Biomolecular Topology: modelling and analysis

Jian Liu[1,2], Kelin Xia[3,*], Jie Wu[2,*], Stephen Shing-Toung Yau[4,*], Guo-Wei Wei[5,6,7,*]
[1]Shool of Mathematical Sciences, Hebei Normal University
[2]Yanqi Lake Beijing Institute of Mathematical Sciences and Applications
[3]School of Physical and Mathematical Sciences, Nanyang Technological University
[4]School of Biological Sciences, Nanyang Technological University
[5]Department of Mathematical Sciences, Tsinghua University
[6]Department of Mathematics, Michigan State University,
[7]Department of Biochemistry and Molecular Biology, Michigan State University
[8]Department of Electrical and Computer Engineering, Michigan State University

May 27, 2022

Abstract

With the great advancement of experimental tools, a tremendous amount of biomolecular data has been generated and accumulated in various databases. The high dimensionality, structural complexity, the nonlinearity, and entanglements of biomolecular data, ranging from DNA knots, RNA secondary structures, protein folding configurations, chromosomes, DNA origami, molecular assembly, to others at the macromolecular level, pose a severe challenge in their analysis and characterization. In the past few decades, mathematical concepts, models, algorithms, and tools from algebraic topology, combinatorial topology, computational topology, and topological data analysis, have demonstrated great power and begun to play an essential role in tackling the biomolecular data challenge. In this work, we introduce biomolecular topology, which concerns the topological problems and models originated from the biomolecular systems. More specifically, the biomolecular topology encompasses topological structures, properties and relations that are emerged from biomolecular structures, dynamics, interactions, and functions. We discuss the various types of biomolecular topology from structures (of proteins, DNAs, and RNAs), protein folding, and protein assembly. A brief discussion of databanks (and databases), theoretical models, and computational algorithms, is presented. Further, we systematically review related topological models, including graphs, simplicial complexes, persistent homology, persistent Laplacians, de Rham-Hodge theory, Yau-Hausdorff distance, and the topology-based machine learning models.

Key words: Persistent homology; Topological data analysis; Biomolecular topology; Protein structure; Machine learning.

*Address correspondences to Kelin Xia (xiakelin@ntu.edu.sg), Jie Wu (wujie@bimsa.cn), Stephen Yau (yau@uic.edu), and Guo-Wei Wei (weig@msu.edu)

Contents

1 Introduction

A major trend for biology in the 21st century is its transition from phenomenological and descriptive sciences to quantitative and predictive sciences. This transition happens due to the generation and accumulation of a gigantic amount of data, which are systematically organized and deposited in various databanks, such as GenBank, Protein Data Bank (PDB), Electron Microscopy Data Bank (EMDB), etc. The availability of the huge amount of experimental data provides both unprecedented opportunities and great challenges for mathematicians.[205] Among them, biomolecular structure-function relationships have the most profound impact on bioengineering and biomedicine, and are widely regarded as the "holy grail".[206] Mathematically, geometry plays a very important role in biological sciences. Geometric tools and modeling not only help to visualize biological data,[60,170] but also fill the gap between theoretical models and structural information.[3,30] A prominent example is the virtual screening in drug design,[113,175] where the drug candidates match with the concave regions of the biomolecular targets geometrically, just like a key to a lock. The combination of geometrical models with physical models, particularly quantum mechanics and molecular mechanics, contributes tremendously to biophysics and biochemistry.[50,172]

However, biomolecular geometrical measurements have always been plagued by excessive structural details and are computationally expensive. In contrast, topology studies the fundamental characteristics of spaces - such as connectivity, dimensionality, and continuous transformation - so topological tools can dramatically reduce the irrelevant structural details and preserve only the intrinsic information.[206] In particular, algebraic topology[97,154] has the unique advantage in the quantitatively description of intrinsic network properties. For instance, it has been used in the characterization of the underling functional networks generated from different brain regions based on the neural activity.[85] Topological invariants have also demonstrated great advantages in characterizing biomolecular structures, and have been used in DNA supercoiling,[176] Flp and Cre recombination,[195] assembly of virus capsids,[193,210] and the design of DNA origami.[52,178] In particular, knot invariants have been widely used in the modeling and analysis of DNA packing, recombination, and replication.[9,10,73]

Recently, persistent homology, a new branch of algebraic topology, has been proposed to bridge traditional topology and geometry, and provide a potentially revolutionary approach to complex biomolecular systems.[206] The essential idea is to introduce a filtration process and measure homology groups by their "lifespans" during the process.[44,72,232] Different from traditional topological models, the "lifespan" measurement provides a family of geometric characterizations of the topological invariants. Persistent homology has been successfully applied in the analysis of protein structure, flexibility, dynamics, and function.[35,78,216–218,222] With the ever-increasing data, the exceptional power of topology in dimensionality and complexity reduction has attracted enormous attention.[41,71] Topological data analysis (TDA) has been developed as one of the most promising apparatus for data science.[41,141] In particular, topological machine learning and deep learning models have delivered amazing results in drug design which is one of the most challenging fields in modern biology.[34,36–38,214] Based on topological representation and featurization, these (deep) learning models have achieved state-of-the-art results in the prediction of protein-ligand binding affinity, protein-protein binding free energy, and mutation-induced protein folding stability change. Topological models have consistently

delivered the best results in D3R Grand Challenges, a worldwide annual competition series in computer-aided drug design.[158,160] Their tremendous successes have demonstrated the great potential of topological models in data analysis and biological science.

The current paper offers a brief review of a new area called biomolecular topology. To avoid confusion, biomolecular topology in this paper refer to as the topological properties and relations that are emerged from biomolecular structures, dynamics, and functions. We discuss the challenging topological problems, methods, algorithms, and models originated from/for the biomolecular systems. In particular, our focus is protein topology, including protein structures, protein folding, protein complexes, and protein assemblies. We also discuss the related topological problems, including protein folding pathways, binding affinity predictions, assembly mechanisms, etc, and the various databanks related to these problems. The other focus of the paper is a general introduction of the related topological tools and models, including Gaussian network model, simplicial complexes, persistent homology, persistent Laplacian, de Rham-Homology, Yau-Hausdorff distance, and topology-based machine learning models.

The paper is organized as follows. Section 2 is devoted to biomolecular topological properties. We discuss topological problems in protein structures, protein folding, protein assembly, and DNA/RNA structures. A review of the topological models and topological data analysis is given discussed in Section 3. The paper ends with a conclusion.

2 Topology for biomolecules

Biological sciences are arguably one of the most important subjects that have fundamentally changed our society and world. During the past few decades, the research on biology has undergone a transition from phenomenological and descriptive sciences to quantitative and predictive sciences. Tremendous topological challenges and problems have arisen from biomolecular structures, dynamics, and functions. In this section, we provide a very brief introduction to these problems. Our focuses are protein structures, folding, and assembly, although DNA and RNA are discussed as well.

2.1 Protein structure

Protein is made up of the amino acid, which contains an amine group (NH_2), a carboxyl group (COOH), as well as a unique side chain (known as R group) specific to each of 20 amino acids. The NH_2 group from one amino acid can interact with the COOH group from another, to form a peptide bond. One H_2O molecule is generated in this process, which composes the H atom from NH_2 and the OH group from COOH. With these covalent peptide bonds, a chain or a sequence of amino acids, called a polypeptide, is formed and becomes the primary structure of the protein. However, in order for the protein to perform its function, the polypeptide has to be further arranged or folded into a three dimensional (3D) structure, i.e., the protein conformation. As illustrated in Fig. 1 (A), the amino acids within a segment of polypeptide chain can interact with each other to form two types of secondary structures, i.e., alpha-helix and beta-sheet. These secondary structure components will further fold into a more complicated 3D structure, known as the protein tertiary structure. It should be noticed that formation of the secondary and tertiary structures is largely due to the non-covalent interactions, such as hydrogen bonding, ionic interactions, van der Waals forces, hydrophobic interactions, and electrostatic interaction. However, these structures are often

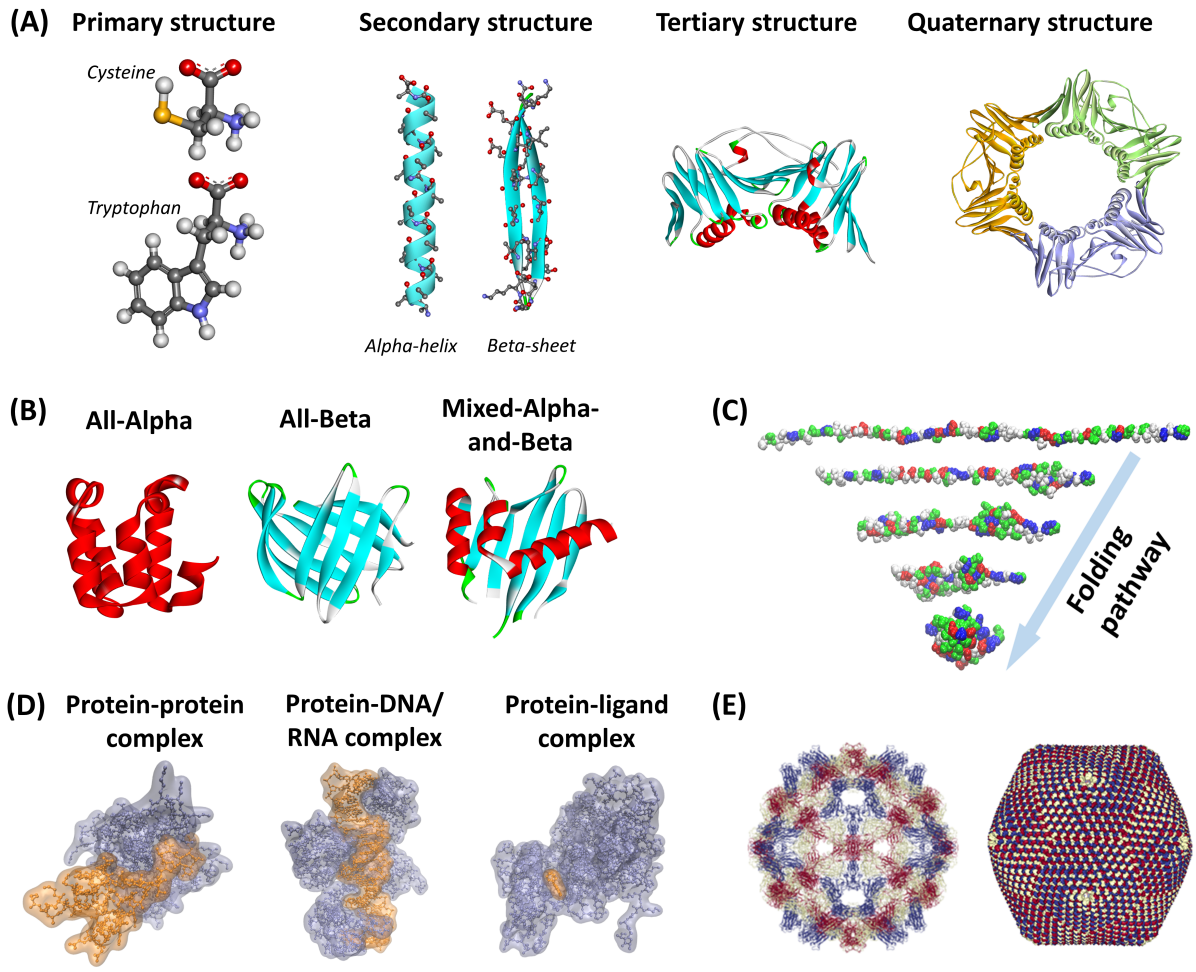


Figure 1: The topology for proteins. (A) A general review of protein structures, including primary structure, secondary structure, tertiary structure and quaternary structure. (B)The secondary-structure-based protein classification into three types, i.e., all-Alpha, all-Beta, mixed-Alpha-and-Beta. (C) The illustration of a simulated protein folding process by using steered molecular dynamics. (D) The illustration of three types of protein complexes, including protein-protein, protein-DNA/RNA, and protein-ligand. (E) The illustration of the icosahedra symmetry structures for two virus capsid structures (PDB IDs: 2WWS and 1M4X).

further stabilized by covalent disulfide bridges. Finally, if more than one polypeptide chains (or subunits) exist simultaneously in a protein-protein complex, it is usually called the protein quaternary structure.

Currently, there are more than 188,000 biomolecular structures in the PDB. The availability of this huge amount of biomolecular 3D structures brings a great opportunity for an in-depth understanding of the protein functions and mechanisms. The detailed comparison and classification is a key step in biomolecular data analysis. Two databanks, i.e., CATH and SCOPe (structural classification of proteins-extended), are built to classify all the available protein structures into hierarchical categories. In both databanks, protein structures are split into individual polypeptide chains. By comparison, structurally or functionally conserved polypeptide regions are identified and are defined as protein domains. The information of secondary structures, domains, and their topological structures are used in the classification. Specifically, in CATH, protein structures are classified into four levels, i.e., class (C), architecture (A), topology (T), and homologous superfamily (H), thus the name CATH. Proteins are divided according to their secondary structures into three types, i.e., all-alpha, all-beta, and mixed-alpha-and-beta. Figure 1 (B) illustrates these three types. The protein structures are classified into many architecture types, including roll, barrel, sandwich, ribbon, trefoil, propeller, clam, etc. The subdivision is mainly based on the secondary structure arrangement. Further description is given to protein fold topology in the biophysics sense. It focuses on the details of how the secondary structure elements are connected, that is the topology of the common structural features. Finally, the evolution information is considered and the highly conserved domains are called homologous at the homologous superfamily level. The SCOPe databank uses a similar way for structural classification.

2.2 Protein folding

Protein folding is a process in which randomly coiled polypeptides fold into their (unique) 3D structures. Figure 1 (C) illustrates a simulated protein folding process. Although exceptions have been found for Anfinsen's dogma,[7] most functional proteins are well-folded with a unique 3D structure. More importantly, misfolded or disordered proteins can usually cause serious diseases, including Mad-cow disease, Parkinson's disease, Alzheimer's disease, etc. In general, these diseases usually involve misfolded intermediate structures, such as Amyloid β, Tau protein, $\alpha-$synuclein, prions, etc. These misfolded structures can be then assembled into toxic oligomers with common amyloid folds.

So far, the prediction of the protein folding pathway remains to be a challenge both theoretically and computationally. In the folding funnel hypothesis, a well-folded protein structure has the global minimum of the Gibbs free energy. In contrast, misfolded configurations or intermediate folding structures, have higher free energies, thus they are kinetically favored but are thermodynamically unstable. Various experimental tools, such as atomic force microscopy, optical tweezers, and bio-membrane force probe, have been devised to study the unfolding force distribution, stable intermediates, and transitional non-native states. Computationally, steered molecular dynamics (SMD) is proposed to explore the inverse process of protein folding.[80,102,104] Despite the progress from experimental and computational works, the folding mechanism remains elusive. This is mainly due to the reason that protein folding is a highly complex and dynamic process, which involves a huge number of degrees of freedom.

Recently, machine learning and deep learning models have demonstrated great power in protein 3D structure prediction.[207] In particular, AlphaFold (and AlphaFold 2) has revolutionized the area of protein-structure prediction, and is widely regarded as one of milestone events in structural biology and biophysics.[110,192]

2.3 Protein complex

Proteins are involved in almost all biological processes. Although some proteins function independently, the vast majority of proteins form protein complexes with other molecules, including proteins, DNA, RNA, ions, and ligands, to achieve their biological functions. Figure 1 (D) illustrates the three types of protein complexes that exist widely. The formation of a protein complex is known as molecular recognition, and the components from the complex demonstrate molecular complementarity. Non-covalent interactions, such as hydrogen bonding, hydrophobic forces, van der Waals forces, $\pi - \pi$ interactions, electrostatic interaction, etc, as well as solvent effects, play the dominant role in the molecular recognition. Various databanks for 3D structures of these complexes are available, for instance, PDBbind (http://www.pdbbind.org.cn/), PDIdb (http://melolab.org/pdidb/web/content/home), and Protein-Protein Interaction Affinity Database (https://bmm.crick.ac.uk/ bmmadmin/Affinity/). The interaction networks from the protein complexes are also available in databanks, such as BioGRID (https://thebiogrid.org/) and STRING (https://string-db.org/).

Protein-DNA complex The protein-DNA complexes play a fundamental role in both the maintenance and regulation of genetic information. In a cell nucleus, the DNA sequence binds with proteins to form histones, which are further packed into highly complicated chromosome structures with the help of various scaffold proteins. The protein-DNA complexes also control and regulate various biological processes, including DNA transcription, DNA repair, DNA expression, etc.

Protein-ligand complex The protein-ligand complexes are important for biological signal transmission. The ligand-mediated signal transmission is essential to all life processes. Biologically, a ligand can bind to a certain specific site of a protein to activate (or deactivate) some biological functions. The binding affinity can vary greatly for different systems, and the binding proteins can have large conformational changes. The protein-ligand binding mechanism is a hot research area in drug discovery. The prediction of potential binding sites (hot spots) and binding affinities is of essential importance for efficient drug design and discovery.

Recently, machine learning-based models have demonstrated their great power in binding affinity prediction with a much higher accuracy than all traditional models[4,27,76,109,125,128,182,186,212]. These learning models can be classified into two classes, i.e., target-based approaches and ligand-based approaches. In target-based approaches, molecular descriptors from proteins and protein-ligand complexes are considered as the input for learning models. In contrast, ligand-based approaches usually make use of features only from ligands. The performance of both two classes highly relies on the selection of molecular descriptors. In fact, molecular descriptors or fingerprints are essential to all learning models for molecular data analysis in materials, chemistry, and biology.[136,166]

Molecular descriptors can be obtained from structural, chemical, physical, and biological properties. Molecular descriptors from structural properties can be one-dimensional (1D), two-

dimensional (2D), three-dimensional (3D), and four-dimensional (4D).[136,166] The 1D molecular descriptors include atom counts, bond counts, molecular weight, fragment counts, functional group counts, and other summarized general properties. The 2D molecular descriptors include topological indices, graph properties, combinatorial properties, molecular profiles, autocorrelation coefficients, etc. There are more than 5,000 types of 2D descriptors that are widely used in QSAR/QSPR models. The 3D molecular descriptors include molecular surface properties, volume properties, autocorrelation descriptors, substitute constants, quantum mechanical descriptors, etc. A related higher computational cost is usually required for the generation of 3D molecular descriptors. The 4D chemical descriptors are a series of 3D descriptors for configurations from a dynamic process.

Recently, topological data analysis (TDA)[72,232] and other advanced mathematics have been considered for biomolecular characterization and description[34,38,157] and achieved great success in various steps of drug design, including protein-ligand binding affinity prediction,[37–39,161,162] protein stability change upon mutation prediction,[34,36] toxicity prediction,[214] solvation free energy prediction,[196,197] partition coefficient and aqueous solubility,[215] binding pocket detection,[228] etc. These models have also demonstrated great advantages over traditional molecular representations in D3R Grand challenges.[158–160] Molecular descriptors can be combined together to form a large-sized vector, known as molecular fingerprints. Based on structural properties, there are various methods and models for the systematical generation of molecular fingerprints, including substructure key-based fingerprints,[69] path-based fingerprints,[95,163] circular fingerprints,[174] pharmacophore fingerprints,[121,184] and encoded fingerprints. Moreover, deep learning models, such as autoencoder, CNN, GNN, and Transformer, have also been used in molecular fingerprint generation.[59,70,146,171,211,221] The molecular descriptors and fingerprints are widely used in QSAR/QSPR models and machine learning models for material, chemical, and biological data analysis.

Protein-protein complex Protein-protein interactions (PPIs) play an essential role in a wide range of biological processes and mechanisms, including cell metabolism, signaling, protein transport, and immune system.[83,86] They can be affected by protein mutations and genetic variations, which may result in disease and drug resistance.[167] The understanding of PPIs, in particular PPIs upon mutations, is significant to various biomedical applications, including disease-associated mutation analysis, drug design, and therapeutic intervention.[83,86] Efficient computational methods and models have been developed for the evaluation PPI binding affinity, in particular, PPI binding affinity changes upon mutations ($\Delta\Delta G$). These models can be grouped into three categories, including molecular dynamic (MD)-based models, statistical energy-based models, and machine learning models. MD-based models, including FoldX,[93] Rosetta,[117] zone equilibration of mutants (ZEMu),[68] single amino acid mutation-induced changes in binding free energy (SAAMBE),[164] and others,[83] usually characterize the binding affinity of PPIs with various physical energy terms, including van der Waals interactions, electrostatic energies, hydrogen bonds, solvation energy, etc. Based on the contacts at atomic, residual, or coarse-grained levels, statistical-energy-based models have been proposed, such as BindProfX,[220] BeAtMuSiC,[64] contact potentials,[150] Profile-score,[187] and Dcomplex.[130] These models extract various intermolecular potentials from experimental structures to study PPI binding affinity. With the great advancements in data accumulation, learning models and

computational power, data-driven machine learning models have been developed and achieved the state-of-the-art results in PPI analysis.[179]

The fast-growing interests in PPI based machine learning models are mainly due to the development of various PPI databases in the past few decades, including Alanine scanning energetics database (ASEdb),[189] protein-protein interactions thermodynamic database (PINT),[120] structural kinetic and energetic database of mutant protein interactions (SKEMPI),[149] a database of binding affinity change upon mutations (DACUM),[81] antibody-bind database (AB-Bind),[180] protein-protein complex mutation thermodynamics (PROXiMATE),[108] kinetic and thermodynamic database of mutant protein interactions (dbMPIKT),[129] and SKEMPI 2.0.[106] SKEMPI 2.0 is the combination of several databases including SKEMPI, AB-Bind, PROXiMATE, and dbMPIKT with manually curated data from the literature. It contains a total of 7,085 mutations on various types of protein complexes, such as protease-inhibitor, antibody-antigen, and TRC-pMHC complexes. More specifically, there are about 3,000 single-point alanine mutations, about 2,000 single-point non-alanine mutations, and roughly 2,000 multi-point mutations.

With the ever-increasing PPI data, a great amount of data-driven learning models have been developed,[83,179] such as mCSM,[173] ELASPIC,[185] BindProf,[28] MutaBind,[227] iSEE,[82] MuPIPR,[230] ProAffiMuSeq,[107] GeoPPI,[134] and so on. In general, these data-driven models can be classified into two types, i.e., featurization-based machine learning models and end-to-end deep learning models. For the machine learning models, different types of PPI information from sequences, inter-residue interactions, evolutionary conservation, dynamic properties, energy terms, pharmacophore descriptors, structure-based descriptors, and others, are used as input features for machine learning models, such as support vector machine (SVM), random forest (RF), gradient boost trees (GBT), etc. Note that these input features are generated by using mathematical, physical, chemical, and biological models. For end-to-end deep learning models, proteins are usually represented as surfaces, graphs, or networks with embedded vectors or one-hot-vectors.[29,77] The intrinsic features for PPIs are automatically learned and implicitly represented in deep learning models. The most commonly used deep learning models for PPIs are graph neural networks (GNN) and geometric learning models. Even with the great advancements, generating highly efficient molecular featurization, which is key to the performance of learning models, is still a challenging problem.[136,166] Recently, advanced mathematical tools, in particular topological data analysis are used in molecular representation and featurization.[34,38,145,157] Their combination with learning models have achieved great success in various steps of drug design. In particular, the TopNetTree model has demonstrated great power in predicting protein-protein binding affinity changes upon mutations.[198] It outperforms all existing models and provides great insights for the SARS-CoV-2 mutations.[54,201]

A detailed summery of topology-based machine learning models can be found in Section 3.8.

2.4 Protein assembly

Protein assembly, which is the protein quaternary structure, is an essential functional form of macromolecules. Biologically, one or several kinds of proteins can assemble under certain symmetric rules into a macromolecular structure. Among the protein assembly structures, the one that has attracted the most attention from mathematicians is virus capsid, which is

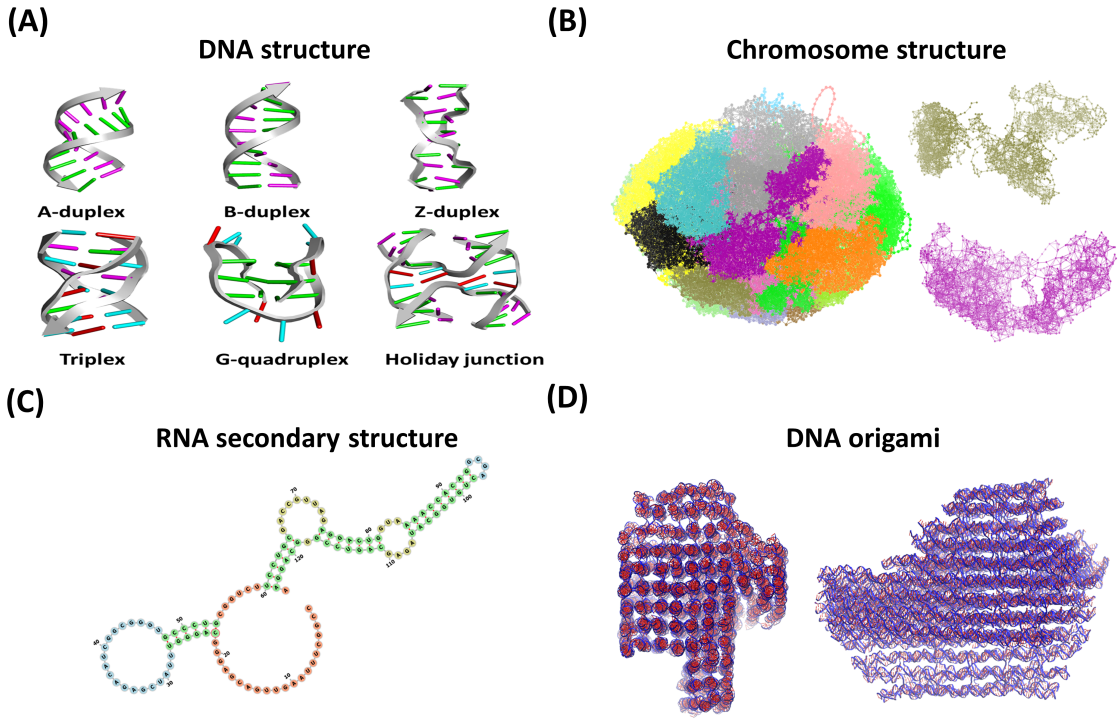


Figure 2: The illustration of DNA/RNA topology. (A) The six types of DNA structures, including A-duplex, B-duplex, Z-duplex, triplex, G-quadruplex, and Holiday junction. (B) The simulated chromosome structures for mammalian genome. (C) A simulated RNA secondary structure by using ViennaRNA. (D) Illustration of a DNA origami (PDB ID:4V5X).

the virus shell structure made from proteins.

Protein assembly is an important topic in mathematical virology, which is a new mathematical area for the study of virus capsid structures and their assembly mechanism[194] and mathematical modeling of viral evolution, transmission, and their impacts on diagnostics, small molecular drugs, antibody therapies, and vaccines.[53,54,199,200] Historically, it has been found that icosahedral symmetry occurs predominantly in virus structures.[226] As illustrated in Figure 1 (E) , viruses usually adopt the icosahedral symmetry for their capsid shell. Proteins on the capsid are highly ordered and aggregated in clusters (or capsomeres) of three, five and six. To explain the organizational principles, Caspar and Klug propose a series of polyhedra that encode the locations of the proteins.[45] Their model is used to reconstruct and classify the viral capsid structures from the experimental data. Recently, Twarock has proposed viral tiling theory by using the Coxeter group theory.[193,210] Other than the family of polyhedra in the Caspar-Klug theory, the viral tiling model can generate a new finite series of polyhedra, called the triacontahedral series, which significantly enhance the performance of traditional models.

2.5 DNA/RNA topology

DNAs and RNAs are molecular repositories of genetic information. Their structures have various interesting topological properties. Other than the common double helix structures, which are usually found in the B-form of DNA, various other forms exist, including A-helix, Z-helix, Triplex, G-quadruplex and Holiday Junction. Figure 2(A) illustrates these different topological structures. Various databases are constructed for the study of DNA structures and their interactions, including nucleic acid databank (http://ndbserver.rutgers.edu/), nucleic acid-ligand database (NALDB)(http://bsbe.iiti.ac.in/bsbe/naldb/HOME.php), and G-quadruplex ligands database (http://www.g4ldb.org/ci2/index.php).

Chromosomes are the physical realization of genetic information. It is composed of DNA-protein complexes and is one of important cellular entities.[24,100] Chromosome is essential for supporting essential biological functions, such as DNA replication, transcription, repair of DNA damage, chromosome translocation, and others.[48,51] More interestingly, chromosomes have highly complicated hierarchical structures, which can be studied by chromosome conformation capture techniques.[63,65,126] Essentially, a contact frequency matrix can be generated and a 3D structure can be reconstructed. Figure 2(B) illustrates the chromosome structure from the mammalian genome.[183]

There are various types of RNAs, including ribosomal RNA(rRNA), messenger RNA(mRNA), transfer RNA(tRNA), and others. Compared with DNAs, RNAs are smaller in their sizes and tend to fold into complicated secondary structures. The understanding of how an RNA sequence folds into a specific secondary structure is key to RNA engineering. Under the energy minimization hypothesis, an RNA sequence will fold into its configuration only when the loop region energies are minimized and their stacked pairs are maximized. Various algorithms and combinatorial models are developed to solve the RNA secondary structure design problem.[98,99,177] An example of RNA secondary structure is demonstrated in Figure 2(C). It is generated by using ViennaRNA Web Service (http://nibiru.tbi.univie.ac.at/forna/).

Other than the generic information, DNAs and RNAs have been used as synthetic materials. Essentially, the sequences of DNAs and RNAs can be specially designed so that they can be engineered into some unique 3D structures. In particular, DNA nanotechnology can be used to assemble nanoscale DNA structures with various structures. One of the most promising DNA nanotechnologies is DNA origami. The transformation of a flat sheet of paper into a highly ordered structure through folding and sculpting techniques is called origami. In DNA origami, specially-designed long DNA strands are folded into highly complicated 3D structures with the help of short staple DNA strands. An example of a DNA origami (PDB ID: 4V5X) is demonstrated in Figure 2(D). Highly complicated topological structures can be achieved by the careful design of DNA sequence information.[52,178]

3 Topological models for biomolecules

Topological representations and modeling are of vital importance for the analysis of biomolecular data. In general, a biomolecule can be characterized from different topological perspectives based on its intrinsic properties that we are interested in. For instance, if one wants to know the knot or supercoil properties of proteins or DNAs, one can consider their backbone structures and topologically characterize them as continuous strings. If one wants to perform a

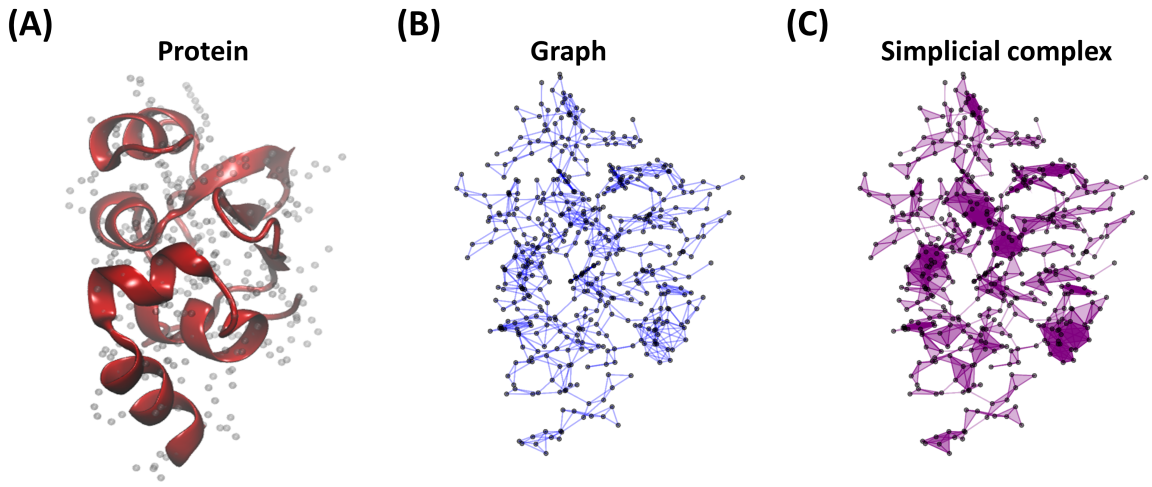


Figure 3: Topological representation of a protein. (A) The chain representation of a protein. The coil chain is composed of Beta-sheets and Alpha-complexes. (B) The graph representation of a protein. (C) The simplicial complex representation of a protein.

molecule dynamic simulation, one must consider all the covalent and non-covalent bonds between the atoms. Mathematically, various topological models can be used in biomolecular data analysis and prediction, including graphs, simplicial complexes, hypergraphs, homology, spectral, Hodge-Laplacian, and others. A briefing introduction of these models is given in this section.

3.1 Graph and network

Graph or network models are the most widely used models in biomolecular representations among all topological representations,.[11,12,105,112,124] For instance, elastic network models, including Gaussian network model (GNM), anisotropic network model (ANN), and others, are popular tools for the study of biomolecular flexibility and normal modes.[61] In molecular dynamic models, biomolecular networks are constructed based on atomic covalent bonds. Other than the characterization of intramolecular interactions, graphs and networks are also key models for the characterization of intermolecular interactions. Figure 3 (B) illustrates a graph representation of a protein.

Mathematically, one can build up various types of matrices based on a graph, and use them to study molecular topological information. One of them is the Laplacian matrix L. For instance, in GNM, for a protein with N number of C_α atoms, an N by N Laplacian matrix can be constructed as,

$$L_{ij} = \begin{cases} -1 & i \neq j \text{ and } r_{ij} \leq r_c \\ 0 & i \neq j \text{ and } r_{ij} > r_c \\ -\sum_{i \neq j}^{N} L_{ij} & i = j \end{cases}. \tag{1}$$

Here r_{ij} is the distance between ith- and jth-atoms, and r_c is a predefined cutoff distance. In GNM, if two atoms are within the cutoff distance, they are assumed to be "connected".

The spectral information from the Laplacian matrix is used for the characterization of protein flexibility.[61]

Another important connection matrix is Hessian matrix.[61] If one defines the coordinate of the ith-atom as (x_i, y_i, z_i), a $3N \times 3N$ Hessian matrix can be constructed with local 3×3 off-diagonal matrix H_{ij} as,

$$H_{ij} = -\frac{1}{r_{ij}^2} \begin{bmatrix} (x_j - x_i)(x_j - x_i) & (x_j - x_i)(y_j - y_i) & (x_j - x_i)(z_j - z_i) \\ (y_j - y_i)(x_j - x_i) & (y_j - y_i)(y_j - y_i) & (y_j - y_i)(z_j - z_i) \\ (z_j - z_i)(x_j - x_i) & (z_j - z_i)(y_j - y_i) & (z_j - z_i)(z_j - z_i) \end{bmatrix} \\ i, j = 1, 2, \cdots, N, i \neq j \text{ and } r_{ij} \leq r_c. \tag{2}$$

The diagonal part is the negative summation of the off diagonal elements:

$$H_{ii} = -\sum_{i \neq j} H_{ij}, \ \forall i = 1, 2, \cdots, N. \tag{3}$$

Hessian matrix is widely used in biomolecular normal mode analysis.[11,61]

3.2 Simplicial complex

Simplicial complex is a very important concept in topology.[62,154] Recently, simplicial complexes have been used in the study of data, in particular, the characterization of complex connection information within/between data.[119] Simplicial complex is able to provide richer information than graph models, which only characterize pair-wise interactions. Computationally, one can construct simplicial complexes from various types of data, including point-clouds, matrices, volumetric functions, networks, graphs, and others. Simplicial complex is one of the essential mathematical tools for data sciences and computer sciences.

Geometrically, 0-simplex means a point, 1-simplex means a line segment, 2-simplex means a triangle, 3-simplex means a tetrahedron, etc. An n-simplex can be regarded as a polyhedron spanned by $n+1$ geometrically independent points (they are not in any hyperplane of dimension n) in the Euclidean space $\mathbb{R}^n$. In mathematical language, a *(geometric) n-simplex* spanned by $n+1$ geometrically independent points $a^0, \ldots, a^n$ is given by

$$\sigma^n = \{\sum_{i=0}^{n} t_i a^i | 0 \leq t_i \leq 1, \sum_{i=0}^{n} t_i = 1\} \subseteq \mathbb{R}^n.$$

Simplices are the basic units of a simplicial complex, and the construction of a simplicial complex from a graph or dataset usually starts with simplices. A *simplicial complex* K in the Euclidean space $\mathbb{R}^n$ is a collection of simplices in $\mathbb{R}^n$ such that

(i) Every face of a simplex of K is in K;

(ii) The intersection of any two simplices of K is either empty, or a common face of them.

Another equivalent description of simplicial complexes is the abstract simplicial complex. Let V be an ordered set. An *abstract simplicial complex* $\mathcal{K}$ is a collection of finite nonempty subsets of V, such that if σ is an element in $\mathcal{K}$, so is every nonempty subset of σ. The abstract

simplicial complex is a combinatorial version of simplicial complex. Figure 3 (C) illustrates a simplicial complex representation for a protein.

One can obtain different simplicial complexes, such as neighborhood complexes, Docker complexes, and Hom complexes from a graph. This gives the possibility of using topological methods to deal with problems in graph theory. The Kneser conjecture was proved in 1978 by L. Lovász, who used the connectivity of neighborhood complexes to give the lower bound of the chromatic number of Kneser graphs.[115]

Example 3.1. Cěch complex and Vietoris-Rips complex.
Given a collection of points $V = \{a^0, a^1, a^2, \dots, a^i, \dots\}$ in Euclidean space $\mathbb{R}^n$, the Cěch complex[44] C_ϵ is the abstract simplicial complex whose k-simplices are determined by unordered $(k+1)$-tuples of points in V whose closed $\epsilon/2$-ball neighborhoods have a point of common intersection.

Given a collection of points $V = \{a^0, a^1, a^2, \dots, a^i, \dots\}$ in the Euclidean space $\mathbb{R}^n$, the Vietoris-Rips complex[41,71] $\mathcal{R}_\epsilon$ is the abstract simplicial complex whose simplices are the set of points in V which are pairwise within distance ϵ.

Example 3.2. Neighborhood complex.
Let G be a graph. The neighborhood complex of G is the (abstract) simplicial complex $\mathcal{N}(G)$ defined as follows: its vertices are all non-isolated vertices of G, and its simplices are all the subsets of $V(G)$ that have a common neighbor.

Example 3.3. Path complex.
A generalization of the notion of simplicial complex is the path complex for directed graphs.[87–89] Let V be a finite set. For any integer $n \geq 0$, an *elementary n-path* is a sequence $v_0 v_1 \cdots v_n$ of vertices in V. A *path complex* on a nonempty finite set V is a nonempty collection $\mathcal{P}$ of elementary paths on V such that if $v_0 v_1 \cdots v_n \in \mathcal{P}$ then also $v_0 v_1 \cdots v_{n-1}$ and $v_1 v_2 \cdots v_n$ belong to $\mathcal{P}$. Let $\mathcal{K}$ be an abstract simplicial complex defined over the finite ordered set V. Each simplex $\sigma \in \mathcal{K}$ can be regarded as an ordered sequence with vertices in V. Recall that the face maps of the simplicial complex $\mathcal{K} = \{\mathcal{K}_n\}_{n\geq 0}$ are defined by

$$d_i : \mathcal{K}_n \to \mathcal{K}_{n-1}, \quad d_i\{v_0, \dots, v_n\} = \{v_0, \dots, v_{i-1}, v_{i+1}, \dots, v_n\}$$

for $\{v_0, \dots, v_n\} \in \mathcal{K}_n$ and $i = 0, 1, \dots, n$. If $\sigma \in \mathcal{K}_n$, then $d_0\sigma, d_n\sigma \in \mathcal{K}_{n-1}$. This implies that a simplicial complex is indeed a path complex.

Let $G = (V, E)$ be a simple digraph, that is, a directed graph having no multiple edges or graph loops. An *allowed p-path* on G is a sequence

$$v_0 v_1 \cdots v_n, \quad v_0, v_1, \dots, v_n \in V$$

such that $(v_{i-1}, v_i) \in E, i = 1, 2, \dots, n$. Denote $\mathcal{P}_n(G)$ the set of all allowed p-paths. In particular, we have $\mathcal{P}_0 = V$ and $\mathcal{P}_1 = E$. It can be directly verified that the family $\{\mathcal{P}_n\}$ of all allowed paths is a path complex. This path complex, denoted by $\mathcal{P}(G)$, is the path complex of the digraph G.

3.3 Homology group

The homology group is a fundamental homotopy invariant.[97,153] In applications, homology groups are more computable relative to other homotopy invariants. The success of persistent homology in data analysis and machine learning is due to the characteristics of homology groups. Firstly, as a topological invariant, the homology group captures the overall structure and stable features of data. Secondly, homotopy groups can describe the date-sets which are independent of metric. However, homotopy groups are flexible as features, a more feasible method is the persistent homology which will be introduced later.

Let G be an abelian group. Recall that a *p-chain* on a simplicial complex K is the Abelian group generated by the p-simplices in K with coefficients in G, denoted by $C_p(K;G)$. Each element in $C_p(K;G)$ can be written as

$$x = \sum_{i\in I} g_i\sigma_i, \quad g_i \in G, \sigma_i \in K_p,$$

where I is a finite set and K_p is the set of p-simplices in K. A chain group of K is a collection of p-chains $\{C_p(K;G)\}_{p\geq 0}$.

Let $V(K)$ be a ordered set of the vertices in a simplicial complex K. Then each p-simplex of K can be written as

$$\sigma_p = \{v_0, \ldots, v_p\}, \quad v_i \in V(K), \quad v_0 < v_1 < \cdots < v_p,$$

and we sometimes write $\sigma_p = v_0 \cdots v_p$ for convenience. The face maps are defined by

$$\begin{aligned} d_i : K_p &\to K_{p-1}, \\ v_0 \cdots v_p &\mapsto v_0 \cdots v_{i-1} v_{i+1} \cdots v_p \end{aligned}$$

for $0 \leq i \leq p$. We have a *boundary operator*

$$\partial_p : C_p(K;G) \to C_{p-1}(K;G)$$

given by $\partial_p\sigma_p = \sum_{i=0}^{p}(-1)^i d_i\sigma_p$. It can be verified that $\partial_{p-1}\partial_p = 0$. Then there is a long sequence

$$\cdots \to C_p(K;G) \xrightarrow{\partial_p} C_{p-1}(K;G) \xrightarrow{\partial_{p-1}} \cdots \xrightarrow{\partial_1} C_0(K;G) \xrightarrow{\partial_0} 0,$$

which is called the *simplicial chain complex* of K, denoted by $\{C_p(K;G), \partial_p\}_{p\geq 0}$. The *$p$-th homology group of K* is defined as the quotient group

$$H_p = \ker \partial_p / \text{im } \partial_{p+1},$$

where $\ker \partial_p = \{x \in C_p(K;G) | \partial_p(x) = 0\}$ and $\text{im } \partial_{p+1} = \{\partial_{p+1}x | x \in C_{p+1}(K;G)\}$. The p-th Betti number β_p of K is the rank of the p-th homology group of K.

Note that the 0-th homology group of K describes the number of path connected components of K, i.e.,

$$H_0(K;G) \cong \overbrace{G \oplus \cdots \oplus G}^{r},$$

where r is he number of path connected components of K. It is obvious that $\beta_0(K) = r$. Especially, if K is path connected, we have

$$H_0(K; G) \cong G.$$

The homology group is a topological invariant that can describe the intrinsic connection information of the data. However, sometimes the features of data described by homology groups are too flexible and rough. Besides, the simplicial complex obtained from a data-set also depends on the scale of our observation. To solve the problem, persistent homology that can describe the topological structures at different scales has been proposed.

Graph (or network) data play an important role in data sciences. Graph-based topology models have attracted great attentions recently. Among them is the path homology model.

Let $\mathcal{P}$ be a path complex over a set V. Let $\mathbb{F}$ be a field. We denote Λ_n the $\mathbb{F}$-linear space generated by all the elementary n-paths, that is,

$$\Lambda_n = \{ \sum_{v_0,\dots,v_n \in V} a^{v_0 v_1 \cdots v_n} e_{v_0 v_1 \cdots v_n} | a^{v_0 v_1 \cdots v_n} \in \mathbb{F} \}.$$

Here $\{e_{v_0 v_1 \cdots v_n}, v_0, \dots, v_n \in V\}$ is the basis of Λ_n. Then $\Lambda = \{\Lambda_n\}$ is a chain complex with the differential $\partial_n : \Lambda_n \to \Lambda_{n-1}$ given by

$$\partial_n e_{v_0 v_1 \cdots v_n} = \sum_{k=0}^{n} (-1)^k e_{v_0 \cdots \hat{v}_k \cdots v_n}, \quad n \geq 1,$$

where $v_0 \cdots \hat{v}_k \cdots v_n$ means omission of the index v_k. For $n = 0$, we set $\partial_0 e_v = 0, e_v \in \Lambda_0$. Now, we denote $\mathcal{A}_n$ the $\mathbb{F}$-linear space generated by all the elementary n-paths from $\mathcal{P}$, that is,

$$\mathcal{A}_n = \mathcal{A}_n(\mathcal{P}) = \{ \sum_{v_0,\dots,v_n \in V} a^{v_0 v_1 \cdots v_n} e_{v_0 v_1 \cdots v_n} | v_0 v_1 \cdots v_n \in \mathcal{P}_n, a^{v_0 v_1 \cdots v_n} \in \mathbb{F} \}.$$

It is obvious that $\mathcal{A} = \{\mathcal{A}_n\}$ is a subspace of Λ. However, $\mathcal{A}$ does not have to be a chain complex. For example, let $G = (V, E)$ with $V = \{0, 1, 2\}$ and $E = \{(0, 1), (1, 2)\}$. Then

$$\mathcal{A}_0 = \text{span}\{e_0, e_1, e_2\}, \quad \mathcal{A}_1 = \text{span}\{e_{01}, e_{12}\}, \quad \mathcal{A}_2 = \text{span}\{e_{012}\}.$$

Note that $\partial e_{012} = e_{01} - e_{02} + e_{12}$. But $e_{02} \notin \mathcal{A}_1$. One motivation is to get a chain complex from $\mathcal{A}$. Let

$$\Omega_0 = \Omega_0(\mathcal{P}) = \mathcal{A}_0, \quad \Omega_n = \Omega_n(\mathcal{P}) = \{\sigma \in \mathcal{A}_n | \partial\sigma \in \mathcal{A}_{n-1}\}, \quad n \geq 1.$$

It can be verified that $\partial\Omega_n \subseteq \Omega_{n-1}$. Then $\Omega = \{\Omega_n\}$ is a chain complex with differential $\partial : \Omega_n \to \Omega_{n-1}$. The path homology of the path complex $\mathcal{P}$ is defined as

$$H_n(\mathcal{P}) = H_n(\Omega_*(\mathcal{P})) = \frac{\ker \partial|_{\Omega_n}}{\text{im}\partial|_{\Omega_{n+1}}}.$$

In particular, if $\mathcal{P}$ is a simplicial complex, then the path homology of $\mathcal{P}$ coincides with the usual homology of $\mathcal{P}$. One interesting interpretation of path homology is as follows. By abuse

of language, it can be verified that Ω is the largest chain complex contained in $\mathcal{A}$. Let Ξ be the minimal chain complex containing $\mathcal{A}$ in Λ. Then we have

$$H_n(\Omega_*) \cong H_n(\Xi_*), \quad n \geq 0.$$

The path homology is an important invariant of digraphs which shows many excellent properties. The reader may refer to the works by A. Grigor'yan, Yu. Muranov, and S.-T. Yau et al[87,90–92] for more details on path homology.

3.4 Persistent homology

Persistent homology has been used to characterize biomolecular structures, flexibility, dynamics, and functions.[35,44,78,216–218,222] Essentially, persistent homology provides a representation that retains the geometric information of the topological invariants, so that it works as a bridge between geometry and topology. It captures the intrinsic topological structure properties and discards all the other irrelevant information. For data analysis, persistent homology delivers a topological simplification of the high-dimensional and highly complicated data.

The filtration is one of the fundamental objects to build persistent homology. Let $\mathbb{R}^*$ be the category of real number with morphisms given by $a \to b$ for $a \leq b$. Commonly, a *filtration (of simplicial complexes)* is a functor $\mathcal{F} : \mathbb{R}^* \to \mathbf{Simp}$ from the category of real numbers to the category of simplicial complexes satisfying

$$\mathcal{F}(a) \subseteq \mathcal{F}(b), \quad \text{for } a \leq b.$$

Two typical examples of filtration are the Cěch complexes and the Vietoris-Rips. For example, the filtration Vietoris-Rips complex is given by $\mathcal{F}(\epsilon) = \mathcal{R}_\epsilon$ for $\epsilon \in \mathbb{R}$. It is obviously that $\mathcal{R}_\epsilon \hookrightarrow \mathcal{R}_{\epsilon'}$ is an inclusion of simplicial complexes for any $\epsilon \leq \epsilon'$.

The persistence module is a classical tool to describe persistent homology.[232] Let R be a commutative ring with unit. A *persistence module* is a functor $\mathcal{M} : \mathbb{R}^* \to \mathbf{Mod_R}$ from the category of real numbers to the category of R-modules. More precisely, a persistence module is a family of R-modules $\{M_p\}_{p\in\mathbb{R}}$ together with R-module morphisms $M_p \to M_q$ for $p \leq q$ such that:

(i) $f_{p\to p} = \mathrm{id}$;

(ii) $f_{q\to r} \circ f_{p\to q} = f_{p\to r}$ for $p \leq q \leq r$.

Let $\mathcal{F} : \mathbb{R}^* \to \mathbf{Simp}$ be a filtration. For real numbers $a \leq b$, we have an inclusion of simplicial complexes

$$\mathcal{F}(a) \hookrightarrow \mathcal{F}(b),$$

it induces a morphism of homology groups

$$H_*(\mathcal{F}(a); \mathbb{F}) \to H_*(\mathcal{F}(b); \mathbb{F})$$

with coefficients $\mathbb{F}$. It can be verified that $H_*(\mathcal{F}(-); \mathbb{F})$ is a persistence module. The (a, b)-*persistent homology* of $\mathcal{F}$, denoted by $H_*^{a\to b}(\mathcal{F})$, is the image of the induced map $H_*(\mathcal{F}(a); \mathbb{F}) \to H_*(\mathcal{F}(b); \mathbb{F})$. The (a, b)-persistent homology characterizes the generators that

are born at time a and survive to time b. Compared with the usual homology, persistent homology gives rigid and computable features of data. Recently, multidimensional or multi-parameter persistent homology have been extensively studied.[123] The essential idea is to consider multidimensional filtration and persistence modules.

The persistent path homology[56] is another new variant of the persistent homology theory. Let $\mathfrak{D}$ be the category of digraphs, the morphisms are given by the digraph maps. A *filtration of digraphs* is a functor $\mathcal{G} : \mathbb{R}^* \to \mathfrak{D}$ from the category of real numbers to the category of digraphs, that is,

(i) $\mathcal{G}_a \in \mathfrak{D}$;

(ii) Denote $f_{a\to b} : \mathcal{G}_a \to \mathcal{G}_b$ for $a \leq b$. Then $f_{b\to c} \circ f_{a\to b} = f_{a\to c}$ for $a \leq b \leq c$.

The (a, b)-*persistent path homology* of $\mathcal{G}$, denoted by $H_*^{a\to b}(\mathcal{G})$, is the image of the map $H_*(\mathcal{G}(a); \mathbb{F}) \to H_*(\mathcal{G}(b); \mathbb{F})$ induced by $a \to b$. Here, $\mathbb{F}$ is the coefficient field. More generally, let $\mathcal{F} : \mathbb{R}^* \to \mathbf{Path}$ be a *filtration of path complexes*, that is, a functor from the category of real numbers to the category of path complexes. Then the (a, b)-*persistent path homology* of $\mathcal{F}$ is defined to be

$$\mathrm{im}(H_*(\mathcal{F}(a); \mathbb{F}) \to H_*(\mathcal{F}(b); \mathbb{F})),$$

where $H_*(\mathcal{F}(a); \mathbb{F}) \to H_*(\mathcal{F}(b); \mathbb{F})$ is induced by $a \to b$.

Example 3.4 (Edge-based filtration)**.** Given a weighted directed network, that is, a digraph $G = (V, E)$ with a weight function $w : E \to \mathbb{R}$. Let $E_a = \{(u, v) \in E | w(u, v) \leq a\}$, and let $G_a = (V, E_a)$. Then $\{G_a\}_{a\in\mathbb{R}}$ gives a filtration of digraphs. The persistent path homology of (G, w) is given by $H_*^{a\to b}(\mathcal{G}) = \mathrm{im}(H_*(\mathcal{G}_a; \mathbb{F}) \to H_*(\mathcal{G}_b; \mathbb{F}))$.

Example 3.5 (Path-based filtration)**.** Let P be a filtration of path complexes, and let w be a weighted function $w : P \to \mathbb{R}$. Then $P_a = \{x \in P | w(x) \leq a\}$ defines a filtration of path complexes. The persistent path homology of (P, w) is given by $H_*^{a\to b}(\mathcal{F}) = \mathrm{im}(H_*(\mathcal{F}_a; \mathbb{F}) \to H_*(\mathcal{F}_b; \mathbb{F}))$.

The filtration process is vital to all the persistent homology models. Other than the above approaches, network-based filtration models can be obtained from time-dependent network growing processes. With the strong connections to graph data, the persistent path homology will have great potential in various applications.

3.5 Persistent Laplacian

Persistent Laplacian is a recent promotion of topological data analysis derived from a series of methods such as spectral graph theory, circuit theory and persistent homology together with a topological view of the combinatorial Laplacian on simplicial complexes.[144]

In spectral graph theory,[57] the graph Laplacian plays a fundamental role in optimization problems on graphs,[122,137,156] the efficient solution of equations[118,135,181] and the network circuit theory.[23,67,138] Coincidentally, the Laplacian is not only a typical operator on graph but also an important concept in the de Rham Hodge theory on manifolds.[26] The Laplacian on de Rham complex is derived from the coboundary operators and the inner product on

differential forms. Lim discussed Hodge Laplacians on graphs.[127] The similarity and difference between Hodge Laplacians and graph Laplacians were studied[169] Omitting the geometry, the natural idea is to consider a (co)chain complex with a specific inner product structure, which leads us to the combinatorial Laplacian. From the perspective of topology, the combinatorial Laplacian could appear in the studying of various objects such as graphs, manifolds, simplicial complexes or even hypergraphs.

In recent years, the combinatorial Laplacian has been widely concerned and studied in theory and application.[94,96,116,139,147] In the de Rham Hodge theory, the Laplacian determines the de Rham cohomology of manifolds, which indicates the potential to follow the persistence of Laplacians. Recently, researchers begin to pay attention to persistent Laplacians.[144] However, the ideas were outlined in earlier papers[55,202] in 2019. Compared to persistent homology, persistent Laplacians offer additional nonharmonic eigenfunctions and eigenvlues for data analysis. Let K be a simplicial complex, and let $C_*(K;\mathbb{R})$ be the chain group of K. We fix an inner product

$$\langle \cdot, \cdot \rangle : C_*(K;\mathbb{R}) \times C_*(K;\mathbb{R}) \to \mathbb{R}$$

on $C_*(K;\mathbb{R})$. Note that the inner product on $C_*(K;\mathbb{R})$ always exists, since we can at least take

$$\langle \sigma, \tau \rangle = \begin{cases} w_\sigma \in \mathbb{R}^+, & \text{if } \sigma = \tau; \\ 0, & \text{otherwise.} \end{cases}$$

Let $\partial_p^* : C_{p-1}(K;\mathbb{R}) \to C_p(K;\mathbb{R})$ be the adjoint operator of the operator $\partial_p : C_p(K;\mathbb{R}) \to C_{p-1}(K;\mathbb{R})$. The *p-th combinatorial Laplacian* of K is defined by

$$\Delta_p^K = \partial_{p+1} \circ \partial_{p+1}^* + \partial_p^* \circ \partial_p.$$

In particular, we consider the case $\langle \sigma, \tau \rangle = \delta_{\sigma\tau}$, or equivalently, the function

$$w : K \to \mathbb{R}^+$$

above by taking $w_\sigma = 1$ for each $\sigma \in K$. Let B_p be the representation matrix of ∂_p respect to fixed bases of C_p and C_{p-1} with a given order. In this case, the coboundary operator $\delta^p : C^{p-1}(K;\mathbb{R}) \to C^p(K;\mathbb{R})$ induced by ∂_p is consistent with the adjoint operator $\partial_p^* : C_{p-1}(K;\mathbb{R}) \to C_p(K;\mathbb{R})$, i.e., the representation matrix of ∂_p^* is exactly B_p with respect to the dual bases of C_p and C_{p-1}. Moreover, the representation matrix of the p-th combinatorial Laplacian is given by

$$\mathcal{L}_p = B_{p+1} B_{p+1}^T + B_p^T B_p.$$

Similar to persistent homology, the persistent Laplacian works on a filtration of simplicial complexes. Let $\mathcal{F} : \mathbb{R}^* \to \mathbf{Simp}$ be a filtration of simplicial complexes of finite type. For real numbers $a < b$, we have an inclusion

$$\mathcal{F}(a) \hookrightarrow \mathcal{F}(b),$$

which induces a morphism of chain complexes

$$C_*(\mathcal{F}(a),\mathbb{R}) \hookrightarrow C_*(\mathcal{F}(b),\mathbb{R}).$$

Let $\mathcal{F}(\infty) = \bigcup_{a\in\mathbb{R}} \mathcal{F}(a)$ and $C_* = C_*(\mathcal{F}(\infty), \mathbb{R})$. We endow an inner product $\langle\cdot,\cdot\rangle$ on C_*. As a subspace of C_*, the chain complex $C_*(\mathcal{F}(a), \mathbb{R})$ can inherit an inner product structure from C_* and a boundary operator given by the restriction

$$\partial_p^a = \partial_p|_{C_p(\mathcal{F}(a),\mathbb{R})} : C_p(\mathcal{F}(a), \mathbb{R}) \to C_{p-1}(\mathcal{F}(a), \mathbb{R}).$$

Here, ∂_* is the boundary operator on C_*. For convenience, we write $C_p^a = C_p(\mathcal{F}(a), \mathbb{R})$. Denote the subspace

$$C_p^{a,b} = \{x \in C_p^b | \partial_p^b x \in C_{p-1}^a\}$$

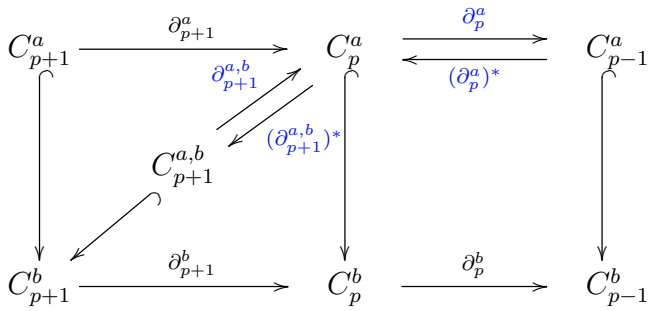
the preimage of C_{p-1}^a in C_{p-1}^b under ∂_p^b. Then we have a linearly operator

$$\partial_p^{a,b} = \partial_p^b|_{C_p^{a,b}} : C_p^{a,b} \to C_{p-1}^a,$$

which induces an adjoint operator

$$(\partial_p^{a,b})^* : C_{p-1}^a \to C_p^{a,b}$$

with respect to the inner product $\langle\cdot,\cdot\rangle$. Consider the following diagram:



The *p-th persistent Laplacian*[144] is defined by

$$\Delta_p^{a,b} = \partial_{p+1}^{a,b} \circ (\partial_{p+1}^{a,b})^* + (\partial_p^a)^* \circ \partial_p^a.$$

In particular, if $a = b$, we have $\Delta_p^{a,b} = \Delta_p^a$, which is exactly the combinatorial Laplacian on C_p^a. The following result shows that the persistent Laplacians indicate persistent Betti numbers of persistent homology.

Theorem 3.6. [144] *For each integer $p \geq 0$, we have that $\beta_p^{a,b} = \text{nullity}(\Delta_p^{a,b})$. Here, $\text{nullity}(\Delta_p^{a,b})$ denotes the number of zero eigenvalues of $\Delta_p^{a,b}$.*

The persistence of the operator $\Delta_p^{a,b}$ is worth considering for many reasons. An interesting story is the persistent Hodge decomposition theorem of combinatorial version.

Theorem 3.7 (**Persistent Hodge decomposition theorem**). *Let $\mathcal{F} : \mathbb{R}^* \to$ **Simp** be a filtration of simplicial complexes of finite type. Then*

$$C_p^a(\mathcal{F}) = \ker \Delta_p^{a,b} \oplus \text{im}\partial_{p+1}^{a,b} \oplus \text{im}(\partial_p^a)^*.$$

Proof. We will first prove

$$\ker \Delta_p^{a,b} = \ker(\partial_{p+1}^{a,b})^* \cap \ker \partial_p^a. \tag{4}$$

Indeed, it can be directly verified that $\ker \Delta_p^{a,b} \supseteq \ker(\partial_{p+1}^{a,b})^* \cap \ker \partial_p^a$. On the other hand, if $\omega \in \ker(\Delta_p^{a,b})$, then we have

$$\begin{aligned}
0 =& \langle \Delta_p^{a,b}\omega, \Delta_p^{a,b}\omega\rangle \\
=& \langle \partial_{p+1}^{a,b} \circ (\partial_{p+1}^{a,b})^*\omega + (\partial_p^a)^* \circ \partial_p^a\omega, \partial_{p+1}^{a,b} \circ (\partial_{p+1}^{a,b})^*\omega + (\partial_p^a)^* \circ \partial_p^a\omega\rangle \\
=& \langle \partial_{p+1}^{a,b} \circ (\partial_{p+1}^{a,b})^*\omega, \partial_{p+1}^{a,b} \circ (\partial_{p+1}^{a,b})^*\omega\rangle + \langle (\partial_p^a)^* \circ \partial_p^a\omega, (\partial_p^a)^* \circ \partial_p^a\omega\rangle
\end{aligned}$$

since $\partial_p^a \circ \partial_{p+1}^{a,b} \circ (\partial_{p+1}^{a,b})^*\omega = \partial_p^b \circ \partial_{p+1}^b \circ (\partial_{p+1}^{a,b})^*\omega = 0$. The positive definiteness of the inner product implies that

$$\partial_{p+1}^{a,b} \circ (\partial_{p+1}^{a,b})^*\omega = 0, \quad (\partial_p^a)^* \circ \partial_p^a\omega = 0.$$

Then we have

$$0 = \langle \partial_{p+1}^{a,b} \circ (\partial_{p+1}^{a,b})^*\omega, \omega\rangle = \langle (\partial_{p+1}^{a,b})^*\omega, (\partial_{p+1}^{a,b})^*\omega\rangle,$$

which implies that $(\partial_{p+1}^{a,b})^*\omega = 0$ in view of the definiteness of the inner product. Similarly, we have $\partial_p^a\omega = 0$. Thus $\omega \in \ker(\partial_{p+1}^{a,b})^* \cap \ker \partial_p^a$.

For any $\alpha \in C_{p+1}^{a,b}, \beta \in C_{p-1}^a$, we have

$$\langle \partial_{p+1}^{a,b}\alpha, (\partial_p^a)^*\beta\rangle = \langle \partial_p^a \circ \partial_{p+1}^{a,b}\alpha, \beta\rangle = 0,$$

which shows that $\mathrm{im}\partial_{p+1}^{a,b}$ and $\mathrm{im}(\partial_p^a)^*$ are orthogonal. Now, we will prove

$$\ker \Delta_p^{a,b} = (\mathrm{im}\partial_{p+1}^{a,b})^\perp \cap (\mathrm{im}(\partial_p^a)^*)^\perp. \tag{5}$$

By Formula (4), we have $\ker \Delta_p^{a,b} \subseteq (\mathrm{im}\partial_{p+1}^{a,b})^\perp \cap (\mathrm{im}(\partial_p^a)^*)^\perp$. If $\omega \in (\mathrm{im}\partial_{p+1}^{a,b})^\perp \cap (\mathrm{im}(\partial_p^a)^*)^\perp$, then we have

$$\langle (\partial_{p+1}^{a,b})^*\omega, x\rangle = \langle \omega, \partial_{p+1}^{a,b}x\rangle = 0, \quad \text{for any } x \in C_{p+1}^{a,b}.$$

It follows that $(\partial_{p+1}^{a,b})^*\omega = 0$ in $C_{p+1}^{a,b}$. Similarly, we have $\partial_p^a\omega = 0$ in C_{p-1}^a. By Fomular (4), we obtain

$$\omega \in \ker(\partial_{p+1}^{a,b})^* \cap \ker \partial_p^a = \ker \Delta_p^{a,b},$$

which leads to the desired result.

Let $\omega \in C_p^a$, and let $v_1, \ldots, v_n$ be an orthogonal basis of $\ker \Delta_p^{a,b}$. Denote $\omega_0 = \sum_{i=0}^{n} \langle \omega, v_i\rangle v_i$. Then we have

$$\langle \omega - \omega_0, v_i\rangle = 0, \quad i = 1, \ldots, n,$$

which implies that $\omega' = \omega - \omega_0 \in \mathrm{im}\partial_{p+1}^{a,b} + \mathrm{im}(\partial_p^a)^* = \mathrm{im}\partial_{p+1}^{a,b} \oplus \mathrm{im}(\partial_p^a)^*$ in terms of Formular (5). Similarly, ω' can be written as

$$\omega' = \omega_1 + \omega_2, \quad \omega_1 \in \mathrm{im}\partial_{p+1}^{a,b}, \omega_2 \in \mathrm{im}(\partial_p^a)^*.$$

Then we have $\omega = \omega_0 + \omega_1 + \omega_2$, $\omega_0 \in \ker \Delta_p^{a,b}$, $\omega_1 \in \mathrm{im}\partial_{p+1}^{a,b}$, $\omega_2 \in \mathrm{im}(\partial_p^a)^*$. Moreover, a straightforward calculation shows that the decomposition is unique. □

A notable result is that the persistent Laplacian indicates the persistent homology.

Theorem 3.8. *Let $\mathcal{F} : \mathbb{R}^* \to \mathbf{Simp}$ be a filtration of simplicial complexes of finite type. Then the morphism*

$$\rho : \ker(\Delta_p^{a,b}) \to H_p^{a\to b}(\mathcal{F}), \quad \omega \mapsto [\omega]. \tag{6}$$

is an isomorphism of abelian groups for all $a < b, p \in \mathbb{N}$.

Proof. By Formula (4), the map ρ is well defined. Let $\rho(\omega) = 0$ for $\omega \in \ker(\Delta_p^{a,b})$. Then we have

$$\omega = \partial_{p+1}^b \alpha, \quad \alpha \in C_{p+1}^b.$$

Noting that $\omega \in C_p^a$, we can write $\omega = \partial_{p+1}^{a,b}\alpha$ and $\alpha \in C_{p+1}^{a,b}$. It follows that

$$0 = (\partial_{p+1}^{a,b})^*\omega = (\partial_{p+1}^{a,b})^*\partial_{p+1}^{a,b}\alpha,$$

which implies that $\omega = \partial_{p+1}^{a,b}\alpha = 0$ by the positive definiteness of inner product. Thus ρ is an injection. For any nonzero element $z \in H_p^{a\to b}(\mathcal{F})$, we choose a representative element $\omega \in \ker \partial_p^a \cap \ker \partial_p^b$ such that $z = [\omega]$. By Theorem 3.7, we have a decomposition

$$\omega = \omega_0 + \omega_1, \quad \omega_0 \in \ker \Delta_p^{a,b}, \omega_1 \in \mathrm{im}\partial_{p+1}^{a,b}.$$

Note that $\mathrm{im}\partial_{p+1}^{a,b} \subseteq \mathrm{im}\partial_{p+1}^b$, we have $[\omega] = [\omega_0]$. Then we have $\rho(\omega_0) = z$, thus ρ is a surjection. □

Remark 3.9. For $a < b$, the inclusion $j : C_*^a \hookrightarrow C_*^b$ of chain complexes induces a morphism of homology groups

$$H_p(j) : H_p(C_*^a) \to H_p(C_*^b).$$

The Hodge theorem says that $H_p(C_*^a) \cong \ker \Delta_p^a$, where $\Delta_p^a = \partial_{p+1}^a \circ (\partial_{p+1}^a)^* + (\partial_p^a)^* \circ \partial_p^a$. Hence, we have a morphism of abelian groups $\ker \Delta_p^a \to \ker \Delta_p^b$ induced by $H_p(j)$. However, it does not always hold that $j(\omega) \in \ker \Delta_p^b$ for $\omega \in \ker \Delta_p^a$. Let $\omega \in C_p^a$. By the Hodge decomposition theorem, we have

$$\omega = \omega_0 + \omega_1, \quad \omega_0 \in \ker \Delta_p^{a,b}, \omega_1 \in \mathrm{im}\partial_{p+1}^{a,b}.$$

Define $h_p^{a,b} : \ker \Delta_p^a \to \ker \Delta_p^b$ by $h_p^{a,b}(\omega) = \omega_0$. The proof of Theorem 3.8 shows that $H_p(j)$ is given by $h_p^{a,b}$. Moreover, we have $\mathrm{im}(h_p^{a,b}) = \ker \Delta_p^{a,b}$.

From the above theorem, we can identify the data $\{\ker(\Delta_p^{a,b})\}_{a,b\in\mathbb{R},p\in\mathbb{N}}$ with the the persistent homology data $\{H_p^{a\to b}(\mathcal{F})\}_{a,b\in\mathbb{R},p\in\mathbb{N}}$. Moveover, these two kinds of data provide the same barcode.

Another interesting aspect is that the image of $\Delta_p^{a,b}$ seems to be stable, or precisely, we have

$$\mathrm{im}(\Delta_p^{a,b}) \subseteq \mathrm{im}(\Delta_p^{a,c}), \quad p \geq 0$$

for any $a \leq b \leq c$. It can be obtained directly by the fact $H_p^{a,b}(\mathcal{F}) \subseteq H_p^{a,c}(\mathcal{F})$ and the persistent Hodge decomposition theorem.

The persistent Laplacian is heavily depend on the up-persistent $\Delta_{p,up}^{a,b} = \partial_{p+1}^{a,b} \circ (\partial_{p+1}^{a,b})^*$, which is regarded as a Schur complement.[40] The monotonicity and stability of up-persistent Laplacian eigenvalues have been proved. Note that the persistent Laplacian provides both topological and spectral information for the characterization of data. Persistent homology based models have been used in molecular data analysis.[203,208,209]

3.6 Persistent de Rham-Hodge theory

The evolutionary de Rham-Hodge method can be regarded as a variant of persistent Laplacian in geometry. In fact, the idea of the evolutionary de Rham-Hodge method is earlier than the persistent Laplacian.[55,229]

The de Rham-Hodge theory establishes the relationship between the de Rham cohomology of an oriented closed Riemannian manifold and the harmonic forms. It is widely used in differential geometry, algebraic geometry, partial differential equations, algebraic topology, etc.[8,13,46,47,152] Additionally, it also appears frequently in physics, data science and computer science.[21,33,79,140,151]

The Laplacian operator plays a fundamental role in the de Rham-Hodge theory. We establish the notations to introduce the Laplacian on de Rham complexes. Let M be a compact oriented Riemannian n-manifold without boundary. Let $x_1, \ldots, x_n$ be an orthogonal basis of M, and let $dx_1, \ldots, dx_n$ be the dual basis of $x_1, \ldots, x_n$. Denote $\Omega^p(M)$ the space of all the differential p-forms on M. The de Rham complex $\Omega^*(M)$ is a cochain complex with differential operators $d^p : \Omega^p(M) \to \Omega^{p+1}(M), p \geq 0$ given by

$$d^p(f dx_{\alpha_1} \wedge \cdots \wedge dx_{\alpha_p}) = \sum_{i=1}^{n} \frac{\partial f}{\partial x_i} dx_i \wedge dx_{\alpha_1} \wedge \cdots \wedge dx_{\alpha_p}.$$

Note that there is a codifferential operator

$$(d^p)^* : \Omega^{p+1}(M) \to \Omega^p(M)$$

adjoint to d^p with respect to the L_2 inner product given by

$$(\alpha, \beta) = \int_M \alpha \wedge \star\beta.$$

Here, $\star$ is the *Hodge star operator*.[142] The Laplacian operator $\Delta^p : \Omega^p(M) \to \Omega^p(M)$ on de Rham complex is defined by

$$\Delta^p = (d^p)^* \circ d^p + d^{p-1} \circ (d^{p-1})^*.$$

This definition is consistent with the definition of combinatorial Laplacian. Recall that the space of harmonic p-forms $\mathcal{H}^p = \{\omega \in \Omega^p(M) | \Delta^p \omega = 0\}$ on a manifold M is isomorphic to the cohomology of M,[26] i.e.,

$$\mathcal{H}^p \cong H^p_{DR}(M) \cong H^p(M, \mathbb{R}),$$

where $H^p_{DR}(M) = H^p(\Omega^*(M))$. By the Hodge composition theorem, we have

$$\Omega^p(M) = \mathcal{H}^p \oplus \mathrm{im} d^p \oplus \mathrm{im}(d^p)^*, \quad p \in \mathbb{N},$$

and each element $\omega \in \Omega^p(M)$ has a unique decomposition. Let $h = \{h^p\}_{p \in \mathbb{N}}$ be a family of linear maps given by

$$h^p : \Omega^p(M) \to \mathcal{H}^p, \quad h^p(\omega) = \omega_0,$$

where $\omega = \omega_0 + \omega_1 + \omega_2$, $\omega_0 \in \mathcal{H}^p$, $\omega_1 \in \mathrm{im}d^p$, $\omega_2 \in \mathrm{im}(d^p)^*$.

From now on, the manifolds considered are always assumed to be compact oriented Riemannian manifolds without boundary[!]. The evolutionary de Rham-Hodge method is to consider the Laplacian with respect to a (smooth) filtration of manifolds, the evolution of manifolds. Let $\mathbb{R}^*$ be the category of real number with morphisms given by $a \to b$ for $a \leq b$. Generally, an *evolution (or filtration) of manifolds* is a functor $\mathcal{M} : \mathbb{R}^* \to \mathbf{Mani}$ from the category of real numbers to the category of smooth manifolds such that

$$j_{a,b} : \mathcal{M}(a) \hookrightarrow \mathcal{M}(b), \quad \text{for } a \leq b$$

is an immersion. For real numbers $a \leq b$, we have a morphism of de Rham complexes

$$j_{a,b}^* = \Omega^*(j_{a,b}) : \Omega^*(\mathcal{M}(b)) \to \Omega^*(\mathcal{M}(a)),$$

which leads to the following commutative diagram.

$$\begin{array}{ccccccccc}
\Omega^0(\mathcal{M}(a)) & \xrightarrow{d^0} & \cdots & \longrightarrow & \Omega^p(\mathcal{M}(a)) & \xrightarrow{d^p} & \Omega^{p+1}(\mathcal{M}(a)) & \xrightarrow{d^{p+1}} & \cdots \\
\uparrow{\scriptstyle j^0_{a,b}} & & & & \uparrow{\scriptstyle j^p_{a,b}} & & \uparrow{\scriptstyle j^{p+1}_{a,b}} & & \\
\Omega^0(\mathcal{M}(b)) & \xrightarrow{d^0} & \cdots & \longrightarrow & \Omega^p(\mathcal{M}(b)) & \xrightarrow{d^p} & \Omega^{p+1}(\mathcal{M}(b)) & \xrightarrow{d^{p+1}} & \cdots
\end{array}$$

It induces a morphism of cohomology groups

$$j_{a,b}^p = H^p(j_{a,b}) : H^p_{DR}(\mathcal{M}(b)) \to H^p_{DR}(\mathcal{M}(a)).$$

Denote $\mathcal{H}^p_a$ the space of harmonic p-forms on $\mathcal{M}(a)$. By the Hodge theorem,[26] the above morphism can be identified with

$$h \circ j^p_{a,b} : \mathcal{H}^p_b \to \mathcal{H}^p_a.$$

The *p-th persistent harmonic space* is defined by

$$\mathcal{H}^p_{a\to b} = \mathrm{im}(\mathcal{H}^p_b \to \mathcal{H}^p_a).$$

The data of persistent harmonic spaces $\{\mathcal{H}^p_{a\to b}\}_{a,b\in\mathbb{R},p\in\mathbb{N}}$ can be identified with the data of persistent cohomology $\{H^p_{a\to b}\}_{a,b\in\mathbb{R},p\in\mathbb{N}}$.

The de Rham complex $\Omega^*(M(a)), a \in \mathbb{R}$ with the inner product $(\cdot,\cdot)$ can also produce a persistent Laplacian similar as the combinatorial version. One notable difference is that the morphism of de Rham complexes

$$j^*_{a,b} : \Omega^*(\mathcal{M}(b)) \to \Omega^*(\mathcal{M}(a)), \quad a \leq b$$

is a projection. The inner product and coboundary operators on $\Omega^*(M(a))$ can be induced by that on $\Omega^*(M(b))$.

[!] The evolutionary de Rham-Hodge method also works for manifolds with boundaries that are leaded to applications.[55] For simplicity, we only highlight the results for manifolds without boundaries.

Now, denote $\Omega_a^p = \Omega^p(M(a)), a \in \mathbb{R}$ for the sake of simplicity. Let $(j_{a,b}^p)^*$ be the adjoint of $j_{a,b}^p$, then $(j_{a,b}^p)^*$ is injective. Indeed, if $(j_{a,b}^p)^* x = 0$, we have

$$0 = ((j_{a,b}^p)^* x, y) = (x, j_{a,b}^p y), \quad \text{for any } y \in \Omega_b^p.$$

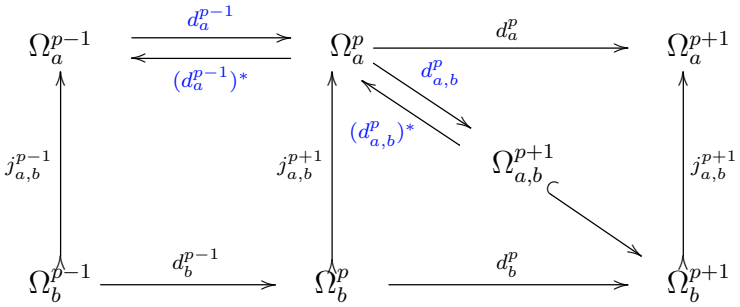
Since $j_{a,b}^p$ is surjective, we have $(x, z) = 0$ for any $z \in \Omega_a^p$, which implies that $x = 0$. Now, we denote

$$\Omega_{a,b}^{p+1} = \{\omega \in \Omega_b^{p+1} | (d_b^p)^* \omega \in (j_{a,b}^p)^* \Omega_a^p\}.$$

Define $(d_{a,b})^* : \Omega_{a,b}^{p+1} \to \Omega_a^p$ as follows: let $\omega \in \Omega_{a,b}^{p+1}$, then $(d_b^p)^* \omega = (j_{a,b}^p)^* \alpha$ for a unique $\alpha \in \Omega_a^p$ since $(j_{a,b}^p)^*$ is an injection. And then we let $(d_{a,b}^p)^* \omega = \alpha$. Note that

$$(j_{a,b}^{p-1})^* (d_a^{p-1})^* (d_{a,b}^p)^* \omega = (j_{a,b}^{p-1})^* (d_a^{p-1})^* \alpha = (d_b^{p-1})^* (j_{a,b}^p)^* \alpha = (d_b^{p-1})^* (d_b^p)^* \omega = 0.$$

Since $(j_{a,b}^p)^*$ is injective, we have $(d_a^{p-1})^* (d_{a,b}^p)^* = 0$. Let $(d_{a,b})^p$ be the adjoint of $(d_{a,b}^p)^*$. Then we have the following diagram.



The *p-th persistent Laplacian* on the evolution of manifolds is defined as

$$\Delta_{a,b}^p = d_a^{p-1} \circ (d_a^{p-1})^* + (d_{a,b}^p)^* \circ d_{a,b}^p.$$

Similar as Formula (4), we have

$$\ker \Delta_{a,b}^p = \ker (d_a^{p-1})^* \cap \ker d_{a,b}^p. \tag{7}$$

And we define the *p-evolution harmonic space*[!] by $\mathcal{H}_{a,b}^p = \ker \Delta_{a,b}^p$.

Moreover, the persistent Hodge decomposition theorem for the evolution of manifolds is also established.

Theorem 3.10 (**Persistent Hodge decomposition theorem**). *Let* $\mathcal{M} : \mathbb{R}^* \to$ **Mani** *be an evolution of manifolds. Then*

$$\Omega_a^p(\mathcal{M}) = \ker \Delta_{a,b}^p \oplus \mathrm{im} d_a^{p-1} \oplus \mathrm{im} (d_{a,b}^p)^*.$$

The proof of Theorem 3.10 is similar to that of Theorem 3.7. The following lemma indicates the persistence of $\ker \Delta_{a,b}^p$ in a certain meaning and will be used in the proof of further theorem.

[!] The definition is somewhat different from the original.[55] The ideas are originally from Guo-Wei Wei et al, and we unify the notations here.

Lemma 3.11. $\ker \Delta^p_{a,b} \subseteq \ker \Delta^p_a$.

Proof. Let $\alpha \in \ker \Delta^p_{a,b}$. By the Hodge decomposition theorem, we have

$$\alpha = \alpha_0 + \alpha_1 + \alpha_2, \quad \alpha_0 \in \ker \Delta^p_a, \alpha_1 \in \mathrm{im} d^{p-1}_a, \alpha_2 \in \mathrm{im}(d^p_a)^*.$$

By formula (7), we have $\alpha = \alpha_0 + \alpha_2$. Let $\alpha_2 = (d^p_a)^*\beta_2$. For any $u \in \Omega^{p+1}_{a,b}$, let $v = (d^p_{a,b})^* u$. It follows that $(d^p_b)^* u = (j^p_{a,b})^* v$ by definition. Then we have

$$0 = (d^p_{a,b}\alpha, u) = (\alpha_0 + (d^p_a)^*\beta_2, (d^p_{a,b})^* u) = (\alpha_0, v) + ((d^p_a)^*\beta_2, v)$$

for any v satisfying $(j^p_{a,b})^* v \in \mathrm{im}(d^p_b)^*$. Since $(j^p_{a,b})^*(d^p_a)^*\beta_2 = (d^p_b)^*(j^{p+1}_{a,b})^*\beta_2$, we choose $v = (d^p_a)^*\beta_2$. Note that $(\alpha_0, (d^p_a)^*\beta_2) = (d^p_a\alpha_0, \beta_2) = 0$. We have

$$((d^p_a)^*\beta_2, (d^p_a)^*\beta_2) = 0,$$

which implies $(d^p_a)^*\beta_2 = 0$. Hence, we obtain $\alpha = \alpha_0 \in \ker \Delta^p_a$. This completes the proof of this lemma. □

An interesting result says that the persistent harmonic space coincides with the persistent evolution harmonic space.

Theorem 3.12. *Let $\mathcal{M} : \mathbb{R}^* \to$ **Mani** be an evolution of manifolds. Then*

$$\mathcal{H}^p_{a,b} = \mathcal{H}^p_{a\to b}, \quad a < b, \quad p \in \mathbb{N}.$$

Proof. It is equivalent to proving $\ker \Delta^p_{a,b} = h^p j^p_{a,b}(\ker \Delta^p_b)$. Remember that

$$\ker \Delta^p_{a,b} = \ker(d^{p-1}_a)^* \cap \ker d^p_{a,b}, \quad \ker \Delta^p_b = \ker(d^{p-1}_b)^* \cap \ker d^p_b.$$

(i) "$\supseteq$". Let $\beta \in \ker \Delta^p_b$. Indeed, for any $u \in \Omega^{p+1}_{a,b}$, let $v = (d^p_{a,b})^* u$. By definition, we have $(d^p_b)^* u = (j^p_{a,b})^* v$. It follows that

$$(d^p_{a,b} j^p_{a,b}(\beta), u) = (j^p_{a,b}(\beta), (d^p_{a,b})^* u) = (j^p_{a,b}(\beta), v) = (\beta, (j^p_{a,b})^* v) = (\beta, (d^p_b)^* u).$$

Since $(\beta, (d^p_b)^* u) = (d^p_b \beta, u) = 0$, we have $(d^p_{a,b} j^p_{a,b}(\beta), u) = 0$ for any $u \in \Omega^{p+1}_{a,b}$, which implies that $d^p_{a,b} j^p_{a,b}(\beta) = 0$. Thus we have

$$j^p_{a,b}(\beta) \in \ker d^p_{a,b}. \tag{8}$$

Since $d^p_a j^p_{a,b}(\beta) = j^p_{a,b} d^p_b(\beta) = 0$, we have $j^p_{a,b}(\beta) \in \ker d^p_a$. By the Hodge decomposition theorem, we obtain

$$j^p_{a,b}(\beta) = h^p j^p_{a,b}(\beta) + \gamma, \quad \gamma \in \mathrm{im} d^{p-1}_a. \tag{9}$$

Recall that $(d^{p-1}_a)^*(d^p_{a,b})^* = 0$, we have $d^p_{a,b} d^{p-1}_a = 0$. It follows that $d^p_{a,b}\gamma = 0$. So we have

$$h^p j^p_{a,b}(\beta) = j^p_{a,b}(\beta) - \gamma \in \ker d^p_{a,b}.$$

It is obvious that $h^p j^p_{a,b}(\beta) \in \ker(d^p_a)^*$. Thus $h^p j^p_{a,b}(\beta) \in \ker \Delta^p_{a,b}$.

(ii) "$\subseteq$". Let $\alpha \in \ker \Delta^p_{a,b}$. Since $j^p_{a,b}$ is surjective, there is an element $\beta \in \Omega^p_{a,b}$ such that $j^p_{a,b}(\beta) = \alpha$. Consider the Hodge decomposition of β given by

$$\beta = \beta_0 + \beta_1 + \beta_2, \quad \beta_0 \in \ker \Delta^p_b, \beta_1 \in \mathrm{im} d^{p-1}_b, \beta_2 \in \mathrm{im}(d^p_b)^*.$$

Let $\beta_1 = d^{p-1}_b \gamma_1$ and $\beta_2 = (d^p_b)^* \gamma_2$ for some $\gamma_1 \in \Omega^{p-1}_b, \gamma_2 \in \Omega^p_b$. Then we have

$$j^p_{a,b}(\beta_1) = j^p_{a,b}(d^{p-1}_b \gamma_1) = d^{p-1}_a j^{p-1}_{a,b} \gamma_1 \in \mathrm{im} d^{p-1}_a. \tag{10}$$

Since $d^p_{a,p}\alpha = 0$ and $d^p_{a,b} d^{p-1}_a = 0$, we have $d^p_{a,b}(j^p_{a,b}(\beta_0) + j^p_{a,b}(\beta_2)) = 0$. For any $u \in \Omega^{p+1}_{a,b}$, let $v = (d^p_{a,b})^* u$ with $(d^p_b)^* u = (j^p_{a,b})^* v$ by definition. It follows that

$$\begin{aligned} 0 =& (d^p_{a,b}(j^p_{a,b}(\beta_0) + j^p_{a,b}(\beta_2)), u) \\ =& ((j^p_{a,b}(\beta_0) + j^p_{a,b}(\beta_2)), (d^p_{a,b})^* u) \\ =& (j^p_{a,b}(\beta_0), v) + (j^p_{a,b}(\beta_2), v). \end{aligned}$$

Note that $(j^p_{a,b}(\beta_0), v) = (\beta_0, (j^p_{a,b})^* v) = (\beta_0, (d^p_b)^* u) = (d^p_b \beta_0, u) = 0$. We have $(j^p_{a,b}(\beta_2), v) = 0$ for any v satisfying $(j^p_{a,b})^* v \in \mathrm{im}(d^p_b)^*$. Choose $v = (d^p_a)^* d^p_a j^p_{a,b}(\beta_2)$, then we have

$$(j^p_{a,b}(\beta_2), (d^p_a)^* d^p_a j^p_{a,b}(\beta_2)) = (d^p_a j^p_{a,b}(\beta_2), d^p_a j^p_{a,b}(\beta_2)) = 0,$$

which implies that

$$d^p_a j^p_{a,b}(\beta_2) = 0. \tag{11}$$

Now, to get the desired result, we need to make further use of the inner product on the de Rham complex. Note that $j^p_{a,b} : \Omega^p_b \to \Omega^p_a$ is given by

$$j^p_{a,b}\beta = \beta \circ j_{a,b} = \beta|_{\mathcal{M}(a)}, \quad \beta \in \Omega^p_b,$$

where $j_{a,b} : \mathcal{M}(a) \hookrightarrow \mathcal{M}(b)$. Now, we will show

$$d^p_b (j^p_{a,b})^* = (j^{p+1}_{a,b})^* d^p_a.$$

It is equivalent to proving $(d^p_a)^* j^p_{a,b} = j^{p+1}_{a,b} (d^p_b)^*$. Let $x_1, \ldots, x_n$ be an orthogonal basis of $\mathcal{M}(b)$. It suffices to show

$$(d^p_a)^* j^p_{a,b} f dx_{i_1} \wedge \ldots dx_{i_p} = j^{p+1}_{a,b} (d^p_b)^* f dx_{i_1} \wedge \ldots dx_{i_p}, \quad f \in C^\infty(\mathcal{M}(b)).$$

Recall that $j_{a,b}$ is an inclusion, we have

$$\frac{\partial (f \circ j_{a,b})}{\partial x_{i_k}} = \frac{\partial f|_{\mathcal{M}(a)}}{\partial x_{i_k}} = \left(\frac{\partial f}{\partial x_{i_k}} \right) \Big|_{\mathcal{M}(a)} = \frac{\partial f \circ j_{a,b}}{\partial x_{i_k}}, \quad k = 1, \ldots, p.$$

A straightforward calculation shows the desired result. Thus we have

$$d^p_b (j^p_{a,b})^* \alpha = (j^{p+1}_{a,b})^* d^p_a \alpha, \quad \text{for any } \alpha \in \Omega^p_a.$$

By formula (11), we obtain that

$$0 = d_a^p j_{a,b}^p(\beta_2) = d_a^p j_{a,b}^p (d_b^p)^* \gamma_2 = d_a^p (d_a^p)^* j_{a,b}^p \gamma_2.$$

In view of the positive definiteness of the inner product, we have

$$(d_a^p)^* j_{a,b}^{p+1} \gamma_2 = 0.$$

It follows that

$$j_{a,b}^p(\beta_2) = j_{a,b}^p (d_b^p)^* \gamma_2 = (d_a^p)^* j_{a,b}^{p+1} \gamma_2 = 0.$$

Finally, by Lemma 3.11, we obtain that

$$\alpha = h^p \alpha = h^p j_{a,b}^p \beta_0,$$

which completes the proof. □

The following corollary is a direct result of Theorem 3.12, which is essentially similar to the case of combinatorial Laplacian shown in Theorem 3.8.

Corollary 3.13 (**Persistent de Rham theorem**)**.** Let $\mathcal{M} : \mathbb{R}^* \to \mathbf{Mani}$ be an evolution of manifolds. Then we have an isomorphism

$$\mathcal{H}_{a,b}^p \cong H_{a\to b}^p(\mathcal{F}), \quad a \le b, \quad p \in \mathbb{N}.$$

It is obvious that

$$\mathrm{im}(\Delta_{a,b}^p) \subseteq \mathrm{im}(\Delta_{a,c}^p)$$

for any $a \le b \le c$, which reveals that the stability of the persistent Hodge Laplacian in some sense.

Quite a few examples such as the multibody systems, benzene (C_6H_6), and buckminsterfullerene (C_{60}) show the practical application potential of the evolutionary Hodge Laplacian in data analysis and molecular biology.[55] Compared with the usual persistent homology, the evolutionary Hodge Laplacian can describe the geometric progression together with the topological persistence of data. In fact, persistent homology is not sensitive to homotopical progression.

3.7 Yau-Hausdorff distance

A systematic comparison of DNA or protein sequences is one of the most important topics in genome data analysis. Many classical mathematical models have been proposed, such as multiple sequence alignment (MSA),[6] moment vectors[224] and feature vectors[101] and so on, to describe gene sequences. Among these methods, Yau-Hausdorff distance, as a metric, provides an accurate and efficient representation of DNA or protein sequences by the careful consideration of "translations" and "rotations".[190,191] The Yau-Hausdorff distance is based on the graphical representation method,[223,224] which gives a one-to-one mapping of DNA sequence or protein sequence to a curves in a two-dimensional plane.

A sequence s of length L over a finite set F can be written as $N_1N_2\cdots N_L$, where $N_l \in F, l = 1, 2, \ldots, L$. The graphical representation of the sequence s is a piecewise function in broken line patterns given by

$$f(0) = 0, \quad f(l) = f(l-1) + y(N_l), \quad l = 1, \cdots, L,$$

where $y : F \to [-1, 1]$ is a one-to-one function. For example, a DNA sequence can be written as $N_1N_2\cdots N_L$, where $N_l \in \{A, C, G, T\}$ representing the four nucleotides A, G, C and T. The function y is chosen by $y(A) = 1/3$, $y(C) = -1/3$, $y(G) = -2/3$, $y(T) = 2/3$.

To compare DNA or protein sequences, it is crucial to find a metric to match two-dimensional curves under translation and rotation. The Hausdorff distance[103,114] is a usual criterion to measure the similarity between two-dimensional point sets. Let $(X, \|\cdot\|)$ be a metric space. For point sets $A, B \subseteq X$, the Hausdorff distance is defined by

$$h(A, B) = \max\{\sup_{a\in A}\inf_{b\in B}\|a-b\|, \sup_{b\in B}\inf_{a\in A}\|a-b\|\}.$$

In early years, S. Yau et al began to find biological distance to characterize genetic sequences.[66] The Hausdorff distance and modified Hausdorff distance was introduced to describe viral genome phylogeny later.[225] Let X be an Euclidean space, the minimum Hausdorff distance between A and B under translation is given by

$$H(A, B) = \inf_{t\in X}\max\{\sup_{a\in A+t}\inf_{b\in B}\|a-b\|, \sup_{b\in B}\inf_{a\in A+t}\|a-b\|\}.$$

The *Yau-Hausdorff distance*, as a criterion that measures the similarity between two curves to characterize similarity of sequences, is defined by

$$D(A, B) = \max\{\sup_{\theta}\inf_{\varphi} H(P_x(A^\theta), P_x(B^\varphi)), \sup_{\varphi}\inf_{\theta} H(P_x(A^\theta), P_x(B^\varphi))\}.$$

Here, $P_x(A^\theta) = \{x\cos\theta - y\sin\theta | (x, y) \in A\}$ denotes the one-dimensional point set given by the projection of A on x-axis after being rotated counterclockwise by θ.

Remark 3.14. Recall that the Gromov-Hausdorff distance[111] between two metric spaces is

$$d_{GH}(X, Y) = \inf_{(Z,d_Z)} h(X, Y),$$

where (Z, d_Z) runs over all the metric spaces that X, Y can be isometrically embedded in. The Mazur-Ulam Theorem says that an isometry between real normed linear spaces must be an affine transformation, or more precisely, a composition of rotation and translation transformations. The definition of Yau-Hausdorff distance has the consideration of Gromov-Hausdorff distance in some sense.

The Yau-Hausdorff distance, proved as a metric, is so far one of the most accurate criteria for comparing the gene sequences in view of the following aspects:

(i) It inherits the advantage of Hausdorff distance, which is one of the most useful criteria to measure the similarity between two-dimensional point sets.

Input data	Topological representation	Topological analysis	Topological features	Machine learning
Function Point cloud Digital image Matrix Network Graph, etc	Vietoris-Rips complex Alpha complex Morse complex Cech complex Clique complex Cubical complex Path complex Neighborhood complex Dock complex Hom complex Hypergraph, etc	JavaPlex PerSeus Dionysus Phat Dipha Gudhi Ripser(++) R-TDA Jhole CTL Phom Eirene, etc	Persistent Betti number Persistent landscapes Persistent surface Persistent image Wasserstein distance Bottleneck distance Persistent kernel, etc	Dimensionality Reduction (PCA, Isomap, Diffusion map, K-means, Spectral clustering, etc) Support vector machine Random Forest Navie-Bayes Convolutional neural network Recurrent neural network Transformer BERT, etc

Figure 4: The general process of topology data analysis (TDA)-based machine learning. There are general four steps, including topological representation, topological analysis, topological feature, and topology-based machine learning.

(ii) All rotation and translation transformations are taken into account to reduce the impact of local differences on the overall differences.

(iii) It avoids the calculation of Hausdorff distance of two-dimensional sets, and can be computed more efficiently.

Yau-Hausdorff method performs very well on the gene sequence analysis and validates many known biological classifications such as DNA barcode, H1N1 virus and influenza virus.[190] In addition, a three-dimensional Yau-Hausdorff method was introduced to compare protein structures.[191]

3.8 Topology-based machine learning

Recently, TDA-based machine learning and deep learning models have achieved great successes in drug design.[34,36–38,214] The essential idea for TDA-based machine learning is to extract topological information from the data with persistent homology, convent the topological information into feature vectors, and input feature vectors into machine learning or deep learning algorithms.[165] Since persistent homology can generate unique topological features, it can be used in both supervised learning and unsupervised learning approaches, including PCA, Isomap, K-means, KNN, Naive-Bayes, spectral clustering, SVM, CNN, Tree-based models, etc. As illustrated in Figure 4, TDA-based machine learning can be roughly divided into four steps, i.e., topological representation, topological analysis, topological feature, and

topology-based machine learning. For each step, various algorithms and softwares are available. For persistent homology analysis, we have JavaPlex,[188] Perseus,[155] Dipha,[15] Dionysus,[1] jHoles,[22] GUDHI,[143] Ripser,[14] PHAT,[17] DIPHA,[16] R-TDA package,[74] HERMES,[204] etc. For topological feature representation, we have persistent diagram (PD),[148] persistent barcode (PB),[84] persistent landscape,[31,32] persistent image,[2] etc. To convert topological information into topological features, we have barcode statistical,[35] binning approach,[34,38] image representation,[34,38] persistent codebooks,[25,231] etc.

It is worth mentioning that for TDA-based machine learning, a great promise comes from new ways of topological representations that can incorporate more structure information, including persistent local homology,[5,18–20,75] element specific PH,[34,36–38,214] weighted PH,[168,213] multidimensional PH,[42,43,49,58,217,219] etc. Although many mathematical tools have been widely used in various fields of biomolecular, more methods are under developing. The topology methods have great development potential in biomolecular in view of the internal relationships between topological structures and biomolecular functions. Some mathematical objects such as neighborhood complex, Hom complex, hypergraph have been used in the modeling of the biomolecules. The persistent theories based on hypergraphs perform well in molecular representations for drug design and protein-ligand binding affinity prediction.[131,132] The neighborhood complex has been introduced and applied to drug design.[133] The weighted hypergraph and Hom-complex are more general mathematical models for datasets and biomolecular structure. New mathematical tools and new applications of mathematical tools in biomolecules remain to be constantly discovered and developed.

4 Conclusion

In this paper, we introduce a new interdisciplinary area—biomolecular topology. Biomolecular topology encompasses topological structures, properties, and relations that are emerged from biomolecular structures, dynamics, interactions, and assemblies. With the availability of the gigantic amount of biomolecular data from experiments, the topological data analysis (TDA) and various topological models will become more and more important. Biomolecular topology will emerge as an important research area in the transition of biology from phenomenological and descriptive sciences to quantitative and predictive sciences.

Acknowledgments

Kelin Xia is supported by Nanyang Technological University Startup Grant M4081842 and Singapore Ministry of Education Academic Research fund Tier 1 RG109/19, MOE-T2EP20120-0013 and MOE-T2EP20220-0010. Guo-Wei Wei is supported by NIH grants R01GM126189 and R01AI164266, and NSF grants DMS-2052983, DMS-1761320, and IIS-1900473. Jie Wu and Jian Liu are supported by Natural Science Foundation of China (NSFC) grant (11971144), High-level Scientific Research Foundation of Hebei Province and the start-up research fund from Yanqi Lake Beijing Institute of Mathematical Sciences and Applications. Stephen Shing-Toung Yau is supported by National Natural Science Foundation of China (NSFC) grant (12171275), Tsinghua University Spring Breeze Fund (2020Z99CFY044), Tsinghua University start-up fund, and Tsinghua University Education Foundation fund (042202008). Professor Stephen Shing-Toung Yau is grateful to the National Center for Theoretical Sciences (NCTS) for providing an excellent research environment while part of this research was done.

References

[1] Dionysus: the persistent homology software. Software available at http://www.mrzv.org/software/dionysus.

[2] H. Adams, T. Emerson, M. Kirby, R. Neville, C. Peterson, P. Shipman, S. Chepushtanova, E. Hanson, F. Motta, and L. Ziegelmeier. Persistence images: A stable vector representation of persistent homology. *The Journal of Machine Learning Research*, 18(1):218–252, 2017.

[3] S. A. Adcock and J. A. McCammon. Molecular dynamics: Survey of methods for simulating the activity of protein. *Chemical Reviews*, 106(5):1589–615, 2006.

[4] K. Afifi and A. F. Al-Sadek. Improving classical scoring functions using random forest: The non-additivity of free energy terms' contributions in binding. *Chemical biology & drug design*, 92(2):1429–1434, 2018.

[5] M. Ahmed, B. T. Fasy, and C. Wenk. Local persistent homology based distance between maps. In *Proceedings of the 22nd ACM SIGSPATIAL International Conference on Advances in Geographic Information Systems*, pages 43–52. ACM, 2014.

[6] Miller W Myers EW Lipman DJ Altschul SF, Gish W. Basic local alignment search tool (BLAST). *Journal of Molecular Biology*, 215:403–410, 10 1990.

[7] C. B. Anfinsen. Einfluss der configuration auf die wirkung den. *Science*, 181:223–230, 1973.

[8] Donu Arapura. *Deformations and Hodge Theory*, pages 293–304. 01 2012.

[9] J. Arsuaga, M. Vázquez, S. Trigueros, J. Roca, et al. Knotting probability of DNA molecules confined in restricted volumes: DNA knotting in phage capsids. *Proceedings of the National Academy of Sciences*, 99(8):5373–5377, 2002.

[10] Javier Arsuaga, Mariel Vazquez, Paul McGuirk, Sonia Trigueros, De Witt Sumners, and Joaquim Roca. Dna knots reveal a chiral organization of dna in phage capsids. *Proceedings of the National Academy of Sciences*, 102(26):9165–9169, 2005.

[11] A. R. Atilgan, S. R. Durrell, R. L. Jernigan, M. C. Demirel, O. Keskin, and I. Bahar. Anisotropy of fluctuation dynamics of proteins with an elastic network model. *Biophys. J.*, 80:505–515, 2001.

[12] I. Bahar, A. R. Atilgan, and B. Erman. Direct evaluation of thermal fluctuations in proteins using a single-parameter harmonic potential. *Folding and Design*, 2:173–181, 1997.

[13] Matthew Baker. Hodge theory in combinatorics. *Bulletin of the American Mathematical Society*, 55, 05 2017.

[14] U. Bauer. Ripser: a lean C++ code for the computation of Vietoris-Rips persistence barcodes. *Software available at https://github. com/Ripser/ripser*, 2017.

[15] U. Bauer, M. Kerber, and J. Reininghaus. Distributed computation of persistent homology. *Proceedings of the Sixteenth Workshop on Algorithm Engineering and Experiments (ALENEX)*, 2014.

[16] U. Bauer, M. Kerber, and J. Reininghaus. Distributed computation of persistent homology. In *2014 proceedings of the sixteenth workshop on algorithm engineering and experiments (ALENEX)*, pages 31–38. SIAM, 2014.

[17] U. Bauer, M. Kerber, J. Reininghaus, and H. Wagner. PHAT–persistent homology algorithms toolbox. In *International Congress on Mathematical Software*, pages 137–143. Springer, 2014.

[18] P. Bendich, D. Cohen-Steiner, H. Edelsbrunner, J. Harer, and D. Morozov. Inferring local homology from sampled stratified spaces. In *Foundations of Computer Science, 2007. FOCS'07. 48th Annual IEEE Symposium on*, pages 536–546. IEEE, 2007.

[19] P. Bendich, E. Gasparovic, J. Harer, R. Izmailov, and L. Ness. Multi-scale local shape analysis and feature selection in machine learning applications. In *Neural Networks (IJCNN), 2015 International Joint Conference on*, pages 1–8. IEEE, 2015.

[20] P. Bendich, B. Wang, and S. Mukherjee. Local homology transfer and stratification learning. In *Proceedings of the twenty-third annual ACM-SIAM symposium on Discrete Algorithms*, pages 1355–1370. SIAM, 2012.

[21] Harsh Bhatia, Gregory Norgard, Valerio Pascucci, and Peer-Timo Bremer. The Helmholtz-Hodge Decomposition-a survey. *IEEE Transactions on Visualization and Computer Graphics*, 19:1386–1404, 11 2012.

[22] J. Binchi, E. Merelli, M. Rucco, G. Petri, and F. Vaccarino. jHoles: A tool for understanding biological complex networks via clique weight rank persistent homology. *Electronic Notes in Theoretical Computer Science*, 306:5–18, 2014.

[23] Bela Bollobas. *Modern Graph Theory*, volume 184. 01 1998.

[24] A. Bolzer, G. Kreth, I. Solovei, D. Koehler, K. Saracoglu, C. Fauth, S. Müller, R. Eils, C. Cremer, M. R. Speicher, et al. Three-dimensional maps of all chromosomes in human male fibroblast nuclei and prometaphase rosettes. *PLoS biology*, 3(5):e157, 2005.

[25] T. Bonis, M. Ovsjanikov, S. Oudot, and F. Chazal. Persistence-based pooling for shape pose recognition. In *International Workshop on Computational Topology in Image Context*, pages 19–29. Springer, 2016.

[26] Raoul Bott and Loring Tu. *Differential Forms in Algebraic Topology*, volume 82. 01 1982.

[27] F. Boyles, C. M. Deane, and G. M. Morris. Learning from the ligand: using ligand-based features to improve binding affinity prediction. *Bioinformatics*, 36(3):758–764, 2020.

[28] Jeffrey R Brender and Yang Zhang. Predicting the effect of mutations on protein-protein binding interactions through structure-based interface profiles. *PLoS computational biology*, 11(10):e1004494, 2015.

[29] Michael M Bronstein, Joan Bruna, Taco Cohen, and Petar Veličković. Geometric deep learning: Grids, groups, graphs, geodesics, and gauges. *arXiv preprint arXiv:2104.13478*, 2021.

[30] B. R. Brooks, R. E. Bruccoleri, B. D. Olafson, D.J. States, S. Swaminathan, and M. Karplus. Charmm: A program for macromolecular energy, minimization, and dynamics calculations. *J. Comput. Chem.*, 4:187–217, 1983.

[31] P. Bubenik. Statistical topological data analysis using persistence landscapes. *The Journal of Machine Learning Research*, 16(1):77–102, 2015.

[32] P. Bubenik and P. T. Kim. A statistical approach to persistent homology. *Homology, Homotopy and Applications*, 19:337–362, 2007.

[33] Jean-Paul Caltagirone. A Discrete Hodge-Helmholtz Decomposition of momentum equation. 01 2015.

[34] Z. X. Cang, L. Mu, and G. W. Wei. Representability of algebraic topology for biomolecules in machine learning based scoring and virtual screening. *PLoS computational biology*, 14(1):e1005929, 2018.

[35] Z. X. Cang, L. Mu, K. D. Wu, K. Opron, K. L. Xia, and G.W. Wei. A topological approach to protein classificationy. *Molecular based Mathematical Biologys*, 3:140–162, 2015.

[36] Z. X. Cang and G. W. Wei. Analysis and prediction of protein folding energy changes upon mutation by element specific persistent homology. *Bioinformatics*, 33(22):3549–3557, 2017.

[37] Z. X. Cang and G. W. Wei. Integration of element specific persistent homology and machine learning for protein-ligand binding affinity prediction. *International journal for numerical methods in biomedical engineering*, page 10.1002/cnm.2914, 2017.

[38] Z. X. Cang and G. W. Wei. TopologyNet: Topology based deep convolutional and multi-task neural networks for biomolecular property predictions. *PLOS Computational Biology*, 13(7):e1005690, 2017.

[39] Z. X. Cang and G. W. Wei. Integration of element specific persistent homology and machine learning for protein-ligand binding affinity prediction. *International journal for numerical methods in biomedical engineering*, 34(2):e2914, 2018.

[40] David Carlson, Emilie Haynsworth, and Thomas Markham. A generalization of the schur complement by means of the moore–penrose inverse. *Siam Journal on Applied Mathematics - SIAMAM*, 26, 01 1974.

[41] G. Carlsson. Topology and data. *Am. Math. Soc*, 46(2):255–308, 2009.

[42] G. Carlsson, G. Singh, and A. Zomorodian. Computing multidimensional persistence. In *Algorithms and computation*, pages 730–739. Springer, 2009.

[43] G. Carlsson and A. Zomorodian. The theory of multidimensional persistence. *Discrete Computational Geometry*, 42(1):71–93, 2009.

[44] G. Carlsson, A. Zomorodian, A. Collins, and L. J. Guibas. Persistence barcodes for shapes. *International Journal of Shape Modeling*, 11(2):149–187, 2005.

[45] D. Caspar and A. Klug. Physical principles in the construction of regular viruses. In *Cold Spring Harbor symposia on quantitative biology*, volume 27, pages 1–24. Cold Spring Harbor Laboratory Press, 1962.

[46] Mark Cataldo and Luca Migliorini. The Hodge theory of algebraic maps. *Annales Scientifiques de l'École Normale Supérieure*, 38:693–750, 07 2003.

[47] Gil Cavalcanti. Hodge theory of SKT manifolds. *Advances in Mathematics*, 374:107270, 07 2020.

[48] G. Cavalli and T. Misteli. Functional implications of genome topology. *Nature structural & molecular biology*, 20(3):290–299, 2013.

[49] A. Cerri and C. Landi. The persistence space in multidimensional persistent homology. In *Discrete Geometry for Computer Imagery*, pages 180–191. Springer, 2013.

[50] C. J. Chen, Rishu Saxena, and G. W. Wei. Differential geometry based multiscale models for virus formation and evolution. *Int. J. Biomed. Imaging*, 2010:308627, 2010.

[51] H. M. Chen, J. Chen, L. A. Muir, S. Ronquist, W. Meixner, M. Ljungman, T. Ried, S. Smale, and I. Rajapakse. Functional organization of the human 4D nucleome. *Proceedings of the National Academy of Sciences*, 112(26):8002–8007, 2015.

[52] J. Chen, N. Jonoska, and G. Rozenberg. *Nanotechnology: Science and Computation*. Springer, 2006.

[53] Jiahui Chen, Rui Wang, Nancy Benovich Gilby, and Guo-Wei Wei. Omicron variant (b. 1.1. 529): Infectivity, vaccine breakthrough, and antibody resistance. *Journal of chemical information and modeling*, 2022.

[54] Jiahui Chen, Rui Wang, Menglun Wang, and Guo-Wei Wei. Mutations strengthened SARS-CoV-2 infectivity. *Journal of molecular biology*, 432(19):5212–5226, 2020.

[55] Jiahui Chen, Rundong Zhao, Yiying Tong, and Guo-Wei Wei. Evolutionary de Rham-Hodge method. *arXiv preprint arXiv:1912.12388*, 2019.

[56] S. Chowdhury and F. Mémoli. Persistent path homology of directed networks. *arXiv preprint arXiv: 1701.00565*, 2022.

[57] F. Chung. *Spectral graph theory*. American Mathematical Society, 1997.

[58] D. Cohen-Steiner, H. Edelsbrunner, and D. Morozov. Vines and vineyards by updating persistence in linear time. In *Proceedings of the twenty-second annual symposium on Computational geometry*, pages 119–126. ACM, 2006.

[59] Connor W Coley, Regina Barzilay, William H Green, Tommi S Jaakkola, and Klavs F Jensen. Convolutional embedding of attributed molecular graphs for physical property prediction. *Journal of chemical information and modeling*, 57(8):1757–1772, 2017.

[60] R. B. Corey and L. Pauling. Molecular models of amino acids, peptides and proteins. *Rev. Sci. Instr.*, 24:621–627, 1953.

[61] Q. Cui and I. Bahar. *Normal mode analysis: theory and applications to biological and chemical systems*. Chapman and Hall/CRC, 2010.

[62] Edward Curtis. Simplicial homotopy theory. *Advances in Mathematics - ADVAN MATH*, 6:107–209, 04 1971.

[63] E. de Wit and W. de Laat. A decade of 3C technologies: insights into nuclear organization. *Genes & development*, 26(1):11–24, 2012.

[64] Yves Dehouck, Jean Marc Kwasigroch, Marianne Rooman, and Dimitri Gilis. BeAtMuSiC: prediction of changes in protein–protein binding affinity on mutations. *Nucleic acids research*, 41(W1):W333–W339, 2013.

[65] J. Dekker, K. Rippe, M. Dekker, and N. Kleckner. Capturing chromosome conformation. *science*, 295(5558):1306–1311, 2002.

[66] Mo Deng, Chenglong Yu, Qian Liang, Rong He, and S.s.-T Yau. A novel method of characterizing genetic sequences: Genome space with biological distance and applications. *PloS one*, 6:e17293, 03 2011.

[67] Florian Dörfler and Francesco Bullo. Kron reduction of graphs with applications to electrical networks. *Computing Research Repository - CORR*, 60, 02 2011.

[68] Daniel FAR Dourado and Samuel Coulbourn Flores. A multiscale approach to predicting affinity changes in protein–protein interfaces. *Proteins: Structure, Function, and Bioinformatics*, 82(10):2681–2690, 2014.

[69] Joseph L Durant, Burton A Leland, Douglas R Henry, and James G Nourse. Reoptimization of mdl keys for use in drug discovery. *Journal of chemical information and computer sciences*, 42(6):1273–1280, 2002.

[70] David K Duvenaud, Dougal Maclaurin, Jorge Iparraguirre, Rafael Bombarell, Timothy Hirzel, Alán Aspuru-Guzik, and Ryan P Adams. Convolutional networks on graphs for learning molecular fingerprints. In *Advances in neural information processing systems*, pages 2224–2232, 2015.

[71] H. Edelsbrunner and J. Harer. *Computational topology: an introduction*. American Mathematical Soc., 2010.

[72] H. Edelsbrunner, D. Letscher, and A. Zomorodian. Topological persistence and simplification. *Discrete Comput. Geom.*, 28:511–533, 2002.

[73] C. Ernst and D. Sumners. A calculus for rational tangles: applications to DNA recombination. In *Mathematical Proceedings of the Cambridge Philosophical Society*, volume 108, pages 489–515. Cambridge University Press, 1990.

[74] B. T. Fasy, J. Kim, F. Lecci, and C. Maria. Introduction to the R package TDA. *arXiv preprint arXiv:1411.1830*, 2014.

[75] B. T. Fasy and B. Wang. Exploring persistent local homology in topological data analysis. In *Acoustics, Speech and Signal Processing (ICASSP), 2016 IEEE International Conference on*, pages 6430–6434. IEEE, 2016.

[76] E. N. Feinberg, D. Sur, Z. Q. Wu, B. E. Husic, H. H. Mai, Y. Li, S. S. Sun, J. Y. Yang, B. Ramsundar, and V. S. Pande. Potentialnet for molecular property prediction. *ACS central science*, 4(11):1520–1530, 2018.

[77] Pablo Gainza, Freyr Sverrisson, Frederico Monti, Emanuele Rodola, D Boscaini, MM Bronstein, and BE Correia. Deciphering interaction fingerprints from protein molecular surfaces using geometric deep learning. *Nature Methods*, 17(2):184–192, 2020.

[78] M. Gameiro, Y. Hiraoka, S. Izumi, M. Kramar, K. Mischaikow, and V. Nanda. Topological measurement of protein compressibility via persistence diagrams. *preprint*, 2013.

[79] Hengzhen Gao, Mrinal Mandal, Gencheng Guo, and Jianwei Wan. Singular point detection using Discrete Hodge Helmholtz Decomposition in fingerprint images. pages 1094–1097, 01 2010.

[80] M. Gao, D. Craig, V. Vogel, and K. Schulten. Identifying unfolding intermediates of $FN-III_{10}$ by steered molecular dynamics. *J. Mol. Biol.*, 323:939–950, 2002.

[81] Cunliang Geng, Anna Vangone, and Alexandre MJJ Bonvin. Exploring the interplay between experimental methods and the performance of predictors of binding affinity change upon mutations in protein complexes. *Protein Engineering, Design and Selection*, 29(8):291–299, 2016.

[82] Cunliang Geng, Anna Vangone, Gert E Folkers, Li C Xue, and Alexandre MJJ Bonvin. iSEE: interface structure, evolution, and energy-based machine learning predictor of binding affinity changes upon mutations. *Proteins: Structure, Function, and Bioinformatics*, 87(2):110–119, 2019.

[83] Cunliang Geng, Li C Xue, Jorge Roel-Touris, and Alexandre MJJ Bonvin. Finding the $\Delta\Delta$G spot: Are predictors of binding affinity changes upon mutations in protein–protein interactions ready for it? *Wiley Interdisciplinary Reviews: Computational Molecular Science*, 9(5):e1410, 2019.

[84] R. Ghrist. Barcodes: the persistent topology of data. *Bulletin of the American Mathematical Society*, 45(1):61–75, 2008.

[85] C. Giusti, E. Pastalkova, C. Curto, and V. Itskov. Clique topology reveals intrinsic geometric structure in neural correlations. *Proceedings of the National Academy of Sciences*, 112(44):13455–13460, 2015.

[86] Mileidy W Gonzalez and Maricel G Kann. Chapter 4: Protein interactions and disease. *PLoS computational biology*, 8(12):e1002819, 2012.

[87] A. Grigor'yan, R. Jimenez, Yu. Muranov, and S.-T. Yau. On the path homology theory and eilenberg-steenrod axioms. *Homology, Homotopy and Appl.*, 20:179–205, 2018.

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

[89] A. Grigor'yan, Y. Lin, Yu. Muranov, and S.-T. Yau. Path complexes and their homologies. *J. Math. Sciences*, 248(5):564–599, 2020.

[90] A. Grigor'yan, Yu. Muranov, and S.-T. Yau. Cohomology of digraphs and (undirected) graphs. *Asian J. Math.*, 19:887–932, 2015.

[91] A. Grigor'yan, Yu. Muranov, and S.-T. Yau. On a cohomology of digraphs and hochschild cohomology. *J. Homotopy Relat. Struct.*, 11(2):209–230, 2016.

[92] A. Grigor'yan, Yu. Muranov, and S.-T. Yau. Homologies of digraphs and künneth formulas. *Comm. Anal. Geom.*, 25(5):969–1018, 2017.

[93] Raphael Guerois, Jens Erik Nielsen, and Luis Serrano. Predicting changes in the stability of proteins and protein complexes: a study of more than 1000 mutations. *Journal of molecular biology*, 320(2):369–387, 2002.

[94] Anna Gundert and May Szedlak. Higher dimensional discrete cheeger inequalities. *Proceedings of the Annual Symposium on Computational Geometry*, 01 2014.

[95] Lowell H Hall and Lemont B Kier. Electrotopological state indices for atom types: a novel combination of electronic, topological, and valence state information. *Journal of Chemical Information and Computer Sciences*, 35(6):1039–1045, 1995.

[96] Jakob Hansen and Robert Ghrist. Toward a spectral theory of cellular sheaves. *Journal of Applied and Computational Topology*, 3, 12 2019.

[97] Allen Hatcher. *Algebraic Topology*. Cambridge University Press, 2001.

[98] C. Heitsch and S. Poznanovic. Combinatorial insights into RNA secondary structure, in N. Jonoska and M. Saito, editors. *Discrete and Topological Models in Molecular Biology*, Chapter 7:145–166, 2014.

[99] C. E. Heitsch, A. E. Condon, and H. H. Hoos. From RNA secondary structure to coding theory: A combinatorial approach. In *International Workshop on DNA-Based Computers*, pages 215–228. Springer, 2002.

[100] C. H. Hou, L. Li, Z. S. Qin, and V. G. Corces. Gene density, transcription, and insulators contribute to the partition of the Drosophila genome into physical domains. *Molecular cell*, 48(3):471–484, 2012.

[101] Guohua Huang, Houqing Zhou, Yongfan Li, and Lixin Xu. Alignment-free comparison of genome sequences by a new numerical characterization. *Journal of theoretical biology*, 281:107–12, 07 2011.

[102] L. Hui, B. Isralewitz, A. Krammer, V. Vogel, and K. Schulten. Unfolding of titin immunoglobulin domains by steered molecular dynamics simulation. *Biophysical Journal*, 75:662–671, 1998.

[103] D.P. Huttenlocher, G.A. Klanderman, and W.J. Rucklidge. Comparing images using the Hausdorff distance. *Pattern Analysis and Machine Intelligence, IEEE Transactions on*, 15:850–863, 10 1993.

[104] B. Isralewitz, M. Gao, and K. Schulten. Steered molecular dynamics and mechanical functions of proteins. *Curr Opin Struct Biol*, 11(2):224–230, 2001.

[105] D. J. Jacobs, A. J. Rader, L. A. Kuhn, and M. F. Thorpe. Protein flexibility predictions using graph theory. *Proteins-Structure, Function, and Genetics*, 44(2):150–165, AUG 1 2001.

[106] Justina Jankauskaitė, Brian Jiménez-García, Justas Dapkūnas, Juan Fernández-Recio, and Iain H Moal. SKEMPI 2.0: an updated benchmark of changes in protein–protein binding energy, kinetics and thermodynamics upon mutation. *Bioinformatics*, 35(3):462–469, 2019.

[107] Sherlyn Jemimah, Masakazu Sekijima, and M Michael Gromiha. ProAffiMuSeq: sequence-based method to predict the binding free energy change of protein–protein complexes upon mutation using functional classification. *Bioinformatics*, 36(6):1725–1730, 2020.

[108] Sherlyn Jemimah, K Yugandhar, and M Michael Gromiha. PROXiMATE: a database of mutant protein–protein complex thermodynamics and kinetics. *Bioinformatics*, 33(17):2787–2788, 2017.

[109] J. Jiménez, M. Skalic, G. Martinez-Rosell, and G. De Fabritiis. K_{DEEP}: Protein–ligand absolute binding affinity prediction via 3D-convolutional neural networks. *Journal of chemical information and modeling*, 58(2):287–296, 2018.

[110] John Jumper, Richard Evans, Alexander Pritzel, Tim Green, Michael Figurnov, Olaf Ronneberger, Kathryn Tunyasuvunakool, Russ Bates, Augustin Žídek, Anna Potapenko, et al. Highly accurate protein structure prediction with AlphaFold. *Nature*, 596(7873):583–589, 2021.

[111] David Kerr and Hanfeng Li. On Gromov-Hausdorff convergence for operator metric spaces. *Journal of Operator Theory*, 62, 12 2004.

[112] O. Keskin, I. Bahar, D. Flatow, D. G. Covell, and R. L. Jernigan. Molecular mechanisms of chaperonin GroEL-GroES function. *Biochem.*, 41:491–501, 2002.

[113] D. B. Kitchen, H. Decornez, J. R. Furr, and J. Bajorath. Docking and scoring in virtual screening for drug discovery: methods and applications. *Nature reviews Drug discovery*, 3(11):935, 2004.

[114] C. Klauck, H. Muller, L.Paul Chew, Michael Goodrich, Daniel Huttenlocher, Klara Kedem, Jon Kleinberg, and Dina Kravets. Geometric pattern matching under Euclidean motion. *Computational Geometry*, 7, 01 1997.

[115] M. Kneser. Aufgabe 360. Jahresbericht der Deutschen Mathematiker-Vereinigung, 2. *Abteilung*, 58, 1955.

[116] Woong Kook and Kang-Ju Lee. Simplicial networks and effective resistance. *Advances in Applied Mathematics*, 100:71–86, 09 2018.

[117] Tanja Kortemme and David Baker. A simple physical model for binding energy hot spots in protein–protein complexes. *Proceedings of the National Academy of Sciences*, 99(22):14116–14121, 2002.

[118] Ioannis Koutis, Gary Miller, and Richard Peng. A fast solver for a class of linear systems. *Communications of the ACM*, 55:99–107, 10 2012.

[119] Dmitry Kozlov. *Combinatorial Algebraic Topology*, volume 68. 2008.

[120] MD Shaji Kumar and M Michael Gromiha. PINT: protein–protein interactions thermodynamic database. *Nucleic acids research*, 34(suppl_1):D195–D198, 2006.

[121] G. Landrum. RDKit: Open-source cheminformatics, 2006.

[122] James Lee, Shayan Oveis Gharan, and Luca Trevisan. Multiway spectral partitioning and higher-order cheeger inequalities. *Proceedings of the Annual ACM Symposium on Theory of Computing*, 61, 11 2011.

[123] Michael Lesnick. *Multidimensional interleaving and applications to topological inference*. PhD thesis, Stanford University, 2012.

[124] G. H. Li and Q. Cui. A coarse-grained normal mode approach for macromolecules: an efficient implementation and application to Ca(2+)-ATPase. *Bipohys. J.*, 83:2457–2474, 2002.

[125] H. J. Li, K. S. Leung, M. H. Wong, and P. J. Ballester. Improving AutoDock Vina using random forest: the growing accuracy of binding affinity prediction by the effective exploitation of larger data sets. *Molecular informatics*, 34(2-3):115–126, 2015.

[126] E. Lieberman-Aiden, N. L. Van Berkum, L. Williams, M. Imakaev, T. Ragoczy, A. Telling, I. Amit, B. R. Lajoie, P. J. Sabo, M. O. Dorschner, et al. Comprehensive mapping of long-range interactions reveals folding principles of the human genome. *science*, 326(5950):289–293, 2009.

[127] Lek-Heng Lim. Hodge laplacians on graphs. *Siam Review*, 62(3):685–715, 2020.

[128] J. Liu and R. X. Wang. Classification of current scoring functions. *Journal of chemical information and modeling*, 55(3):475–482, 2015.

[129] Quanya Liu, Peng Chen, Bing Wang, and Jinyan Li. dbMPIKT: a web resource for the kinetic and thermodynamic database of mutant protein interactions. *arXiv preprint arXiv:1708.01857*, 2017.

[130] Song Liu, Chi Zhang, Hongyi Zhou, and Yaoqi Zhou. A physical reference state unifies the structure-derived potential of mean force for protein folding and binding. *Proteins: Structure, Function, and Bioinformatics*, 56(1):93–101, 2004.

[131] Xiang Liu, Huitao Feng, Jie Wu, and Kelin Xia. Persistent spectral hypergraph based machine learning (PSH-ML) for protein-ligand binding affinity prediction. *Briefings in bioinformatics*, 22, 04 2021.

[132] Xiang Liu, Xiangjun Wang, Jie Wu, and Kelin Xia. Hypergraph-based persistent cohomology (HPC) for molecular representations in drug design. *Briefings in Bioinformatics*, 22, 01 2021.

[133] Xiang Liu and Kelin Xia. Neighborhood complex based machine learning (NCML) models for drug design. *Interpretability of Machine Intelligence in Medical Image Computing, and Topological Data Analysis and Its Applications for Medical Data*, 12929:87–97, 09 2021.

[134] Xianggen Liu, Yunan Luo, Pengyong Li, Sen Song, and Jian Peng. Deep geometric representations for modeling effects of mutations on protein-protein binding affinity. *PLoS computational biology*, 17(8):e1009284, 2021.

[135] Oren Livne and Achi Brandt. Lean algebraic multigrid (LAMG): Fast graph laplacian linear solver. *arXiv preprint arXiv:1108.0123v1*, 07 2011.

[136] Y. C. Lo, S. E. Rensi, W. Torng, and R. B. Altman. Machine learning in chemoinformatics and drug discovery. *Drug discovery today*, 23(8):1538–1546, 2018.

[137] Ulrike Luxburg. A tutorial on spectral clustering. *Statistics and Computing*, 17:395–416, 01 2004.

[138] R. Lyons and Yuval Peres. *Probability on Trees and Networks*. 01 1996.

[139] Slobodan Maletić and Milan Rajkovic. Combinatorial laplacian and entropy of simplicial complexes associated with complex networks. *The European Physical Journal Special Topics*, 212, 09 2012.

[140] Ravikanth Malladi, James Sethian, and Baba Vemuri. Shape modeling with front propagation: A level set approach. *Pattern Analysis and Machine Intelligence, IEEE Transactions on*, 17:158–175, 03 1995.

[141] Farid Manuchehrfar, Wei Tian, Hubert Wagner, Herbert Edelsbrunner, and Jie Liang. Topological data analysis of high dimensional probability landscapes of biochemical reaction networks using persistent homology. *Biophysical Journal*, 121(3):127a, 2022.

[142] Marvin Marcus and Herbert Robinson. A note on the Hodge star operator. *Linear Algebra and its Applications*, 10:85–87, 02 1975.

[143] C. Maria. Filtered complexes. In *GUDHI User and Reference Manual*. GUDHI Editorial Board, 2015.

[144] Facundo Mémoli, Zhengchao Wan, and Yusu Wang. Persistent Laplacians: properties, algorithms and implications. *arXiv preprint arXiv:2012.02808*, 12 2020.

[145] Zhenyu Meng and Kelin Xia. Persistent spectral–based machine learning (perspect ml) for protein-ligand binding affinity prediction. *Science Advances*, 7(19):eabc5329, 2021.

[146] Christian Merkwirth and Thomas Lengauer. Automatic generation of complementary descriptors with molecular graph networks. *Journal of chemical information and modeling*, 45(5):1159–1168, 2005.

[147] R. Meshulam and Nathan Wallach. Homological connectivity of random k-complexes. *Random Structures and Algorithms*, 34, 05 2009.

[148] K. Mischaikow and V. Nanda. Morse theory for filtrations and efficient computation of persistent homology. *Discrete and Computational Geometry*, 50(2):330–353, 2013.

[149] Iain H Moal and Juan Fernández-Recio. SKEMPI: a structural kinetic and energetic database of mutant protein interactions and its use in empirical models. *Bioinformatics*, 28(20):2600–2607, 2012.

[150] Iain H Moal and Juan Fernandez-Recio. Intermolecular contact potentials for protein–protein interactions extracted from binding free energy changes upon mutation. *Journal of Chemical Theory and Computation*, 9(8):3715–3727, 2013.

[151] Yoshihiko Mochizuki and Atsushi Imiya. Spatial reasoning for robot navigation using the Helmholtz-Hodge Decomposition of omnidirectional optical flow. *2009 24th International Conference Image and Vision Computing New Zealand, IVCNZ 2009 - Conference Proceedings*, pages 1–6, 11 2009.

[152] Stefan Müller-Stach. Hodge theory and algebraic cycles. *Global Aspects of Complex Geometry*, 01 2006.

[153] J. R. Munkres. *Elements of algebraic topology*, volume 2. Addison-Wesley Menlo Park, 1984.

[154] J. R. Munkres. *Elements of algebraic topology*. CRC Press, 2018.

[155] Vidit Nanda. Perseus: the persistent homology software. Software available at http://www.sas.upenn.edu/~vnanda/perseus.

[156] A. Y. Ng, M. I. Jordan, and Y. Weiss. On spectral clustering: Analysis and an algorithm. *Advances in neural information processing systems*, 2:849–856, 2002.

[157] D. D. Nguyen, Z. X. Cang, and G. W. Wei. A review of mathematical representations of biomolecular data. *Physical Chemistry Chemical Physics*, 2020.

[158] D. D. Nguyen, Z. X. Cang, K. D. Wu, M. L. Wang, Y. Cao, and G. W. Wei. Mathematical deep learning for pose and binding affinity prediction and ranking in D3R Grand Challenges. *Journal of computer-aided molecular design*, 33(1):71–82, 2019.

[159] D. D. Nguyen, Z. X. Cang, K. D. Wu, M. L. Wang, Y. Cao, and G. Wei. Wei. Mathematical deep learning for pose and binding affinity prediction and ranking in D3R Grand Challenges. *Journal of computer-aided molecular design*, 33(1):71–82, 2019.

[160] D. D. Nguyen, K. F. Gao, M. L. Wang, and G. W. Wei. MathDL: Mathematical deep learning for D3R Grand Challenge 4. *Journal of computer-aided molecular design*, pages 1–17, 2019.

[161] D. D. Nguyen and G. W. Wei. AGL-Score: Algebraic graph learning score for protein-ligand binding scoring, ranking, docking, and screening. *Journal of chemical information and modeling*, 59(7):3291–3304, 2019.

[162] D. D. Nguyen, T. Xiao, M. L. Wang, and G. W. Wei. Rigidity strengthening: A mechanism for protein–ligand binding. *Journal of chemical information and modeling*, 57(7):1715–1721, 2017.

[163] N. M. O'Boyle, M. Banck, C. A. James, C. Morley, T. Vandermeersch, and G. R. Hutchison. Open Babel: An open chemical toolbox. *Journal of cheminformatics*, 3(1):33, 2011.

[164] Marharyta Petukh, Luogeng Dai, and Emil Alexov. Saambe: webserver to predict the charge of binding free energy caused by amino acids mutations. *International journal of molecular sciences*, 17(4):547, 2016.

[165] C. S. Pun, K. L. Xia, and S. X. Lee. Persistent-homology-based machine learning and its applications–a survey. *arXiv preprint arXiv:1811.00252*, 2018.

[166] T. Puzyn, J. Leszczynski, and M. T. Cronin. *Recent advances in QSAR studies: methods and applications*, volume 8. Springer Science & Business Media, 2010.

[167] Manuele Rebsamen, Richard Kumaran Kandasamy, and Giulio Superti-Furga. Protein interaction networks in innate immunity. *Trends in immunology*, 34(12):610–619, 2013.

[168] Shiquan Ren, Chengyuan Wu, and Jie Wu. Weighted persistent homology. *arXiv preprint arXiv:1708.06722*, 2017.

[169] E. Ribando-Gros, Rui Wang, Jiahui Chen, Guo-Wei Wei, and Yiying Tong. Graph and hodge laplacians: Similarity and difference. 2022.

[170] F. M. Richards. Areas, volumes, packing, and protein structure. *Annual Review of Biophysics and Bioengineering*, 6(1):151–176, 1977.

[171] Alexander Rives, Joshua Meier, Tom Sercu, Siddharth Goyal, Zeming Lin, Jason Liu, Demi Guo, Myle Ott, C Lawrence Zitnick, Jerry Ma, et al. Biological structure and function emerge from scaling unsupervised learning to 250 million protein sequences. *Proceedings of the National Academy of Sciences*, 118(15), 2021.

[172] W. Rocchia, S. Sridharan, A. Nicholls, E. Alexov, A. Chiabrera, and B. Honig. Rapid grid-based construction of the molecular surface and the use of induced surface charge to calculate reaction field energies: Applications to the molecular systems and geometric objects. *Journal of computational chemistry*, 23(1):128–137, 2002.

[173] Carlos HM Rodrigues, Yoochan Myung, Douglas EV Pires, and David B Ascher. mCSM-PPI2: predicting the effects of mutations on protein–protein interactions. *Nucleic acids research*, 47(W1):W338–W344, 2019.

[174] David Rogers and Mathew Hahn. Extended-connectivity fingerprints. *Journal of chemical information and modeling*, 50(5):742–754, 2010.

[175] BK Schichet. Virtual screening of chemical libraries. *Nature*, 432(7019):862–865, 2004.

[176] T. Schlick and W. K. Olson. Trefoil knotting revealed by molecular dynamics simulations of supercoiled DNA. *Science*, 257(5073):1110–1115, 1992.

[177] P. Schuster, W. Fontana, P. F. Stadler, and I. L. Hofacker. From sequences to shapes and back: a case study in RNA secondary structures. *Proceedings: Biological Sciences*, pages 279–284, 1994.

[178] N. C Seeman. Nucleic acid nanostructures and topology. *Angewandte Chemie International Edition*, 37(23):3220–3238, 1998.

[179] Qiang Shi, Weiya Chen, Siqi Huang, Yan Wang, and Zhidong Xue. Deep learning for mining protein data. *Briefings in bioinformatics*, 22(1):194–218, 2021.

[180] Sarah Sirin, James R Apgar, Eric M Bennett, and Amy E Keating. AB-Bind: antibody binding mutational database for computational affinity predictions. *Protein Science*, 25(2):393–409, 2016.

[181] Daniel Spielman and Shang-Hua Teng. Nearly-linear time algorithms for graph partitioning, graph sparsification, and solving linear systems. *Conference Proceedings of the Annual ACM Symposium on Theory of Computing*, 10 2003.

[182] M. M. Stepniewska-Dziubinska, P. Zielenkiewicz, and P. Siedlecki. Development and evaluation of a deep learning model for protein–ligand binding affinity prediction. *Bioinformatics*, 34(21):3666–3674, 2018.

[183] T. J. Stevens, D. Lando, S. Basu, L. P. Atkinson, Y. Cao, et al. 3D structures of individual mammalian genomes studied by single-cell Hi-C. *Nature*, 544(7648):59, 2017.

[184] Nikolaus Stiefl, Ian A Watson, Knut Baumann, and Andrea Zaliani. ErG: 2D pharmacophore descriptions for scaffold hopping. *Journal of chemical information and modeling*, 46(1):208–220, 2006.

[185] Alexey Strokach, Tian Yu Lu, and Philip M Kim. ELASPIC2 (EL2): combining contextualized language models and graph neural networks to predict effects of mutations. *Journal of molecular biology*, 433(11):166810, 2021.

[186] M. Y. Su, Q. F. Yang, Y. Du, G. Q. Feng, Z. H. Liu, Y Li, and R. X. Wang. Comparative assessment of scoring functions: The CASF-2016 update. *Journal of chemical information and modeling*, 59(2):895–913, 2018.

[187] Andras Szilagyi and Yang Zhang. Template-based structure modeling of protein–protein interactions. *Current opinion in structural biology*, 24:10–23, 2014.

[188] Andrew Tausz, Mikael Vejdemo-Johansson, and Henry Adams. Javaplex: A research software package for persistent (co)homology. Software available at http://code.google.com/p/javaplex, 2011.

[189] Kurt S Thorn and Andrew A Bogan. ASEdb: a database of alanine mutations and their effects on the free energy of binding in protein interactions. *Bioinformatics*, 17(3):284–285, 2001.

[190] Kun Tian, Xiaoqian Yang, Qin Kong, Changchuan Yin, Rong He, and Stephen Yau. Two dimensional Yau-Hausdorff distance with applications on comparison of dna and protein sequences. *PloS one*, 10:e0136577, 09 2015.

[191] Kun Tian, Xin Zhao, Yuning Zhang, and Stephen Yau. Comparing protein structures and inferring functions with a novel three-dimensional Yau-Hausdorff method. *Journal of Biomolecular Structure and Dynamics*, 37(16):4151–4160, 2019. PMID: 30518311.

[192] Kathryn Tunyasuvunakool, Jonas Adler, Zachary Wu, Tim Green, Michal Zielinski, Augustin Žídek, Alex Bridgland, Andrew Cowie, Clemens Meyer, Agata Laydon, et al. Highly accurate protein structure prediction for the human proteome. *Nature*, 596(7873):590–596, 2021.

[193] R. Twarock. A tiling approach to virus capsid assembly explaining a structural puzzle in virology. *Journal of Theoretical Biology*, 226(4):477–482, 2004.

[194] R. Twarock. Mathematical virology: a novel approach to the structure and assembly of viruses. *Philosophical Transactions: Mathematical, Physical and Engineering Sciences*, pages 3357–3373, 2006.

[195] Alexandre A Vetcher, Alexander Y Lushnikov, Junalyn Navarra-Madsen, Robert G Scharein, Yuri L Lyubchenko, Isabel K Darcy, and Stephen D Levene. Dna topology and geometry in flp and cre recombination. *Journal of molecular biology*, 357(4):1089–1104, 2006.

[196] B. Wang, C. Z. Wang, K. D. Wu, and G. W. Wei. Breaking the polar-nonpolar division in solvation free energy prediction. *Journal of computational chemistry*, 39(4):217–233, 2018.

[197] B. Wang, Z. X. Zhao, and G. W. Wei. Automatic parametrization of non-polar implicit solvent models for the blind prediction of solvation free energies. *The Journal of chemical physics*, 145(12):124110, 2016.

[198] Menglun Wang, Zixuan Cang, and Guo-Wei Wei. A topology-based network tree for the prediction of protein–protein binding affinity changes following mutation. *Nature Machine Intelligence*, 2(2):116–123, 2020.

[199] Rui Wang, Jiahui Chen, and Guo-Wei Wei. Mechanisms of sars-cov-2 evolution revealing vaccine-resistant mutations in europe and america. *The journal of physical chemistry letters*, 12(49):11850–11857, 2021.

[200] Rui Wang, Yuta Hozumi, Changchuan Yin, and Guo-Wei Wei. Decoding sars-cov-2 transmission and evolution and ramifications for covid-19 diagnosis, vaccine, and medicine. *Journal of chemical information and modeling*, 60(12):5853–5865, 2020.

[201] Rui Wang, Yuta Hozumi, Changchuan Yin, and Guo-Wei Wei. Mutations on COVID-19 diagnostic targets. *Genomics*, 112(6):5204–5213, 2020.

[202] Rui Wang, Duc Duy Nguyen, and Guo-Wei Wei. Persistent spectral graph. *arXiv preprint arXiv:1912.04135*, 2019.

[203] Rui Wang, Rundong Zhao, Emily Ribando-Gros, Jiahui Chen, Yiying Tong, and Guo-Wei Wei. Hermes: Persistent spectral graph software. *Foundations of Data Science*, 3, 03 2021.

[204] Rui Wang, Rundong Zhao, Emily Ribando-Gros, Jiahui Chen, Yiying Tong, and Guo-Wei Wei. Hermes: Persistent spectral graph software. *Foundations of data science (Springfield, Mo.)*, 3(1):67, 2021.

[205] G. W. Wei. Mathematics at the eve of a historic transition in biology. *Computational and Mathematical Biophysics*, 5(1), 2017.

[206] G. W. Wei. Persistent homology analysis of biomolecular data. *SIAM News*, Dec 1, 2017.

[207] Guo-Wei Wei. Protein structure prediction beyond alphafold. *Nature Machine Intelligence*, 1(8):336–337, 2019.

[208] Xiaoqi Wei and Guo-Wei Wei. Homotopy continuation for the spectra of persistent Laplacians. *Foundations of Data Science*, 01 2021.

[209] Xiaoqi Wei and Guo-Wei Wei. Persistent sheaf Laplacians. *arXiv preprint arXiv:2112.10906v2*, 12 2021.

[210] Peter Weiss. Armor-plated puzzle: Deciphering the code of viral geometry. *Science News*, 168(10):152–154, 2005.

[211] Robin Winter, Floriane Montanari, Frank Noé, and Djork-Arné Clevert. Learning continuous and data-driven molecular descriptors by translating equivalent chemical representations. *Chemical science*, 10(6):1692–1701, 2019.

[212] M. Wójcikowski, M. Kukiełka, M. M. Stepniewska-Dziubinska, and P. Siedlecki. Development of a protein–ligand extended connectivity (PLEC) fingerprint and its application for binding affinity predictions. *Bioinformatics*, 35(8):1334–1341, 2019.

[213] Chengyuan Wu, Shiquan Ren, Jie Wu, and Kelin Xia. Weighted (co)homology and weighted Laplacian. *arXiv preprint arXiv:1804.06990*, 2018.

[214] K. D. Wu and G. W. Wei. Quantitative toxicity prediction using topology based multi-task deep neural networks. *Journal of chemical information and modeling*, page 10.1021/acs.jcim.7b00558, 2018.

[215] K. D. Wu, Z. X. Zhao, R. X. Wang, and G. W. Wei. TopP–S: Persistent homology-based multi-task deep neural networks for simultaneous predictions of partition coefficient and aqueous solubility. *Journal of computational chemistry*, 39(20):1444–1454, 2018.

[216] K. L. Xia and G. W. Wei. Persistent homology analysis of protein structure, flexibility and folding. *International Journal for Numerical Methods in Biomedical Engineerings*, 30:814–844, 2014.

[217] K. L. Xia and G. W. Wei. Multidimensional persistence in biomolecular data. *Journal Computational Chemistry*, 36:1502–1520, 2015.

[218] K. L. Xia and G. W. Wei. Persistent topology for cryo-EM data analysis. *International Journal for Numerical Methods in Biomedical Engineering*, 31:e02719, 2015.

[219] K. L. Xia, Z. X. Zhao, and G. W. Wei. Multiresolution topological simplification. *Journal Computational Biology*, 22:1–5, 2015.

[220] Peng Xiong, Chengxin Zhang, Wei Zheng, and Yang Zhang. Bindprofx: assessing mutation-induced binding affinity change by protein interface profiles with pseudo-counts. *Journal of molecular biology*, 429(3):426–434, 2017.

[221] Youjun Xu, Jianfeng Pei, and Luhua Lai. Deep learning based regression and multiclass models for acute oral toxicity prediction with automatic chemical feature extraction. *Journal of chemical information and modeling*, 57(11):2672–2685, 2017.

[222] Y. Yao, J. Sun, X. H. Huang, G. R. Bowman, G. Singh, M. Lesnick, L. J. Guibas, V. S. Pande, and G. Carlsson. Topological methods for exploring low-density states in biomolecular folding pathways. *The Journal of Chemical Physics*, 130:144115, 2009.

[223] S.s.-T Yau, Jiasong Wang, Amir Niknejad, Chaoxiao Lu, Ning Jin, and Yee-Kin Ho. Dna sequence representation without degeneracy. *Nucleic acids research*, 31:3078–80, 07 2003.

[224] S.s.-T Yau, Chenglong Yu, and Rong He. A protein map and its application. *DNA and cell biology*, 27:241–50, 06 2008.

[225] Chenglong Yu, Rong He, and Stephen Yau. Viral genome phylogeny based on Lempel-Ziv complexity and Hausdorff distance. *Journal of theoretical biology*, 348, 01 2014.

[226] R. Zandi and D. Reguera. Mechanical properties of viral capsids. *Physical Review E*, 72(2):021917, 2005.

[227] Ning Zhang, Yuting Chen, Haoyu Lu, Feiyang Zhao, Roberto Vera Alvarez, Alexander Goncearenco, Anna R Panchenko, and Minghui Li. MutaBind2: predicting the impacts of single and multiple mutations on protein-protein interactions. *Iscience*, 23(3):100939, 2020.

[228] R. D. Zhao, Z. X. Cang, Y. Y. Tong, and G. W. Wei. Protein pocket detection via convex hull surface evolution and associated Reeb graph. *Bioinformatics*, 34(17):i830–i837, 2018.

[229] Rundong Zhao, Menglun Wang, Jiahui Chen, Yiying Tong, and Guo-Wei Wei. The de Rham-Hodge analysis and modeling of biomolecules. *Bulletin of Mathematical Biology*, 82, 08 2020.

[230] Guangyu Zhou, Muhao Chen, Chelsea JT Ju, Zheng Wang, Jyun-Yu Jiang, and Wei Wang. Mutation effect estimation on protein–protein interactions using deep contextualized representation learning. *NAR genomics and bioinformatics*, 2(2):lqaa015, 2020.

[231] B. Zielinski, M. Juda, and M. Zeppelzauer. Persistence codebooks for topological data analysis. *arXiv preprint arXiv:1802.04852*, 2018.

[232] A. Zomorodian and G. Carlsson. Computing persistent homology. *Discrete Comput. Geom.*, 33:249–274, 2005.