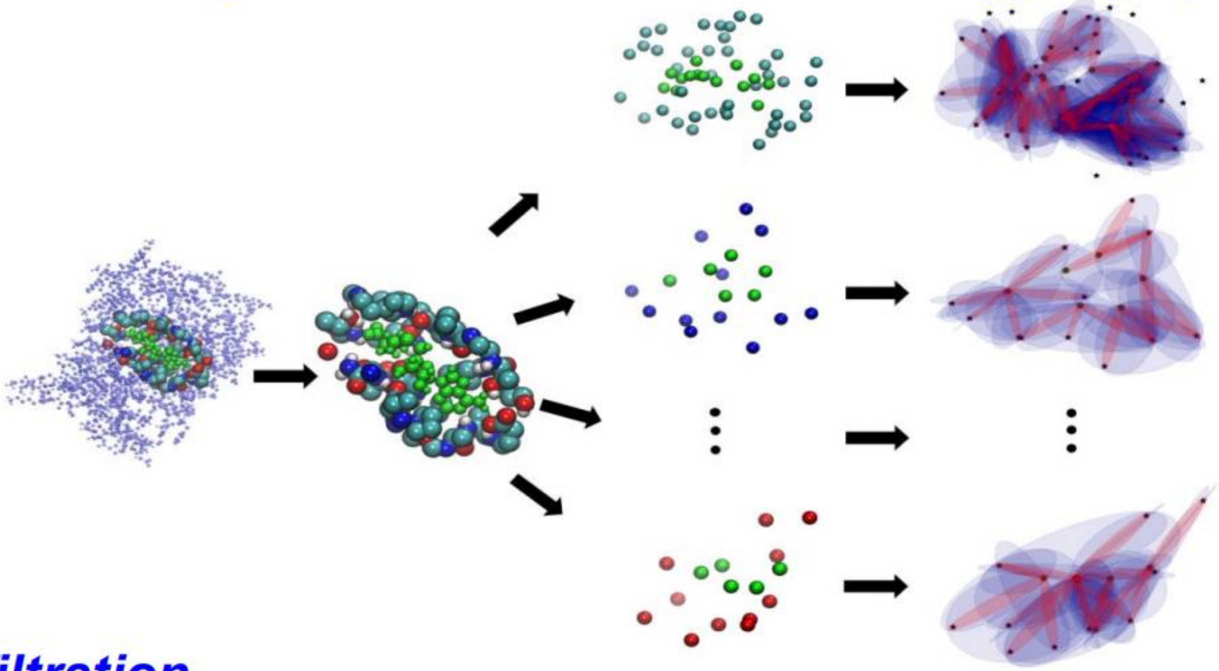
$$H_1^s = Ker(\partial_1^s) / \mathrm{Im}(\partial_2^s)$$
$$= Z\{\{3,4\}-\{2,4\}+\{2,3\},\partial\{0,1,2\}\} / Z\{\partial\{0,1,2\}\}$$
$$= Z\{\{3,4\}-\{2,4\}+\{2,3\}\}$$
$$= Ker(\partial_1^i) / \mathrm{Im}(\partial_2^i)$$
$$= H_1^i$$

$$H_2^s = Ker(\partial_2^s) / \mathrm{Im}(\partial_3^s)$$
$$= Ker(\partial_2^s)$$
$$= 0$$
$$= Ker(\partial_2^i) / \mathrm{Im}(\partial_3^i)$$
$$= H_2^i$$

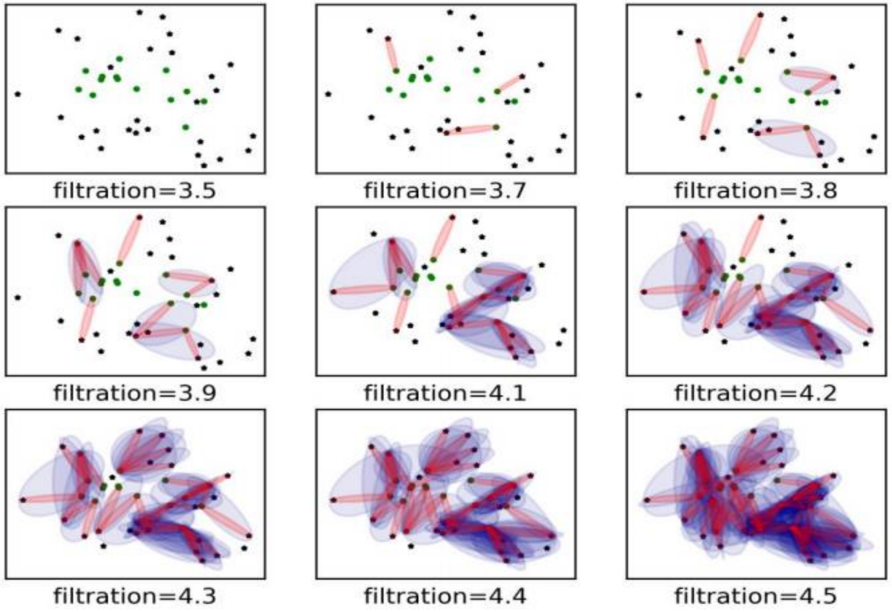
Liu, Wang, Wu, Xia, BIB, 2021

Hypergraph-based models

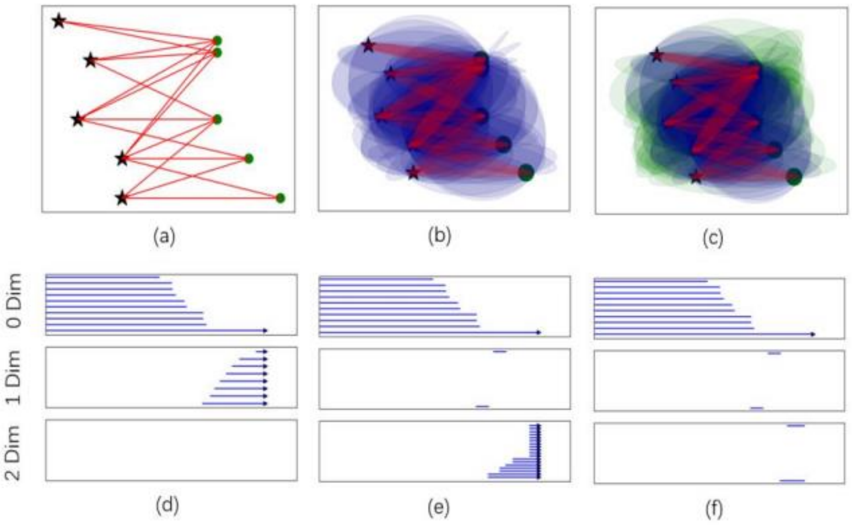
Protein-ligand interaction modeled as hypergraph



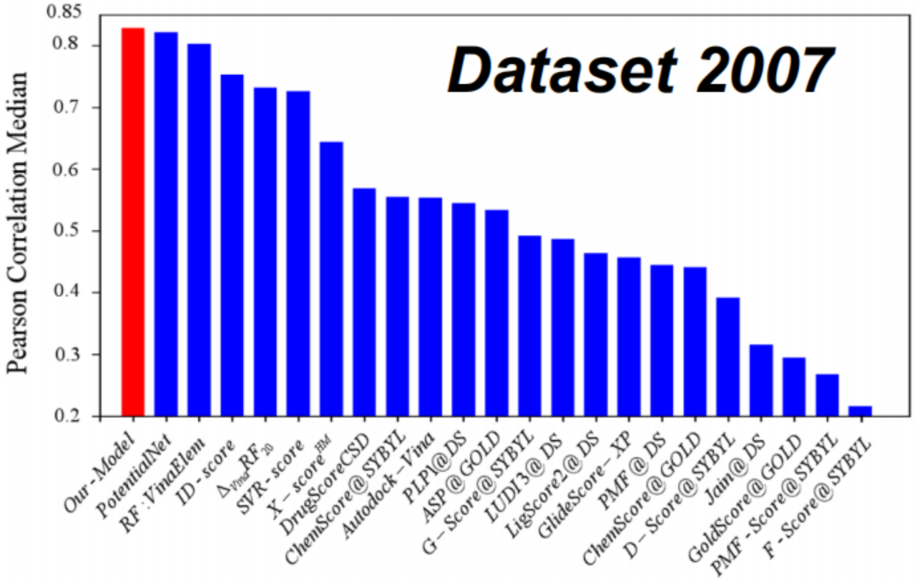
Hypergraph-based filtration



Bipartite graph VS Hypergraph

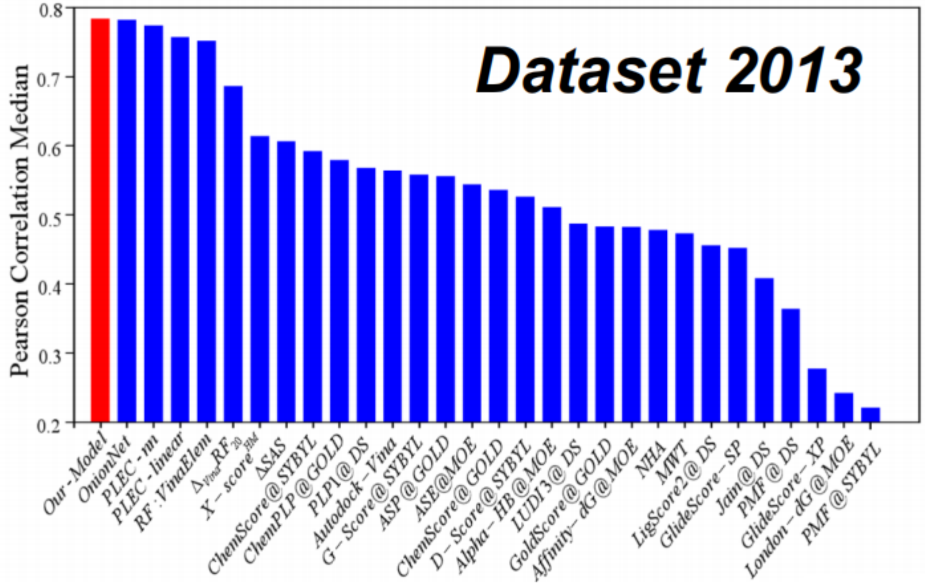


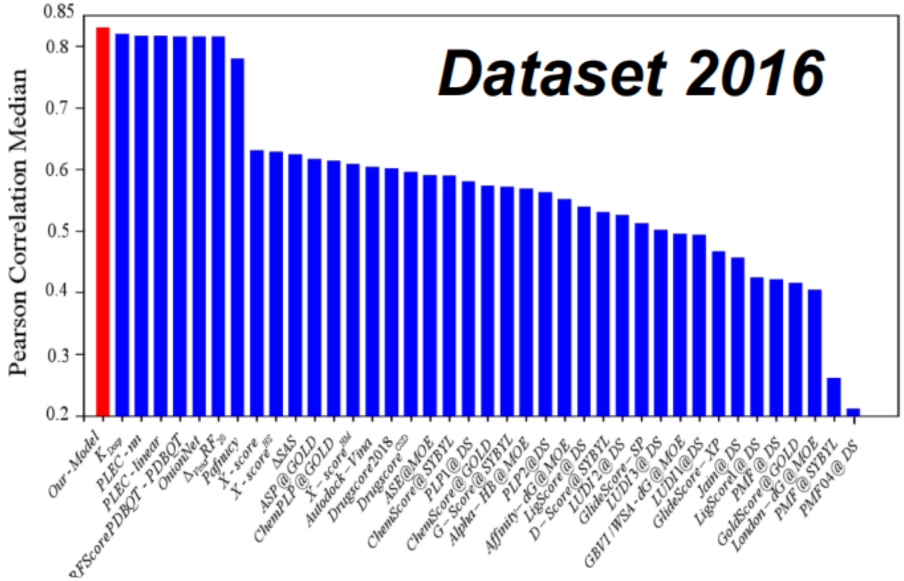
Benchmark testing with PDBbind datasets



Model setting:

homology vectors
+
Gradientboostingtree

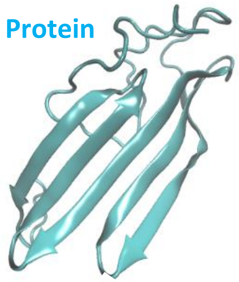




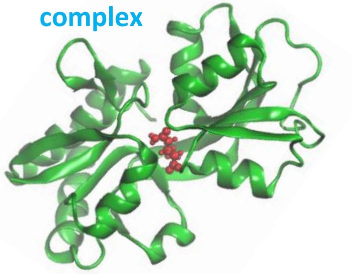
Persistent function based machine learning

Data

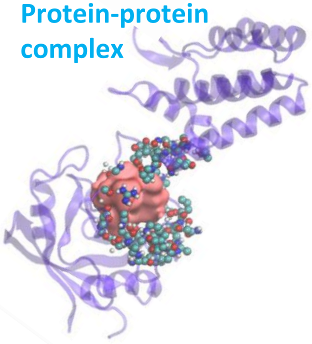
Protein



Protein-ligand complex

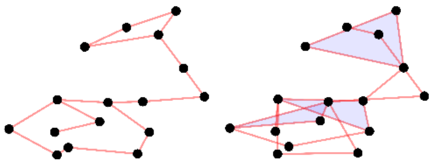


Protein-protein complex

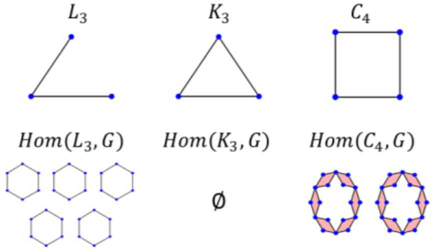


Representation

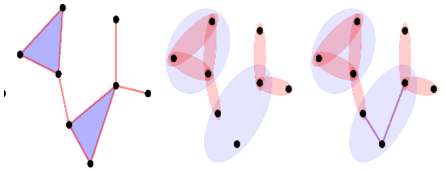
Simplicial complex: Neighborhood complex, Dowker complex,...



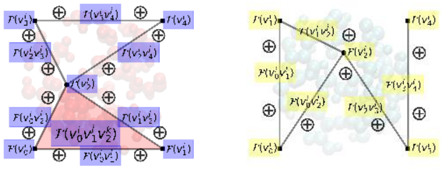
Polyhedral complex: Hom complex...



Hypergraph, Super-hypergraph ...

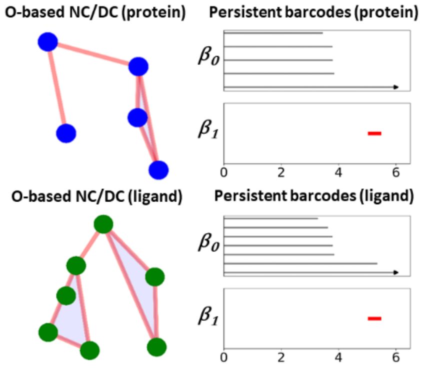


Algebraic representation:face ring...

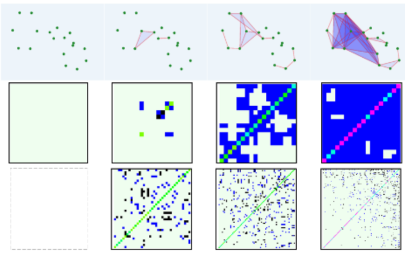


Featurization

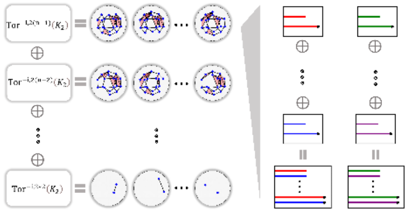
Persistent homology



Persistent Spectral

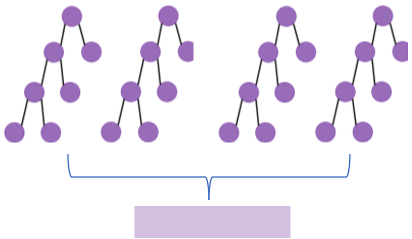


Persistent Tor-algebra

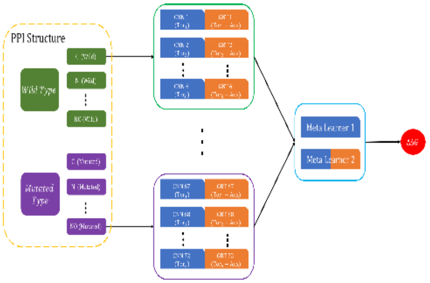


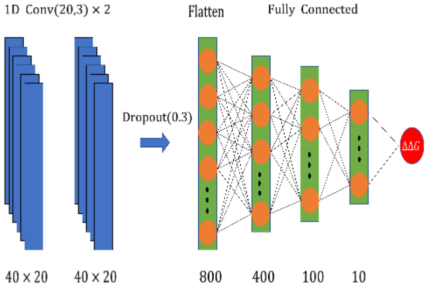
Learning

Machine learning: Random Forest, GBT, SVM,...



Deep learning: Convolution neural network,...





Thank You!