Topological Data Analysis and Topological Approaches to Drug Design and Discovery

Southern University of Science and Technology
October 2020

Jie Wu

Hebei Normal University

October 23, 2020

Topological Data Analysis and Topological Approaches to Drug Design and Discovery

Topological Data Analysis

Topological Approach to Drug Design and Discovery

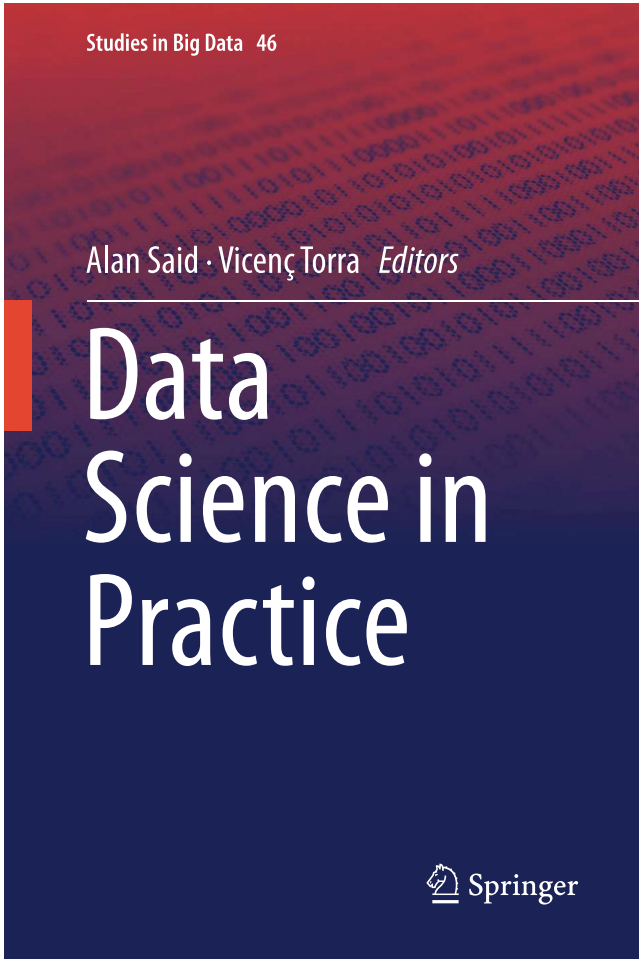
Our Work

Importance of Topology in Data Science

According to **Book published in 2019:** Alan Said and Vicenç Torra Editors, **Data Science in Practice**, 195pp, Studies in Big Data **46**, Springer Nature Switzerland AG 2019. ISBN 978-3-319-97555-9 ISBN 978-3-319-97556-6 (eBook), the following is a list of **tools that are commonly used** within data science.

- Optimization. Probability theory. Linear algebra. Graphs.
- **Topology**.
- Visual analytics. Programming languages and software.
- Other mathematical tools.

Data Science



1.3 Tools

In the previous section we have connected data science with the three related areas of statistics, machine learning, and big data technologies. In this section we will list some of the tools that are commonly used within data science.

- **Optimization**. Quite a few methods for modeling can be formulated in terms of an optimization problem [4]. That is, there is an objective function to be maximized (or minimized) and a set of constraints to be considered. The goal is to find an object or a combination of objects that satisfy the constraints, and are optimal in terms of the objective function. Optimization methods study approaches to solve this type of problems. Metaheuristics is a related area, and is about finding good heuristics to solve effectively optimization problems.
- **Probability theory**. Quite a few ways to model data are based on probability theory. Graphical models, and Bayesian networks, are some of them.
- **Linear algebra**. A simple multivariate linear regression model can be better (or more easily) represented and solved using matrices and vectors, and solved using linear algebra. Optimization problems are typically formulated using linear algebra. E.g., linear equality constraints are represented as the product of a matrix and a vector of variables equal to a vector. Some other machine learning and statistical models are also represented and solved (at least for some instances) using linear algebra. This is the case of support vector machines.
- **Graphs**. Some of the information available is conveniently represented in terms of graphs. This is the case of social networks. Graph theory provides concepts and tools to analyze this type of data. Complex networks is the term to denote a network with non-trivial topological structure. Trees are also graphs with the constraint that they should not contain cycles. In addition, some of the tools for data modeling, as the graphical models, also rely on graphs for the representation of knowledge.
- **Topology**. The field of topological data analysis [3, 5] has emerged recently as a way to extract relevant characteristics from data. Chazal and Michel [3] outlines a pipeline that stresses the role of topology and geometry in the analysis. This pipeline consists of (i) input data consisting on a finite set of points coming with a notion of distance; (ii) a "continuous shape" is built on top of the data: this results into an structure over the data; (iii) topological and geometric information is extracted from the structures; (iv) the topological and geometric information are the output of the approach and correspond to the new features of the data.
- **Visual analytics**. It is difficult to understand big data. Data visualization provides tools for a more effective understanding of the data, and visual analytics additionally provides tools for analyzing large data sets and helping in decision making processes.
- **Programming languages and software**. Appropriate programming languages for big data include R, Scala, and Python. Programming language frameworks commonly used in this field include Apache Spark, MapReduce, Hadoop, Flink,

Geometric simplicial complex

A **geometric simplicial complex** K is a collection of simplices, all contained in some Euclidean space $\mathbb{R}^J$ for some index set J such that

1. if σ^n is a simplex in K and τ^p is a face of σ^n, then τ^p is in K; and
2. if σ^n and τ^p are simplices of K, then $\sigma^n \cap \tau^p$ is either empty, or a common face of σ^n and τ^p.

There is a theory on weighted simplcial homology[1][2] with successful applications in biomolecular data analysis[3].

---

[1] Ren, Shiquan; Wu, Chengyuan; Wu, Jie *Weighted persistent homology*. Rocky Mountain J. Math. 48 (2018), no. 8, 2661-2687.

[2] Wu, Chengyuan; Ren, Shiquan; Wu, Jie; Xia, Kelin *Discrete Morse theory for weighted simplicial complexes*. Topology Appl. 270 (2020), 107038, 19 pp.

[3] Zhenyu Meng, D Vijay Anand, Yunpeng Lu, Jie Wu, Kelin Xia, *Weighted persistent homology for biomolecular data analysis*, Scientific Reports, 10, 2079 (2020).

Abstract simplicial complex, and Hypergraph

An **abstract simplicial complex** $\mathcal{K}$ is a collection of finite nonempty sets of a given set, such that if A is an element in $\mathcal{K}$, so is every nonempty subset of A.

A **hypergraph** is a collection of finite nonempty subsets of a given set.

Intuitively, a hypergraph is a simplicial complex with same faces missed. **Examples:** coauthorship network, protein-ligand binding network.

Simplicial homology can be extended to hypergraphs[4].

[4] Bressan, Stephane; Li, Jingyan; Ren, Shiquan; Wu, Jie *The embedded homology of hypergraphs and applications*. Asian J. Math. 23 (2019), no. 3, 479-500.

Čech's Covers, Nerve Complex, and Clustering

Let $\mathcal{U} = \{U_i\}_{i \in I}$ be a locally finite open cover of X. **Nerve complex** $\mathcal{N}(\mathcal{U})$ is the abstract simplicial complex whose vertices are U_i and simplices given by nonempty intersection of U_i's, i.e $U_{i_0}, \ldots, U_{i_n}$ forms an n-simplex iff their intersection is nonempty.

Čech cover is a locally finite cover such that any nonempty intersection is contractible.

Classical Čech Theorem. Let $\mathcal{U}$ be a Čech cover of X such that there exists a partition of unit subordinate to $\mathcal{U}$. Then X is weakly homotopy equivalent to $\mathcal{N}(\mathcal{U})$.

Covers $\longleftrightarrow$ clustering in data science.

Mapper from Data to Simplicial Complexes

Given a data as a finite subset X in $\mathbb{R}^n$, we consider X as a vertex set and draw a ball of radius r centered at each point $x \in X$. Then we get a collection of balls of radius r (centered at points in X), $B(x, r)$, $x \in X$.

We obtain a simplicial complex $K_r(X)$ (**depending on radius** r), called **Vietoris-Rips complex**, whose vertices are the balls and simplices given by **pairwise nonempty intersection** of these balls, i.e the balls $B_0(x_0, r), \ldots, B_n(x_n, r)$ forms an n-simplex (with $x_0, \ldots, x_n \in X$) iff $B_i(x_i, r) \cap B_j(x_j, r) \neq \emptyset$ for $i \neq j$.

Note. One can also use nerve complex of these balls to get **Čech complex**. Vietoris-Rips complex is simpler in terms of computation.

Persistent Homology

Given a data as a finite subset X in $\mathbb{R}^n$, we take homology with coefficients in a field of its Vietoris-Rips complex $K_r(X)$.

For each integer $k \geq 0$, homology group $H_k(K_r(X))$ is a vector space depending on parameter $r \geq 0$, called a **persistence module**.

Structure Theorem. Each (finite dimensional) persistence module admits a **factorization** in terms of **irreducible persistence modules**.

Each irreducible persistence module is an **interval persistence module**, V_t, from a to b, with $V_t = 0$ for $t < a$ or $t > b$, and $\dim(V_t) = 1$. Here a is called **birth** of V_t, and b is called **death** of V_t.

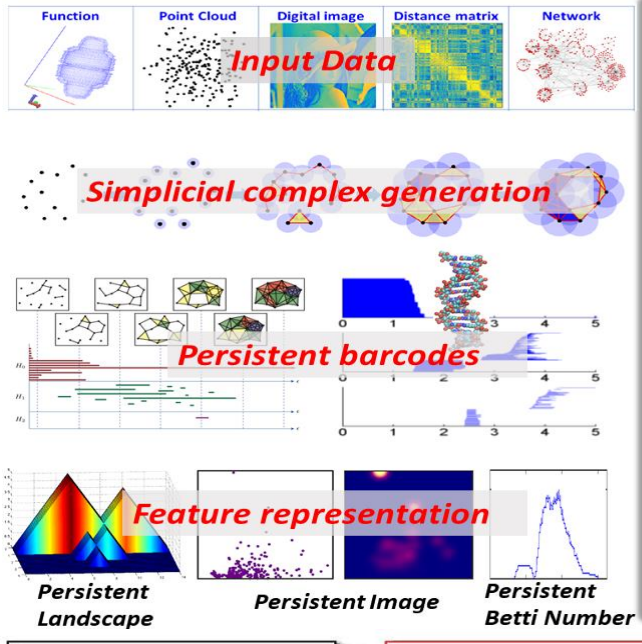
Topological Feature of Data

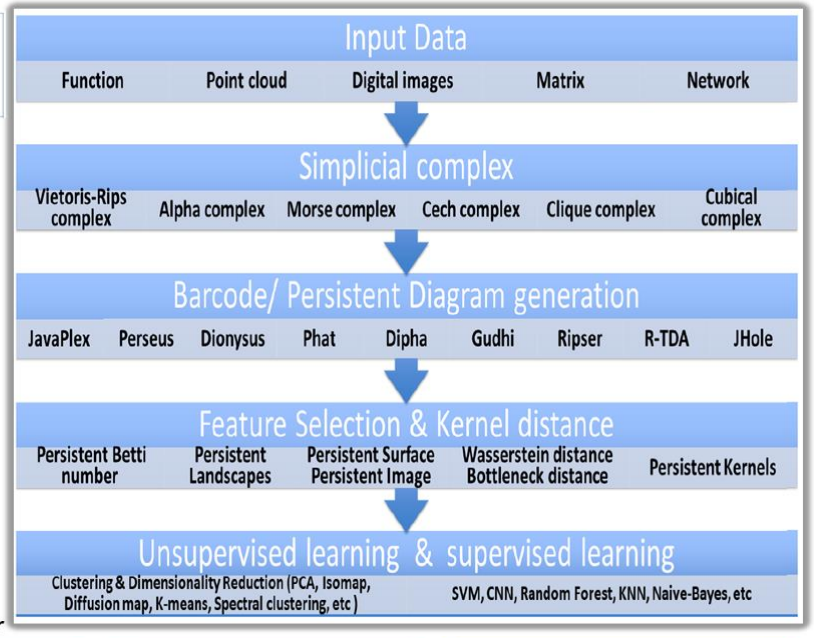
Given a data as a finite subset X in $\mathbb{R}^n$, fixing an integer $k \geq 0$, the homology group $H_k(K_r(X))$ (briefly speaking, **persistent Betti numbers**) induces **topological barcodes**, which is a **multi-set** in the plane $\mathbb{R}^2$ consisting of (a, b)'s with a (b) the births (deaths) of irreducible factors of the persistence module $H_k(K_r(X))$.

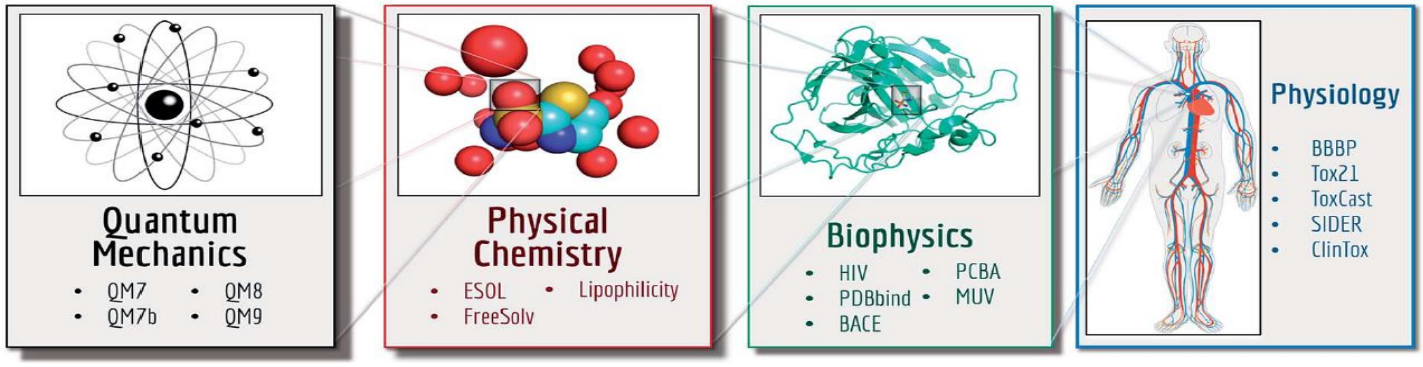
Topological barcodes can be also intuitively drawn as fingerprint, called **topological fingerprint**.

Then, what to do? Well, you do machine learning and data analysis!

Topology Based Learning Models





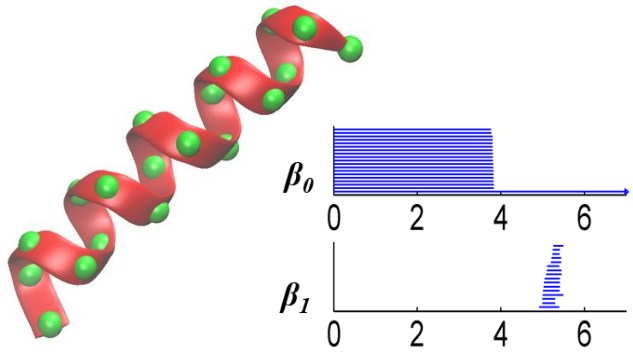


Some References on TDA and its Applications

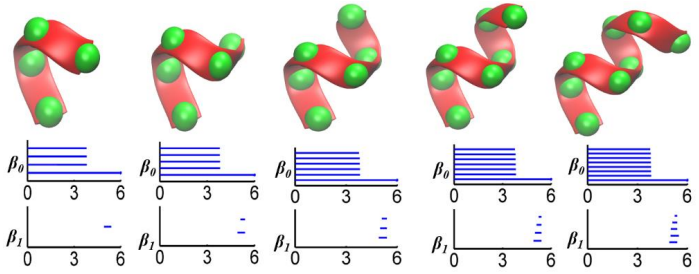
1. Edelsbrunner, H., Letscher, D., and Zomorodian, A. (2002). *Topological persistence and simplification*. **Discrete Comput. Geom.**, 28:511-533.
2. Carlsson, G. (2009). *Topology and data*. **AMS Bulletin**, 46(2):255-308.
3. Lee, Y., et al (... Hess, K. ...) (2017). *Quantifying similarity of pore-geometry in nanoporous materials*. **Nature Communications**, 8.
4. Menglun Wang, Z. X. Cang, and Guo-Wei Wei, *Topology-based network tree for the prediction of antibody-antigen binding free energy changes upon mutation*, **Nature Machine Intelligence**, 2, 116-123 (2020).
5. M. W. Reimann, et al (... Ran Levi, Kathryn Hess and H. Markram), *Cliques of Neurons Bound into Cavities Provide a Missing Link between Structure and Function*, **Frontiers in Computational Neuroscience** 11(4) (2017).

Biomolecular Topological Fingerprint

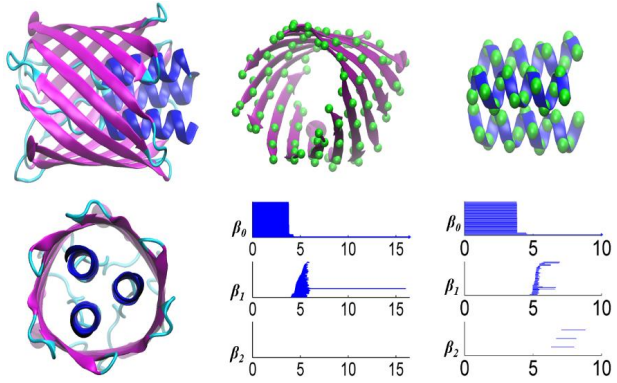
TF for alpha helix



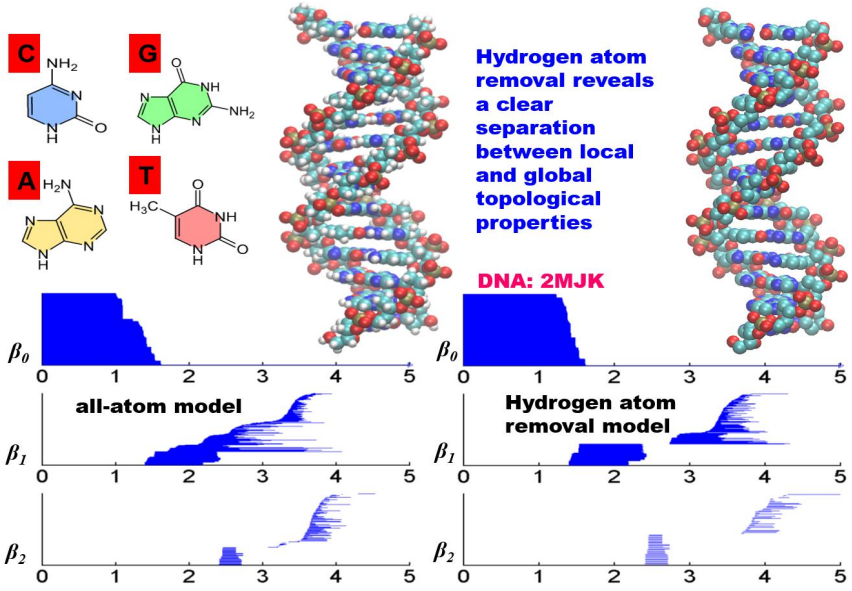
Slicing method



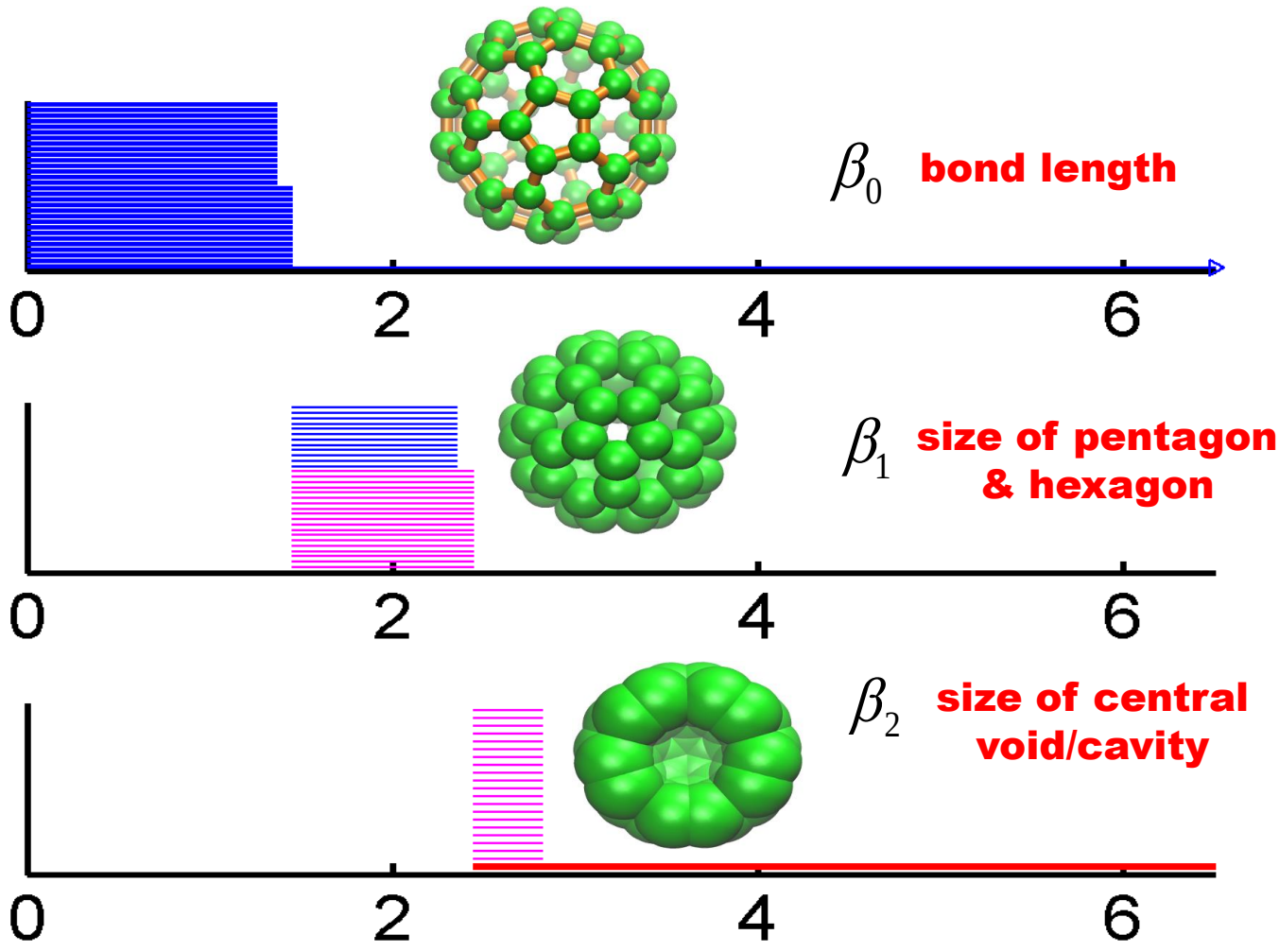
TF for beta barrel



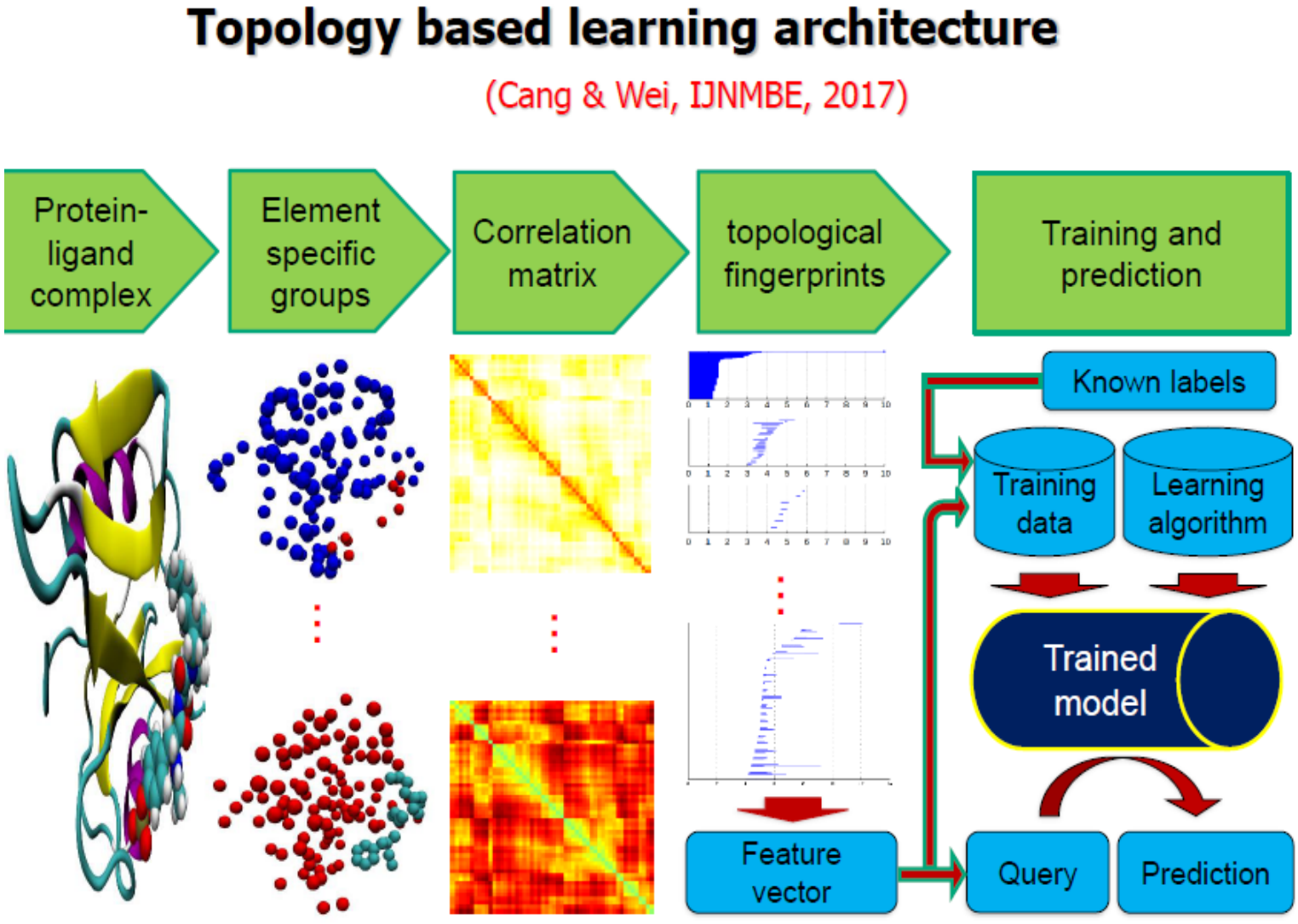
Topological fingerprints of DNA



PHA of fullerene C_{60}—(Xia, Feng, Tong & Wei, JCC, 2015)



Topology Based Learning



Topological approach to molecular biology

Guowei Wei group's works

Professor
Mathematics,
Electrical & Computer Engineering,
Biochemistry & Molecular Biology,
Michigan State University , USA

SIAM NEWS **DECEMBER 2017**

Research | December 01, 2017

Persistent Homology Analysis of Biomolecular Data

By Guo-Wei Wei

SIAM NEWS **SEPTEMBER 2016**

Get Involved | September 01, 2016

Mathematical Molecular Bioscience and Biophysics

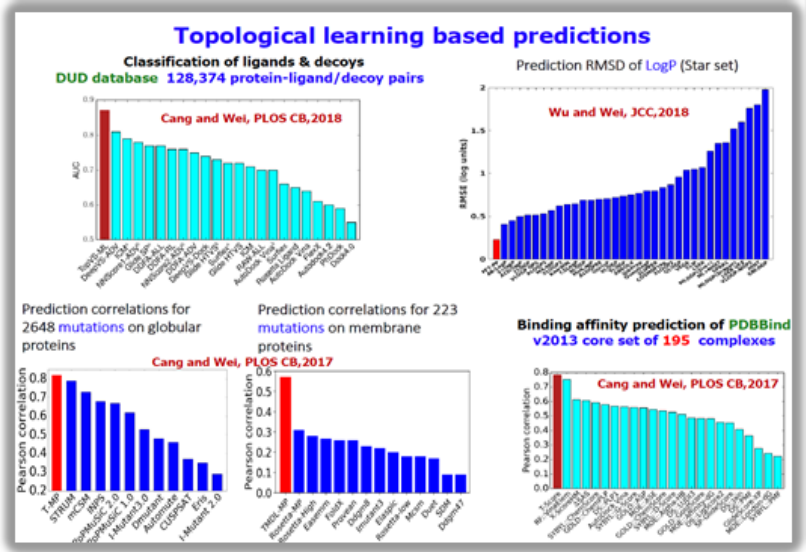
A Recurring Theme at the SIAM Conference on the Life Sciences

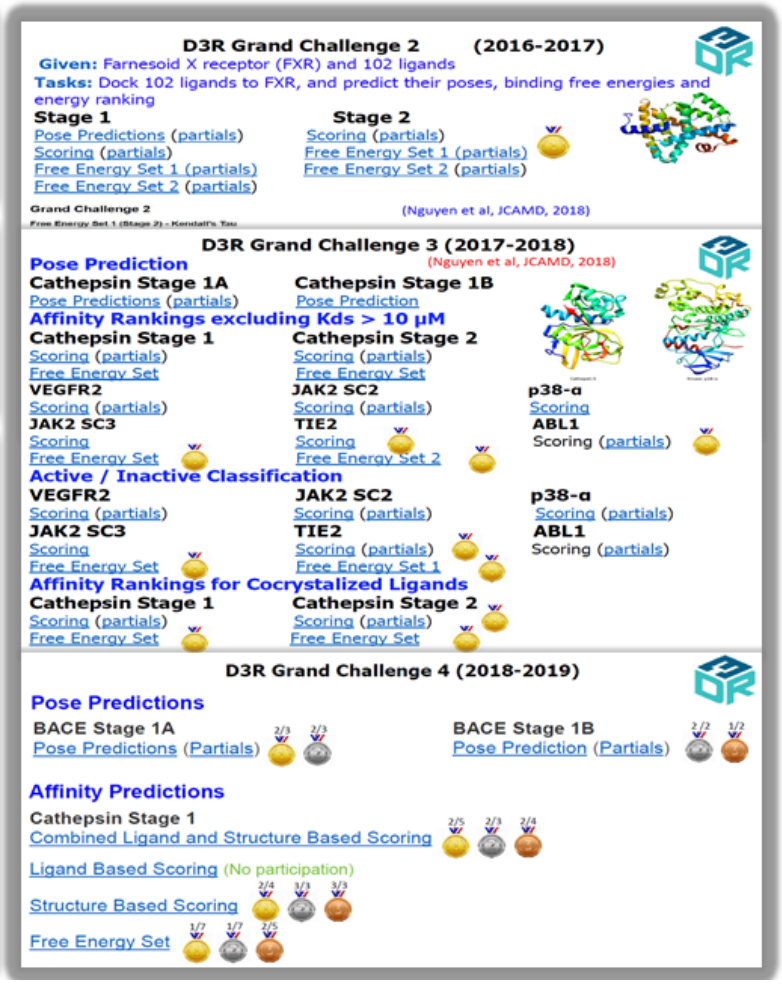
By Guo-Wei Wei

Software packages:

- MIBPB: Online server for electrostatic analysis using the second-order accurate Poisson-Boltzmann solver.
- ESES: Open-source online server for the generation of Eulerian solvent excluded surface.
- PPD: Online server for Protein Pocket Detection.
- FRI: Online server for the flexibility analysis of biomolecules based on flexibility and rigidity index.
- RI-Score: Online server for geometric graph theory or rigidity index (RI) based scoring function for protein ligand binding affinity prediction.
- TML-BP: Online server for topological learning for protein-ligand binding affinity prediction.
- TML-MP: Online server for topology based machine learning for the prediction of protein folding stability change upon mutation.
- TDL-BP: Online server for topological deep learning for protein-ligand binding affinity prediction.
- TDL-MP: Online server for topological deep learning for the prediction of protein folding stability change upon mutation.
- TopP-S: Online server for topological learning of partition coefficient (LogP) and aqueous solubility (LogS).
- TopTox: Online server for computing element-specific topological descriptors (ESTDs) for toxicity endpoint predictions.

Topology in Drug design





Topology in Drug design

Guo-Wei Wei, *Persistent homology analysis of biomolecular data*, **SIAM News 50 (10)**, December 1, 2017:

> However, persistent homology neglects chemical and biological information ... and is thus **not as competitive as** geometry or physics-based representation in quantitative predictions. **Element-specific persistent homology**, or multi-component persistent homology built on colored biomolecular network, has been introduced... This approach enciphers biological properties—such as hydrogen bonds, van der Waals interactions, hydrophilicity, and hydrophobicity—into topological invariants, rendering a **potentially revolutionary representation** for biomolecules.

Element-specific=subnetworks only having *C* or *O* or "*C* and *O*"...

Why is topology good in molecular biology?

25. Guo-Wei Wei, Duc Duy Nguyen and Zixuan Cang, System and methods for machine learning for drug design and discovery, United States Patent Application Publication, Pub. No.: US 2019 / 0304568 A1. Pub. Date: Oct. 3, 2019. [005]:

- Theoretical models for the study of structure-function relationships of biomolecules may conventionally be based on **pure geometric modeling techniques**.
- Mathematically, these approaches make use of local geometric information, which may include, but is not limited to, **coordinates, distances, angles, areas and sometimes curvatures** for the physical modeling of biomolecular systems.
- Indeed, geometric modeling may generally be considered to have value for structural biology and biophysics.

Why is topology good in molecular biology?

- However, conventional purely geometry-based models may tend to be inundated with **too much structural detail** and are frequently **computationally intractable**.
- In many biological problems, such as the opening or closing of ion channels, the association or disassociation of binding ligands (or proteins), the folding or unfolding of proteins, the symmetry breaking or formation of virus capsids, there exist topological changes. In fact, **full-scale quantitative** information **may not be needed to understand some physical and biological functions**.
- Put another way, in many biomolecular systems **there are topology-function relationships**, which **cannot be effectively identified using purely geometry-based models**.

Which data would topological approaches would be good for?

- Philosophically, topological approach would become more powerful for **more complicated data**.
- H_0 can quickly detect **density** information of data. So it detects **pattern** of data.
- H_1 can quickly detect **circles** in data.
- H_2 can quickly detect S^2**-shaped** information in data.

Weighted Persistent Homology

- Zhenyu Meng, D Vijay Anand, Yunpeng Lu, Jie Wu, Kelin Xia, *Weighted persistent homology for biomolecular data analysis*, **Scientific Reports**, 10, 2079 (2020).
  - Wu, Chengyuan (my former student), *weighted topological data analysis*, PhD thesis of National University of Singapore, 2019.
  - Ren, Shiquan; Wu, Chengyuan; Wu, Jie *Weighted persistent homology*. Rocky Mountain J. Math. 48 (2018), no. 8, 2661-2687.
  - Wu, Chengyuan; Ren, Shiquan; Wu, Jie; Xia, Kelin *Discrete Morse theory for weighted simplicial complexes*. Topology Appl. 270 (2020), 107038, 19 pp.

Weighted Persistent Homology for biomolecular DA

www.nature.com/scientificreports

SCIENTIFIC REPORTS

natureresearch

OPEN

Weighted persistent homology for biomolecular data analysis

Zhenyu Meng[1], D. Vijay Anand[1], Yunpeng Lu[2], Jie Wu[3] & Kelin Xia[1,4*]

In this paper, we systematically review weighted persistent homology (WPH) models and their applications in biomolecular data analysis. Essentially, the weight value, which reflects physical, chemical and biological properties, can be assigned to vertices (atom centers), edges (bonds), or higher order simplexes (cluster of atoms), depending on the biomolecular structure, function, and dynamics properties. Further, we propose the first localized weighted persistent homology (LWPH). Inspired by

Embedded Homology of Hypergraphs

- Xiang Liu, Xiangjun Wang, Jie Wu, and Kelin Xia, *Hypergraph based persistent cohomology (HPC) for machine learning in drug design*, preprint.
  - Bressan, Stephane; Li, Jingyan; Ren, Shiquan; Wu, Jie The embedded homology of hypergraphs and applications. Asian J. Math. 23 (2019), no. 3, 479-500.

The ideas introducing embedded homology of hypergraphs were inspired from the ideas of path homology introduced by S. -T. Yau et al[5]:

- A. Grigor'yan, Y. Lin, Y. Muranov, and S.-T. Yau, Homologies of path complexes and digraphs, Math arXiv: 1207.2834v4, 2013.

---

[5]Thanks **Prof. S. -T. Yau**  for his elegant ideas so that we can generalize simplicial homology to hypergraphs in a natural way.

Hypergraph homology in drug design

Hypergraph based persistent cohomology (HPC) for machine learning in drug design

Xiang Liu[a,b], Xiangjun Wang[b], Jie Wu[c], and Kelin Xia[a]

[a]Division of Mathematical Sciences, School of Physical and Mathematical Sciences, Nanyang Technological University, Singapore 637371; [b]School of Mathematical Science and LPMC, Nankai University, Tianjin 300071, China; [c]School of Mathematical Sciences, Hebei Normal University, Hebei 050024, China

This manuscript was compiled on September 25, 2020

Artificial intelligence (AI) based drug design has demonstrated great potential to fundamentally change the drug design and drug discovery. However, a key issue in AI-based drug design is efficient molecular descriptors or fingerprints. Here, we present hypergraph-based molecular topological representation, hypergraph-based (weighted) persistent cohomology (HPC/HWPC), and HPC-based molecular fingerprints for machine learning models in drug design. Molecular structures and their atomic interactions are highly complicated and pose great challenges for efficient mathematical representations. We develop the first hypergraph-based topological framework to characterize detailed molecular structures and interactions at atomic level. Inspired by the elegant path complex model, hypergraph-based em-

development of highly-efficient learning algorithms, will pave the way for AI-based drug design to fundamentally change the landscape of drug design and drug discovery (5, 6).

With the excitement and opportunities come challenges. Currently, one of the central challenges for machine learning models in drug design is molecular featurization, which is to identify or design appropriate molecular descriptors or fingerprints (16–19). In fact, featurization is a long-standing issue for chemical informatics and bioinformatics (14, 15). Traditional molecular/chemical descriptors are structural and physical properties obtained from structural geometry, chemical conformation, chemical graph, structure topology, as well as

HPC in drug design—Red is our result

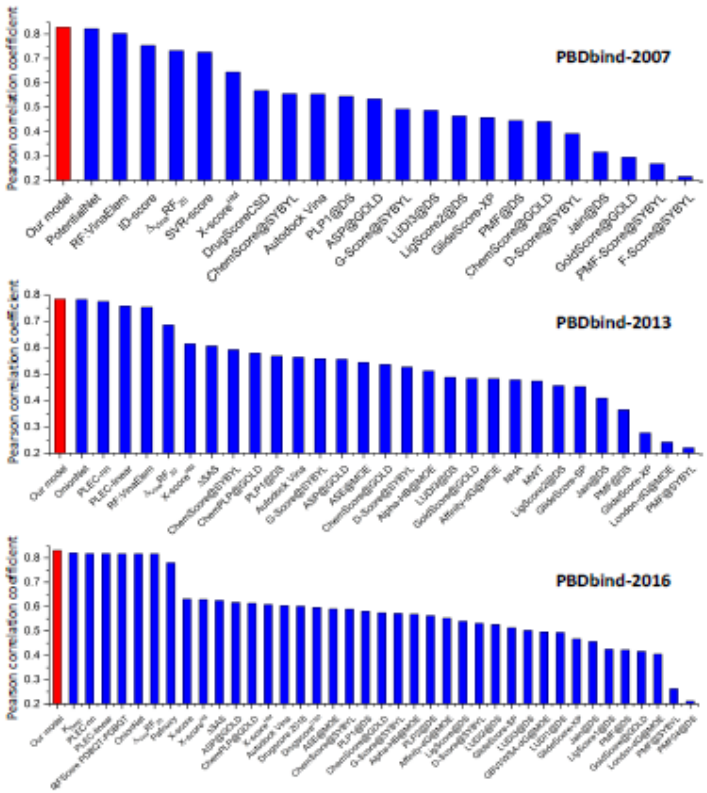


Fig. 3. The comparison of PCCs between our combined HPC/HWPC-GBT model and traditional molecular descriptor based models, for the prediction of protein-ligand binding affinity. The PCCs are calculated based on the core set (test set) of PDBbind-2007, PDBbind-2013 and PDBbind-2016.

A unified topological approach to data science[6]

Our new work will give a generalization of current persistent homology by introducing **super persistent homology**, which would give a unified topological approach to graphic data/network as well as point cloud data.

- Jelena Grbic and Jie Wu, *a unified topological approach to data science*, work in progress.

[6]The ideas were motivated from communications with Prof **Liu, Jianya**, Shandong University, about the possibility of the applications of TDA to social network, finance and management. **Thanks to Jianya**  **for his deep insights on mathematics!**

Our Goal/Hope—provide an unified approach suitable for both point cloud data and graphic data

- In our setting, we explore topological structures on graphic data with scoring schemes.
- The current persistent homology can be obtained as special cases of our more general theory from a natural transformation from point cloud data to graphic data with scoring schemes.
- This is a **theoretical research** on topological approaches in data science for hoping to make a tunnel between topology and data science.

Our Approaches

A. Homology Theory on **any collection** of subgraphs of a working graph. In theory, you choose whatever collection of subgraphs, you get homology on this collection of subgraphs.

B. Assign **any scoring scheme** on the working graph so that there is a **score** for any subgraph in the collection of subgraphs on your hand. Then it creates **persistent homology** as **new feature** for you.

C. Of course the **current persistent homology on point cloud data** should be answered from **A and B**.

Answer to C for Vietoris-Rips persistent homology

Let X be a point cloud data in $\mathbb{R}^N$. Mathematically, X is a finite set located in $\mathbb{R}^N$.

- **Step 1.** The **working graph** G is a **complete graph** by joining one edge for each pair of points in X.—**simple!**
- **Step 2.** The **collection of subgraphs**: any clique (complete subgraph) of G.—**simple!**
- **Step 3.** The **scoring scheme**: Let G' be a subgraph. Define its score

$$\mathfrak{M}^{VR}(G') = \frac{1}{2}\max\{d(v,w) \mid v,w \in V(G')\},$$

the half of the maximal embedded distance in the Euclidean space between pairwise vertices.—**natural!**

Answer to C for Čech persistent homology

Let X be a point cloud data in $\mathbb{R}^N$. Mathematically, X is a finite set located in $\mathbb{R}^N$.

- **Step 1.** The **working graph** G is a **complete graph** by joining one edge for each pair of points in X.—**simple!**
- **Step 2.** The **collection of subgraphs**: any clique (complete subgraph) of G.—**simple!**
- **Step 3.** The **scoring scheme**: Let G' be a subgraph. Define its score

$$\mathfrak{M}^C(G') = \inf_{x \in \mathbb{R}^N} \max\{d(x, v) \mid v \in V(G')\},$$

—also natural!

Answer to C for Witness persistent homology

- The first two steps are the same.

- Only **re-define scoring schemes**: Let $G' \leq G$ be a subgraph of G embedded in $\mathbb{R}^N$.
  1. **Strong Witness Scoring**

$$\mathfrak{M}^{W^s}(G') = \inf_{x \in \mathbb{R}^N} \{ \sup_{y \in V(G')} d(x,y) - \inf_{z \in V(G)} d(x,z) \}.$$

  2. Similarly, there are **Vietoris-Rips Strong witness scoring**, **Weak witness scoring**, **Vietoris-Rips weak witness scoring** by translating Carlsson's setting for witness complexes into scoring.

Can we get anything new by looking scoring? Quick example 1

Let G be a graph located in $\mathbb{R}^m$. (e.g. graph data on $3D$ objects, data on protein structure.) Take VR-scoring on G. In stead of **complete graph on vertices of** G, we take **clique complex** $\mathrm{Clique}(G)$.

- $\Longrightarrow$ persistent homology converges to $H_*(\mathrm{Clqiue}(G))$.
- **Comparison.** *VR* persistent homology on point cloud data $V(G)$ converges to trivial homology.

- **Why is** $\mathrm{Clique}(G)$ **good?** Let $X = |K|$ be a polyhedron with K simplicial complex. Take G=1-skeleton of bary-centric subdivision of K. Then $|\mathrm{Clique}(G)| \cong |X|$.

Anything new? Quick example 2

Let us consider **pull-back scoring from a non-injective function from the vertex set to a Euclidean space**.

Let $p\colon E \to B$ be a fibration or fibre bundle with E, B polyhedra. Take triangulations on E and B to make p simplicial up to homotopy. Take graphs $G(E)$ and $G(B)$ as 1-skeletons of the bary-centric subdivisions of simplicial models for E and B.

Take scoring scheme on $G(E)$ as the **pull-back** of

$$V(G(E)) \xrightarrow{\text{proj}} V(G(B)) \xhookrightarrow{\text{embedding}} \mathbb{R}^m$$

Consider clique complexes $\mathrm{Clique}(G(E))$ and $\mathrm{Clique}(G(B)) \Longrightarrow$ **persistent Leray-Serre spectral sequence**.

The mathematical question

Let G be a working graph. Let $\mathcal{H}$ be a family of finite subgraphs.

Question. What is a natural way to define homology of $\mathcal{H}$?

Rationality of Question: Abstract simplicial complex is a family of (finite) subsets that is closed under subset-operation. There is a well-established **simplicial homology theory**.

New Situation:

1) $\mathcal{H}$ is a family of finite subgraphs, rather than a family of finite sets; and
2) **no hypothesis** that $\mathcal{H}$ is closed under subgraph-operation.

High-dimensional structures

Clique complex (also named as **flag complex**) and **independence complex** (the clique complex of the complementary graph) are widely used notions in mathematics and practical applications.

The *clique complex* of a simple graph G is the abstract simplicial complex $\mathrm{Clique}(G)$ whose simplices consist of all cliques of G.

Let $G = (V, E)$ be a multi-graph. Then the set of cliques $\mathrm{Clique}(G)$ is **no longer** a simplicial complex in general[7]. The correct notion for describing the topological structure of the set $\mathrm{Clique}(G)$ is Δ**-set** (also called **semi-simplicial set**).

[7] let G be a graph with two vertices v and w and two edges e_1 and e_2 joining with them. Then $\mathrm{Clique}(G) = \{ve_1w, ve_2w, v, w\}$, which is not a simplicial complex.

Neighborhood complex—introduced by Lovász

Neighborhood complex $\mathcal{N}(G)$ of a graph G is a simplicial complex whose vertices are the vertices of G and whose simplices are those subsets of the vertex set $V(G)$ which have a **common neighbor**—landmark work on topological combinatorics of L. Lovász's solution to Kneser conjecture:[8].

- If we split the n-subsets of a $(2n+k)$-element set into $k+1$ classes, one of the classes will contain two disjoint n-subsets

The topology on the geometric realization of $\mathcal{N}(G)$ can be quite different from that of $\mathrm{Clique}(G)$ in general[9]. Namely, one could have different higher dimensional structures.

[8] Lovász, L. *Kneser's conjecture, chromatic number, and homotopy*, J. Combin. Theory Ser. A **25** (1978), no. **3**, 319-324.

[9] For instance, let G be a graph with three vertices a, b, c and two edges given by ab and bc. Then $\mathcal{N}(G) = \{\{a,c\},\{a\},\{b\},\{c\}\}$, which is not connected, and $\mathrm{Clique}(G) = \{\{a,b\},\{b,c\},\{a\},\{b\},\{c\}\}$ which is connected.

High-dimensional structures—Other complexes

- **Hom complexes**, a generalization of neighborhood complex introduced by Lovász[10].
- **Graph complex:** abstract simplicial complex on the edge set.— Jacob Jonsson, book in 2008[11][12].
- **Path complexes**—first introduced by **Shing-Tung Yau** and his collaborators[13], which was a mathematization of the work motivated from physical applications.
  - Recently introduced **magnitude homology** (of graphs) is related to path homology.
- **Tournaplexes**—in the paper of Ran Levi, Kathryn Hess et al.

[10] recent work: Eric, Babson and Dmitry N. Kozlov, *Proof of the Lovász conjecture*, Ann. of Math. (2) 165 (2007), no. 3, 965-1007.

[11] Jonsson, Jakob *Simplicial complexes of graphs. Lecture Notes in Mathematics*, **1928**. Springer-Verlag, Berlin, 2008. xiv+378 pp. ISBN: 978-3-540-75858-7.

[12] **Kontsevich**'s graph complex is a different notion.

[13] A. Grigor'yan, Y. Lin, Y. Muranov, and S.-T. Yau, Homologies of path complexes and digraphs, Math arXiv: 1207.2834v4, 2013.

Two types of topology on $\mathcal{H}$

For creating topology, we regard a subgraph in $\mathcal{H}$ as a **simplex** in certain dimension. It requires a **face-operation** so that we can "**glue**" together.

Face-operation 1. Vertex-deletion: clique complex, neighborhood complex, path complexes.

Face-operation 2. Edge-deletion: graph complex.

Edge-deletion Topology—Need homology of hypergraphs

Let G be a working graph. Let $\mathcal{H}$ be a family of finite subgraphs.

Consider $\mathcal{H}$ as a family of finite subsets of the edge set $E(G)$.

Each subgraph is determined by its edge set.

$\mathcal{H}$ is a **hypergraph** under **edge-deletion operation**.

There is a homology theory (as extension of simplicial homology theory) on hypergraphs: [14]

[14] Stephane Bressan, Jingyan Li, Shiquan Ren, Jie Wu, *The Embedded Homology of Hypergraphs and Applications*, Asia J. Math. **23** (2019), no. 3, 479-500.

Vertex-deletion Topology—Need homology of super-hypergraphs

Let G be a working graph. Let $\mathcal{H}$ be a family of finite subgraphs.

Consider $\mathcal{H}$ as a family of finite subsets of the vertex set $V(G)$.

Each subgraph **may not be** determined by its vertex set.

Example. Let G be a multi-graph with vertices a and b and two edges f_1, f_2 between a and b. Then af_1b and af_2b are two subgraphs having the same vertices.

If we want to explore topology of subgraphs, the notion of hypergraph is insufficient.

We need a new notion. We call it **super-hypergraph**.

Δ-set

A Δ**-set** means a sequence of sets $X = \{X_n\}_{n\geq 0}$ with *faces* $d_i : X_n \to X_{n-1}$, $0 \leq i \leq n$, such that

$$d_i d_j = d_j d_{i+1}$$

for $i \geq j$, which is called the Δ*-identity*.

The notion of Δ-set is a generalization of (abstract) simplicial complex by ruling out **face-operation**.

Simplicial homology can be defined using the notion of Δ-set.

Super-hypergraph

A **super-hypergraph** is a pair $(\mathcal{H}, X)$, where X is a Δ-set and $\mathcal{H}$ is a graded subset of X.

We call $\mathcal{H}$ a **super-hypergraph born from** X, and X is called **a parental Δ-set** of $\mathcal{H}$.

Example. Let G be a multi-graph with vertices a and b and two edges f_1, f_2 between a and b. Let $\mathcal{H} = \{af_1b, af_2b\}$ be two simple subgraphs. Then $\mathcal{H}$ can be viewed as a super-hypergraph with two 1-simplex with sharing the same missing vertices.

Algebraic Lemmas

Let G_* be a chain complex of groups and let D_* be a graded subgroup of G_*. Here we do not assume that G_n is commutative. Define

- $\sup_*^{G_*}(D_*)$ is the intersection of subcomplexes C_* of G_* with property that $D_n \leq C_n$ for $n \in \mathbb{Z}$.
- $\inf_*^{G_*}(D_*)$ is the product of subcomplexes E_* of G_* with property that $E_n \leq D_n$ for $n \in \mathbb{Z}$.

We briefly denote $\sup_*(D_*)$ for $\sup_*^{G_*}(D_*)$ and $\inf_*(D_*)$ for $\inf_*^{G_*}(D_*)$ if the embedding of $D_* \subseteq G_*$ is clear.

Algebraic Lemmas

Proposition. Let G_* be a chain complex of groups and let D_* be a graded subgroup of G_*.

1. The inclusion
$$\inf_*(D_*) \longrightarrow \sup_*(D_*)$$
induces an injective mapping on homology.

2. Suppose that $\partial_{n+1}^{G_*}(D_{n+1})$ is contained in the normalizer of D_n for each n. Then the inclusion
$$\inf_*(D_*) \longrightarrow \sup_*(D_*)$$
induces an isomorphism on homology. In particular, if D_n is normal in G_n for $n \in \mathbb{Z}$, then the inclusion $\inf_*(D_*) \longrightarrow \sup_*(D_*)$ induces an isomorphism on homology.

Embedded Homology of Super-hypergraphs

Let $(\mathcal{H}, X)$ be a super-hypergraph. Let A be an abelian group. The **embedded homology** $H_*^{\mathrm{emb},X}(\mathcal{H}; A)$ **with coefficients in** A of $(\mathcal{H}, X)$ is defined by the homology of the chain complex of $\inf_*$ and $\sup_*$ of the graded subgroup $\mathbb{Z}(\mathcal{H}) \otimes A$ in the chain complex $C_*(X; A)$.

Note. The **gap complex** $\sup_*(\mathbb{Z}(\mathcal{H}) \otimes A)/\inf_*(\mathbb{Z}(\mathcal{H}) \otimes A)$ is contractible. If there are some additional information, one may get further information on the gap complex. For instance, if there is a group G-action, then one may look at homology $H_*((\sup_*/\inf_*) \otimes_{\mathbb{Z}(G)} M)$ for G-modules M.

Remark

Embedded Homology of a hypergraph/super-hypergraph **may not be equal to** homology of a simplicial complex in general.

Let $\mathcal{H}$ be the boundary of a 2-simplex with **removing all three vertices**. Let X be the boundary of the 2-simplex. Then $H_1(X) = \mathbb{Z}$, $H_0(X) = \mathbb{Z}$, and $H_1(\mathcal{H}) = \mathbb{Z}$, $H_0(\mathcal{H}) = 0$.

No nonempty space whose unreduced 0-th homology is 0.

hypergraphs/superhypergraphs seem **geometry-like** objects.

Geometric gap complex: Let $\Delta\mathcal{H}$ be the minimal Δ-subset of X containing $\mathcal{H}$, and let $\delta\mathcal{H}$ be the maximal Δ-subset of X contained in $\mathcal{H}$. The inclusion $\delta\mathcal{H} \to \Delta\mathcal{H}$ may not be homotopy equivalent.

Our work on other applications of topology

Topological distributed computing:

- Yue, Yunguang; Wu, Jie; Lei, Fengchun *The evolution of non-degenerate and degenerate rendezvous tasks*, Topology Appl. 264 (2019), 187-200.
- Yue, Y.; Lei, F.; Liu, X.; Wu, J. *Asynchronous Computability Theorem in Arbitrary Solo Models*, Mathematics 2020, 8 (5), 757. https://doi.org/10.3390/math8050757

Topological robotics

- Murillo, Aniceto; Wu, Jie *Topological complexity of the work map*, Journal of Topology and Analysis online ready https://doi.org/10.1142/S179352532050003X

Sequence Analysis:

- Wu,Chengyuan; Ren, Shiquan; Wu, Jie; Xia, Kelin, *Magnus representation of genome sequences*, Journal of Theoretical Biology 480 (2019) 104-111.

Acknowledgements

- **core collaborators** of the **applied topology projects**
  - Prof **XIA Kelin**(biomolecular topologist, NTU) 
  - Prof **LIU Xingwu** (computer scientist) 
  - Dr **YUE Yunguang** (first young researcher on topological distributed computing in China) 
  - Mr **LIU Xiang** (extremely strong master student) 
  - Profs **Stephen Theriault** (leading topologist), **Jelena Grbic**(leading topologist)...
  - Prof **WANG Zhigang**  and Dr **LI Yanling** 
  - My former PhDs: Dr **WU Chengyuan**, Dr **REN Shiquan**, Dr **Fedor Pavutnitskiy**
  - Topology community in China
  - Topology team in Hebei.
- **Collaborators**: Prof **WEI Guo-Wei** (leading biomolecular topologist)...

Center for Topology and Geometry Based Technology

河北師范大学 HEBEI NORMAL UNIVERSITY

CTGT

拓扑与几何技术创新中心 Center for Topology and Geometry based Technology

CTGT Scientific Advisory Board

Buchstaber, Victor Matveevich
Steklov Mathematical Institute, Russian

Cohen, Frederick
University of Rochester, USA

Duan, Haibao
Chinese Academy of Sciences, China

Fang, Fuquan
Capital Normal University, China

Feng, Jianfeng
Fudan University, China

Grbic, Jelena
University of Southampton, UK

Jiang, Chunlan
Hebei Normal University, China

Liu, Jianya
Shandong University, China

Liu, Jingze
Hebei Normal University, China

Liu, Kefeng
University of California, Los Angeles, USA

Liu, Ying
Hebei Normal University, China

Mi, Jusheng
Hebei Normal University, China

Miller, Haynes
Massachusetts Institute of Mathematics, USA

Orlov, Dmitri Olegovich
Steklov Mathematical Institute, Russian

Prasad, Dipendra
Tata Institute of Fundamental Research, India

Vesnin, Andrei Yurievich
Tomsk State University, Russian

Wei, Guo-Wei
Michigan State University, USA

Wang, Yanying
Hebei Normal University, China

Wu, Jie
Hebei Normal University, China

Thank You for Your Attention!