ASPECTS OF COMPUTATIONAL TOPOLOGY AND MATHEMATICAL VIROLOGY

By

Rui Wang

A DISSERTATION

Submitted to
Michigan State University
in partial fulfillment of the requirements
for the degree of

Applied Mathematics – Doctor of Philosophy

2022

ABSTRACT

ASPECTS OF COMPUTATIONAL TOPOLOGY AND MATHEMATICAL VIROLOGY

By

Rui Wang

Being able to describe the shape of data is of paramount importance to the fields of biology, physics, chemistry, pharmaceutics, etc. Therefore, in recent years, scientists from the TDA community have been applying advanced mathematical tools to decode the topological structures of data. Methods such as persistent homology, path homology, and de Rham-Hodge theory have become the main workhorse of TDA, which pioneered new branches in algebraic topology and differential geometry. Later, various topological Laplacians such as graph Laplacian, Hodge Laplacian, sheaf Laplacian, and Dirac Laplacian are proposed to preserve topological invariants and geometric shapes simultaneously. However, such Laplacians fail to extract the topological and geometric deformations when one introduces the filtration parameters in. Therefore, we proposed a new topological Laplacians called persistent Laplacians to fully recover the topological persistence and homotopic shape evolution during filtration.

It is worth mentioning that persistent Laplacians are insensitive to asymmetry or directed relations, which limits their power to preserve the directional information of structures in practical applications. Therefore, we proposed persistent path Laplacians to overcome this issue. Similar to the persistent Laplacians, one can also extract the topological persistence and geometric deformations during filtration from the persistent path Laplacians by calculating their harmonic and non-harmonic spectra. In addition, the persistent path Laplacians are constructed on the directed graphs or network, which address the importance of directional representation in datasets such as gene regulation datasets in biology.

Versatile mathematical tools have been playing an essential role in various biological

applications. Since the first COVID-19 case was reported in December 2019, researchers worldwide have been pursuing scientific endeavors in the SARS-CoV-2 projects. Instead of designing promising vaccines and antibody therapies that required wet lab resources, we proposed a new mathematical-AI model called TopNetmAb to systematically analyze the mutation-induced impacts on the SARS-CoV-2 infectivity, vaccines, and antibody drugs. In this dissertation, the topological data analysis (including the persistent Laplacians mentioned above), artificial intelligence, various network models, and genomics analysis are all included in our SARS-CoV-2-related projects to provide comprehensive representations for the understanding of the transmission and evolution of SARS-CoV-2.

This thesis is dedicated to my family.

ACKNOWLEDGEMENTS

Especially, I would like to thank Prof. Guowei Wei for being an energetic, meticulous, and long-headed advisor throughout my Ph.D. journey. He is an outstanding advisor with enough foresight in various research topics, which steered me to tackle projects that yielded fruitful results. Furthermore, his deep thinking and intuitive explanations have enhanced my comprehension of how mathematical tools will benefit life sciences, which has led to my firm interest in the field of mathematical biology. I genuinely appreciate him for providing valuable opportunities and constant support during my graduate studies. Without his encouragement, I probably would not have chosen to pursue a career in academia.

Importantly, I would like to express gratitude to my committee members, Professors Chichia Chiu, Moxun Tang, Yiying Tong, and Jeanne Wald, for their precious comments and feedback. Furthermore, I want to thank Professors Yiying Tong and Ping Wang for welcoming me into their research projects, which have been invaluable to my interdisciplinary research experience. Also, I would like to thank Professor Changchuan Yin for teaching me multiple techniques in genetic biology that are indispensable to the applications of my thesis. Besides, I genuinely appreciated my lab mates Drs. Jiahui Chen, Kaifu Gao, and Duc Nguyen for helping me in all our collaboration projects against SARS-CoV-2 since 2020. Leading by Prof. Wei, we had a quite tacit collaboration and it was probably one of the best collaborative experience I have ever had.

Particularly, I would like to thank Prof. Jeanne Wald and her husband, Mark, for the cherished time we spent together. I deeply appreciated their experienced and valuable advice on my career and life. In addition, I would like to thank my lab mates, colleagues, and friends for their helpful comments and suggestions. Further, I want to thank Drs. Jiahui Chen, Duc Nguyen, Kaifu Gao, Menglun Wang, Timothy Szocinski, Zixuan Cang, Dong Chen, and Mrs. Jing Huang for their friendly conversations and helpful discussions

of diverse topics during my time at MSU.

Undoubtedly, my husband, Shihao Liu, deserves all my thanks. His supportive words of encouragement have been keeping me at ease and focused during this journey. Thank you, Shihao, for all the companionship and understanding you have given to enrich my research career. Most importantly, I should thank my parents, Ling Lu and Yunru Wang, for the enduring sacrifices they have made for me. My Ph.D. journey would have never started without their support. I do sincerely thank them for their endless and unrequited support and love. Foremost, I am grateful to my grandparents, Guojun Gao and Weizhong Lu, who raised me and gave me a great childhood with plenty of family activities, indoor and outdoor. They shaped my character and set an example for me to be a gentle and decent person. My late grandfather Weizhong Lu, passed away in 2020 at the age of 81. I will always miss him.

Lastly, I would like to acknowledge that all of the work included in this thesis was supported in part by NIH grants R01GM126189 and R01AI164266, NSF grants DMS-2052983, DMS-1761320, IIS-1900473, and NASA grant 80NSSC21M0023.

TABLE OF CONTENTS

LIST OF TABLES

LIST OF FIGURES

CHAPTER 1

INTRODUCTION

1.1 Topological Laplacian

Persistent homology (PH) is one of the most popular tools in topological data analysis (TDA), which is constrained to purely topological persistence obtained from its persistent betti numbers. PH has had tremendous success in various fields such as biology [4], chemistry [5], drug discovery [6], and 3D shape analysis [7]. Inspired by the success of PH, multiple advanced mathematical tools in TDA have emerged, and one of the new rising stars in TDA is the de Rham-Hodge theory in differential geometry. De Rham-Hodge theory aims to use the differential forms to represent the cohomology of an oriented closed Riemannian manifold with boundary in terms of a topological Laplacian named Hodge Laplacian [8]. Similar to homology, the de Rham-Hodge theory fails to give an in-depth analysis of data through Hodge Laplacians. Therefore, the evolutionary de Rham-Hodge theory [9] was introduced to alleviate or heal problems arising in the de Rham-Hodge. A persistent Hodge Laplacian was developed to offer a multiscale-level analysis on a family of evolutionary manifolds. Such a method provides an answer to the old question “can one hear the shape of a drum" [10]. One can decode the topological persistence and the homotopic shape evolution of data during filtration by calculating the harmonic and non-harmonic spectra of persistent Hodge Laplacians.

Nonetheless, one main concern we should address in evolutionary de Rham-Hodge theory is that it is set up on the Riemannian manifold, which is quite computational-consuming in real applications. Therefore, seeking a method that can reduce the computational complexity is indeed needed. One natural idea to overcome this issue is to set up a similar system on the discrete points instead of the Riemannian manifold. Hence, a multiscaled-based topological Laplacian, namely persistent spectral graph (PSG) [11],

was introduced by creating low-dimensional multiscale representations (i.e., persistent combinatorial graph Laplacians, , persistent Laplacians) on graphs. In PSG theory, families of persistent Laplacian matrices (PLMs) corresponding to various topological dimensions are constructed via filtration to sample a given dataset at multiple scales. The harmonic spectra from the null spaces of PLMs offer the same topological invariants, namely persistent Betti numbers, at various dimensions as those provided by PH, while the non-harmonic spectra of PLMs give rise to additional geometric analysis of the shape of the data. Meanwhile, we developed an open-source software package called highly efficient robust multidimensional evolutionary spectra (HERMES), to enable broad applications of PSGs in science, engineering, and technology. To ensure the reliability and robustness of HERMES, we have validated the software with simple geometric shapes and complex datasets from three-dimensional (3D) protein structures. We found that the smallest non-zero eigenvalues are very sensitive to data abnormality.

It is noticed that the persistent Laplacians are insensitive to asymmetry or directed relations (i.e, they treat all data points equally). That is to say, each point does not carry any labeled information such as the type, mass, color, etc. Therefore, they fail to represent the structures that have directional information. Undoubtedly, we need a method that has a flavor to deal with asymmetry structures. Notably, the path homology [12] proposed by Grigor'yan, Yong Lin, Yuri Muranov, and S.-T.Yau provides a powerful tool to analyze datasets with asymmetric structures. To encode richer information, Chowdhury and Mémoli extended path homology to a persistent framework on a directed network [13] call persistent path homology (PPH). Such methods are perfect tools for us to fix the aforementioned issue in the persistent Laplacian. Similar to the PH, PPH also decodes purely topological persistence and cannot track the homotopic shape evolution of data during filtration. To overcome the limitation of PPH, persistent path Laplacian (PPL) is introduced to capture the shape evolution of data. PPL's harmonic spectra fully recover PPH's topological persistence and its non-harmonic spectra reveal the homotopic shape

evolution of data during filtration.

Topological Laplacians are powerful tools to extract both topological invariants and geometric deformation of a given system. In this dissertation, we mainly discuss two new multiscale-based topological Laplacians: persistent Laplacians and persistent path Laplacians, and their applications in life science, especially in the fields of molecular biology.

1.2 Mathematical Modeling of Virology

Since its first case was identified in Wuhan, China, in December 2019, coronavirus disease 2019 (COVID-19) caused by severe acute respiratory syndrome coronavirus 2 (SARS-CoV-2) has expeditiously spread to as many as 226 countries and territories worldwide and led to over 541 million confirmed cases and over 6.3 million fatalities as of June 2022. This pandemic has also brought a massive economic recession globally. The countries all around the world have implemented a variety of policies to tackle the COVID-19 pandemic.

Many SARS-CoV-2 vaccines and monoclonal antibodies (mAbs) have already obtained the use authorization worldwide (See Coronavirus Vaccine Tracker). Additionally, U.S. Food and Drug Administration (FDA) has given the emergency use authorization to the oral SARS-CoV-2 Mpro inhibitor PAXLOVID (PF-07321332) developed by Pfizer[14, 15]. However, COVID-19 has a high infection rate, high prevalence, long incubation period [16], asymptomatic transmission [17, 18, 19], and potential seasonal pattern [20]. SARS-CoV-2 keeps involving into new infectious and antibody resistant variants [21, 22, 23]. Therefore, it is imperative to understand its viral molecular mechanism [24], track its genetic evolution [25], and continuously improve the efficacy of antiviral drugs and antibody therapies.

Belonging to the β-coronavirus genus and coronaviridae family, SARS-CoV-2 is an unsegmented positive-sense single-stranded RNA (+ssRNA) virus with a compact 29,903

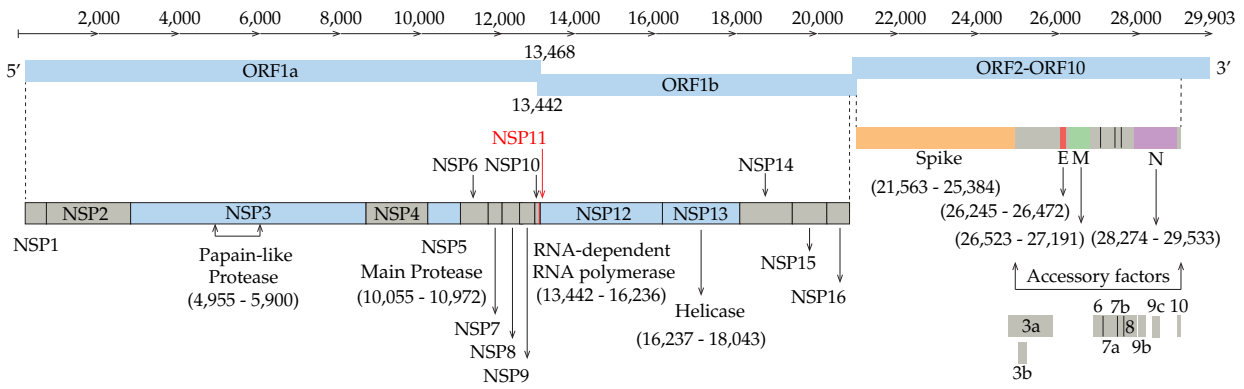

Figure 1.1: Genomics organization of SARS-CoV-2.

nucleotide-long genome and the diameter of each SARS-CoV-2 virion is about 50-200 nm [26]. In the first 20 years of the 21st century, β-coronaviruses have triggered three major outbreaks of deadly pneumonia: SARS-CoV (2002), Middle East respiratory syndrome coronavirus (MERS-CoV) (2012), and SARS-CoV-2 (2019) [27]. Like SARS-CoV and MERS-CoV, SARS-CoV-2 also causes respiratory infections, but at a much higher infection rate [28, 29]. The complete genome of SARS-CoV-2 comprises 15 open reading frames (ORFs), which encodes 29 structural and non-structural proteins (nsps). The 16 non-structural proteins nsp1-nsp16 get expressed by protein-coding genes ORF1a and ORF1b, while four canonical 3′ structural proteins: spike (S), envelope (E), membrane (M), and nucleocapsid (N) proteins, as well as accessory factors, are encoded by other four major ORFs, namely ORF2, ORF4, ORF5, and ORF9 (see Figure 1.1) [30, 31, 32, 33].

The viral structure of SARS-CoV-2 can be found in Figure 1.1. This structure is formed by the four structural proteins: the N protein holds the RNA genome, the S protein helps virus enter into the host cell, and M and E proteins define the shape of the viral envelope [34]. The studies on SARS-CoV-2 as well as previous SARS-CoV and other coronaviruses have mostly identified the functions of these structural proteins, nonstructural proteins as well as accessory proteins. Their 3D structures are also largely known from experiments or predictions.

With these SARS-CoV-2 proteins, the intracellular viral life cycle of SARS-CoV-2 can

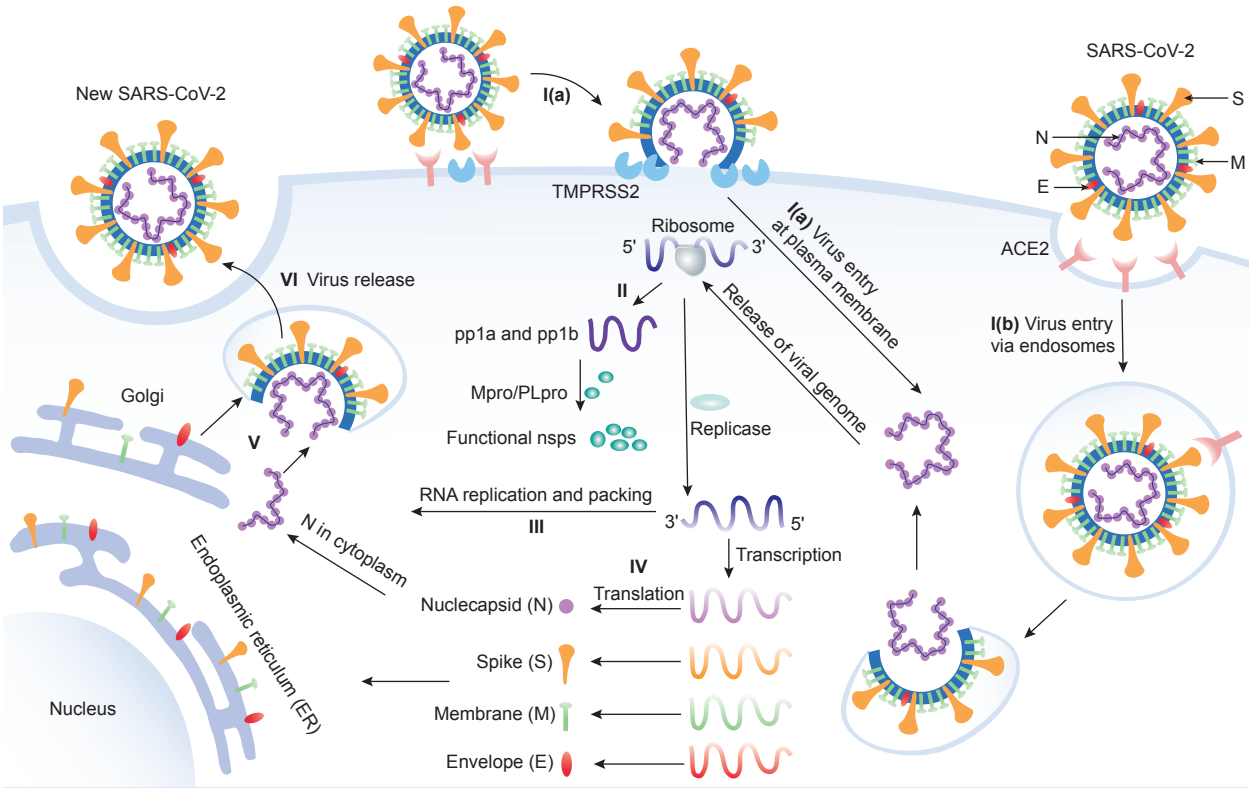

Figure 1.2: Six stages of the SARS-CoV-2 life cycle. Stage **I**: Virus entry. **I**(a) Virus can enter the host cell via plasma membrane fusion. **I**(b) Virus can enter the host cell via endosomes. Stage **II**: Translation of viral replication. Stage **III**: Replication. Here, nsp12 (RdRp) and nsp13 (helicase) cooperate to perform the replication of the viral genome. Stage **IV**: Translation of viral structure proteins. Stage **V**: Virion assembly. Stage **VI**: Release of a virus.

be realized [35]. This life cycle has six stages as shown in Figure 1.2. The first stage is the entry of the virus. SARS-CoV-2 enters the host cell either via endosomes or plasma membrane fusion. In both ways, the S protein of SARS-CoV-2 first attaches to the host cell-surface protein, angiotensin-converting enzyme 2 (ACE2). Then, the cell's protease, TMPRSS2, cuts and opens the S protein of the virus, exposing a fusion peptide in the S2 subunit of the S protein [36]. After fusion, an endosome forms around the virion, separating it from the rest of the host cell. The virion escapes when the pH of the endosome drops or when cathepsin, a host cysteine protease, cleaves it. The virion then releases its RNA into the cell [37]. After the RNA release, polyproteins pp1a and pp1ab are translated. Notably, facilitated by viral papain-like protease (PLpro), nsp1, nsp2, nsp3, and the

amino terminus of nsp4 from the pp1a and pp1ab are released. Moreover, nsp5-nsp16 are also cleaved proteolytically by the main protease [38]. The next stage of the life cycle is the replication process, where nsp12 (RdRp) and nsp13 (helicase) cooperate to perform the replication of the viral genome. Stages IV and V are the translation of viral structural proteins and the virion assembly process. In these stages, structural proteins S, E, and M are translated by ribosomes and then present on the surface of the endoplasmic reticulum (ER), which is transported from the ER through the Golgi apparatus for the preparation of virion assembly. Meanwhile, multiple copies of N protein package the genomics RNA in cytoplasm, which interacts with other 3 structural proteins to direct the assembly of virions. Finally, virions will be secreted from the infected cell through exocytosis.

Since the initial outbreak of the COVID-19, the raging pandemic caused by SARS-CoV-2 has lasted over two years. We do have many promising vaccines, but they might have side effects and their full side effects, particularly, long-term side effects, remain unknown. To make things worse, near 29260 unique mutations have been recorded for SARS-CoV-2 as shown by Mutation Tracker (https://users.math.msu.edu/users/weig/SARS-CoV-2_Mutation_Tracker.html). All of these reveal the sad reality that our current understanding of life science, virology, epidemiology, and medicine is severely limited. Ultimately, the core of challenges is the lack of molecular mechanistic understandings of many aspects, namely coronavirus RNA proofreading, virus-host cell interactions, antibody-antigen interactions, protein-protein interactions, protein-drug interactions, viral regulation of host cell functions, including autophagocytosis and apoptosis, and irregular host immune response behavior such as cytokine storm and antibody-dependent enhancement. Molecular-level experiments on SARS-CoV-2 are both expensive and time-consuming and require to take heavy safety measures. Moreover, disparities among reported experimental binding affinities can be more than 100 fold for the receptor-binding domain (RBD) of S protein binding to ACE2 or antibodies (see Table 1 of Ref. [39]). All these complicated realities make the understanding of viral evolution and transmission

mechanism some of the most challenging tasks.

On the other hand, computational tools provide alternative approaches in understanding viral evolution and transmission with higher efficiency and lower costs. The increasing computer power, the accumulation of molecular data, the availability of artificial intelligence (AI) algorithms, and the development of new mathematical tools have paved the road for mechanistic understanding from molecular modeling, simulations, and predictions.

In May 2020, we developed an intensively validated topology-based neural network model [40] called TopNetmAb to predict certain RBD mutations. It showed that RBD residues 452 and 501 were predicted to "have very high chances to mutate into significantly more infectious COVID-19 strains" in summer 2020 [41] and were later confirmed in prevailing SARS-CoV-2 variants Alpha, Beta, Gamma, Delta, Theta, Epsilon, Kappa, Lambda, Mu, and Omicron. These predictions [41], achieved via the integration of deep learning, biophysics, genotyping, and advanced mathematics, are some of the most remarkable events.

Additionally, 3,696 possible RBD mutations were classified into three categories with different appearance likelihoods, namely, 1149 most likely, 1912 likely, and 625 unlikely [41]. The predicted "most likely" partition successfully contained all the newly observed RBD mutations, until the recent appearance of S371L from Omicron BA.1. Most remarkably, the mechanism governing SARS-CoV-2 evolution and transmission, i.e., natural selection via mutation-strengthened infectivity, was discovered in July 2020 [41] when there were only 89 RBD mutations with the highest observed frequency of merely 50 globally [41].

In April 2021, this mechanism was confirmed beyond any doubt. By using 506,768 sequences isolated from patients, the authors demonstrated that the predicted binding free energy (BFE) changes of the 100 most observed RBD mutations out of 651 existing RBD mutations are all above the BFE change of -0.28 kcal/mol, indicating evolution fa-

vors variants having higher infectivity [2]. Moreover, using network-based modeling for drug repurposing, it was found out Baricitinib as a potential treatment for COVID-19[42]. These extraordinary results prove that mathematical modeling of virology spearhead the discovery of new drugs and the mechanisms of SARS-CoV-2 evolution and transmission.

1.3 Outline

In Chapter 2, we provide a mathematical background in two topological Laplacians: persistent Laplacians and persistent path Laplacians. Also, vital examples are involved to illustrate how we construct two types of topological Laplacians on a given point-cloud dataset. In Chapter 3, we review the theoretically details in the mathematical modeling of virology, including the methods in the genomics analysis and the structure of the math-AI models that we used in the SARS-CoV-2 studies. In Chapter 4, we mainly discuss the applications in the PL and PPL, and their advantages compared to other topological Laplacians. We further introduce an open-source package called HERMES, which is designed to extract the harmonic and non-harmonic spectra of persistent Laplacians. In addition, the validation of the HERMES is also discussed in the Chapter 5 to show its accuracy, robustness, and reliability on standard test datasets and multiple complex protein structures. Chapter 6 includes several applications in the study of SARS-CoV-2, including the mutational impacts on the SARS-CoV-2 diagnostic targets, vaccines, antibodies, along with the discussion about the mechanisms of SARS-CoV-2 evolution and transmission. The dissertation contribution is summarized in Chapter 7.

CHAPTER 2

METHODS ON TOPOLOGICAL LAPLACIANS

2.1 Persistent Laplacians

2.1.1 Simplex

Let $\{v_0, v_1, \cdots, v_q\}$ be a set of points in $\mathbb{R}^n$. A point $v = \sum_{i=0}^{q} \lambda_i v_i, \lambda_i \in \mathbb{R}$ is an affine combination of v_i if $\sum_{i=0}^{q} \lambda_i = 1$. An affine hull is the set of affine combinations. Here, $q+1$ points $v_0, v_1, \cdots, v_q$ are affinely independent if $v_1 - v_0, v_2 - v_0, \cdots, v_q - v_0$ are linearly independent. A q-plane is well-defined if the $q+1$ points are affinely independent. In $\mathbb{R}^n$, one can have at most n linearly independent vectors. Therefore, there are at most $n+1$ affinely independent points. An affine combination $v = \sum_{i=0}^{q} \lambda_i v_i$ is a convex combination if all λ_i are non-negative. The convex hull is the set of convex combinations.

A (geometric) q-simplex denoted as σ_q is the convex hull of $q+1$ affinely independent points in $\mathbb{R}^q$ with dimension $\dim(\sigma_q) = q$. A 0-simplex is a vertex, a 1-simplex is an edge, a 2-simplex is a triangle, and a 3-simplex is a tetrahedron, as shown in Figure 2.1. The convex hull of each nonempty subset of $q+1$ points forms a subsimplex and is regraded as a face of σ_q denoted τ. The p-face of a q-simplex is the subset $\{v_{i1}, \cdots, v_{ip}\}$ of the q-simplex.

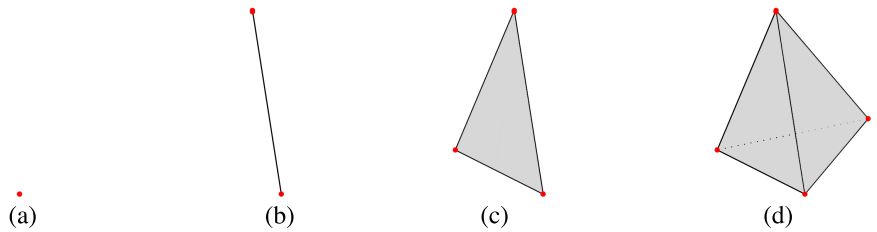

Figure 2.1: Illustration of simplices. (a) 0-simplex (a vertex), (b) 1-simplex (an edge), (c) 2-simplex (a triangle), and (d) 3-simplex (a tetrahedron).

2.1.2 Simplicial Complex

A simplicial complex is a powerful algebraic topology tool that has wide applications in graph theory, topological data analysis [43], and many physical fields [44]. We briefly review simplicial complexes to generate notation and provide essential preparation for introducing persistent spectral graphs. A (finite) simplicial complex K is a (finite) collection of simplices in $\mathbb{R}^n$ satisfying the following conditions

(1) If $\sigma_q \in K$ and σ_p is a face of σ_q, then $\sigma_p \in K$.

(2) The non-empty intersection of any two simplices $\sigma_q, \sigma_p \in K$ is a face of both of σ_q and σ_p.

Each element $\sigma_q \in K$ is a q-simplex of K. The dimension of K is defined as $\dim(K) = \max\{\dim(\sigma_q) : \sigma_q \in K\}$. To distinguish topological spaces based on the connectivity of simplicial complexes, one uses Betti numbers. The k-th Betti number, β_k, counts the number of k-dimensional holes on a topological surface. The geometric meaning of Betti numbers in $\mathbb{R}^3$ is the following: β_0 represents the number of connected components, β_1 counts the number of one-dimensional loops or circles, and β_2 describes the number of two-dimensional voids or holes. In a nutshell, the Betti number sequence $\{\beta_0, \beta_1, \beta_2, \cdots\}$ reveals the intrinsic topological property of the system. To illustrate the simplicial complex and its corresponding Betti number, we have designed two simple models as is shown in Figure 2.2. [1]

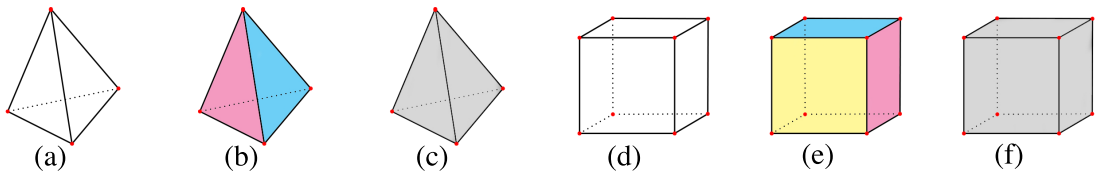

Figure 2.2: Illustrations of simplicial complexes.

[1] These examples show an intuitive way to count Betti numbers. However, it is impossible to generate structures (b), (e), and (f) in Rips complex.

Table 2.1: The Betti number of simplicial complexes in Figure 2.2. Each color represents different faces. The tetrahedron-shaped simplicial complexes are demonstrated in (a)-(c), and the cube-shaped simplicial complexes are depicted in (d) - (f). (a) and (d) only has 0-simplices and 1-simplices, (b) has four 2-simplices, and (c) has one more 4-simplex. (e) and (f) do not have any 2-simplex.

Betti number	Fig. 3 (a)	Fig. 3 (b)	Fig. 3 (c)	Fig. 3 (d)	Fig. 3 (e)	Fig. 3 (f)
β_0	1	1	1	1	1	1
β_1	3	0	0	5	0	0
β_2	0	1	0	0	1	0

Recall that in graph theory, the degree of a vertex (0-simplex) v is the number of edges that are adjacent to the vertex, denoted as $\deg(v)$. However, once we generalize this notion to q-simplex, problem arouse since a q-simplex can have $(q-1)$-simplices and $(q+1)$-simplices adjacent to it at the same time. Therefore, the upper adjacency and lower adjacency are required to define the degree of a q-simplex for $q > 0$ [45, 46].

Defination 2.1.1 *Two q-simplices σ_q^i and σ_q^j of a simplicial complex K are lower adjacent if they share a common $(q-1)$-face, denoted $\sigma_q^i \overset{L}{\sim} \sigma_q^j$. The lower degree of q-simplex, denoted $deg_L(\sigma_q)$, is the number of nonempty $(q-1)$-simplices in K that are faces of σ_q, which is always $q+1$.*

Defination 2.1.2 *Two q-simplices σ_q^i and σ_q^j of a simplicial complex K are upper adjacent if they share a common $(q+1)$-face, denoted $\sigma_q^i \overset{U}{\sim} \sigma_q^j$. The upper degree of q-simplex, denoted $deg_U(\sigma_q)$, is the number of $(q+1)$-simplices in K of which σ_q is a face.*

Then, the degree of a q-simplex ($q > 0$) is defined as:

$$\deg(\sigma_q) = \deg_L(\sigma_q) + \deg_U(\sigma_q) = \deg_U(\sigma_q) + q + 1. \tag{2.1}$$

2.1.2.1 Delaunay Triangulation and Alpha Shapes

In this section, we provide the details on a practical construction of filtration for persistent spectral graph theory based on the alpha complex. The alpha complex can be regarded as a simplicial complex, which is a homotopy equivalent to the nerve of balls around data points. Its geometric realization built as the union of convex hulls of points in each

simplex is called the alpha shape. First proposed in 1983, t he alpha shape defined the shape associated with a finite set of points in the plane controlled by one parameter [47].

In the following, we first describe how to construct the alpha shape, and then provide some necessary concepts for the implementation of the alpha complex in PSG theory. Let P be a finite set of points in qD Euclidean space $\mathbb{R}^q$ ($q = 2$ or 3 in most applications), and α be a positive real number. Denote an open ball with radius α as an alpha ball (α-ball). We say that an α-ball is empty if it contains no point of P, and the alpha hull (α-hull) of P is the set of points that do not belong to any empty α-ball. For any subset $T \subseteq P$ with size $|T| = k + 1, 0 \leq k \leq q$, the geometric realization of k-simplex σ_T is the convex hull of T. We say that a k-simplex σ_T is α-exposed if there exists an empty α-ball $\mathbf{b}$ such that $T = \partial\mathbf{b} \cap P$ for $0 \leq k \leq q-1$. Denoting the collection of α-exposed k-simplices as $F_{k,\alpha}$ for $0 \leq k \leq q-1$, the alpha shape (α-shape) of P is the polytope whose boundary consists of the k-simplices in $F_{k,\alpha}$. The alpha complex is just the simplicial complex that is the collection of the simplices in the alpha shape.

There are two structures that are closely related to the alpha shape and helpful in efficient implementation of alpha shape and alpha complex. One is the Voronoi diagram [48] and the other is its dual structure, the Delaunay tessellation [49]. The latter is the alpha complex for sufficiently large α, e.g., when α is greater than the diameter of P. Thus, the Delaunay tessellation is the final complete simplicial complex in the filtration that we use.

For a given set of points $P = \{p_1, p_2, \cdots, p_n\} \subseteq \mathbb{R}^q$, the Voronoi cell V_i of a point $p_i \in P$ contains all of the points for which p_i is the closest among all the points in P,

$$V_i = \{x \in \mathbb{R}^q \mid \quad \|x - p_i\| \leq \|x - p_j\|, \quad \forall p_j \in P\}. \tag{2.2}$$

The Voronoi diagram of P is the set of Voronoi cells, which is defined as

$$\text{Vor}P = \{V_i \mid \forall i \in \{1, 2, \cdots, |P|\}\}. \tag{2.3}$$

The Delaunay tessellation for a given set P in general position (i.e., no $q+1$ ponits are in

a $(q-1)$-D linear subspace, and no $q+2$ points share the same circumsphere) is the dual simplicial complex to the Voronoi diagrams. For instance, a Delaunay tessellation for a given set P in 2D is a triangulation DT(P) such that no point in P is inside the circumcircle of any triangle in DT(P) [50, 51]. A formal way to define the Delaunay tessellation is to use the nerve of the collection of Voronoi cells (Nrv(Vor$P$)), which can be expressed as

$$\text{DT}(P) = \text{Nrv}(\text{Vor}P) = \{J \subseteq \{1, 2, ..., |P|\} \mid \bigcap_{i \in J} V_i \neq \emptyset\}, \tag{2.4}$$

under the condition that the points in P are general position. Note that, in practice, a set of points that are not in general position can be symbolically perturbed to general position.

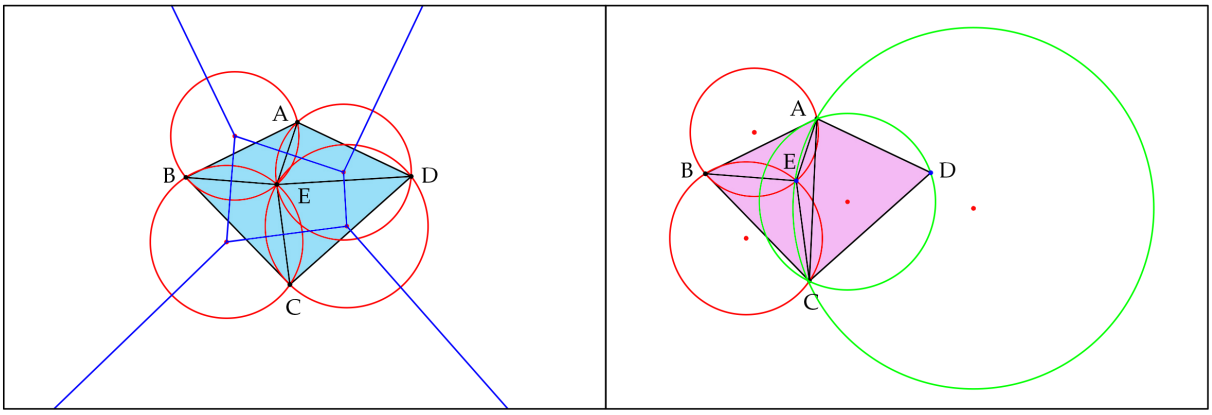

Figure 2.3: Illustration of Voronoi diagram, Delaunay triangulation, and Non-Delaunay triangulation. **Left chart:** The Voronoi diagram and its dual Delaunay triangulation. The points set is P = {A,B,C,D,E} and the Delaunay is defined as DT(P). The blue lines tessellate the plane into Voronoi cells. The red circle are the circumcircles of triangles in DT(P). **Right chart:** A Non-Delaunay triangulation. Vertices E and D are in the green circumcircles, implying the right chart is an example of Non-Delaunay triangulation.

Next, we introduce the mathematical description of the construction of alpha complex through the union of balls centered at points in P, which is essentially a van der Waals surface for atoms positioned at P with the same radius α. For a given set of points $P = \{p_1, p_2, \cdots, p_n\}$ in $\mathbb{R}^q$ and a positive real number α, we can denote the closed ball centered at p_i as $B_i(\alpha) = p_i + \alpha\mathbb{B}^q$, where $\mathbb{B}^q$ is a qD unit ball around the origin. The union of these

balls can be expressed as

$$U(\alpha) = \{x \in \mathbb{R}^q \mid \exists p_i \in P \text{ s.t. } \|x - p_i\| \leq \alpha\}. \tag{2.5}$$

To ensure that we obtain a subcomplex of the Delaunay tessellation, we intersect $B_i(\alpha)$ with its corresponding Voronoi cell,

$$R_i(\alpha) = B_i(\alpha) \cap V_i. \tag{2.6}$$

It can be observed that $U(\alpha) = \cup_{p_i \in P} R_i(\alpha)$, so the R_i's is a covering of $U(\alpha)$. The alpha complex K_α is the simplicial complex representing the nerve of this covering,

$$K_\alpha = \{J \subseteq \{1, 2, ..., |P|\} \mid \bigcap_{i \in J} R_i(\alpha) \neq \emptyset\}. \tag{2.7}$$

The equivalence to the original definition can be readily checked. The union of all simplices in the alpha complex forms the alpha shape. Figure 2.3 illustrates the Voronoi diagram, Delaunay triangulation, and non-Delaunay triangulation. The point set is P = {A,B,C,D,E}, and the blue lines in the left chart of Figure 2.3 separate the plane into the Voronoi cells. The red circles are the empty circumcircles for triples of points in P. We can notice that no four points are on the same red circle, which satisfies the uniqueness condition for constructing the Delaunay triangulation. In the right chart of Figure 2.3, the green circumcircle of ACD contains E and the green circumcirlce of AEC contains D, indicating that those two triangles do not belong to the Delaunay triangulation.

Figure 2.4 illustrates the standard filtration of alpha complexes. The top left figure is the Delaunay triangulation of six 2D points A, B, C, D, E, and F. With an ever-growing radius α centered at these points, a family of sub-complexes of the Delaunay triangulation can be constructed. Figure 2.5 shows the persistence barcode of these 6 points. It can be seen that when $\alpha = 0.2$, all six points are disconnected, indicating that 6 0-cycles (connected components) existed, which matches with Figure 2.5, where there are a total of 6 bars when $\alpha = 0.2$. With the radius α continually increasing, a 1-cycle will be formed, and the associated alpha shape are shown in the bottom left chart of Figure 2.4. One

can notice that in Figure 2.5, when $\alpha = 0.6$, $\beta_1^{\alpha,0} = 1$. When α reaches 0.83, the 1-cycle disappears and $\beta_1^{\alpha,0} = 0$ as shown in the bottom left panel of Figure 2.4. Table 2.2 and Table 2.3 show how we construct the qth-order persistent Laplacian $\mathcal{L}_q^{t,p}$ and calculate the harmonic ($\beta_q^{t,p}$) and non-harmonic persistent spectra of $\mathcal{L}_q^{t,p}$ from the simplicial complexes $K_{0.2}$ to $K_{0.6}$ and $K_{0.6}$ to $K_{0.6}$.

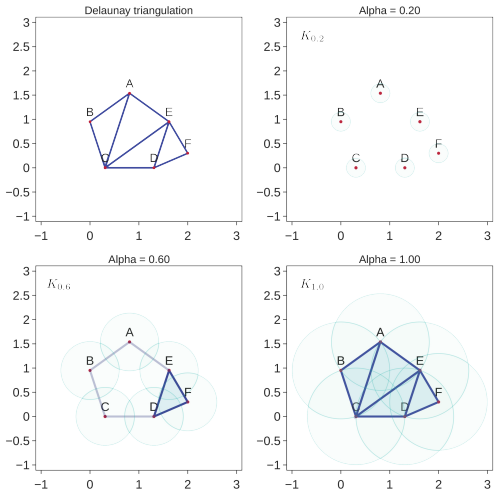

Figure 2.4: Illustration of 2D Delaunay triangulation, alpha shapes, and alpha complexes for a set of 6 points A, B, C, D, E, and F. **Top left**: The 2D Delaunay triangulation. **Top right**: The alpha shape and alpha complex at filtration value $\alpha = 0.2$. **Bottom left** : The alpha shape and alpha complex at filtration value $\alpha = 0.6$. **Bottom right**: The alpha shape and alpha complex at filtration value $\alpha = 1.0$. Here, we use dark blue color to fill the alpha shape.

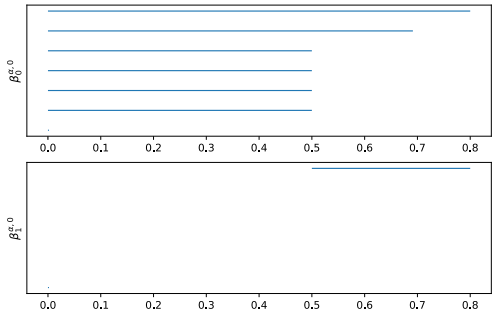

Figure 2.5: The persistent barcode for a set of points as illustrated in Figure 2.4 that are generated from Gudhi and DioDe.

Table 2.2: The matrix representation of q-boundary operator and its qth-order persistent Laplacian with corresponding dimension, rank, nullity, and spectra from alpha complex $K_{0.6} \to K_{0.6}$.

q	$q=0$	$q=1$	$q=2$
$\mathcal{B}_{q+1}^{0.6,0}$	$\begin{array}{c} \\ A \\ B \\ C \\ D \\ E \\ F \end{array}\begin{array}{c}\begin{array}{ccccccc} AB & BC & CD & DE & EF & DF & AE \end{array}\\ \begin{bmatrix} -1 & 0 & 0 & 0 & 0 & 0 & -1 \\ 1 & -1 & 0 & 0 & 0 & 0 & 0 \\ 0 & 1 & -1 & 0 & 0 & 0 & 0 \\ 0 & 0 & 1 & -1 & 0 & -1 & 0 \\ 0 & 0 & 0 & 1 & -1 & 0 & 1 \\ 0 & 0 & 0 & 0 & 1 & 1 & 0 \end{bmatrix}\end{array}$	$\begin{array}{c} \\ AB \\ BC \\ CD \\ DE \\ EF \\ DF \\ AE \end{array}\begin{array}{c}\begin{array}{c} DEF \end{array}\\ \begin{bmatrix} 0 \\ 0 \\ 0 \\ 1 \\ 1 \\ -1 \\ 0 \end{bmatrix}\end{array}$	/
$\mathcal{B}_q^{0.6}$	$\begin{array}{c}\begin{array}{cccccc} A & B & C & D & E & F \end{array}\\ \begin{bmatrix} 0 & 0 & 0 & 0 & 0 & 0 \end{bmatrix}\end{array}$	$\begin{array}{c} \\ A \\ B \\ C \\ D \\ E \\ F \end{array}\begin{array}{c}\begin{array}{ccccccc} AB & BC & CD & DE & EF & DF & AE \end{array}\\ \begin{bmatrix} -1 & 0 & 0 & 0 & 0 & 0 & -1 \\ 1 & -1 & 0 & 0 & 0 & 0 & 0 \\ 0 & 1 & -1 & 0 & 0 & 0 & 0 \\ 0 & 0 & 1 & -1 & 0 & -1 & 0 \\ 0 & 0 & 0 & 1 & -1 & 0 & 1 \\ 0 & 0 & 0 & 0 & 1 & 1 & 0 \end{bmatrix}\end{array}$	$\begin{array}{c} \\ AB \\ BC \\ CD \\ DE \\ EF \\ DF \\ AE \end{array}\begin{array}{c}\begin{array}{c} DEF \end{array}\\ \begin{bmatrix} 0 \\ 0 \\ 0 \\ 1 \\ 1 \\ -1 \\ 0 \end{bmatrix}\end{array}$
$\mathcal{L}_q^{0.6,0}$	$\begin{bmatrix} 2 & -1 & 0 & 0 & -1 & 0 \\ -1 & 2 & -1 & 0 & 0 & 0 \\ 0 & -1 & 2 & -1 & 0 & 0 \\ 0 & 0 & -1 & 3 & -1 & -1 \\ -1 & 0 & 0 & -1 & 3 & -1 \\ 0 & 0 & 0 & -1 & -1 & 2 \end{bmatrix}$	$\begin{bmatrix} 2 & -1 & 0 & 0 & 0 & 0 & 1 \\ -1 & 2 & -1 & 0 & 0 & 0 & 0 \\ 0 & -1 & 2 & -1 & 0 & -1 & 0 \\ 0 & 0 & -1 & 3 & 0 & 0 & 1 \\ 0 & 0 & 0 & 0 & 3 & 0 & -1 \\ 0 & 0 & -1 & 0 & 0 & 3 & 0 \\ 1 & 0 & 0 & 1 & -1 & 0 & 2 \end{bmatrix}$	[3]
$\beta_q^{0.6,0}$	1	1	0
$\dim(\mathcal{L}_q^{0.6,0})$	6	7	1
$\text{rank}(\mathcal{L}_q^{0.6,0})$	5	6	1
$\text{nullity}(\mathcal{L}_q^{0.6,0})$	1	1	0
$\text{Spec}(\mathcal{L}_q^{0.6,0})$	$\{0, 1, 1.5858, 3, 4, 4.4142\}$	$\{0, 1, 1.5858, 3, 3, 4, 4.4142\}$	$\{3\}$

2.1.2.2 Vietoris-Rips Complex

Vietoris-Rips complex is an abstract simplicial complex. It is commonly used in various applications. For a given set of points $P = \{p_1.p_2, \cdots, p_n\}$ in a metric space and a real value $r > 0$, a k-simplex $\sigma_k = [p_{i0}, \cdots, p_{ik}]$ is in the Vietoris-Rips complex if and only if $\mathbb{B}(p_{ij,r}) \cap \mathbb{B}(p_{ij',r}) \neq \emptyset, \forall j, j' \in [0, k]$.

2.1.3 Chain Complex

Chain complex is an important concept in topology, geometry, and algebra. A q-chain is a formal sum of q-simplices in simplicial complex K with $\mathbb{Z}_2$ coefficients. The set of all q-chains has a basis which the set of q-simplices in K, thus forming a finitely generated free abelian group denoted as $C_q(K)$. The boundary operator is a group homomorphism

Table 2.3: The matrix representation of q-boundary operator and its qth-order persistent Laplacian with corresponding dimension, rank, nullity, and spectra from alpha complex $K_{0.2} \to K_{0.6}$.

q	$q=0$	$q=1$	$q=2$
$\mathcal{B}_{q+1}^{0.2,0.4}$	$\begin{array}{c} \\ \text{A} \\ \text{B} \\ \text{C} \\ \text{D} \\ \text{E} \\ \text{F} \end{array}\begin{array}{c}\begin{array}{ccccccc}\text{AB} & \text{BC} & \text{CD} & \text{DE} & \text{EF} & \text{DF} & \text{AE}\end{array}\\ \begin{bmatrix} -1 & 0 & 0 & 0 & 0 & 0 & -1 \\ 1 & -1 & 0 & 0 & 0 & 0 & 0 \\ 0 & 1 & -1 & 0 & 0 & 0 & 0 \\ 0 & 0 & 1 & -1 & 0 & -1 & 0 \\ 0 & 0 & 0 & 1 & -1 & 0 & 1 \\ 0 & 0 & 0 & 0 & 1 & 1 & 0 \end{bmatrix}\end{array}$	/	/
$\mathcal{B}_q^{0.2}$	$\begin{array}{ccccccc}\text{A} & \text{B} & \text{C} & \text{D} & \text{E} & \text{F}\end{array}$ $\begin{bmatrix} 0 & 0 & 0 & 0 & 0 & 0 \end{bmatrix}$	/	/
$\mathcal{L}_q^{0.2,0.4}$	$\begin{bmatrix} 2 & -1 & 0 & 0 & -1 & 0 \\ -1 & 2 & -1 & 0 & 0 & 0 \\ 0 & -1 & 2 & -1 & 0 & 0 \\ 0 & 0 & -1 & 3 & -1 & -1 \\ -1 & 0 & 0 & -1 & 3 & -1 \\ 0 & 0 & 0 & -1 & -1 & 2 \end{bmatrix}$	/	/
$\beta_q^{0.2,0.4}$	1	/	/
$\dim(\mathcal{L}_q^{0.2,0.4})$	6	/	/
$\text{rank}(\mathcal{L}_q^{0.2,0.4})$	5	/	/
$\text{nullity}(\mathcal{L}_q^{0.2,0.4})$	1	/	/
$\text{Spec}(\mathcal{L}_q^{0.2,0.4})$	$\{0, 1, 1.5858, 3, 4, 4.4142\}$	/	/

defined by $\partial_q : C_q(K) \to C_{q-1}(K)$ to relate the chain groups. More specifically, denoting q-simplex as $\sigma_q = [v_0, v_1, \cdots, v_q]$ by its vertices v_i, the boundary operator is defined through its action on the basis,

$$\partial_q \sigma_q = \sum_{i=0}^{q} (-1)^i \sigma_{q-1}^i. \tag{2.8}$$

Here, $\sigma_{q-1}^i = [v_0, \cdots, \hat{v_i}, \cdots, v_q]$ is the $(q-1)$-simplex with v_i omitted. The following sequence of chain groups connected by boundary operators is a *chain complex* (defined as a set of abelian groups connected by homomorphisms such that the composite of any two consecutive homomorphisms is zero, $\partial_q \partial_{q+1} = 0$.)

$$\cdots \xrightarrow{\partial_{q+2}} C_{q+1}(K) \xrightarrow{\partial_{q+1}} C_q(K) \xrightarrow{\partial_q} C_{q-1}(K) \xrightarrow{\partial_{q-1}} \cdots$$

2.1.4 Combinatorial Laplacians

Combinatorial Laplacians[52] offer both spectral analysis and topological analysis [53]. One central role played by the chain complex associated with a simplicial complex is to define its q-th homology group ($H_q = \ker \partial_q / \operatorname{im} \partial_{q+1}$), which is a topological invariant of the simplicial complex. The dimension of H_q is denoted by $\beta_q = \dim H_q$, the q-th Betti number, which, roughly speaking, measures the number of q-dimensional holes in the simplicial complex, or the geometric object tessellated into the simplicial complex.

A dual chain complex can be defined on any chain complex through the adjoint operator of ∂_q defined on the dual spaces $C^q(K) = C_q^*(K)$. The q-coboundary operator $\partial_q^* : C^{q-1}(K) \to C^q(K)$ is defined as:

$$\partial^* \omega^{q-1}(c_q) \equiv \omega^{q-1}(\partial c_q), \tag{2.9}$$

where $\omega^{q-1} \in C^{q-1}(K)$ is a $(q-1)$-*cochain*, which is a homomorphism mapping a chain to the coefficient group, and $c_q \in C_q(K)$ is a q-chain. The homology of the dual chain complex is often called *cohomology*.

If we denote by $\mathcal{B}_q$ the matrix representation of a q-boundary operator with respect to the standard basis for $C_q(K)$ and $C_{q-1}(K)$, the number of rows and the number of columns in $\mathcal{B}_q$ correspond to the number of $(q-1)$-simplices and that of q-simplices in K, respectively. Moreover, the matrix representation of q-coboundary operator is denoted $\mathcal{B}_q^T$.

In de Rham-Hodge theory, homology and cohomology are often studied through their correspondences to the q-combinatorial Laplacian operator, defined as the linear operator $\Delta_q : C^q(K) \to C^q(K)$ as follows,

$$\Delta_q := \partial_{q+1} \partial_{q+1}^* + \partial_q^* \partial_q, \tag{2.10}$$

where the isomorphism $C^q(K) \cong C_q(K)$ is assumed, where each q-simplex is mapped to its own dual, i.e., the isomorphism keeps the coefficients of chains and cochains in the

standard simplicial basis. Correspondingly, the matrix representation of Δ_q is the qth-order Laplacian, which is denoted $\mathcal{L}_q(K)$,

$$\mathcal{L}_q(K) = \mathcal{B}_{q+1}\mathcal{B}_{q+1}^T + \mathcal{B}_q^T\mathcal{B}_q. \tag{2.11}$$

Assume the number of q-simplices existing in K to be N_q, then $\mathcal{L}_q(K)$ is an $N_q \times N_q$-matrix. Since the qth-order Laplacian $\mathcal{L}_q(K)$ is symmetric and positive semi-definite, its spectrum consists of only real and non-negative eigenvalues. We denote the spectrum of $\mathcal{L}_q(K)$ as

$$\text{Spec}(\mathcal{L}_q(K)) = \{\lambda_{1,q}, \lambda_{2,q}, \cdots, \lambda_{N_q,q}\}.$$

The multiplicity of zero in the spectrum (also called the harmonic spectrum) reveals the topological information β_q, whereas the non-harmonic spectrum encodes further geometric information. The correspondence between the multiplicity of zero spectra of $\mathcal{L}_q(K)$ and the qth Betti number defined in the homology is an important result in de Rham-Hodge theory, [54, 55, 56]

$$\beta_q = \dim\ker\partial_q - \dim\text{im}\,\partial_{q+1} = \dim\ker\mathcal{L}_q(K) = \#0 \text{ eigenvalues of } \mathcal{L}_q(K). \tag{2.12}$$

Intuitively, β_0 represents the number of connected components in K, β_1 reveals the number of 1D noncontractible loops or circles in K, and β_2 shows the number of 2D voids or cavities in K.

2.1.5 Persistent Laplacian

Both topological and geometric information can be derived from analyzing the spectra of qth-order Laplacian. However, the information is restricted to those pieces contained in the connectivity of the simplicial complex. A single simplicial complex produces insufficient information for practical problems such as feature extraction for machine learning analysis. To enrich the spectral information, persistent spectral graph (PSG) is proposed by creating a sequence of simplicial complexes induced by varying a filtration parameter,

which is inspired by persistent homology as well as our earlier multiscale graph Laplacians [57].

First, we consider a filtration of simplicial complex K which is a nested sequence of subcomplexes $(K_t)_{t=0}^m$ of the final complex K:

$$\emptyset = K_0 \subseteq K_1 \subseteq K_2 \subseteq \cdots \subseteq K_m = K. \tag{2.13}$$

For each subcomplex K_t, we denote its corresponding chain group to be $C_q(K_t)$, and the q-boundary operator will be denoted by $\partial_q^t : C_q(K_t) \to C_{q-1}(K_t)$. As conventionally done, we define $C_q(K_t)$ for $q < 0$ as the zero group $\{0\}$ and ∂_q^t as a zero map. [2] If $0 < q \leq \dim K_t$, then

$$\partial_q^t(\sigma_q) = \sum_i^q (-1)^i \sigma_{q-1}^i, \quad \forall \sigma_q \in K_t, \tag{2.14}$$

with $\sigma_q = [v_0, \cdots, v_q]$ being any q-simplex, and $\sigma_{q-1}^i = [v_0, \cdots, \hat{v}_i, \cdots, v_q]$ being the $(q-1)$-simplex constructed by removing v_i . The adjoint operator of ∂_q^t is the coboundary operator $\partial_q^{t^*} : C^{q-1}(K_t) \to C^q(K_t)$, which can be regarded as a map from $C_{q-1}(K_t)$ to $C_q(K_t)$ through the isomorphisms $C^q(K_t) \cong C_q(K_t)$ between cochain groups and chain groups.

Similar to the persistent homology, a sequence of chain complexes can be defined as below:

$$\begin{array}{ccccccccccccccc}
\cdots & C_{q+1}^1 & \underset{\partial_{q+1}^{1*}}{\overset{\partial_{q+1}^1}{\rightleftarrows}} & C_q^1 & \underset{\partial_q^{1*}}{\overset{\partial_q^1}{\rightleftarrows}} & \cdots & \underset{\partial_3^{1*}}{\overset{\partial_3^1}{\rightleftarrows}} & C_2^1 & \underset{\partial_2^{1*}}{\overset{\partial_2^1}{\rightleftarrows}} & C_1^1 & \underset{\partial_1^{1*}}{\overset{\partial_1^1}{\rightleftarrows}} & C_0^1 & \underset{\partial_0^{1*}}{\overset{\partial_0^1}{\rightleftarrows}} & C_{-1}^1 = \{0\} \\
 & \rotatebox{90}{$\subseteq$} & & \rotatebox{90}{$\subseteq$} & & & & \rotatebox{90}{$\subseteq$} & & \rotatebox{90}{$\subseteq$} & & \rotatebox{90}{$\subseteq$} & & \\
\cdots & C_{q+1}^2 & \underset{\partial_{q+1}^{2*}}{\overset{\partial_{q+1}^2}{\rightleftarrows}} & C_q^2 & \underset{\partial_q^{2*}}{\overset{\partial_q^2}{\rightleftarrows}} & \cdots & \underset{\partial_3^{2*}}{\overset{\partial_3^2}{\rightleftarrows}} & C_2^2 & \underset{\partial_2^{2*}}{\overset{\partial_2^2}{\rightleftarrows}} & C_1^2 & \underset{\partial_1^{2*}}{\overset{\partial_1^2}{\rightleftarrows}} & C_0^2 & \underset{\partial_0^{2*}}{\overset{\partial_0^2}{\rightleftarrows}} & C_{-1}^2 = \{0\} \\
 & \vdots & & \vdots & & & & \vdots & & \vdots & & \vdots & & \vdots \\
 & \rotatebox{90}{$\subseteq$} & & \rotatebox{90}{$\subseteq$} & & & & \rotatebox{90}{$\subseteq$} & & \rotatebox{90}{$\subseteq$} & & \rotatebox{90}{$\subseteq$} & & \\
\cdots & C_{q+1}^m & \underset{\partial_{q+1}^{m*}}{\overset{\partial_{q+1}^m}{\rightleftarrows}} & C_q^m & \underset{\partial_q^{m*}}{\overset{\partial_q^m}{\rightleftarrows}} & \cdots & \underset{\partial_3^{m*}}{\overset{\partial_3^m}{\rightleftarrows}} & C_2^m & \underset{\partial_2^{m*}}{\overset{\partial_2^m}{\rightleftarrows}} & C_1^m & \underset{\partial_1^{m*}}{\overset{\partial_1^m}{\rightleftarrows}} & C_0^m & \underset{\partial_0^{m*}}{\overset{\partial_0^m}{\rightleftarrows}} & C_{-1}^m = \{0\}
\end{array} \tag{2.15}$$

[2] We define the boundary matrix $\mathcal{B}_0^t$ for the boundary operator ∂_0^t as a zero matrix. The number of columns of $\mathcal{B}_0^t$ is the number of 0-simplices in K_t, the number of rows will be 1.

For simplicity, we use C_q^t to denote the chain group $C_q(K_t)$.

Next, we introduce persistence to the Laplacian spectra. We define the subset of C_q^{t+p} whose boundary is in C_{q-1}^t as $\mathbb{C}_q^{t,p}$, assuming the natural inclusion map from C_{q-1}^t to C_{q-1}^{t+p},

$$\mathbb{C}_q^{t,p} := \{\beta \in C_q^{t+p} \mid \partial_q^{t+p}(\beta) \in C_{q-1}^t\}. \tag{2.16}$$

On this subset, one may define the p-persistent q-boundary operator denoted by $\eth_q^{t,p} : \mathbb{C}_q^{t,p} \to C_{q-1}^t$. Its corresponding adjoint operator is $(\eth_q^{t,p})^* : C_{q-1}^t \to \mathbb{C}_q^{t,p}$, again through the identification of cochains with chains. We then define the q-order p-persistent Laplacian operator $\Delta_q^{t,p} : C_q^t \to C_q^t$ associated with the filtration as

$$\Delta_q^{t,p} = \eth_{q+1}^{t,p} \left(\eth_{q+1}^{t,p}\right)^* + \partial_q^{t^*} \partial_q^t. \tag{2.17}$$

The matrix representation of $\Delta_q^{t,p}$ in the simplicial basis is

$$\mathcal{L}_q^{t,p} = \mathcal{B}_{q+1}^{t,p} (\mathcal{B}_{q+1}^{t,p})^T + (\mathcal{B}_q^t)^T \mathcal{B}_q^t, \tag{2.18}$$

where $\mathcal{B}_{q+1}^{t,p}$ is the matrix representation of $\eth_{q+1}^{t,p}$.

We denote the spectrum of $\mathcal{L}_q^{t,p}$ as

$$\text{Spec}(\mathcal{L}_q^{t,p}) = \{\lambda_{1,q}^{t,p}, \lambda_{2,q}^{t,p}, \cdots, \lambda_{N_q^t,q}^{t,p}\},$$

where $N_q^t = \dim C_q^t$ is the number of q-simplices in K_t, and the eigenvalues are listed in the ascending order. Thus, the smallest non-zero eigenvalue of $\mathcal{L}_q^{t,p}$ is denoted as $\lambda_{2,q}^{t,p}$. We may recognize the multiplicity of zero in the spectrum of $\mathcal{L}_q^{t,p}$ as the qth order p-persistent Betti number $\beta_q^{t,p}$, which counts the number of (independent) q-dimensional holes in K_t that still exists in K_{t+p}. The relation can be observed in

$$\beta_q^{t,p} = \dim \ker \partial_q^t - \dim \text{im}\, \eth_{q+1}^{t,p} = \dim \ker \mathcal{L}_q^{t,p} = \#0 \text{ eigenvalues of } \mathcal{L}_q^{t,p}. \tag{2.19}$$

In this paper, we focus on the $0, 1, 2$th-order persistent Laplacians, which depict the relations among vertices, edges, triangles, and tetrahedra, as we target 3D real-world applications.

For instance, given a set of vertices $V = \{v_0, v_1, \cdots, v_{N_0-1}\}$, N_0 embedded in $\mathbb{R}^3$, we consider a nested family of simplicial complexes that may be created for a positive real number α. Denoting the simplicial complex generated for α by K_α, the traditional qth-order Laplacian is just a special case of qth-order 0-persistent Laplacian at K_α

$$\mathcal{L}_q^{\alpha,0} = \mathcal{B}_{q+1}^{\alpha,0}(\mathcal{B}_{q+1}^{\alpha,0})^T + (\mathcal{B}_q^{\alpha})^T\mathcal{B}_q^{\alpha}. \tag{2.20}$$

The spectrum of $\mathcal{L}_q^{\alpha,0}$ is simply associated with a snapshot of the filtration,

$$\text{Spec}(\mathcal{L}_q^{\alpha,0}) = \{\lambda_{1,q}^{\alpha,0}, \lambda_{2,q}^{\alpha,0}, \cdots, \lambda_{N_q^\alpha,q}^{\alpha,0}\}. \tag{2.21}$$

Correspondingly, the q-th 0-persistent Betti number $\beta_q^{\alpha,0} = \beta_q^{\alpha}$. In addition to the traditional homology information, and persistent homology information, our proposed persistent spectral graph theory, through the nonzero eigenvalues in the spectrum of the persistent Laplacian operator, provide richer spatial information induced by varying the filtration parameters. Thus it provides a powerful tool to encode high-dimensional datasets into various topological and geometric features in a coherent fashion.[3]

Figure 2.6 demonstrates an example of a standard filtration process. Here the initial setup K_1 consists of five 0-simplices (vertices). We construct Vietoris-Rips complexes by using an ever-growing circle centered at each vertex with radius r. Once two circles overlapped with each other, an 1-simplex (edge) is formed. A 2-simplex (triangle) will be created when 3 circles contact with one another, and a 3-simplex will be generated once 4 circles get overlapped one another. As Figure 2.6 shows, we can attain a series of simplicial complexes from K_1 to K_6 with the radius of circles increasing. To fully illustrate how to construct p-persistent q-combinatorial Laplacian matrices by the boundary operator and determine persistent Betti numbers, we analyze 6 p-persistent q-combinatorial Laplacian matrices and their corresponding harmonic persistent spectra (i.e., persistent Betti numbers) and non-harmonic persistent spectra. Additional matrices are analyzed in Appendix Section A.1.

[3]In this work, we use notations $\mathbb{C}_q^{t,p}, \eth_q^{t,p}, \Delta_q^{t,p}, \mathcal{L}_q^{t,p}$, and $\beta_q^{t,p}$ instead of $\mathbb{C}_q^{t+p}, \eth_q^{t+p}, \Delta_q^{t+p}, \mathcal{L}_q^{t+p}$, and β_q^{t+p} used in Ref. [11].

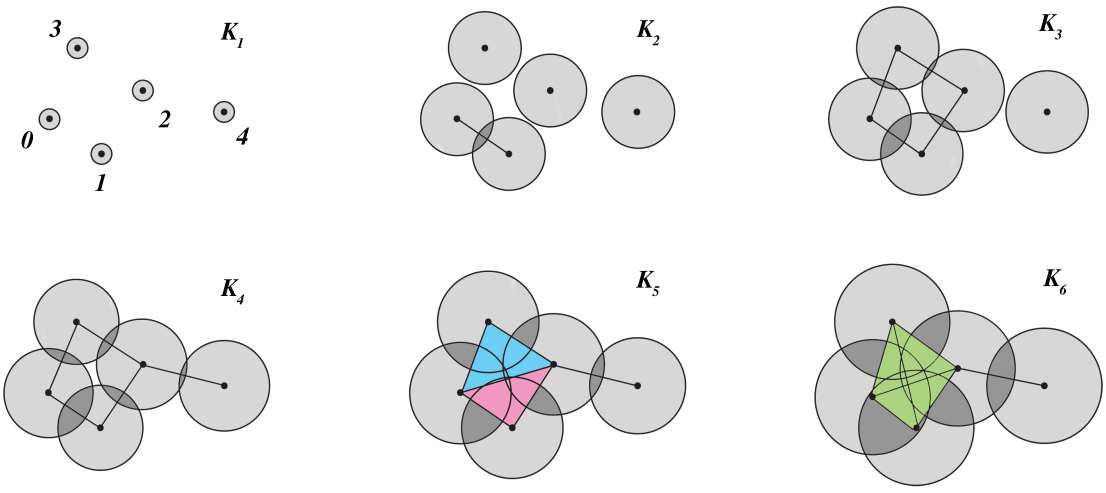

Figure 2.6: Illustration of filtration. We use $0, 1, 2, 3,$ and 4 to stand for 0-simplices, $01, 12, 23, 03, 24, 02,$ and 13 for 1-simplices, $012, 023, 013,$ and 123 for 2-simplices, and 0123 for the 3-simplex. Here, K_1 has five 0-cycles, K_2 has four 0-cycles, K_3 has two 0-cycles and a 1-cycle, K_4 has a 0-cycle and a 1-cycle, K_5 has one 0-cycle, and K_6 has a 0-cycle.

Table 2.4: The number of q-cycles of simplicial complexes demonstrated in Figure 2.6.

# of q-cycles	K_1	K_2	K_3	K_4	K_5	K_6
$q=0$	5	4	2	1	1	1
$q=1$	0	0	1	1	0	0
$q=2$	0	0	0	0	0	0

Case 1. In this case, the initial setup is K_1 and the end status is K_3. Therefore, $t=1$ and $p=2$ in Eq. (2.18). We will calculate $\mathcal{L}_0^{1+2}, \mathcal{L}_1^{1+2}$, and $\mathcal{L}_2^{1+2}$ first and find out their corresponding persistent spectra.

The 2-persistent $0, 1, 2$-combinatorial Laplacian operators are:

$$\Delta_0^{1+2} = \eth_1^{1+2}\left(\eth_1^{1+2}\right)^* + \partial_0^{1^*}\partial_0^1,$$
$$\Delta_1^{1+2} = \eth_2^{1+2}\left(\eth_2^{1+2}\right)^* + \partial_1^{1^*}\partial_1^1,$$
$$\Delta_2^{1+2} = \eth_3^{1+2}\left(\eth_3^{1+2}\right)^* + \partial_2^{1^*}\partial_2^1,$$

Since 2-simplex and 3-simplex do not exist in K_1 and K_3, $\eth_2^{1+2}, \partial_1^1, \eth_3^{1+2}$, and ∂_2^1 do not exist and ∂_0^1 is a zero map. Then, there is only one per-

sistent combinatorial Laplacian matrix

$$\mathcal{L}_0^{1+2} = \mathcal{B}_1^{1+2}(\mathcal{B}_1^{1+2})^T + (\mathcal{B}_0^1)^T\mathcal{B}_0^1.$$

It can be seen in Figure 2.6 that two 0-cycles (connected components) in K_1 are still alive in K_3, while no 1-cycle and 2-cycle exist in the initial set up K_1, which perfectly match the calculations in Table 2.5: $\beta_0^{1+2} = 2$.

Table 2.5: $K_1 \to K_3$.

q	$q=0$	$q=1$	$q=2$
$\mathcal{B}_{q+1}^{1+2}$	$\begin{array}{c} \\ 0 \\ 1 \\ 2 \\ 3 \\ 4 \end{array}\begin{array}{c} \begin{array}{cccc} 01 & 12 & 23 & 03 \end{array} \\ \begin{bmatrix} -1 & 0 & 0 & -1 \\ 1 & -1 & 0 & 0 \\ 0 & 1 & -1 & 0 \\ 0 & 0 & 1 & 1 \\ 0 & 0 & 0 & 0 \end{bmatrix} \end{array}$	/	/
$\mathcal{B}_q^1$	$\begin{array}{c} \begin{array}{ccccc} 0 & 1 & 2 & 3 & 4 \end{array} \\ / \begin{bmatrix} 0 & 0 & 0 & 0 & 0 \end{bmatrix} \end{array}$	/	/
$\mathcal{L}_q^{1+2}$	$\begin{bmatrix} 2 & -1 & 0 & -1 & 0 \\ -1 & 2 & -1 & 0 & 0 \\ 0 & -1 & 2 & -1 & 0 \\ -1 & 0 & -1 & 2 & 0 \\ 0 & 0 & 0 & 0 & 0 \end{bmatrix}$	/	/
β_q^{1+2}	2	/	/
$\dim(\mathcal{L}_q^{1+2})$	5	/	/
$\text{rank}(\mathcal{L}_q^{1+2})$	3	/	/
$\text{nullity}(\mathcal{L}_q^{1+2})$	2	/	/
$\text{Spectrum}(\mathcal{L}_q^{1+2})$	$\{0, 0, 2, 2, 4\}$	/	/

Case 2. The initial setup is K_3 and the end status is K_4. The 1-persistent

$0,1,2$-combinatorial Laplacian operators are

$$\Delta_0^{3+1} = \eth_1^{3+1}\left(\eth_1^{3+1}\right)^* + \partial_0^{3^*}\partial_0^3,$$
$$\Delta_1^{3+1} = \eth_2^{3+1}\left(\eth_2^{3+1}\right)^* + \partial_1^{3^*}\partial 1^3,$$
$$\Delta_2^{3+1} = \eth_3^{3+1}\left(\eth_3^{3+1}\right)^* + \partial_2^{3^*}\partial_2^3,$$

Since 2-simplex and 3-simplex do not exist in K_4, $\partial_2^3, \partial_2^{3+1}$, and ∂_2^3 do not exist, then

$$\mathcal{L}_0^{3+1} = \mathcal{B}_1^{3+1}\left(\mathcal{B}_1^{3+1}\right)^T + (\mathcal{B}_0^3)^T\mathcal{B}_0^3,$$
$$\mathcal{L}_1^{3+1} = (\mathcal{B}_1^3)^T\mathcal{B}_1^3.$$

From Table 2.6, one can see that $\beta_0^{3+1} = 0$ and $\beta_1^{3+1} = 1$, which reveals only one 0-cycle and one 1-cycle in K_3 are still alive in K_4.

Table 2.6: $K_3 \to K_4$.

q	$q=0$	$q=1$	$q=2$
$\mathcal{B}_{q+1}^{3+1}$	$\begin{array}{c} \\ 0 \\ 1 \\ 2 \\ 3 \\ 4 \end{array}\begin{array}{c}\begin{array}{ccccc} 01 & 12 & 23 & 03 & 24 \end{array}\\ \left[\begin{array}{ccccc} -1 & 0 & 0 & -1 & 0 \\ 1 & -1 & 0 & 0 & 0 \\ 0 & 1 & -1 & 0 & -1 \\ 0 & 0 & 1 & 1 & 0 \\ 0 & 0 & 0 & 0 & 1 \end{array}\right]\end{array}$	/	/
$\mathcal{B}_q^3$	$\begin{array}{c}\begin{array}{ccccc} 0 & 1 & 2 & 3 & 4 \end{array}\\ \left[\begin{array}{ccccc} 0 & 0 & 0 & 0 & 0 \end{array}\right]\end{array}$	$\begin{array}{c} \\ 0 \\ 1 \\ 2 \\ 3 \\ 4 \end{array}\begin{array}{c}\begin{array}{cccc} 01 & 12 & 23 & 03 \end{array}\\ \left[\begin{array}{cccc} -1 & 0 & 0 & -1 \\ 1 & -1 & 0 & 0 \\ 0 & 1 & -1 & 0 \\ 0 & 0 & 1 & 1 \\ 0 & 0 & 0 & 0 \end{array}\right]\end{array}$	/
$\mathcal{L}_q^{3+1}$	$\begin{bmatrix} 2 & -1 & 0 & -1 & 0 \\ -1 & 2 & -1 & 0 & 0 \\ 0 & -1 & 3 & -1 & -1 \\ -1 & 0 & -1 & 2 & 0 \\ 0 & 0 & -1 & 0 & 1 \end{bmatrix}$	$\begin{bmatrix} 2 & -1 & 0 & 1 \\ -1 & 2 & -1 & 0 \\ 0 & -1 & 2 & 1 \\ 1 & 0 & 1 & 2 \end{bmatrix}$	/
β_q^{3+1}	1	1	/
dim($\mathcal{L}_q^{3+1}$)	5	4	/
rank($\mathcal{L}_q^{3+1}$)	4	3	/
nullity($\mathcal{L}_q^{3+1}$)	1	1	/
Spectra($\mathcal{L}_q^{3+1}$)	$\{0, 0.8299, 2, 2.6889, 4.4812\}$	$\{0, 2, 2, 4\}$	/

Case 3. The initial setup is K_4 and the end status is K_4. Similarly,

$$\mathcal{L}_0^{4+0} = \mathcal{B}_1^{4+0}\left(\mathcal{B}_1^{4+0}\right)^T + (\mathcal{B}_0^4)^T\mathcal{B}_0^4,$$
$$\mathcal{L}_1^{4+0} = (\mathcal{B}_1^4)^T\mathcal{B}_1^4,$$

and $\mathcal{L}_2^{4+0}$ does not exist. In this case, the 0-persistent q-combinatorial Laplacian matrix is actually the q-combinatorial Laplacian matrix defined in Eq. (2.11). Therefore, $\beta_0^{4+0}, \beta_1^{4+0}$, and β_2^{4+0} actually represent the number of $0, 1, 2$-cycles in K_4. With the filtration parameter r increasing, all the circles overlapped with at least another circle in K_4, which results in $\beta_0^{4+0} = 1$. Since only one 1-cycle formed in K_4, one has $\beta_1^{4+0} = 1$.

Case 4. The initial setup is K_4 and the end status is K_5. Using similar analysis as in previous cases, we have

$$\mathcal{L}_0^{4+1} = \mathcal{B}_1^{4+1}\left(\mathcal{B}_1^{4+1}\right)^T + (\mathcal{B}_0^4)^T\mathcal{B}_0^4,$$
$$\mathcal{L}_1^{4+1} = \mathcal{B}_2^{4+1}\left(\mathcal{B}_2^{4+1}\right)^T + (\mathcal{B}_1^4)^T\mathcal{B}_1^4,$$

and $\mathcal{L}_2^{4+1}$ does not exist. Notice that two 2-simplices 012 and 023 are created under the filtration process. The appearance of these two newborns results in the 1-cycle that was alive in K_4 being killed. Therefore $\beta_1^{4+1} = 0$ and $\beta_0^{4+1} = 1$ because only one connected component keeps alive until K_5.

Case 5. The initial setup is K_5 and the end status is K_6. The 1-persistent $0, 1, 2$-combinatorial Laplacian matrices are

$$\mathcal{L}_0^{5+1} = \mathcal{B}_1^{5+1}\left(\mathcal{B}_1^{5+1}\right)^T + (\mathcal{B}_0^5)^T\mathcal{B}_0^5,$$
$$\mathcal{L}_1^{5+1} = \mathcal{B}_2^{5+1}\left(\mathcal{B}_2^{5+1}\right)^T + (\mathcal{B}_1^5)^T\mathcal{B}_1^5,$$
$$\mathcal{L}_2^{5+1} = \mathcal{B}_3^{5+1}\left(\mathcal{B}_3^{5+1}\right)^T + (\mathcal{B}_2^5)^T\mathcal{B}_2^5.$$

Table 2.7: $K_4 \to K_4$.

q	$q=0$	$q=1$	$q=2$
$\mathcal{B}_{q+1}^{4+0}$	$\begin{array}{c} \\ 0 \\ 1 \\ 2 \\ 3 \\ 4 \end{array}\begin{array}{c} \begin{array}{ccccc} 01 & 12 & 23 & 03 & 24 \end{array} \\ \begin{bmatrix} -1 & 0 & 0 & -1 & 0 \\ 1 & -1 & 0 & 0 & 0 \\ 0 & 1 & -1 & 0 & -1 \\ 0 & 0 & 1 & 1 & 0 \\ 0 & 0 & 0 & 0 & 1 \end{bmatrix} \end{array}$	/	/
$\mathcal{B}_q^4$	$/ \begin{array}{c} \begin{array}{ccccc} 0 & 1 & 2 & 3 & 4 \end{array} \\ \begin{bmatrix} 0 & 0 & 0 & 0 & 0 \end{bmatrix} \end{array}$	$\begin{array}{c} \\ 0 \\ 1 \\ 2 \\ 3 \\ 4 \end{array}\begin{array}{c} \begin{array}{ccccc} 01 & 12 & 23 & 03 & 24 \end{array} \\ \begin{bmatrix} -1 & 0 & 0 & -1 & 0 \\ 1 & -1 & 0 & 0 & 0 \\ 0 & 1 & -1 & 0 & -1 \\ 0 & 0 & 1 & 1 & 0 \\ 0 & 0 & 0 & 0 & 1 \end{bmatrix} \end{array}$	/
$\mathcal{L}_q^{4+0}$	$\begin{bmatrix} 2 & -1 & 0 & -1 & 0 \\ -1 & 2 & -1 & 0 & 0 \\ 0 & -1 & 3 & -1 & -1 \\ -1 & 0 & -1 & 2 & 0 \\ 0 & 0 & -1 & 0 & 1 \end{bmatrix}$	$\begin{bmatrix} 2 & -1 & 0 & 1 & 0 \\ -1 & 2 & -1 & 0 & -1 \\ 0 & -1 & 2 & 1 & 1 \\ 1 & 0 & 1 & 2 & 0 \\ 0 & -1 & 1 & 0 & 2 \end{bmatrix}$	/
β_q^{4+0}	1	1	/
$\dim(\mathcal{L}_q^{4+0})$	5	5	/
$\text{rank}(\mathcal{L}_q^{4+0})$	4	4	/
$\text{nullity}(\mathcal{L}_q^{4+0})$	1	1	/
$\text{Spectra}(\mathcal{L}_q^{4+0})$	$\{0, 0.8299, 2, 2.6889, 4.4812\}$	$\{0, 0.8299, 2, 2.6889, 4.4812\}$	/

In this situation, a new 3-simplex is formed in K_6, which means that $\mathcal{B}_3^{5+1}$ is no long a non-zero matrix. From Table 2.9, we can see that $\beta_2^{5+1} = 0$ because K_5 does not own any 2-cycle and thus, there is no 2-cycle keeping alive up to K_6. β_0^{5+1} implies only one 0-cycle preserved along the filtration process.

Case 6. The initial setup is K_6 and the end status is K_6. The 0-persistent

Table 2.8: $K_4 \to K_5$.

q	$q=0$	$q=1$	$q=2$
$\mathcal{B}_{q+1}^{4+1}$	$\begin{array}{c} \\ 0 \\ 1 \\ 2 \\ 3 \\ 4 \end{array} \begin{array}{c} \begin{array}{cccccc} 01 & 12 & 23 & 03 & 24 & 02 \end{array} \\ \begin{bmatrix} -1 & 0 & 0 & -1 & 0 & -1 \\ 1 & -1 & 0 & 0 & 0 & 0 \\ 0 & 1 & -1 & 0 & -1 & 1 \\ 0 & 0 & 1 & 1 & 0 & 0 \\ 0 & 0 & 0 & 0 & 1 & 0 \end{bmatrix} \end{array}$	$\begin{array}{c} \\ 01 \\ 12 \\ 23 \\ 03 \\ 24 \end{array} \begin{array}{c} \begin{array}{cc} 012 & 023 \end{array} \\ \begin{bmatrix} 1 & 0 \\ 1 & 0 \\ 0 & 1 \\ 0 & -1 \\ 0 & 0 \end{bmatrix} \end{array}$	/
$\mathcal{B}_q^4$	/	$\begin{array}{c} \\ 0 \\ 1 \\ 2 \\ 3 \\ 4 \end{array} \begin{array}{c} \begin{array}{ccccc} 01 & 12 & 23 & 03 & 24 \end{array} \\ \begin{bmatrix} -1 & 0 & 0 & -1 & 0 \\ 1 & -1 & 0 & 0 & 0 \\ 0 & 1 & -1 & 0 & -1 \\ 0 & 0 & 1 & 1 & 0 \\ 0 & 0 & 0 & 0 & 1 \end{bmatrix} \end{array}$	/
$\mathcal{L}_q^{4+1}$	$\begin{bmatrix} 3 & -1 & -1 & -1 & 0 \\ -1 & 2 & -1 & 0 & 0 \\ -1 & -1 & 4 & -1 & -1 \\ -1 & 0 & -1 & 2 & 0 \\ 0 & 0 & -1 & 0 & 1 \end{bmatrix}$	$\begin{bmatrix} 3 & 0 & 0 & 1 & 0 \\ 0 & 3 & -1 & 0 & -1 \\ 0 & -1 & 3 & 0 & 1 \\ 1 & 0 & 0 & 3 & 0 \\ 0 & -1 & 1 & 0 & 2 \end{bmatrix}$	/
β_q^{4+1}	1	0	/
$\dim(\mathcal{L}_q^{4+1})$	5	5	/
$\text{rank}(\mathcal{L}_q^{4+1})$	4	5	/
$\text{nullity}(\mathcal{L}_q^{4+1})$	1	0	/
$\text{Spectra}(\mathcal{L}_q^{4+1})$	$\{0, 1, 2, 4, 5\}$	$\{1.2677, 2, 2, 4, 4.7321\}$	/

Table 2.9: $K_5 \to K_6$.

q	$q=0$	$q=1$	$q=2$
$\mathcal{B}_{q+1}^{5+1}$	$\begin{matrix} & 01 & 12 & 23 & 03 & 24 & 02 \\ 0 & -1 & 0 & 0 & -1 & 0 & -1 \\ 1 & 1 & -1 & 0 & 0 & 0 & 0 \\ 2 & 0 & 1 & -1 & 0 & -1 & 1 \\ 3 & 0 & 0 & 1 & 1 & 0 & 0 \\ 4 & 0 & 0 & 0 & 0 & 1 & 0 \end{matrix}$	$\begin{matrix} & 012 & 023 & 013 & 123 \\ 01 & 1 & 0 & 1 & 0 \\ 12 & 1 & 0 & 0 & 1 \\ 23 & 0 & 1 & 0 & 1 \\ 03 & 0 & -1 & -1 & 0 \\ 24 & 0 & 0 & 0 & 0 \\ 02 & -1 & 1 & 0 & 0 \end{matrix}$	$\begin{matrix} & 0123 \\ 012 & -1 \\ 023 & -1 \end{matrix}$
$\mathcal{B}_q^5$	$\begin{matrix} & 0 & 1 & 2 & 3 & 4 \\ / & 0 & 0 & 0 & 0 & 0 \end{matrix}$	$\begin{matrix} & 01 & 12 & 23 & 03 & 24 & 02 \\ 0 & -1 & 0 & 0 & -1 & 0 & -1 \\ 1 & 1 & -1 & 0 & 0 & 0 & 0 \\ 2 & 0 & 1 & -1 & 0 & -1 & 1 \\ 3 & 0 & 0 & 1 & 1 & 0 & 0 \\ 4 & 0 & 0 & 0 & 0 & 1 & 0 \end{matrix}$	$\begin{matrix} & 012 & 023 \\ 01 & 1 & 0 \\ 12 & 1 & 0 \\ 23 & 0 & 1 \\ 03 & 0 & -1 \\ 24 & 0 & 0 \\ 02 & -1 & 1 \end{matrix}$
$\mathcal{L}_q^{5+1}$	$\begin{bmatrix} 3 & -1 & -1 & -1 & 0 \\ -1 & 2 & -1 & 0 & 0 \\ -1 & -1 & 4 & -1 & -1 \\ -1 & 0 & -1 & 2 & 0 \\ 0 & 0 & -1 & 0 & 1 \end{bmatrix}$	$\begin{bmatrix} 4 & 0 & 0 & 0 & 0 & 0 \\ 0 & 4 & 0 & 0 & -1 & 0 \\ 0 & 0 & 4 & 0 & 1 & 0 \\ 0 & 0 & 0 & 4 & 0 & 0 \\ 0 & -1 & 1 & 0 & 2 & -1 \\ 0 & 0 & 0 & 0 & -1 & 4 \end{bmatrix}$	$\begin{bmatrix} 4 & 0 \\ 0 & 4 \end{bmatrix}$
β_q^{5+1}	1	0	0
$\dim(\mathcal{L}_q^{5+1})$	5	6	2
$\text{rank}(\mathcal{L}_q^{5+1})$	4	6	2
$\text{nullity}(\mathcal{L}_q^{5+1})$	1	0	0
$\text{Spectra}(\mathcal{L}_q^{5+1})$	$\{0,1,2,4,5\}$	$\{1,4,4,4,4,5\}$	$\{4,4\}$

$0,1,2$-combinatorial Laplacian operators are

$$\mathcal{L}_0^{6+0} = \mathcal{B}_1^{6+0}(\mathcal{B}_1^{6+0})^T + (\mathcal{B}_0^6)^T\mathcal{B}_0^6,$$
$$\mathcal{L}_1^{6+0} = \mathcal{B}_2^{6+0}(\mathcal{B}_2^{6+0})^T + (\mathcal{B}_1^6)^T\mathcal{B}_1^6,$$
$$\mathcal{L}_2^{6+0} = \mathcal{B}_3^{6+0}(\mathcal{B}_3^{6+0})^T + (\mathcal{B}_2^6)^T\mathcal{B}_2^6,$$

$\beta_0^{6+0} = 1, \beta_1^{6+0} = 0$, and $\beta_2^{6+0} = 0$ imply that only one 0-cycle (connected component) exists in K_6.

Table 2.10: $K_6 \to K_6$.

q	$q=0$	$q=1$	$q=2$
$\mathcal{B}_{q+1}^{6+0}$	$\begin{array}{c} \\ 0 \\ 1 \\ 2 \\ 3 \\ 4 \end{array}\begin{array}{c} \begin{array}{ccccccc} 01 & 12 & 23 & 03 & 24 & 02 & 13 \end{array} \\ \begin{bmatrix} -1 & 0 & 0 & -1 & 0 & -1 & 0 \\ 1 & -1 & 0 & 0 & 0 & 0 & -1 \\ 0 & 1 & -1 & 0 & -1 & 1 & 0 \\ 0 & 0 & 1 & 1 & 0 & 0 & 1 \\ 0 & 0 & 0 & 0 & 1 & 0 & 0 \end{bmatrix} \end{array}$	$\begin{array}{c} \\ 01 \\ 12 \\ 23 \\ 03 \\ 24 \\ 02 \\ 13 \end{array}\begin{array}{c} \begin{array}{cccc} 012 & 023 & 013 & 123 \end{array} \\ \begin{bmatrix} 1 & 0 & 1 & 0 \\ 1 & 0 & 0 & 1 \\ 0 & 1 & 0 & 1 \\ 0 & -1 & -1 & 0 \\ 0 & 0 & 0 & 0 \\ -1 & 1 & 0 & 0 \\ 0 & 0 & 1 & -1 \end{bmatrix} \end{array}$	$\mathcal{B}_3^{6+0}$
$\mathcal{B}_q^6$	$\begin{array}{c} \\ / \end{array}\begin{array}{c} \begin{array}{ccccc} 0 & 1 & 2 & 3 & 4 \end{array} \\ \begin{bmatrix} 0 & 0 & 0 & 0 & 0 \end{bmatrix} \end{array}$	$\begin{array}{c} \\ 0 \\ 1 \\ 2 \\ 3 \\ 4 \end{array}\begin{array}{c} \begin{array}{ccccccc} 01 & 12 & 23 & 03 & 24 & 02 & 13 \end{array} \\ \begin{bmatrix} -1 & 0 & 0 & -1 & 0 & -1 & 0 \\ 1 & -1 & 0 & 0 & 0 & 0 & -1 \\ 0 & 1 & -1 & 0 & -1 & 1 & 0 \\ 0 & 0 & 1 & 1 & 0 & 0 & 1 \\ 0 & 0 & 0 & 0 & 1 & 0 & 0 \end{bmatrix} \end{array}$	$\mathcal{B}_2^6$
$\mathcal{L}_q^{6+0}$	$\begin{bmatrix} 3 & -1 & -1 & -1 & 0 \\ -1 & 2 & -1 & 0 & 0 \\ -1 & -1 & 4 & -1 & -1 \\ -1 & 0 & -1 & 2 & 0 \\ 0 & 0 & -1 & 0 & 1 \end{bmatrix}$	$\begin{bmatrix} 4 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 4 & 0 & 0 & -1 & 0 & 0 \\ 0 & 0 & 4 & 0 & 1 & 0 & 0 \\ 0 & 0 & 0 & 4 & 0 & 0 & 0 \\ 0 & -1 & 1 & 0 & 2 & -1 & 0 \\ 0 & 0 & 0 & 0 & -1 & 4 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & 4 \end{bmatrix}$	$\mathcal{L}_3^{6+0}$
β_q^{6+0}	1	0	0
$\dim(\mathcal{L}_q^{6+0})$	5	7	4
$\text{rank}(\mathcal{L}_q^{6+0})$	4	7	4
$\text{nullity}(\mathcal{L}_q^{6+0})$	1	0	0
$\text{Spectra}(\mathcal{L}_q^{6+0})$	$\{0,1,4,4,5\}$	$\{1,4,4,4,4,4,5\}$	$\{4,4,4,4\}$

with

$$\mathcal{B}_3^{6+0} = \begin{array}{c} \\ 012 \\ 023 \\ 013 \\ 123 \end{array}\begin{array}{c} \begin{array}{c} 0123 \end{array} \\ \begin{bmatrix} -1 \\ -1 \\ 1 \\ 1 \end{bmatrix} \end{array}, \mathcal{B}_2^6 = \begin{array}{c} \\ 01 \\ 12 \\ 23 \\ 03 \\ 24 \\ 02 \\ 13 \end{array}\begin{array}{c} \begin{array}{cccc} 012 & 023 & 013 & 123 \end{array} \\ \begin{bmatrix} 1 & 0 & 1 & 0 \\ 1 & 0 & 0 & 1 \\ 0 & 1 & 0 & 1 \\ 0 & -1 & -1 & 0 \\ 0 & 0 & 0 & 0 \\ -1 & 1 & 0 & 0 \\ 0 & 0 & 1 & -1 \end{bmatrix} \end{array}$$

and

$$\mathcal{L}_3^{6+0} = \begin{bmatrix} 4 & 0 & 0 & 0 \\ 0 & 4 & 0 & 0 \\ 0 & 0 & 4 & 0 \\ 0 & 0 & 0 & 4 \end{bmatrix}.$$

2.1.6 Variants of Persistent Laplacians

The traditional approach in defining the q-boundary operator $\partial_q : C_q(K) \to C_{q-1}(K)$ can be expressed as:

$$\partial_q \sigma_q = \sum_{i=0}^{q} (-1)^i \sigma_{q-1}^i,$$

which leads to the corresponding elements in the boundary matrices being either 1 or -1. However, to encode more geometric information into the Laplacian operator, we add volume information of q-simplex σ_q to the expression of q-boundary operator.

Given a vertex set $V = \{v_0, v_1, \cdots, v_q\}$ with $q+1$ isolated points (0-simplices) randomly arranged in the n-dimensional Euclidean space $\mathbb{R}^n$, often with $n \geq q$. Set d_{ij} to be the distances between v_i and v_j with $0 \leq i \leq j \leq q$ and obviously, $d_{ij} = d_{ji}$. The Cayley-Menger determinant can be expressed as [58]

$$\text{Det}_{\text{CM}}(v_0, v_1, \cdots, v_q) = \begin{vmatrix} 0 & d_{01}^2 & d_{02}^2 & \cdots & d_{0q}^2 & 1 \\ d_{10}^2 & 0 & d_{12}^2 & \cdots & d_{1q}^2 & 1 \\ d_{20}^2 & d_{21}^2 & 0 & \cdots & d_{2q}^2 & 1 \\ \vdots & \vdots & \vdots & \vdots & \ddots & \vdots \\ d_{q0}^2 & d_{q1}^2 & d_{q2}^2 & \cdots & 0 & 1 \\ 1 & 1 & 1 & 1 & 1 & 0 \end{vmatrix} \tag{2.22}$$

The q-dimensional volume of q-simplex σ_q with vertices $\{v_0, v_1, \cdots, v_q\}$ is defined by

$$\text{Vol}(\sigma_q) = \sqrt{\frac{(-1)^{q+1}}{(q!)^2 2^q} \text{Det}_{\text{CM}}(v_0, v_1, \cdots, v_q)}. \tag{2.23}$$

In trivial cases, $\text{Vol}(\sigma_0) = 1$, meaning the 0-dimensional volume of 0-simplex is 1, i.e., there is only 1 vertex in a 0-simplex. Also, the 1-dimensional volume of 1-simplex $\sigma_1 = [v_i, v_j]$ is the distance between v_i and v_j, and the 2-dimensional volume of 2-simplex is the area of a triangle $[v_i, v_j, v_k]$.

The weighted boundary operator equipped with volume, denoted $\hat{\partial}_q$, is given by

$$\hat{\partial}_q \sigma_q = \sum_{i=0}^{q} (-1)^i \text{Vol}(\sigma_q^i) \sigma_{q-1}^i. \tag{2.24}$$

Employed the same concept to the persistent spectral theory, we have the volume-weighted p-persistent q-combinatorial Laplacian operator. We also define

$$\hat{\eth}_q^{t+p}(\sigma_q) \coloneqq \begin{cases} \hat{\partial}_q^{t+p}(\sigma_q), & \text{if } \sigma_q \in \mathbb{C}_q^{t+p} \\ 0, & \text{if } \sigma_q \in C_q^{t+p} \setminus \mathbb{C}_q^{t+p} \end{cases} \tag{2.25}$$

with

$$\mathbb{C}_q^{t+p} \coloneqq \{\sigma_q \in C_q^{t+p} \mid \hat{\partial}_q^{t+p}(\sigma_q) \in C_{q-1}^t\}.$$

Similarly, an inverse-volume weighted boundary operator, denoted $\check{\partial}_q$, is given by

$$\check{\partial}_q \sigma_q = \sum_{i=0}^{q} (-1)^i \frac{1}{\text{Vol}(\sigma_q^i)} \sigma_{q-1}^i. \tag{2.26}$$

To define an inverse-volume weighted p-persistent q-combinatorial Laplacian operator. We define

$$\check{\eth}_q^{t+p}(\sigma_q) \coloneqq \begin{cases} \check{\partial}_q^{t+p}(\sigma_q), & \text{if } \sigma_q \in \mathbb{C}_q^{t+p} \\ 0, & \text{if } \sigma_q \in C_q^{t+p} \setminus \mathbb{C}_q^{t+p} \end{cases} \tag{2.27}$$

with

$$\mathbb{C}_q^{t+p} \coloneqq \{\sigma_q \in C_q^{t+p} \mid \check{\partial}_q^{t+p}(\sigma_q) \in C_{q-1}^t\}.$$

Then volume-weighted and inverse-volume-weighted p-persistent $\boldsymbol{q}$-combinatorial Laplacian operators defined along the filtration can be expressed as

$$\begin{aligned} \hat{\Delta}_q^{t+p} &= \hat{\eth}_{q+1}^{t+p} \left(\hat{\eth}_{q+1}^{t+p}\right)^* + \hat{\partial}_q^{t^*} \hat{\partial}_q^t, \\ \check{\Delta}_q^{t+p} &= \check{\eth}_{q+1}^{t+p} \left(\check{\eth}_{q+1}^{t+p}\right)^* + \check{\partial}_q^{t^*} \check{\partial}_q^t. \end{aligned} \tag{2.28}$$

The corresponding weighted matrix representations of boundary operators $\hat{\eth}_{q+1}^{t+p}$, $\hat{\eth}_{q}^{t}$, $\check{\eth}_{q+1}^{t+p}$, and $\check{\eth}_{q}^{t}$ are denoted $\hat{\mathcal{B}}_{q+1}^{t+p}$, $\hat{\mathcal{B}}_{q}^{t}$, $\check{\mathcal{B}}_{q+1}^{t+p}$, and $\check{\mathcal{B}}_{q}^{t}$, respectively. Therefore, volume-weighted and inverse-volume-weighted p-persistent $\boldsymbol{q}$-combinatorial Laplacian matrices can be expressed as

$$\begin{aligned}\hat{\mathcal{L}}_{q}^{t+p} &= \hat{\mathcal{B}}_{q+1}^{t+p}(\hat{\mathcal{B}}_{q+1}^{t+p})^{T} + (\hat{\mathcal{B}}_{q}^{t})^{T}(\hat{\mathcal{B}}_{q}^{t}),\\ \check{\mathcal{L}}_{q}^{t+p} &= \check{\mathcal{B}}_{q+1}^{t+p}(\check{\mathcal{B}}_{q+1}^{t+p})^{T} + (\check{\mathcal{B}}_{q}^{t})^{T}(\check{\mathcal{B}}_{q}^{t}).\end{aligned} \tag{2.29}$$

Although the expressions of the weighted persistent Laplacian matrices are different from the original persistent Laplacian matrices, some properties of $\mathcal{L}_{q}^{t+p}$ are preserved. The weighted persistent Laplacian operators are still symmetric and positive semi-defined. Additionally, their ranks are the same as $\mathcal{L}_{q}^{t+p}$. With the embedded volume information, weighted PSGs can provide richer topological and geometric information through the associated persistent Betti numbers and non-harmonic spectra (i.e., non-zero eigenvalues).

In real applications, we are more interested in the $0, 1, 2$-combinatorial Laplacian matrices because its more intuitive to depict the relation among vertex, edges, and faces. Given a set of vertices $V = \{v_0, v_2, \cdots, v_N\}$ with $N + 1$ isolated points (0-simplices) randomly arranged in $\mathbb{R}^n$. By varying the radius r of the $(n-1)$-sphere centered at each vertex, a variety of simplicial complexes is created. We denote the simplicial complex generated at radius r to be K_r, then the 0-persistent q-combinatorial Laplacian operator and matrix at initial set up K_r is

$$\mathcal{L}_{q}^{r+0} = \mathcal{B}_{q+1}^{r+0}(\mathcal{B}_{q+1}^{r+0})^{T} + (\mathcal{B}_{q}^{r})^{T}\mathcal{B}_{q}^{r}. \tag{2.30}$$

The volume of any 1-simplex $\sigma_1 = [v_i, v_j]$ is $\mathrm{Vol}(\sigma_1)$ is actually the distance between v_i and v_j denoted d_{ij}. Then the 0-persistent 0-combinatorial Laplacian matrix based on filtration

r can be expressed explicitly as

$$(\mathcal{L}_0^{r+0})_{ij} = \begin{cases} -\sum_j (\mathcal{L}_0^{r+0})_{ij}, & \text{if } i = j \\ -1, & \text{if } i \neq j \text{ and } d_{ij} - 2r < 0 \\ 0, & \text{otherwise.} \end{cases} \tag{2.31}$$

Correspondingly, we can denote the 0-persistent 1-combinatorial Laplacian matrix based on filtration r by $\mathcal{L}_1^{r+0}$, and the 0-persistent 2-combinatorial Laplacian matrix based on filtration r by $\mathcal{L}_2^{r+0}$.

Alternatively, variants of persistent 0-combinatorial Laplacian matrices can be defined by adding the Euclidean distance information. The distance-weight persistent 0-combinatorial Laplacian matrix based on filtration r can be expressed explicitly as

$$(\hat{\mathcal{L}}_0^{r+0})_{ij} = \begin{cases} -\sum_j (\hat{\mathcal{L}}_0^{r+0})_{ij}, & \text{if } i = j \\ -d_{ij}, & \text{if } i \neq j \text{ and } d_{ij} - 2r < 0 \\ 0, & \text{otherwise.} \end{cases} \tag{2.32}$$

Moreover, the inverse-distance-weight persistent 0-combinatorial Laplacian matrix based on filtration r can also be implemented:

$$(\check{\mathcal{L}}_0^{r+0})_{ij} = \begin{cases} -\sum_j (\check{\mathcal{L}}_0^{r+0})_{ij}, & \text{if } i = j \\ -\dfrac{1}{d_{ij}}, & \text{if } i \neq j \text{ and } d_{ij} - 2r < 0 \\ 0, & \text{otherwise.} \end{cases} \tag{2.33}$$

The spectra of the aforementioned 0-persistent 0-combinatorial Laplacian matrices based on filtration are given by

$$\text{Spectra}(\mathcal{L}_0^{r+0}) = \{(\lambda_1)_0^{r+0}, (\lambda_2)_0^{r+0}, \cdots, (\lambda_N)_0^{r+0}\},$$
$$\text{Spectra}(\hat{\mathcal{L}}_0^{r+0}) = \{(\hat{\lambda}_1)_0^{r+0}, (\hat{\lambda}_2)_0^{r+0}, \cdots, (\hat{\lambda}_N)_0^{r+0}\},$$
$$\text{Spectra}(\check{\mathcal{L}}_0^{r+0}) = \{(\check{\lambda}_1)_0^{r+0}, (\check{\lambda}_2)_0^{r+0}, \cdots, (\check{\lambda}_N)_0^{r+0}\},$$

where N is the dimension of persistent Laplacian matrices, $(\hat{\lambda}_j)_0^{r+0}$ and $(\check{\lambda}_j)_0^{r+0}$ are the j-th eigenvalues of $\hat{\mathcal{L}}_0^{r+0}$ and $\check{\mathcal{L}}_0^{r+0}$, respectively. We denote $\hat{\beta}_q^{r+0}$ and $\check{\beta}_q^{r+0}$ the qth Betti for $\hat{\mathcal{L}}_q^{r+0}$ and $\check{\mathcal{L}}_q^{r+0}$, respectively.

The smallest non-zero eigenvalue of $\mathcal{L}_0^{r+0}$, denoted $(\tilde{\lambda}_2)_0^{r+0}$, is particularly useful in many applications. Similarly, the smallest non-zero eigenvalues of $\hat{\mathcal{L}}_0^{r+0}$ and $\check{\mathcal{L}}_0^{r+0}$ are denoted as $(\tilde{\hat{\lambda}}_2)_0^{r+0}$ and $(\tilde{\check{\lambda}}_2)_0^{r+0}$, respectively.

Finally, it is mentioned that using the present procedure, more general weights, such as the radial basis function of the Euclidean distance, can be employed to construct weighted boundary operators and associated persistent combinatorial Laplacian matrices.

2.2 Persistent Path Laplacian

2.2.1 Paths on a Finite Set

Denote set V an arbitrary nonempty finite set, and elements in V are called vertices. For $p \in \mathbb{Z}_0^+$ (i.e., a set with integers $p \geq 0$), an *elementary p-path* on V is any sequence $i_0 \dots i_p$ of $p+1$ vertices in V. An elementary p-path is an empty set $\emptyset$ for $p = -1$. For a fixed field $\mathbb{K}$, a vector space that consists of all formal linear combinations of elementary p-paths with its coefficients in $\mathbb{K}$ is called the space generated by the elementary paths, denoted as $\Lambda_p = \Lambda_p(V, \mathbb{K}) = \Lambda_p(V)$. One says the elements in Λ_p are *p-paths* on V, and an elementary p-path $i_0 \dots i_p \in \Lambda_p$ is denoted by $e_{i_0 \dots i_p}$. By definition, $\forall v \in \Lambda_p$, its unique representation can be given by the basis in Λ_p:

$$v = \sum_{i_0, \dots, i_p \in V} c^{i_0 \dots i_p} e_{i_0 \dots i_p}, \tag{2.34}$$

where $c^{i_0 \dots i_p}$ is the coefficient in $\mathbb{K}$. For instance, Λ_0 contains all linear combination of e_i with $i \in V$, Λ_1 has all linear combination of e_{ij} with $(i, j) \in V \times V$, and so on so forth. Since Λ_{-1} consists of all multiples of e, one has $\Lambda_{-1} \cong \mathbb{K}$.

Additionally, $\forall p \in \mathbb{Z}_0^+$, the linear *boundary operator* from Λ_p to Λ_{p-1} that acts on ele-

mentary paths can be defined as

$$\partial : \Lambda_p \to \Lambda_{p-1} \tag{2.35}$$

with

$$\partial e_{i_0 \dots i_p} = \sum_{q=0}^{p} (-1)^q e_{i_0 \dots \hat{i}_q \dots i_p}, \tag{2.36}$$

where $\hat{i}_q$ denotes the omission of index i_q from the elementary p-path $e_{i_0 \dots i_p}$. One sets $\Lambda_{-2} = \{0\}$, and for $p = -1$, defines $\partial : \Lambda_{-1} \to \Lambda_{-2}$ to be a zero map. Following Lemma 2.1 in [59], one has $\partial^2 = 0$, which indicates that the collection of boundary operator ∂ and space Λ_p can form a chain complex of V denoted as $\Lambda_* = \{\Lambda_p\}$ as

$$\cdots \Lambda_p \xrightarrow{\partial} \Lambda_{p-1} \xrightarrow{\partial} \cdots \xrightarrow{\partial} \Lambda_0 \xrightarrow{\partial} \mathbb{K} \xrightarrow{\partial} 0. \tag{2.37}$$

Next, the concepts of regular path and non-regular path are introduced according to [59]. An elementary path $e_{i_0 \dots i_p}$ on a set V is *regular* if $i_{k-1} \neq i_k$, and *non-regular* if $i_{k-1} = i_k$ for $k = 1, \dots, p$. For any $p \in \mathbb{Z}_0^+ \cup \{-1\}$, let $\mathcal{R}_p$ be the subspace of Λ_p spanned by all regular elementary paths, and $\mathcal{N}_p$ be the subspace of Λ_p spanned by all non-regular elementary paths. Therefore, one has

$$\mathcal{R}_p = \text{span}\{e_{i_0 \dots i_p} : i_0 \dots i_p \text{ is regular}\}$$
$$\mathcal{N}_p = \text{span}\{e_{i_0 \dots i_p} : i_0 \dots i_p \text{ is non-regular}\}.$$

Note that $\mathcal{R}_p = \Lambda_p$ for integers $p = -1, 0$.

Then $\forall p \in \mathbb{Z}_0^+ \cup \{-1\},\ \Lambda_p = \mathcal{R}_p \oplus \mathcal{N}_p$. Therefore,

$$\mathcal{R}_p \cong \Lambda_p / \mathcal{N}_p.$$

According to Section 2.4 in [59], the boundary operator ∂ is well-defined on the quotient space $\Lambda_p / \mathcal{N}_p$. Moreover, $\partial^2 = 0$ and the product rules are satisfied in the quotient space $\Lambda_p / \mathcal{N}_p$ as well. One has an induced *regular boundary operator*:

$$\bar{\partial} : \mathcal{R}_p \to \mathcal{R}_{p-1}, \tag{2.38}$$

where the regular boundary operator $\bar{\partial}$ satisfies (2.36) except that all non-regular terms on the right hand side should be treated as 0. Then a chain complex of V, denoted as $\mathcal{R}_*(V) = (\mathcal{R}_p)_p$ and equipped with $\bar{\partial}$, can be expressed as:

$$\cdots \mathcal{R}_p \xrightarrow{\bar{\partial}} \mathcal{R}_{p-1} \xrightarrow{\bar{\partial}} \cdots \xrightarrow{\bar{\partial}} \mathcal{R}_0 \xrightarrow{\bar{\partial}} \mathbb{K} \xrightarrow{\bar{\partial}} 0. \tag{2.39}$$

It can be verified that $R_p \cong \Lambda_p/N_p$ is an isomorphism of chain complexes [60]. In the following sections, for simplicity, we use ∂ to denote the boundary operator of Eq. (2.39) unless specified differently.

2.2.2 Path Complex

A *path complex* over set V is a nonempty collection P of elementary paths on V for any $n \in \mathbb{Z}_0^+$,

$$\text{if } i_0 \dots i_n \in P, \text{ then } i_0 \dots i_{n-1} \in P, \text{ and } i_1 \dots i_n \in P. \tag{2.40}$$

For a fixed path complex, all the paths from P are called *allowed* (i.e. $i_{k-1} \to i_k$ for any $k = 1, \dots, n$), while the elementary paths on V that are not in P are *non-allowed*. We say a path complex P is *perfect* if any subsequence of any path from P is also in P. We choose P_n to denote all n-paths from P. Then the set P_{-1} has a single empty path e, the set P_0 consists of all the *vertices* of P, and clearly, $V = P_0$. To be noted, a path complex P is a collection $\{P_n\}_{n=-1}^{\infty}$ satisfying Eq. (2.40). Let $\mathcal{K}$ be an abstract simplicial complex defined over a finite vertex set V, satisfying

$$\text{if } \sigma \in \mathcal{K}, \text{ then any subset of } \sigma \text{ is also in } \mathcal{K}.$$

The collection of elementary paths on V is denoted by $P(\mathcal{K})$. Follows from [59] (cf. Example 3.2), the family $P(\mathcal{K})$ is a path complex, and the allowed n-paths are n-simplices.

2.2.3 Path Homology

For any $n \in \mathbb{Z}_0^+$, the $\mathbb{K}$-linear space $\mathcal{A}_n$ is spanned by all the elementary n-paths from a given path complex $P = \{P_n\}_{n=0}^{\infty}$ over a finite set V, i.e.,

$$\mathcal{A}_n = \mathcal{A}_n(P) = \text{span}\{e_{i_0 \dots i_n} : i_0 \dots i_n \in P_n\}.$$

We call the elements of $\mathcal{A}_n$ the *allowed n-paths*. By the definition of $\mathcal{A}_n$, $\mathcal{A}_n \subset \Lambda_n$, and $\mathcal{A}_n = \Lambda_n$ for $n \leq 0$. It is natural that the boundary operator ∂ defined on $\mathcal{R}_n$ can be introduced to $\mathcal{A}_n$ under certain condition: $\partial\mathcal{A}_n \subseteq \mathcal{A}_{n-1}$. For example, for perfect path complexes, we can obtain a chain complex:

$$\cdots \mathcal{A}_n \xrightarrow{\partial} \mathcal{A}_{n-1} \xrightarrow{\partial} \cdots \xrightarrow{\partial} \mathcal{A}_0 \xrightarrow{\partial} \mathbb{K} \xrightarrow{\partial} 0.$$

Next, we consider a general path complex P (i.e., $\partial\mathcal{A}_n$ does not have to be a subspace of $\mathcal{A}_{n-1}$). For any $n \in \mathbb{Z}_0^+ \cup \{-1\}$, we define a subspace of $\mathcal{A}_n$:

$$\Omega_n = \Omega_n(P) = \{v \in \mathcal{A}_n : \partial v \in \mathcal{A}_{n-1}\}. \tag{2.41}$$

The elements of Ω_n are called ∂-invariant n-paths. To be noted, $\partial\Omega_n \subset \Omega_{n-1}$ always satisfies. Moreover, $\partial^2 = 0$ has been established in the previous section. Therefore, the *augmented* chain complex of ∂-invariant paths can be denoted as

$$\cdots \Omega_n \xrightarrow{\partial} \Omega_{n-1} \xrightarrow{\partial} \cdots \xrightarrow{\partial} \Omega_0 \xrightarrow{\partial} \mathbb{K} \xrightarrow{\partial} 0, \tag{2.42}$$

whose homology group $\tilde{H}_n(P)$ of the chain complex in Eq. (2.42) are called the *reduced path homology groups* of the path complex P for $n \in \mathbb{Z}_0^+ \cup \{-1\}$. The truncated version of the chain complex in Eq. (2.42) for $n \in \mathbb{Z}_0^+$ is:

$$\cdots \Omega_n \xrightarrow{\partial} \Omega_{n-1} \xrightarrow{\partial} \cdots \xrightarrow{\partial} \Omega_0 \xrightarrow{\partial} 0, \tag{2.43}$$

whose homology group $H_n(P)$ of the chain complex in Eq. (2.43) are called the *path homology groups* of the path complex P.

2.2.4 Path Homology on Directed Graphs

A directed graph is an ordered pair $G = (V, E)$, where V is a set of all vertices and E is a set of ordered pairs of vertices (i.e. directed edges that satisfy $E \subseteq V \times V$). If $G = (V, E)$ does not contain any loop and multiple edge, then it is called *simple directed graph*. Moreover, for the path homology of *multigraph* or *quiver*, one can refer to Ref. [61]. In the following section of this work, we use $G(V, E)$ to represent the simple directed graphs unless specified differently.

The path complex $P(G)$ is regular if $G = (V, E)$ is a simple directed graph. In this section, we mainly discuss the regular spaces $\Omega_n(G) = \Omega_n(P(G))$ and their associated regular homology groups $H(G) = H_n(P(G))$. Similar to the discussion in Subsection 2.2.3, given a simple digraph $G(V, E)$, for any $n \in \mathbb{Z}_0^+ \cup \{-1\}$, the space of ∂-invariant n-paths on G is defined by the subspace of $\mathcal{A}_n(G) = \mathcal{A}_n(V, E; \mathbb{K})$:

$$\Omega_n = \Omega_n(G) = \{v \in \mathcal{A}_n : \partial v \in \mathcal{A}_{n-1}\},$$

with $\Omega_{-1} = \mathcal{A}_{-1} \cong \mathbb{K}$ and $\Omega_{-2} = \mathcal{A}_{-2} = \{0\}$. Since $\partial(\Omega_n) \subseteq \Omega_{n-1}$ (as $\partial^2 = 0$), then we have the following chain complex of V denoted as $\Omega_*(V) = \{\Omega_n\}$,

$$\cdots \xrightarrow{\partial} \Omega_3 \xrightarrow{\partial} \Omega_2 \xrightarrow{\partial} \Omega_1 \xrightarrow{\partial} \Omega_0 \xrightarrow{\partial} \mathbb{K} \xrightarrow{\partial} 0,$$

and the associated n- *dimensional path homology groups* of $G = (V, E)$ are defined as:

$$H_n(G) = H_n(V, E; \mathbb{K}) := \ker(\partial|_{\Omega_n}) / \operatorname{im}(\partial|_{\Omega_{n+1}}). \tag{2.44}$$

To be noted, the elements of $\ker(\partial|_{\Omega_n})$ are called *n-cycles*, and the elements of $\operatorname{im}(\partial|_{\Omega_{n+1}})$ are referred to as *n-boundaries*. For simplicity, we define $\partial_n = \partial|_{\Omega_n}$, and the chain complex of ∂-invariant paths is written as

$$\cdots \Omega_{n+1} \xrightarrow{\partial_{n+1}} \Omega_n \xrightarrow{\partial_n} \Omega_{n-1} \xrightarrow{\partial_{n-1}} \Omega_{n-2} \cdots .$$

Notably, the path cohomology, introduced in Refs. [60, 62], is isomorphic to the dual space of path homology when the coefficient ring is a field. The associated n- *dimensional*

path homology groups of digraphs are defined as:

$$H^n(G) = H^n(V, E; \mathbb{K}) := \ker(d_{n+1})/\operatorname{im}(d_n), \tag{2.45}$$

where d is called coboundary operator.

Given two simple digraphs $G = (V, E)$ and $G' = (V', E')$. According to the Definition 2.2 in [63], a *morphism of digraphs/digraphs map* from G to G' is a map $f : V \to V'$ such that for any directed edge $i \to j$ in E, one has either $f(i) \to f(j)$ is a directed edge on E' or $f(i) = f(j)$.

Let f be a digraph map from G to G'. For $n \in \mathbb{Z}_0^+ \cup \{-1\}$, one defines a map $(f_{**})_n : \Lambda_n(V) \to \Lambda_n(V')$ such that:

$$(f_{**})_n(e_{i_0 \ldots i_n}) = e_{f(i_0)\ldots f(i_n)}. \tag{2.46}$$

Assume ∂ and ∂' are the boundary operators of chain complexes $\Lambda_*(V)$ and $\Lambda_*(V')$, then for $e_{i_0 \ldots i_n} \in \Lambda_n$, one has

$$((f_{**})_{n-1} \circ \partial)(e_{i_0 \ldots i_n}) = \sum_{q=0}^{n} (-1)^q (f_{**})_{n-1}(e_{i_0 \ldots \hat{i}_q \ldots i_n}) \tag{2.47}$$

$$= \sum_{q=0}^{n} (-1)^q (e_{f(i_0)\ldots \hat{f}(i_q)\ldots f(i_n)}) \tag{2.48}$$

$$= (\partial' \circ (f_*)_n)(e_{i_0 \ldots i_n}). \tag{2.49}$$

Hence f_{**} is a chain map. By the definition of digraph map, $(f_{**})_n$ maps non-regular elementary n-paths on V to non-regular elementary n-paths on V'. Therefore, one has $(f_{**})_n(\mathcal{N}_n(V)) \subseteq \mathcal{N}_n(V')$, and then $(f_{**})_n$ descends to a quotient homomorphism of chain complexes:

$$(\tilde{f}_{**})_n : \Lambda_n(V)/\mathcal{N}_n(V) \to \Lambda_n(V')/\mathcal{N}_n(V'). \tag{2.50}$$

It can be verified that $R_p \cong \Lambda_p/N_p$ is an isomorphism of chain complexes [60], then the map in (2.50) induces a morphism of chain complexes:

$$(f_*)_n : \mathcal{R}_n(V) \to \mathcal{R}_n(V'). \tag{2.51}$$

Since $(f_{**})_n$ maps non-regular paths to non-regular, then similarly to what Eq. (2.47) shows, $(f_*)_n$ is also a chain map that follows:

$$(f_*)_n(e_{i_0\dots i_n}) := \begin{cases} e_{f(i_0)\dots f(i_n)} & \text{if } e_{f(i_0)\dots f(i_n)} \text{ is regular,} \\ 0 & \text{otherwise.} \end{cases} \tag{2.52}$$

Following the Theorem 2.10 in [63], the induced map $(f_*)_n$ induces a morphism of chain complexes:

$$(f_*)_n : \Omega_n(G;\mathbb{K}) \to \Omega_n(G';\mathbb{K}) \tag{2.53}$$

and consequently induces a homomorphism between the path homology groups:

$$(f_*)_n : H_n(G;\mathbb{K}) \to H_n(G';\mathbb{K}), \quad n \geq 0. \tag{2.54}$$

2.2.5 Homologies of Directed Subgraphs

Some interesting propositions on the homologies of subgraphs provide a way to simplify complicated digraphs to relatively simple ones. Following the Section 4.2 in [59], three propositions are discussed.

Proposition 2.2.1 *Given a simple digraph G that has a vertex v with n outcoming arrows $v \to v'_0, v \to v'_1, \dots, v \to v'_{n-1}$. Note that v does not have any incoming arrows. Assume that for all $i \geq 1$, one has $v'_0 \to v'_i$. Denote G' be the subgraph of G by removing the vertex v with all adjacent edges (i.e. $V' = V\backslash\{v\}$ and $E' = E\backslash\{vv'_i\}_{i=0}^{n-1}$). Then, one has $H_*(G) \cong H_*(G')$ (See Figure 2.7* ***a****).*

Proposition 2.2.2 *Given a simple digraph $G = (V, E)$ that has a vertex v with n incoming arrows $v'_0 \to v, v'_1 \to v, \dots, v'_{n-1} \to v$. Note that v does not have any outcoming arrows. Assume that for all $i \geq 1$, one has $v'_i \to v'_0$. Denote $G' = (V', E')$ be the subgraph of G by removing the vertex v with all adjacent edges (i.e. $V' = V\backslash\{v\}$ and $E' = E\backslash\{v'_i v\}_{i=0}^{n-1}$). Then, one has $H_*(G) \cong H_*(G')$ (See Figure 2.7* ***b****).*

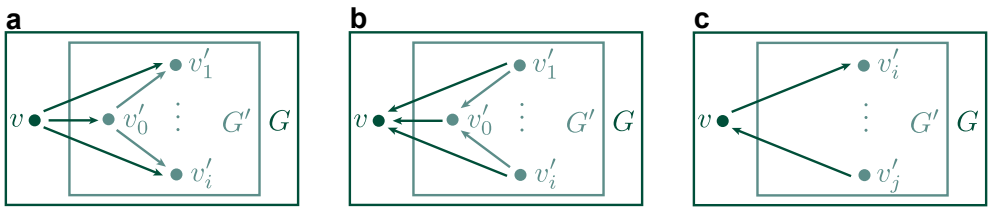

Figure 2.7: Homologies of directed subgraphs. **a**, **b**, and **c** illustrate three subgraphs whose homology groups or homology group dimensions are related to the original digraphs.

Proposition 2.2.3 *Given a simple digraph $G = (V, E)$ that has a vertex v with only one outcoming arrow $v \to v_i'$ and only one incoming arrow $v_j' \to v$, where $i \neq j$. Denote $G' = (V', E')$ be the subgraph of G (See Figure 2.7 **c**) by removing the vertex v and the adjacent edges $v \to v_i'$ and $v_j' \to v$ (i.e. $V' = V \backslash \{v\}$ and $E' = E \backslash \{vv_i', v_j'v\}$). Then,*

(i) $\dim H_p(G) = \dim H_p(G')$ *for* $p \neq 2$ *or for* $p = 0, 1$ *if* $v_j'v_i'$ *is an edge/semi-edge in* G'.

(ii) If $v_j'v_i'$ *is neither an edge or a semi-edge in* G'*, but* v_j' *and* v_i' *are in the same connected component of* G'*, then* $\dim H_1(G) = \dim H_1(G' + 1)$*, and* $\dim H_0(G) = \dim H_0(G')$.

(iii) If v_j' *and* v_i' *are not in the same connected component of* G'*, then* $\dim H_1(G) = \dim H_1(G')$ *and* $\dim H_0(G) = \dim H_0(G') - 1$.

2.2.6 Path Laplacian

Recall that a chain complex of ∂-invariant paths is given by

$$\cdots \Omega_{n+1} \xrightarrow{\partial_{n+1}} \Omega_n \xrightarrow{\partial_n} \Omega_{n-1} \xrightarrow{\partial_{n-1}} \Omega_{n-2} \cdots ,$$

where $\Omega_n = \Omega_n(P) = \{v \in \mathcal{A}_n : \partial v \in \mathcal{A}_{n-1}\}$ and $\partial_n := \partial|_{\Omega_n}$. Alternatively, assume $S_n := S_n(P)$ to be the set of n-th elementary paths in P, then we define an inner product

$$\langle \cdot, \cdot \rangle : S_n \times S_n \to \mathbb{R}$$

such that for any $e_{i_0 \dots i_n}, e_{j_0 \dots j_n} \in S_n$, the following satisfies

$$\langle e_{i_0 \dots i_n}, e_{j_0 \dots j_n} \rangle = \begin{cases} 1 & \text{if } e_{i_0 \dots i_n} = e_{j_0 \dots j_n}, \\ 0 & \text{otherwise.} \end{cases} \tag{2.55}$$

Let M_n be a matrix representation of $\partial : \mathcal{A}_n \to \mathcal{A}_{n-1}$ with respect to the standard basis of $\mathcal{A}_n$ and $\mathcal{A}_{n-1}$. Define an inclusion map $\iota_n : \Omega_n \hookrightarrow \mathcal{A}_n$, then the matrix representation of ι_n with respect to the basis of Ω_n (i.e., the standard basis of $\mathcal{A}_n$ with the removal of generators that are not in Ω_n) and the standard basis of $\mathcal{A}_n$ is denoted as O_n. Denote the boundary matrix representation of ∂_n as B_n, then we have

$$O_{n-1} B_n = \tilde{M}_n O_n. \tag{2.56}$$

If O_{n-1} is a square matrix, then O_n is actually an identity matrix, and we have

$$B_n = O_{n-1}^{-1} \tilde{M}_n O_n = \tilde{M}_n O_n, \tag{2.57}$$

where $\tilde{M}_n$ is M_n with the removal of rows that their basis are not elementary $(n-1)$-paths in P. Otherwise, B_n is the least-square solution to Eq. (2.56).

Note that B_n is the matrix representation of ∂_n with respect to the basis of Ω_n and Ω_{n-1}. Dual space $\Omega^n := \text{Hom}(\Omega_n, \mathbb{K})$ of Ω_n is equipped with dual maps d to form a cochain complex

$$\cdots \Omega^{n+1} \xleftarrow{d_{n+1}} \Omega^n \xleftarrow{d_n} \Omega^{n-1} \xleftarrow{d_{n-1}} \Omega^{n-2} \cdots,$$

where d_n is called a coboundary operator. The inner product on Ω_n induces an inner product $\ll \cdot, \cdot \gg$ on Ω^n such that

$$\ll f, g \gg = \sum_{e \in S_n} f(e) g(e), \quad \forall f, g \in \Omega^n.$$

We denote the adjoint operator of ∂_n be $\partial_n^* : \Omega_{n-1} \to \Omega_n$. Note that similar inner product $\ll \cdot, \cdot \gg$ on Ω^n was defined in the literature [64]. Hence, the coboundary operator d_n is

consistent with the adjoint operator ∂_n^*. Then, for integers $p \geq 0$, the *n-th path Laplacian operator* is a linear operator: $\Delta_n : \Omega_n \to \Omega_n$ given by

$$\Delta_n = \partial_{n+1}\partial_{n+1}^* + \partial_n^*\partial_n, \tag{2.58}$$

and $\Delta_0 = \partial_1\partial_1^*$. The *n-th path Laplacian matrix* corresponding to Δ_n is expressed by

$$L_n = B_{n+1}B_{n+1}^T + B_n^T B_n. \tag{2.59}$$

Since L_n is positive semi-definite and symmetric, its eigenvalues are all real and non-negative. Additionally, recall that the Betti number β_n of path complex P satisfies

$$\beta_n = \dim\ker\partial_n - \dim\mathrm{im}\,\partial_{n+1} = \dim\ker\Delta_n. \tag{2.60}$$

It is easy to show that

$$\beta_n = \text{nullity}(L_n) = \text{the number of zero eigenvalues of } L_n. \tag{2.61}$$

Moreover, assume the dimension of L_n is N, then the set of spectra of L_n is denoted as

$$\text{Spectra}(L_n) = \{(\lambda_1)_n, (\lambda_2)_n, \cdots, (\lambda_N)_n\}.$$

Figure 2.8 shows 5 digraphs with multiple vertices and directed edges. Here, we take them as examples to give a detailed illustration of L_n matrix constructions.

Construction of L_0 – Figure 2.8a Since $L_0 = B_1B_1^T$, then we first construct B_1, where $B_1 = O_0^{-1}\tilde{M}_1O_1$ according to Eq. (2.57), we have $O_0 = \begin{array}{c} \\ e_1 \\ e_2 \\ e_3 \end{array}\begin{array}{c} \begin{array}{ccc} e_1 & e_2 & e_3 \end{array} \\ \begin{pmatrix} 1 & 0 & 0 \\ 0 & 1 & 0 \\ 0 & 0 & 1 \end{pmatrix} \end{array}$, and $M_1 = \begin{array}{c} \\ e_1 \\ e_2 \\ e_3 \end{array}\begin{array}{c} \begin{array}{ccc} e_{12} & e_{23} & e_{31} \end{array} \\ \begin{pmatrix} -1 & 0 & 1 \\ 1 & -1 & 0 \\ 0 & 1 & -1 \end{pmatrix} \end{array}$, and $O_1 = \begin{array}{c} \\ e_{12} \\ e_{23} \\ e_{31} \end{array}\begin{array}{c} \begin{array}{ccc} e_{12} & e_{23} & e_{31} \end{array} \\ \begin{pmatrix} 1 & 0 & 0 \\ 0 & 1 & 0 \\ 0 & 0 & 1 \end{pmatrix} \end{array}$. Since e_1, e_2, and e_3 are all elemen-

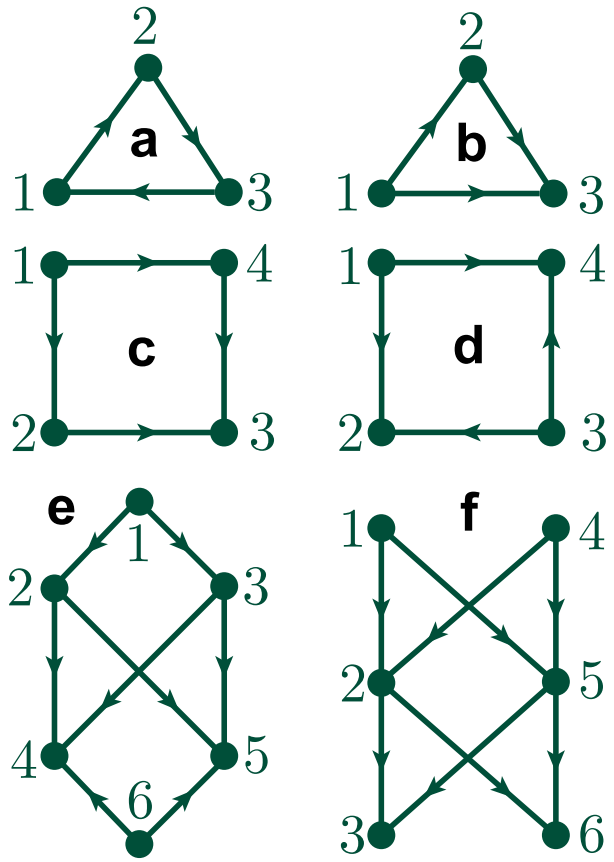

Figure 2.8: Five digraphs. **a** and **b** Digraphs with 3 vertices and 3 directed edges. **c** and **d** Digraphs with 4 vertices and 4 directed edges. **e** A digraph with 6 vertices and 8 directed edges. **f** A digraph with 6 vertices and 8 directed edges.

tary 0-paths (vertices), $M_1 = \tilde{M}_1$. We have $B_1 = O_0^{-1}\tilde{M}_1 O_1 = \begin{array}{c} \\ e_1 \\ e_2 \\ e_3 \end{array} \begin{array}{c} \begin{array}{ccc} e_{12} & e_{23} & e_{31} \end{array} \\ \begin{pmatrix} -1 & 0 & 1 \\ 1 & -1 & 0 \\ 0 & 1 & -1 \end{pmatrix} \end{array}$. Then $L_0 = B_1 B_1^T = \begin{pmatrix} 2 & -1 & -1 \\ -1 & 2 & -1 \\ -1 & -1 & 2 \end{pmatrix}$, which gives Spectra$(L_0) = \{0, 3, 3\}$ and thus, one finally has $\beta_0 = 1$.

Construction of L_1 – Figure 2.8a We have $L_1 = B_2 B_2^T + B_1^T B_1$, where B_1 has been formed, so we focus on the construction of $B_2 = O_1^{-1}\tilde{M}_2 O_2$ according to Eq. (2.57). Since

$$O_1 = \begin{array}{c} \\ e_{12} \\ e_{23} \\ e_{31} \end{array} \begin{array}{c} \begin{array}{ccc} e_{12} & e_{23} & e_{31} \end{array} \\ \begin{pmatrix} 1 & 0 & 0 \\ 0 & 1 & 0 \\ 0 & 0 & 1 \end{pmatrix} \end{array}, \text{ and } M_2 = \begin{array}{c} \\ e_{11} \\ e_{12} \\ e_{13} \\ e_{21} \\ e_{22} \\ e_{23} \\ e_{31} \\ e_{32} \\ e_{33} \end{array} \begin{array}{c} \begin{array}{ccc} e_{123} & e_{231} & e_{312} \end{array} \\ \begin{pmatrix} 0 & 0 & 0 \\ 1 & 0 & 1 \\ -1 & 0 & 0 \\ 0 & -1 & 0 \\ 0 & 0 & 0 \\ 1 & 1 & 0 \\ 0 & 1 & 1 \\ 0 & 0 & -1 \\ 0 & 0 & 0 \end{pmatrix} \end{array}, \text{ and } O_2 \text{ is a } 3 \times 0 \text{ empty}$$

matrix since $\Omega_2 = \{0\}$. Therefore, $B_2 = O_1^{-1}\tilde{M}_2 O_2$ is a 3×0 empty matrix. Additionally, $L_1 = B_2 B_2^T + B_1^T B_1 = \begin{pmatrix} 2 & -1 & -1 \\ -1 & 2 & -1 \\ -1 & -1 & 2 \end{pmatrix}$, where Spectra$(L_1) = \{0, 3, 3\}$ and thus, one finally has $\beta_1 = 1$.

Construction of L_2 – Figure 2.8a We have $L_2 = B_3 B_3^T + B_2^T B_2$, where B_2 is an empty matrix. Hence, we focus on the construction of $B_3 = O_2^{-1}\tilde{M}_3 O_3$ according to Eq. (2.57). We have $\mathcal{A}_2 = \text{span}\{e_{123}, e_{231}, e_{312}\}$ and $\mathcal{A}_1 = \text{span}\{e_{12}, e_{23}, e_{31}\}$. Note that $\partial_2(e_{123}) = e_{23} - e_{13} + e_{12}$ where e_{13} is not in $\mathcal{A}_1$. Hence, e_{123} is not in Ω_2. The same conclusion can be deduced for e_{231} and e_{312}. Therefore, we have $\Omega_2 = \{0\}$, and it is straightforward to get that L_2 is an empty matrix.

Construction of L_0 – Figure 2.8b Since $L_0 = B_1 B_1^T$, then we should first construct B_1, where $B_1 = O_0^{-1}\tilde{M}_1 O_1$ according to Eq. (2.57). Since $O_0 = \begin{array}{c} \\ e_1 \\ e_2 \\ e_3 \end{array} \begin{array}{c} \begin{array}{ccc} e_1 & e_2 & e_3 \end{array} \\ \begin{pmatrix} 1 & 0 & 0 \\ 0 & 1 & 0 \\ 0 & 0 & 1 \end{pmatrix} \end{array}$,

$$M_1 = \begin{array}{c} \\ e_1 \\ e_2 \\ e_3 \end{array} \begin{array}{c} \begin{array}{ccc} e_{12} & e_{13} & e_{23} \end{array} \\ \begin{pmatrix} -1 & -1 & 0 \\ 1 & 0 & -1 \\ 0 & 1 & 1 \end{pmatrix} \end{array}, \text{ and } O_1 = \begin{array}{c} \\ e_{12} \\ e_{13} \\ e_{23} \end{array} \begin{array}{c} \begin{array}{ccc} e_{12} & e_{13} & e_{23} \end{array} \\ \begin{pmatrix} 1 & 0 & 0 \\ 0 & 1 & 0 \\ 0 & 0 & 1 \end{pmatrix} \end{array}.$$ Since $e_1, e_2,$ and e_3 are all elementary 0-paths (vertices). Therefore, $M_1 = \tilde{M}_1$, and we have $B_1 = O_0^{-1}\tilde{M}_1 O_1 =$

$$\begin{array}{c} \\ e_1 \\ e_2 \\ e_3 \end{array} \begin{array}{c} \begin{array}{ccc} e_{12} & e_{13} & e_{23} \end{array} \\ \begin{pmatrix} -1 & -1 & 0 \\ 1 & 0 & -1 \\ 0 & 1 & 1 \end{pmatrix} \end{array}.$$ Then $L_0 = B_1 B_1^T = \begin{pmatrix} 2 & -1 & -1 \\ -1 & 2 & -1 \\ -1 & -1 & 2 \end{pmatrix}$, which gives the Spectra$(L_0) =$ $\{0, 3, 3\}$ and thus, one finally has $\beta_0 = 1$.

Construction of L_1 – Figure 2.8b We have $L_1 = B_2 B_2^T + B_1^T B_1$, where B_1 has been formed, so we focus on the construction of $B_2 = O_1^{-1}\tilde{M}_2 O_2$ according to Eq. (2.57). First, $\mathcal{A}_2 = \text{span}\{e_{123}\}$ and $\mathcal{A}_1 = \text{span}\{e_{12}, e_{13}, e_{23}\}$. Note that $\partial_2(e_{123}) = e_{23} - e_{13} + e_{12}$ where $e_{12}, e_{23},$ and e_{13} are all in $\mathcal{A}_1$. Hence, $\Omega_2 = \mathcal{A}_2 = \text{span}\{e_{123}\}$. Note that $O_1 =$

$$\begin{array}{c} \\ e_{12} \\ e_{13} \\ e_{23} \end{array} \begin{array}{c} \begin{array}{ccc} e_{12} & e_{13} & e_{23} \end{array} \\ \begin{pmatrix} 1 & 0 & 0 \\ 0 & 1 & 0 \\ 0 & 0 & 1 \end{pmatrix} \end{array}, M_2 = \begin{array}{c} \\ e_{11} \\ e_{12} \\ e_{13} \\ e_{21} \\ e_{22} \\ e_{23} \\ e_{31} \\ e_{32} \\ e_{33} \end{array} \begin{array}{c} \begin{array}{c} e_{123} \end{array} \\ \begin{pmatrix} 0 \\ 1 \\ -1 \\ 0 \\ 0 \\ 1 \\ 0 \\ 0 \\ 0 \end{pmatrix} \end{array}, \text{ and } O_2 = \begin{array}{c} \\ e_{123} \end{array} \begin{array}{c} \begin{array}{c} e_{123} \end{array} \\ \begin{pmatrix} 1 \end{pmatrix} \end{array}.$$ The $e_{11}, e_{21}, e_{22}, e_{31}, e_{32},$

and e_{33} are not elementary 1-paths in P. Hence, $\tilde{M}_2 = \begin{matrix} \\ e_{12} \\ e_{13} \\ e_{23} \end{matrix} \begin{matrix} e_{123} \\ \begin{pmatrix} 1 \\ -1 \\ 1 \end{pmatrix} \end{matrix}$, and then $B_2 = O_1^{-1}\tilde{M}_2 O_2 = \begin{matrix} \\ e_{12} \\ e_{13} \\ e_{23} \end{matrix} \begin{matrix} e_{123} \\ \begin{pmatrix} 1 \\ -1 \\ 1 \end{pmatrix} \end{matrix}$. Therefore, $L_1 = B_2 B_2^T + B_1^T B_1 = \begin{pmatrix} 3 & 0 & 0 \\ 0 & 3 & 0 \\ 0 & 0 & 3 \end{pmatrix}$, where Spectra$(L_1) = \{3, 3, 3\}$ and thus, we finally have $\beta_1 = 0$.

Construction of L_2 – Figure 2.8b According to Eq. (2.59), we have $L_2 = B_3 B_3^T + B_2^T B_2$ and $B_3 = O_2^{-1}\tilde{M}_3 O_3$. Since there is no 3-path existing, so the M_3 and O_3 are both empty matrix. Hence $L_2 = (3)$, Spectra$(L_2) = \{3\}$, and thus, one has $\beta_2 = 0$.

In the following section, we will omit the detailed construction steps of boundary matrix B_n. Table 2.11, Table 2.12, Table 2.13, and Table 2.14 list the boundary matrix B_n and the n-th path Laplacian matrix L_n for with its corresponding Betti numbers β_n and spectrum Spectra(L_n) for Figure 2.8 **c**, **d**, **e**, and **f**. It is worth to mention that β_n can distinguish the same graph with different paths assigned. For example, Figure 2.8 **c** and **d** have the same undirected graph structure with different paths assigned. We have $\beta_1 = 0$ for Figure 2.8 **c** and $\beta_1 = 1$ for Figure 2.8 **d**.

2.2.7 Persistent Path Laplacian

From Section 2.2.6, the way to calculate both harmonic spectra (topological invariants) and non-harmonic spectra of n-th path Laplacian matrix is genuinely free of metrics or coordinates, which contains too little information to fully describe the object. Therefore, inspired by the idea of the persistent spectral graph (PSG), persistent path Laplacian (PPL) is proposed to create a sequence of digraphs induced by varying a filtration parameter to encode more geometric or structural information.

Table 2.11: Illustration of digraph **c** in Figure 2.8.

n	$n=0$	$n=1$	$n=2$
Ω_n	span$\{e_1, e_2, e_3, e_4\}$	span$\{e_{12}, e_{14}, e_{23}, e_{43}\}$	span$\{e_{143} - e_{123}\}$
B_{n+1}	$\begin{array}{c} \\ e_1 \\ e_2 \\ e_3 \\ e_4 \end{array} \begin{array}{c} \begin{array}{cccc} e_{12} & e_{14} & e_{23} & e_{43} \end{array} \\ \begin{pmatrix} -1 & -1 & 0 & 0 \\ 1 & 0 & -1 & 0 \\ 0 & 0 & 1 & 1 \\ 0 & 1 & 0 & -1 \end{pmatrix} \end{array}$	$\begin{array}{c} \\ e_{12} \\ e_{14} \\ e_{23} \\ e_{43} \end{array} \begin{array}{c} \begin{array}{c} e_{143} - e_{123} \end{array} \\ \begin{pmatrix} -1 \\ 1 \\ -1 \\ 1 \end{pmatrix} \end{array}$	1×0 empty matrix
L_n	$\begin{pmatrix} 2 & -1 & 0 & -1 \\ -1 & 2 & -1 & 0 \\ 0 & -1 & 2 & -1 \\ -1 & 0 & -1 & 2 \end{pmatrix}$	$\begin{pmatrix} 3 & 0 & 0 & -1 \\ 0 & 3 & -1 & 0 \\ 0 & -1 & 3 & 0 \\ -1 & 0 & 0 & 3 \end{pmatrix}$	$\begin{pmatrix} 4 \end{pmatrix}$
β_n	1	0	0
Spectra(L_n)	$\{0, 2, 2, 4\}$	$\{2, 2, 4, 4\}$	{4}

Table 2.12: Illustration of digraph **d** in Figure 2.8.

n	$n=0$	$n=1$	$n=2$
Ω_n	span$\{e_1, e_2, e_3, e_4\}$	span$\{e_{12}, e_{14}, e_{32}, e_{34}\}$	$\{0\}$
B_{n+1}	$\begin{array}{c} \\ e_1 \\ e_2 \\ e_3 \\ e_4 \end{array} \begin{array}{c} \begin{array}{cccc} e_{12} & e_{14} & e_{32} & e_{34} \end{array} \\ \begin{pmatrix} -1 & -1 & 0 & 0 \\ 1 & 0 & 1 & 0 \\ 0 & 0 & -1 & -1 \\ 0 & 1 & 0 & 1 \end{pmatrix} \end{array}$	4×0 empty matrix	$\begin{pmatrix} \ / \ \end{pmatrix}$
L_n	$\begin{pmatrix} 2 & -1 & 0 & -1 \\ -1 & 2 & -1 & 0 \\ 0 & -1 & 2 & -1 \\ -1 & 0 & -1 & 2 \end{pmatrix}$	$\begin{pmatrix} 2 & 1 & 1 & 0 \\ 1 & 2 & 0 & 1 \\ 1 & 0 & 2 & 1 \\ 0 & 1 & 1 & 2 \end{pmatrix}$	$\begin{pmatrix} \ / \ \end{pmatrix}$
β_n	1	1	0
Spectra(L_n)	$\{0, 2, 2, 4\}$	$\{0, 2, 4, 4\}$	/

First, we consider a *filtration of digraphs* $\mathcal{G} : \mathbb{R} \to \mathcal{D}$, which is a morphism $f_{s,t} : H_p(\mathcal{G}_t; \mathbb{K}) \to H_p(\mathcal{G}_s; \mathbb{K})$ from the category of real number $\mathbb{R}$ to the category of digraphs $\mathcal{D}$ that satisfies:

$$\mathcal{G}(t) \subseteq \mathcal{G}(s), \forall t \leq s,$$

Table 2.13: Illustration of digraph **e** in Figure 2.8.

n	$n=0$	$n=1$	$n=2$
Ω_n	span$\{e_1,e_2,e_3,e_4,e_5,e_6\}$	span$\{e_{12},e_{13},e_{24},e_{25},e_{34},e_{35},e_{64},e_{65}\}$	span$\{e_{134}-e_{124},e_{135}-e_{125}\}$
B_{n+1}	$\begin{array}{c} \\ e_1\\e_2\\e_3\\e_4\\e_5\\e_6\end{array}\begin{array}{c}\begin{array}{cccccccc} e_{12}&e_{13}&e_{24}&e_{25}&e_{34}&e_{35}&e_{64}&e_{65}\end{array}\\ \begin{pmatrix} -1&-1&0&0&0&0&0&0\\ 1&0&-1&-1&0&0&0&0\\ 0&1&0&0&-1&-1&0&0\\ 0&0&1&0&1&0&1&0\\ 0&0&0&1&0&1&0&1\\ 0&0&0&0&0&0&-1&-1\end{pmatrix}\end{array}$	$\begin{array}{c} \\ e_{12}\\e_{13}\\e_{24}\\e_{25}\\e_{34}\\e_{35}\\e_{64}\\e_{65}\end{array}\begin{array}{c}\begin{array}{cc} e_{134}-e_{124}&e_{135}-e_{125}\end{array}\\ \begin{pmatrix} -1&-1\\ 1&1\\ -1&0\\ 0&-1\\ 1&0\\ 0&1\\ 0&0\\ 0&0\end{pmatrix}\end{array}$	2×0 empty matrix
L_n	$\begin{pmatrix} 2&-1&-1&0&0&0\\ -1&3&0&-1&-1&0\\ -1&0&3&-1&-1&0\\ 0&-1&-1&3&0&-1\\ 0&-1&-1&0&3&-1\\ 0&0&0&-1&-1&2\end{pmatrix}$	$\begin{pmatrix} 4&-1&0&0&-1&-1&0&0\\ -1&4&-1&-1&0&0&0&0\\ 0&-1&3&1&0&0&1&0\\ 0&-1&1&3&0&0&0&1\\ -1&0&0&0&3&1&1&0\\ -1&0&0&0&1&3&0&1\\ 0&0&1&0&1&0&2&1\\ 0&0&0&1&0&1&1&2\end{pmatrix}$	$\begin{pmatrix} 4&2\\2&4\end{pmatrix}$
β_n	1	1	0
Spectra(L_n)	$\{0,1.4384,3,3,3,5\}$	$\{0,1.4384,2,3,3,3,5.5616,6\}$	{2,6}

Table 2.14: Illustration of digraph **f** in Figure 2.8.

n	$n=0$	$n=1$	$n=2$
Ω_n	span$\{e_1,e_2,e_3,e_4,e_5,e_6\}$	span$\{e_{12},e_{15},e_{23},e_{26},e_{42},e_{45},e_{53},e_{56}\}$	span$\{e_{153}-e_{123}$, $e_{156}-e_{126}$, $e_{453}-e_{423}$, $e_{456}-e_{426}\}$
B_{n+1}	$\begin{array}{c} \\ e_1\\e_2\\e_3\\e_4\\e_5\\e_6\end{array}\begin{array}{c}\begin{array}{cccccccc} e_{12}&e_{15}&e_{23}&e_{26}&e_{42}&e_{45}&e_{53}&e_{56}\end{array}\\ \begin{pmatrix} -1&-1&0&0&0&0&0&0\\ 1&0&-1&-1&1&0&0&0\\ 0&0&1&0&0&0&1&0\\ 0&0&0&0&-1&-1&0&0\\ 0&1&0&0&0&1&-1&-1\\ 0&0&0&1&0&0&0&1\end{pmatrix}\end{array}$	$\begin{array}{c} \\ e_{12}\\e_{15}\\e_{23}\\e_{26}\\e_{42}\\e_{45}\\e_{53}\\e_{56}\end{array}\begin{array}{c}\begin{array}{cccc} e_{153}-e_{123}&e_{156}-e_{126}&e_{453}-e_{423}&e_{456}-e_{426}\end{array}\\ \begin{pmatrix} -1&-1&0&0\\ 1&1&0&0\\ -1&0&-1&0\\ 0&-1&0&-1\\ 0&0&-1&-1\\ 0&0&1&1\\ 1&0&1&0\\ 0&1&0&1\end{pmatrix}\end{array}$	4×0 empty matrix
L_n	$\begin{pmatrix} 2&-1&0&0&-1&0\\ -1&4&-1&-1&0&-1\\ 0&-1&2&0&-1&0\\ 0&-1&0&2&-1&0\\ -1&0&-1&-1&4&-1\\ 0&-1&0&0&-1&2\end{pmatrix}$	$\begin{pmatrix} 4&-1&0&0&1&0&-1&-1\\ -1&4&-1&-1&0&1&0&0\\ 0&-1&4&1&0&-1&-1&0\\ 0&-1&1&4&0&-1&0&-1\\ 1&0&0&0&4&-1&-1&-1\\ 0&1&-1&-1&-1&4&0&0\\ -1&0&-1&0&-1&0&4&1\\ -1&0&0&-1&-1&0&1&4\end{pmatrix}$	$\begin{pmatrix} 4&2&2&0\\2&4&0&2\\2&0&4&2\\0&2&2&4\end{pmatrix}$
β_n	1	0	1
Spectra(L_n)	$\{0,2,2,2,4,6\}$	$\{2,2,2,4,4,4,6,8\}$	{0,4,4,8}

where $G_t := \mathcal{G}(t) \in \mathcal{D}$ and $G_s := \mathcal{G}(s) \in \mathcal{D}$. Consider a sequence of finitely many positive

integers $1, 2, \ldots, m$, we have a sequence of digraphs

$$G_1 \subseteq G_2 \subseteq \cdots \subseteq G_m.$$

For each digraph G_t, we denote its corresponding chain group to be $\Omega_n(G_t)$, and the n-boundary operator of G_t is denoted by $\partial_n^t : \Omega_n(G_t) \to \Omega_{n-1}(G_t), \forall n \geq 0$.

Similarly, as in persistent homology, a sequence of chain complexes can be denoted as

$$\begin{array}{ccccccccccccc}
\cdots & \Omega_{n+1}^1 & \xrightarrow{\partial_{n+1}^1} & \Omega_n^1 & \xrightarrow{\partial_n^1} & \cdots & \xrightarrow{\partial_3^1} & \Omega_2^1 & \xrightarrow{\partial_2^1} & \Omega_1^1 & \xrightarrow{\partial_1^1} & \Omega_0^1 & \xrightarrow{\partial_0^1} & \Omega_{-1}^1 \\
& \downarrow & & \downarrow & & & & \downarrow & & \downarrow & & \downarrow & & \\
\cdots & \Omega_{n+1}^2 & \xrightarrow{\partial_{q+1}^2} & \Omega_n^2 & \xrightarrow{\partial_n^2} & \cdots & \xrightarrow{\partial_3^2} & \Omega_2^2 & \xrightarrow{\partial_2^2} & \Omega_1^2 & \xrightarrow{\partial_1^2} & \Omega_0^2 & \xrightarrow{\partial_0^2} & \Omega_{-1}^2 \\
& \vdots & & \vdots & & & & \vdots & & \vdots & & \vdots & & \\
& \downarrow & & \downarrow & & & & \downarrow & & \downarrow & & \downarrow & & \\
\cdots & \Omega_{n+1}^m & \xrightarrow{\partial_{q+1}^m} & \Omega_n^m & \xrightarrow{\partial_n^m} & \cdots & \xrightarrow{\partial_3^m} & \Omega_2^m & \xrightarrow{\partial_2^m} & \Omega_1^m & \xrightarrow{\partial_1^m} & \Omega_0^m & \xrightarrow{\partial_0^m} & \Omega_{-1}^m
\end{array} \tag{2.62}$$

For the sake of simplicity, we use Ω_n^t to represent $\Omega_n(G_t)$. Suppose a subset of Ω_n^s whose boundary is in Ω_{n-1}^t as:

$$\Omega_n^{t,s} \coloneqq \{\alpha \in \Omega_n^s \mid \partial_n^s \alpha \in \Omega_{n-1}^t\}. \tag{2.63}$$

The persistent n-boundary operator is denoted as $\eth_n^{t,s} : \Omega_n^{t,s} \to \Omega_{n-1}^t$, and its corresponding adjoint operator is $(\eth_n^{t,s})^* : \Omega_{n-1}^t \to \Omega_n^{t,s}$. Therefore, the persistent n-th path Laplacian operator $\Delta_n^{t,s} : \Omega_n^t \to \Omega_n^t$ defined along the filtration is:

$$\Delta_n^{t,s} = \eth_{n+1}^{t,s} \left(\eth_{n+1}^{t,s}\right)^* + \partial_n^{t^*} \partial_n^t. \tag{2.64}$$

Since $\Delta_n^{t,s}$ inherits the inner product from $\eth_{n+1}^{t,s}$, then the adjoint map $\left(\eth_{n+1}^{t,s}\right)^*$ is well defined. Intuitively, the matrix representation of $\Delta_n^{t,s}$ is

$$L_n^{t,s} = B_{n+1}^{t,s} P^{-1} (B_{n+1}^{t,s})^T + (B_n^t)^T B_n^t, \tag{2.65}$$

where P^{-1} is the associated inner product matrix of $\Omega_{n+1}^{t,s}$ with arbitrary basis. Moreover, assume the dimension of $L_n^{t,s}$ is N, then the spectra of $L_n^{t,s}$ that are arranged in ascending

order can be displayed as:

$$\text{Spectra}(L_n^{t,s}) = \{(\lambda_1)_n^{t,s}, (\lambda_2)_n^{t,s}, \cdots, (\lambda_N)_n^{t,s}\}.$$

Note that the smallest non-harmonic spectra of $L_n^{t,s}$ is denoted as $(\tilde{\lambda}_2)_n^{t,s}$. We call the multiplicity of zero spectra of $L_q^{t,s}$ to be persistent n-th Betti number $\beta_n^{t,s}$ from G_t to G_s.

$$\beta_n^{t,s} = \text{nullity}(L_n^{t,s}) = \text{the number of zero eigenvalues (i.e., harmonic eigenvalues) of } L_n^{t,s}. \tag{2.66}$$

Distanced-based filtration Specifically, suppose $G(w) = (V, E, w)$ is a weighted digraph, where V is the set of the vertices and E is the set of the directed edges. Assume w is a weight function $w : E \to \mathbb{R}$. For example, if V is in the Euclidean space, then a digraph $G(w)$ is a geometric digraph (a geometric digraph is a digraph in which the vertices are embedded as points in the Euclidean space, and the edges are embedded as non-crossing directed line segments). For any $(i, j) \in E$ where $i, j \in V$, we define $w(i, j) = \|i - j\|$, where $\|\cdot\|$ is a Euclidean metric. Hence, for every $\delta \in \mathbb{R}$, a digraph can be described as $G^\delta = (V, E^\delta) = (V, \{e \in E : w(e) \leq \delta\})$, and a filtration of digraphs can be described as $\{G^\delta \hookrightarrow G^{\delta'}\}_{\delta \leq \delta'}$.

Therefore, the persistent n-th path Laplacian matrix defined on the filtration is

$$L_n^{\delta,\delta'} = B_{n+1}^{\delta,\delta'} P^{-1} (B_{n+1}^{\delta,\delta'})^T + (B_n^\delta)^T B_n^\delta, \tag{2.67}$$

where its corresponding Betti numbers and spectra can be expressed as:

$$\beta_n^{\delta,\delta'} = \text{nullity}(L_n^{\delta,\delta'}) = \text{the number of zero eigenvalues (i.e., harmonic eigenvalues) of } L_n^{\delta,\delta'}. \tag{2.68}$$

$$\text{Spectra}(L_n^{\delta,\delta'}) = \{(\lambda_1)_n^{\delta,\delta'}, (\lambda_2)_n^{\delta,\delta'}, \cdots, (\lambda_N)_n^{\delta,\delta'}\}. \tag{2.69}$$

Notably, the Fiedler value (i.e., spectral gap) of $L_n^{\delta,\delta'}$ is widely used in many other areas such as physics and geography, which is denoted as $\tilde{\lambda}_n^{\delta,\delta'}$. As shown below, it is sensitive to both topological and geometric changes.

Moreover, it is worth to mention that isolated points (vertices) can be either included in the digraphs (under the distance-based filtration) or removed from the digraphs (under the distanced-based filtration with removal of isolated points).

CHAPTER 3

METHODS ON MATHEMATICAL MODELING OF VIROLOGY

3.1 Genomics Analysis

3.1.1 Sequence Alignment

Sequence alignment is a method in which one can arrange DNA, RNA, or amino acid sequences to identify their similar regions [65]. Such similar regions may arise from functional, structural, geometrical, or evolutionary similarities. Though sequence alignment offers the best accuracy, it is not practical to be used for a large sample size. There are two main categories of sequence alignment, namely pair-wise sequence alignment and multiple sequence alignment. The former only compares two sequences at a time, while the latter compares many sequences. There are many popular tools for sequence alignment such as BLAST (Basic Local Alignment Search Tool) for pair-wise alignment and MAFFT, Clustal Omega, ClustalW, and MUSCLE, for multiple sequence alignment. The following section describes BLAST first followed by several multiple sequence alignment tools.

3.1.1.1 Pairwise Sequence Alignment

One of the popular pair-wise sequence alignment tools is BLAST. BLAST is a local similarity search tool that is commonly used to find similar DNA, RNA, and amino acid sequences to the sequence in question. BLAST was created in 1990 based on the k-tuple method, and has since been implemented in the GenBank, and had numerous updates to increase efficiency and accuracy. k-tuple method [66] is a fast heuristic method for pairwise alignment and is commonly used as an initial step for a large sample size. Similarity score, S_{ij} between sequences i and j is defined as the number of k-tuple matches in the best pairwise alignment minus a fixed gap penalty term. For DNA and RNA, k usually

ranges from 2 to 4, and for amino acids, k is 1 or 2. S_{ij} is calculated as the number of identities divided by the number of residues compared between i and j. The distance is defined as,

$$d_{ij} = 1 - \frac{S_{ij}}{100}. \tag{3.1}$$

Note that this method does not guarantee optimal alignment, but it is a fast heuristic method and can be used for the initialization of BLAST and multiple sequence alignment.

BLAST begins by first creating a list of k-letter words. It then searches for possible matching k-letter words in the databank and scores them, and any words that score above a threshold are kept. The high-scoring words are kept in a search tree. This process is then extended to high scoring pairs (HSPs), which also looks for similar words, rather than only looking at exact matching words. After searching for HSPs, the significance of the HSPs score is considered by utilizing Gumbel extreme value distribution (EVD). Further details can be found in the literature [67, 68]. The GenBank tutorial can be found in Ref. [69]. As a basic tool for sequence alignment, it is utilized to detect, identify, or search sequences in a database. For example, similar coronavirus strands in other organisms, such as that of pangolins [70, 71] and bats[72] were found. This tool is also used to detect SARS-CoV-2 virus in the environment[73, 74] such as waste waters[75, 76].

3.1.1.2 Multiple Sequence Alignment (MSA)

Unlike pair-wise sequence alignment, MSA arranges 3 or more DNA, RNA, or protein sequences by identical regions. Through multiple sequence alignment, one can further analyze sequence homology to find evolutionary origins. In many cases, one uses a reference sequence, which is the first sequenced data, to observe mutation in SARS-CoV-2 genome [77]. There are several popular tools, Clustal[78], MUSCLE[79], MAFFT[80, 81], etc.

Clustal Clustal is a series of multiple sequence alignment tools for sequence analysis. With the first version Clustal released in 1988[78], its package has been developed for several generations based on different methods. ClustalW is the third generation and is updated to ClustalW2 currently, which aligns sequences with the best similarity score first, and progressively aligns more distant scores[82, 83]. This is achieved by first obtaining a rough pairwise sequence alignment using the k-tuple method [66], followed by a neighbor-joining method [84], which uses midpoint rooting to create a guided tree. ClustalW2 is used as the basis for global alignment.

As for Clustal Omega, unlike the ClustalW, it uses a guided tree approach, rather than a progressive alignment method. Clustal Omega begins with first producing a pairwise alignment using the k-tuple method. This, however, does not guarantee finding optimal alignment, but it is time-efficient. Then, the sequences are clustered using the mBed method [85], which calculates pairwise distance using the embedding method. Afterward, K-means clustering is used to further cluster the sequence. Then, a guided tree is formed utilizing the UPGMA method [86]. Lastly, MSA is produced using the HHAlign package from HH-Suite [86]. Clustal Omega's advantage comes from the large-scale MSA. The accuracy and time complexity are average for a low number of samples. For a large number of samples with a long sequence, Clustal Omega produces high accuracy and is time-efficient. ClustalW is the updated version of the original Clustal MSA tool.

Multiple alignment using fast Fourier transform (MAFFT) MAFFT is a MSA package based on fast Fourier transform (FFT). Given two sequences v_1 and v_2, the correlation $c_v(s)$ of volume between the two sequences with positional lag of s sites can be defined as

$$c_v(s) = \sum_{1 \le n \le N, 1 \le n+s \le M} \hat{v}_1(n)\hat{v}_2(n+s)$$

where $\hat{v}_1$ and $\hat{v}_2$ are the FFT of the two sequences. If homologous regions exists, through Fourier analysis, there will be a peak in similar region. For amino acid sequences, MAFFT

also calculates correlation between polarity:

$$c_\rho(s) = \sum_{1 \le n \le N, 1 \le n+s \le M} \hat{\rho}_1(n)\hat{\rho}_2(n+s)$$

where $\rho(s)$ is the polarity of each amino acid, N is the length of v_1, and M is the length of v_2. Then, a scoring function can be calculated through the sum of the two correlations

$$c(s) = c_v(s) + c_\rho(s).$$

To reduce the computational complexity, only peaks above some threshold are considered. Note that the peak does not tell the location of the homologous region directly, and only shows the lag. Therefore, neighboring regions at the peak must be analyzed carefully. Further details of MAFFT can be found in the literature [80, 81].

3.1.2 Single Nucleotide Polymorphism Calling

Single nucleotide polymorphism (SNP) calling measures the genetic variations between different members of a species. Establishing the SNP calling method to the investigation of the genotype changes during the transmission and evolution of SARS-CoV-2 is of great importance [21, 25]. By analyzing the rearranged genome sequences, SNP profiles, which record all of the SNP positions in teams of the nucleotide changes and their corresponding positions, can be constructed. The SNP profiles of a given SARS-CoV-2 genome isolated from a COVID-19 patient capture all the differences from a complete reference genome sequence and can be considered as the genotype of the individual SARS-CoV-2.

3.1.3 Jaccard Distance of SNP profiles

In this work, we use the Jaccard distance to measure the similarity between SNP profiles and compare the difference between the SNP variant profiles of SARS-CoV-2 genomes.

The Jaccard similarity coefficient is defined as the intersection size divided by the union of two sets A and B [87]:

$$J(A,B) = \frac{|A \cap B|}{|A \cup B|} = \frac{|A \cap B|}{|A| + |B| - |A \cap B|}. \tag{3.2}$$

The Jaccard distance of two sets A and B is scored as the difference between one and the Jaccard similarity coefficient and is a metric on the collection of all finite sets:

$$d_J(A,B) = 1 - J(A,B) = \frac{|A \cup B| - |A \cap B|}{|A \cup B|}. \tag{3.3}$$

Therefore, the genetic distance of two genomes corresponds to the Jaccard distance of their SNP profiles.

In principle, the Jaccard distance of SNP profiles takes account of the ordering of SNP positions, i.e., transmission trajectory, when an appropriate reference sample is selected. However, one may fail to identify the infection pathways from the mutual Jaccard distances of multiple samples. In this case, the dates of the sample collection provide key information. Additionally, clustering techniques, such as k-means, UMAP, and t-distributed stochastic neighbor embedding (t-SNE), enable us to characterize the spread of COVID-19 onto the communities.

3.1.4 k-nearest Neighbors

The k-nearest neighbors algorithm (k-NN) is a non-parametric technique proposed by Thomas Cover and P. E. Hart in 1967 [88]. k-NN can be used for solving both regression and classification problems [89], and it is sensitive to the local structure of the data. The flowchart of the k-NN algorithm can be found in Figure 3.1. The features of the training set is $\{\mathbf{x}_i\}_{i=1}^n$ with $\mathbf{x}_i \in \mathbb{R}^m$, k shows the number of the nearest neighbors, and $\mathbf{x} \in \mathbb{R}^m$ is a feature representation of the training set. Different distance metrics can be employed in the k-NN algorithm, such as Euclidean distance, Manhattan distance, Minkowski distance, Chebyshev distance, natural log distance, generalized exponential distance, gener-

alized Lorentzian distance, Canberra distance, quadratic distance, and Mahalanobis distance.

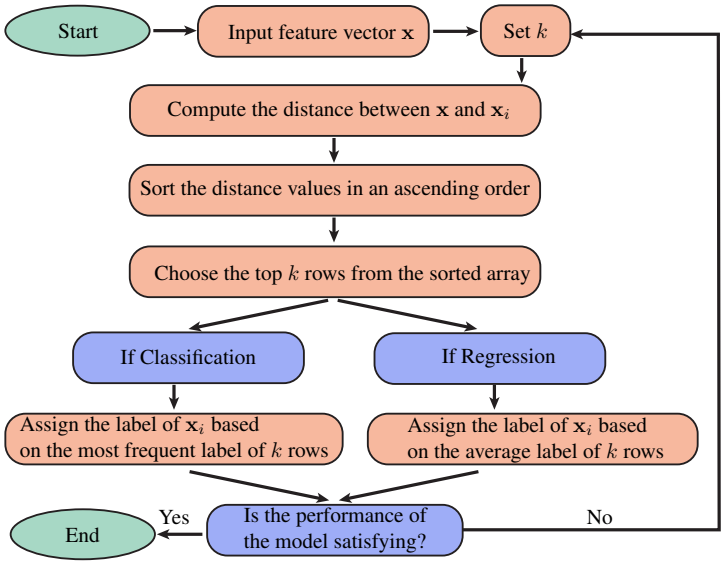

Figure 3.1: The flowchart of k-NN algorithm. The features of the training set is $\{\mathbf{x}_i\}_{i=1}^n$ with $\mathbf{x}_i \in \mathbb{R}^m$, k shows the number of the nearest neighbors, and $\mathbf{x} \in \mathbb{R}^m$ is a feature representation of the training set.

3.1.5 k-means Clustering

k-means clustering is an unsupervised learning algorithm, aiming to partition a set of observations into k subsets or clusters. It typically partitions a given dataset

$$X = \{x_1, x_2, \cdots, x_n, \cdots, x_N\}, x_n \in \mathbb{R}^d$$

into k different clusters $\{C_1, C_2, \cdots, C_k\}, k \leq N$ such that the specific clustering criteria are optimized. The standard procedure of k-means clustering method aims to obtain the optimal partition for a fixed number of clusters. First, we randomly pick k points as the cluster centers and then assign each data to its nearest cluster. Next, we calculate the within-cluster sum of squares (WCSS) defined below to update the cluster centers iteratively.

$$\sum_{i=1}^{k} \sum_{x_i \in C_k} \|x_i - \mu_k\|_2^2, \tag{3.4}$$

where μ_k is the mean value of the points located in the k-th cluster C_k. Here, $\|\cdot\|_2$ denotes the L_2 distance. It is noted that the k-mean clustering method described above aims to find the optimal partition for a fixed number of clusters. However, seeking the best number of clusters for the SNP profiles is essential as well. In this work, by varying the number of clusters k, a set of WCSS with its corresponding number of clusters can be plotted. The location of the elbow in this plot will be taken as the optimal number of clusters. Such a procedure is called the Elbow method which is frequently applied in the k-means clustering problem.

Specifically, in this work we apply the k-means clustering with the Elbow method for the analysis of the optimal number of the subtypes of SARS-CoV-2 SNP profiles. The pairwise Jaccard distances between different SNP profiles are considered as the input features for the k-means clustering method.

3.2 Mathematical-assisted Machine Learning Models in SARS-CoV-2

In this section, the workflow of the deep learning-based BFE change predictions of protein-protein interactions induced by mutations for the present SARS-CoV-2 variant analysis and prediction will be firstly introduced, which includes three steps as shown in Figure 3.2: (1) Data collection and pre-processing; (2) training data preparation; (3) feature generations of protein-protein interaction complexes; (4) predictive models of protein-protein interactions.

3.2.1 Data Collection and Pre-processing

The first step is to pre-process the original SARS-CoV-2 sequences data. In this step, a total of 1,983,328 complete SARS-CoV-2 genome sequences with high coverage and exact collection date are downloaded from the GISAID database [90] (https://www.gisaid.org/) as of August 05, 2021. Complete SARS-CoV-2 genome sequences are available from the GISAID database [90]. Next, the 1,983,328 complete SARS-CoV-2 genome se-

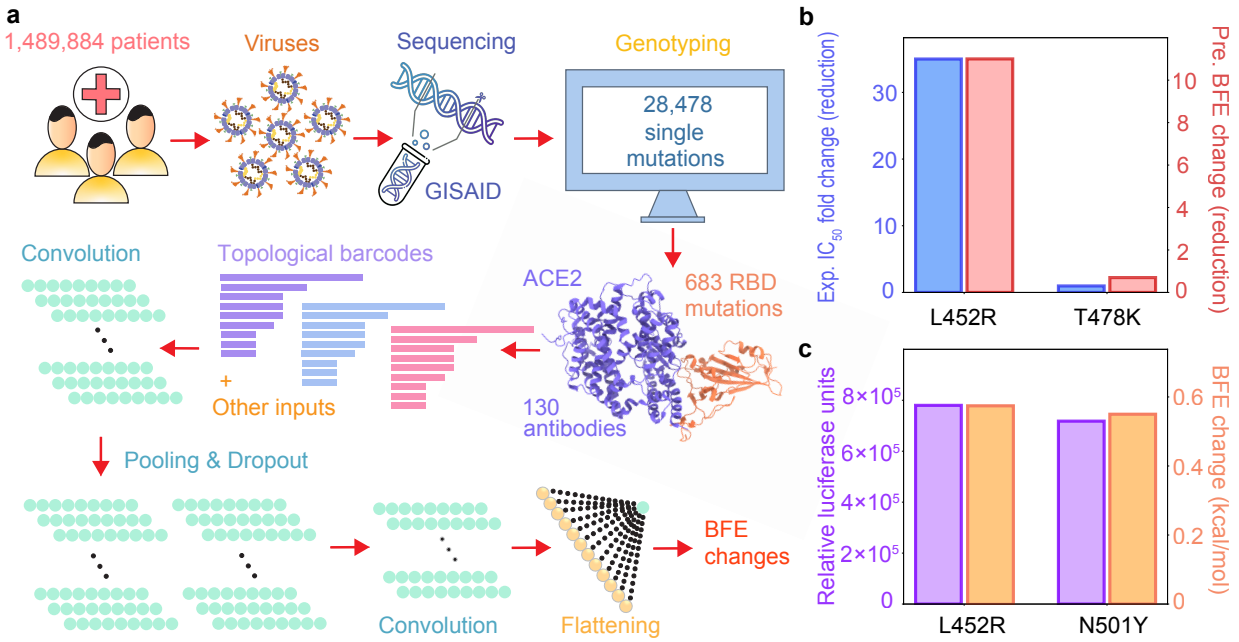

Figure 3.2: Illustration of genome sequence data pre-processing and BFE change predictions.

quences were rearranged according to the reference genome downloaded from the GenBank (NC_045512.2)[91], and multiple sequence alignment (MSA) is applied by using Cluster Omega with default parameters. Then, single nucleotide polymorphism (SNP) genotyping is applied to measure the genetic variations between different isolates of SARS-CoV-2 by analyzing the rearranged sequences [21, 92], which is of paramount importance for tracking the genotype changes during the pandemic. The SNP genotyping captures all of the differences between patients' sequences and the reference genome, which decodes a total of 28,865 unique single mutations from 1,983,328 complete SARS-CoV-2 genome sequences. Among them, 724 non-degenerate mutations on the S protein RBD (S protein residues from 329 to 530) are detected. In this work, the co-mutation analysis is more crucial than the unique single mutation analysis. Notably, the SARS-CoV-2 unique single mutations in the world are available at Mutation Tracker. The analysis of RBD mutations is available at Mutation Analyzer.

3.2.2 Preparation of Machine learning Datasets

Dataset is important to train accurate machine learning models. Both the BFE changes and enrichment ratios describe the effects on the binding affinity of protein-protein interactions. Therefore, integrating both kinds of datasets can improve the prediction accuracy. Especially, due to the urgency of COVID-19, the BFE changes of SARS-CoV-2 data are rarely reported, while the enrichment ratio data via high-throughput deep mutations are relatively easy to obtain. The most important dataset that provides the information for binding free energy changes upon mutations is the SKEMPI 2.0 dataset [93]. The SKEMPI 2.0 is an updated version of the SKEMPI database, which contains new mutations and data from other three databases: AB-Bind [94], PROXiMATE[95], and dbMPIKT [96]. There are 7,085 elements, including single- and multi-point mutations in SKEMPI 2.0. 4,169 variants in 319 different protein complexes are filtered as single-point mutations are used for our TopNetTree model training. Moreover, SARS-CoV-2 related datasets are also included to improve the prediction accuracy after a label transformation. They are all deep mutation enrichment ratio data, mutational scanning data of ACE2 binding to the receptor-binding domain (RBD) of the S protein [97], mutational scanning data of RBD binding to ACE2 [98, 3], and mutational scanning data of RBD binding to CTC-445.2 and of CTC-445.2 binding to the RBD [3]. Note that our training datasets used in the validation do not include the test dataset, which is a mutational scanning data of RBD binding to ACE2.

3.2.3 Features Generalization

Once the data pre-processing and SNP genotyping are carried out, we will firstly proceed with the training data preparation process, which plays a key role in reliability and accuracy. A library of 130 antibodies and RBD complexes, as well as an ACE2-RBD complex, are obtained from Protein Data Bank (PDB). RBD mutation-induced BFE changes of these complexes are evaluated by the following machine learning model. According to

the emergency and the rapid change of RNA virus, it is rare to have massive experimental BFE change data of SARS-CoV-2, while, on the other hand, next-generation sequencing data is relatively easy to collect. In the training process, the dataset of BFE changes induced by mutations of the SKEMPI 2.0 dataset [93] is used as the basic training set, while next-generation sequencing datasets are added as assistant training sets. The SKEMPI 2.0 contains 7,085 single- and multi-point mutations and 4,169 elements of that in 319 different protein complexes used for the machine learning model training. The mutational scanning data consists of experimental data of the binding of ACE2 and RBD induced mutations on ACE2[97] and RBD[98, 3], and the binding of CTC-445.2 and RBD with mutations on both protein[3].

Next, the feature generations of protein-protein interaction complexes are performed. The element-specific algebraic topological analysis on complex structures is implemented to generate topological bar codes [99, 100, 101, 4]. In addition, biochemistry and biophysics features such as Coulomb interactions, surface areas, electrostatics, et al., are combined with topological features [102].

3.2.3.1 Generation of Topological Features for PPIs

Algebraic topology [100, 101] has had tremendous success in describing biochemical and biophysical properties [4]. Element-specific and site-specific persistent homology can effectively simplify the structural complexity of protein-protein complex and extract the abstract properties of the vital biological information in PPIs [40, 41]. The algebraic topological analysis on PPIs is constructed based on a series of atom subsets of complex structures, which are atoms of the mutation sites, $\mathcal{A}_{\mathrm{m}}$, atoms in the neighborhood of the mutation site within a cut-off distance r, $\mathcal{A}_{\mathrm{mn}}(r)$, antibody atoms within r of the binding site, $\mathcal{A}_{\mathrm{Ab}}(r)$, antigen atoms within r of the binding site, $\mathcal{A}_{\mathrm{Ag}}(r)$, and atoms in the system that has atoms of element type of {C, N, O}, $\mathcal{A}_{\mathrm{ele}}(\mathrm{E})$. Additionally, a bipartition graph is introduced to describe the antibody and antigen in PPIs. Then, molecular atoms construct point clouds

for simplicial complex, which is a finite collection of sets of linear combinations of points. We apply the Vietoris-Rips (VR) complex for dimension 0 topology, and alpha complex for point cloud of dimensions 1 and 2 topology [4]. Overall, element-specific and site-specific persistent homology is devised to capture the multiscale topological information over different scales along a filtration [100] and is important for our machine learning predictions.

Simplex and simplicial complex Given a set of independent k+1 points $U = \{u_0, u_1, ..., u_k\}$ in $\mathbb{R}^N$, the convex combination is a point $u = \sum_{i=0}^{k} \alpha_i u_i$, where $\sum_i \alpha_i = 1$ and $\alpha_i \geq 0$. The convex hull of U is the collection of convex combinations of U, and a k-simplex σ is the convex hull of k+1 independent points U. For example, a 0-simplex is a point, a 1-simplex is an edge, a 2-simplex is a triangle, and a 3-simplex is a tetrahedron. A proper m-face of the k-simplex is a subset of the $k+1$ vertices of a k-simplex with $m+1$ vertices forms a convex hull in a lower dimension and $m < k$. The boundary of a k-simplex σ is defined as a sum of all its $(k-1)$–faces as

$$\partial_k \sigma = \sum_{i=1}^{k} (-1)^i \langle u_0, ..., \hat{u}_i, ..., u_k \rangle, \tag{3.5}$$

where $\langle u_0, ..., \hat{u}_i, ..., u_k \rangle$ is a convex hull formed by vertices of σ excluding u_i. A simplicial complex denotes by K is a collection of finitely many simplices forms a simplicial complex. Thus, faces of any simplex in K are also simplices in K, and intersections of any 2 simplices are only faces of both or an empty set. A k-simplex $\sigma = \langle u_{i_0}, ..., u_{i_k} \rangle$ is in Vietoris–Rips complex $R^r(U)$ if and only if $\mathbb{B}(u_{i_j}, r) \cap \mathbb{B}(u_{i_{j'}}, r) \neq \emptyset$ for $j, j' \in [0, k]$ and is in alpha complex $A^r(U)$ if and only if $\cap_{u_{i_j} \in \sigma} \mathbb{B}(u_{i_j}, r) \neq \emptyset$.

Homology For a simplicial complex K, a k-chain c_k of K is a formal sum of the k-simplices in K defined as $c_k = \sum \alpha_i \sigma_i$, where σ_i is the k-simplices and α_i is coefficients. α_i can be in different fields such as $\mathbb{R}$, $\mathbb{Q}$, and $\mathbb{Z}$. Typically, α_i is chosen to be $\mathbb{Z}_2$, which is $\{-1, 0, 1\}$ and forms an Abelian group $C_k(K, \mathbb{Z}_2)$. Then, the boundary operator can be

extended to a k-chain c_k as

$$\partial_k c_k = \sum \alpha_i \partial_k \sigma_i, \tag{3.6}$$

such that $\partial_k : C_k \to C_{k-1}$ and satisfies $\partial_{k-1}\partial_k = \emptyset$, follows from that boundaries are boundaryless. The chain complex is defined as a sequence of complexes by boundary maps is called a chain complex

$$\cdots \xrightarrow{\partial_{i+1}} C_i(K) \xrightarrow{\partial_i} C_{i-1}(K) \xrightarrow{\partial_{i-1}} \cdots \xrightarrow{\partial_2} C_1(K) \xrightarrow{\partial_1} C_0(K) \xrightarrow{\partial_0} 0. \tag{3.7}$$

The k-homology group is the quotient group defined by taking k-cycle group module of k-boundary group as

$$H_k = Z_k / B_k, \tag{3.8}$$

where H_k is the k-homology group, and k-cycle group Z_k and the k-boundary group B_k are the subgroups of C_k defined as,

$$\begin{aligned} Z_k &= \ker \partial_k = \{c \in C_k \quad | \quad \partial_k c = \emptyset\}, \\ B_k &= \operatorname{im} \partial_{k+1} = \{\partial_{k+1} c \quad | \quad c \in C_{k+1}\} \end{aligned} \tag{3.9}$$

The Betti numbers are defined by the ranks of kth homology group H_k as $\beta_k = \text{rank}(H_k)$. β_0 reflects the number of connected components, β_1 reflects the number of loops, and β_2 reflects the number of cavities.

Filtration and Persistent Homology A filtration of a topology space K is a nested sequence of K such that

$$\emptyset = K_0 \subseteq K_1 \subseteq \cdots \subseteq K_m = K. \tag{3.10}$$

Then, a sequence of chain complexes and a homology sequence are constructed on the filtration. The pth persistent of kth homology group of K_t are defined as

$$H_k^{t,p} = Z_k^t / (B_k^{t+p} \bigcap Z_k^t), \tag{3.11}$$

and the Betti numbers $\beta_k^{t,p} = \text{rank}(H_k^{t,p})$. These persistent Betti numbers are applied to represent topological fingerprints.

3.2.3.2 Generation of Residue-level Features for PPIs

Mutation site neighborhood amino acid composition Neighbor residues are the residues within 10 Å of the mutation site. Distances between residues are calculated based on residue C_α atoms. Six categories of amino acid residues are counted, which are hydrophobic, polar, positively charged, negatively charged, special cases, and pharmacophore changes. The count and percentage of the 6 amino acid groups in the neighbor site are regrading as the environment composition features of the mutation site. The sum, average, and variance of residue volumes, surface areas, weights, and hydropathy scores are used but only the sum of charges is included.

pKa shifts The pKa values are calculated by the PROPKA software [103], namely the values of 7 ionizable amino acids, namely, ASP, GLU, ARG, LYS, HIS, CYS, and TYR. The maximum, minimum, sum, the sum of absolute values, and the minimum of the absolute value of total pKa shifts are calculated. We also consider the difference of pKa values between a wild type and its mutant. Additionally, the sum and the sum of the absolute value of pKa shifts based on ionizable amino acid groups are included.

Position-specific scoring matrix (PSSM) Features are computed from the conservation scores in the position-specific scoring matrix of the mutation site for the wild type and the mutant as well as their difference. The conservation scores are generated by PSI-BLAST [104].

Secondary structure The SPIDER2 software is used to compute the probability scores for residue torsion angle and residues being in a coil, alpha helix, and beta strand based on the sequences for the wild type and the mutant [105].

3.2.3.3 Generation of Atom-level Features for PPIs

Seven groups of atom types, including C, N, O, S, H, all heavy atoms, and all atoms, are considered when generating the element-type features. Meanwhile, other three atom types, i.e., mutation site atoms, all heavy atoms, and all atoms, are used when generating the general atom-level features.

Surface areas Atom-level solvent excluded surface areas are computed by ESES [106].

Partial changes Partial change of each atom is generated by pdb2pqr software [107] using the Amber force field [108] for wild type and CHARMM force field [109] for mutant. The sum of the partial charges and the sum of absolute values of partial charges for each atomic group are collected.

Atomic pairwise interaction interactions Coulomb energy of the ith single atom is calculated as the sum of pairwise coulomb energy with every other atom as

$$C_i = \sum_{j,j\neq i} k_e \frac{q_i q_j}{r_{ij}}, \tag{3.12}$$

where k_e is the Coulomb's constant, r_{ij} is the distance of ith atom to jth atom, and q_i is the charge of ith atom. The van der Waals energy of the ith atom is modeled as the sum of pairwise Lennard-Jones potentials with other atoms as

$$V_i = \sum_{j,j\neq i} \epsilon \Big[\big(\frac{r_i + r_j}{r_{ij}}\big)^{12} - 2\big(\frac{r_i + r_j}{r_{ij}}\big)^{6} \Big], \tag{3.13}$$

where ϵ is the depth of the potential well, and r_i is van der Waals radii.

In atomic pairwise interaction, 5 groups (C, N, O, S, and all heavy atoms) are counted both for Coulomb interaction energy and van der Waals interaction energy.

Electrostatic solvation free energy Electrostatic solvation free energy of each atom is calculated using the Poisson-Boltzmann equation via MIBPB [110] and are summed up by atom groups.

3.2.4 Models for the Binding Free Energy Change Prediction of Protein-protein Interaction on SARS-CoV-2

3.2.4.1 TopNet Model

In this section, we illustrate the construction of a topology-based network (TopNet) model for the BFE change prediction of protein-protein interactions (PPIs) on SARS-CoV-2 studies. These approaches have been widely applied in studying protein-ligand and protein-protein binding free energy predictions [41, 102]. Firstly, one ensemble method, gradient boosting decision tree (GBDT), is studied as baselines in comparison to deep neural network methods. The ensemble methods naturally handle correlation between descriptors and are robust to redundant features. Therefore, they usually do not depend on a sophisticated feature selection procedure and a complicated grid search of hyper-parameters. The implemented GBDT is a function from the scikit-learn package (version 0.22.2.post1)[111]. The number of estimators and the learning is optimized for ensemble methods as 20000 and 0.01, respectively. For each set, 10 runs (with different random seeds) were done and the average result is reported in this work. Considering a large number of features, the maximum number of features to consider is set to the square root of the given descriptor length for GBDT methods to accelerate the training process. The parameter setting shows that the performance of the average of sufficient runs is decent.

A neural network is a network of neurons that maps an input feature layer to an output layer. The neural network simulates a biological brain solves problems with numerous neuron units by backpropagation to update weights on each layer. To reveal the facts of input features at different levels and abstract more properties, one can construct more layers and more neurons in each layer, which is known as a deep neural network. Optimization methods for feedforward neural networks and dropout methods are applied to prevent overfitting. In 10-fold cross validations, the neural network model has a slightly better performance than the GBDT model, where Pearson correlations for these algorithms are 0.864 and 0.838 and root mean square errors are 1.019 kcal/mol and 1.063

kcal/mol, respectively. Thus, we applied the deep neural network for predictions, validation, and comparison.

Deep learning algorithms A deep neural network is a neural network methods with multi-layers (hidden layer) of neurons between the input and output layers. In each layer, the single neuron gets fully connecting with the neurons in next layer. It should be preserve the consistency of all labels when applying the model for mutation-induced BFE change predictions. The loss function is constructed as following:

$$\underset{W,b}{\operatorname{argmin}} L(W,b) = \underset{W,b}{\operatorname{argmin}} \frac{1}{2} \sum_{i=1}^{N} \big(y_i - f(x_i; \{W,b\})\big)^2 + \lambda \|W\|^2 \tag{3.14}$$

where N is the number of samples, f is a function of the feature vector x_i parameterized by a weight vector W and bias term b, and λ represents a penalty constant.

Optimization The backpropagation is applied to evaluated the loss function start from the output layer and propagates backward through the network structure to update the weight vector W and bias term b. According to that the gradient calculation is required, we apply the stochastic gradient descent method with momentum which only evaluates a small part of training data and can be considered as calculating exponentially weighted averages, which is given as

$$\begin{aligned} V_i &= \beta V_{i-1} + \eta \nabla_{W_i} L(W_i, b_i) \\ W_{i+1} &= W_i - V_i, \end{aligned} \tag{3.15}$$

where W_i is the parameters in the network, $L(W_i, b_i)$ is the objective function, η is the learning rate, X and y are the input and target of the training set, and $\beta \in [0, 1]$ is a scalar coefficient for the momentum term. The momentum term involved accelerates the converging speed.

Dropout Fully connected layers possess a large number of degrees of freedom. This can easily cause an over-fitting issue, while the dropout technique is an easy way of preventing network over-fitting.[112] In the training process, hidden units are randomly set

zero values to their connected neurons in the next layer. Suppose that a percentage of neurons at a certain layer is chosen to be dropped during training. The number of computed neurons of this layer is equal to the neuron number multiplied by a coefficient such as 1-p, where p is the dropout rate. Then, in the testing process, the output of these layers is computed by randomly dropouts the same rate of neurons, to approximate the network in each training step.

3.2.4.2 TopNetmAb Model

In this section, the TopNet model trained with additional experimental data was introduced to predict mAb binding free energy changes [99]. Such a model is called TopNetmAb model. Persistent homology is the main workhorse for TopNetmAb, but auxiliary features inherited from our earlier TopNetTree [40] are utilized. The detailed descriptions of dataset and machine learning model are found in the literature [41, 22, 99] and are available at TopNetmAb.

3.2.5 Other Models

As mentioned above, we constructed a TopNet model for the BFE change prediction of protein-protein interactions (PPIs) on SARS-CoV-2 studies. A topology-based GBT model (TopBGT) is also developed in the present work by replacing Net in the TopNet model with GBT. Both TopNet and TopGBT include a set of auxiliary features inherited from our earlier TopNetTree [40] and TopNetmAb [99] to enhance their performance.

Additionally, to evaluate the performance of persistent Laplacian (Lap) for PPIs, we construct persistent Laplacian-based GBT (LapGBT) and persistent Laplacian-based deep neural network (LapNet). Note that unlike TopNet and TopGBT, LapGBT and LapNet employ only persistent Laplacian features extracted from protein structures. Therefore, their performance depends purely on persistent Laplacian.

Moreover, TopLapGBT and TopLapNet are constructed by adding persistent Laplacian features to TopGBT and TopNet, respectively. Furthermore, the consensus of GBT and Net predictions are also used for validations, denoted as TopNetGBT and LapNetGBT, respectively. Finally, the consensus of TopLapNet and TopLapGBT is called TopLapNetGBT.

CHAPTER 4

APPLICATIONS IN TOPOLOGICAL LAPLACIANS

4.1 Persistent Laplacians

Graph theory, a branch of discrete mathematics, concerns the relationship between objects. These objects can be either simple vertices, i.e., nodes and/or points (zero simplexes), or high-dimensional simplexes. Here, the relationship refers to connectivity with possible orientations. Graph theory has many branches, such as geometric graph theory, algebraic graph theory, and topological graph theory. The study of graph theory draws on many other areas of mathematics, including algebraic topology, knot theory, algebra, geometry, group theory, combinatorics, etc. For example, algebraic graph theory can be investigated by using either linear algebra, group theory, or graph invariants. Among them, the use of learning algebra in graph study leads to spectral graph theory.

Precursors of the spectral theory have often had a geometric flavor. An interesting spectral geometry question asked by Mark Kac was "Can one hear the shape of a drum?" [10]. The Laplace-Beltrami operator on a closed Riemannian manifold has been intensively studied [54]. Additionally, eigenvalues and isoperimetric properties of graphs are the foundation of the explicit constructions of expander graphs [113]. Moreover, the study of random walks and rapidly mixing Markov chains utilized the discrete analog of the Cheeger inequality [114]. The interaction between spectral theory and differential geometry became one of the critical developments [115]. For example, the spectral theory of the Laplacian on a compact Riemannian manifold is a central object of de Rham-Hodge theory [54]. Note that the Hodge Laplacian spectrum contains the topological information of the underlying manifold. Specifically, the harmonic part of the Hodge Laplacian spectrum corresponds to topological cycles. Connections between topology and spectral graph theory also play a central role in understanding the connectivity properties

of graphs [116, 117, 118, 119]. Similarly, as the topological invariants revealing the connectivity of a topological space, the multiplicity of 0 eigenvalues of a 0-combinatorial Laplacian matrix is the number of connected components of a graph. Indeed, the number of q-dimensional holes can also be unveiled from the number of 0 eigenvalues of the q-combinatorial Laplacian [45, 53, 46, 120]. Nonetheless, spectral graph theory offers additional non-harmonic spectral information beyond topological invariants.

The traditional topology and homology are independent of metrics and coordinates and thus, retain little geometric information. This obstacle hinders their practical applicability in data analysis. Recently, persistent homology has been introduced to overcome this difficulty by creating low-dimensional multiscale representations of a given object of interest [121, 101, 122, 43, 123, 124]. Specifically, a filtration parameter is devised to induce a family of geometric shapes for a given initial data. Consequently, the study of the underlying topologies or homology groups of these geometric shapes leads to the so-called topological persistence. Like the de Rham-Hodge theory which bridges differential geometry and algebraic topology, persistent homology bridges multiscale analysis and algebraic topology. Topological persistence is the most important aspect of the popular topological data analysis (TDA) [125, 126, 127, 128] and has had tremendous success in computational biology [129, 44] and worldwide competitions in computer-aided drug design [6].

Graph theory has been applied in various fields [130]. For example, spectral graph theory is applied to the quantum calculation of π-delocalized systems. The Hückel method, or Hückel molecular orbital theory, describes the quantum molecular orbitals of π-electrons in π-delocalized systems in terms of a kind of adjacency matrix that contains atomic connectivity information [131, 132]. Additionally, the Gaussian network model (GNM) [133] and anisotropic network model (ANM) [134] represent protein C_α atoms as an elastic mass-and-spring network by graph Laplacians. These approaches were influenced by the Flory theory of elasticity and the Rouse model [135]. Like traditional topology, tra-

ditional graph theory extracts very limited information from data. In our earlier work, we have proposed multiscale graphs, called multiscale flexibility rigidity index (mFRI), to describe the multiscale nature of biomolecular interactions [136], such as hydrogen bonds, electrostatic effects, van der Waals interactions, hydrophilicity, and hydrophobicity. A multiscale spectral graph method has also been proposed as generalized GNM and generalized ANM [57]. Our essential idea is to create a family of graphs with different characteristic length scales for a given dataset. We have demonstrated that our multiscale weighted colored graph (MWCG) significantly outperforms traditional spectral graph methods in protein flexibility analysis [137]. More recently, we demonstrate that our MWCG outperforms other existing approaches in protein-ligand binding scoring, ranking, docking, and screening [138].

The objective of the present work is to introduce persistent spectral graph as a new paradigm for the multiscale analysis of the topological invariants and geometric shapes of high-dimensional datasets. Motivated by the success of persistent homology [44] and multiscale graphs [138] in dealing with complex biomolecular data, we construct a family of spectral graphs induced by a filtration parameter. In the present work, we consider the radius filtration via the Vietoris-Rips complex while other filtration methods can be implemented as well. As the filtration radius is increased, a family of persistent q-combinatorial Laplacians are constructed for a given point-cloud dataset. The diagonalization of these persistent q-combinatorial Laplacian matrices gives rise to persistent spectra. It is noted that our harmonic persistent spectra of 0-eigenvalues fully recover the persistent barcode or persistent diagram of persistent homology. Additional information is generated from non-harmonic persistent spectra, namely, the non-zero eigenvalues and associated eigenvectors. In a combination with a simple machine learning algorithm, this additional spectral information is found to provide a powerful new tool for the quantitative analysis of molecular data.

4.1.1 Benzene Structure Analysis

In the past few years, we have developed a multiscale spectral graph method such as generalized GNM and generalized ANM [136, 57], to create a family of spectral graphs with different characteristic length scales for a given dataset. Similarly, in our persistent spectral theory, we can construct a family of spectral graphs induced by a filtration parameter. Moreover, we can sum over all the multiscale spectral graphs as an accumulated spectral graph. Specifically, a family of $\mathcal{L}_0^{r+0}$ matrices, as well as the accumulated combinatorial Laplacian matrices, can be generated via the filtration. By analyzing the persistent spectra of these matrices, the topological invariants and geometric shapes can be revealed from the given input point-cloud data.

The spectra of $\mathcal{L}_0^{r+0}$, $\hat{\mathcal{L}}_0^{r+0}$, and $\check{\mathcal{L}}_0^{r+0}$ mentioned above carry similar information on how the topological structures of a graph are changed during the filtration. Benzene molecule (C_6H_6), a typical aromatic hydrocarbon which is composed of six carbon atoms bonded in a planar regular hexagon ring with one hydrogen joined with each carbon atom. It provides a good example to demonstrate the proposed PST. Figure 4.1 illustrates the filtration of the benzene molecule. Here, we label 6 hydrogen atoms by H_1, H_2, H_3, H_4, H_5, and H_6, and the carbon adjacent to the labeled hydrogen atoms are labeled by C_1, C_2, C_3, C_4, C_5, and C_6, respectively. Figure Figure 4.1 **b** depicts that when the radius of the solid sphere reaches 0.54 Å, each carbon atom in the benzene ring is overlapped with its joined hydrogen atom, resulting in the reduction of β_0^{r+0} to 6. Moreover, once the radius of solid spheres is larger than 0.70 Å, all the atoms in the benzene molecule will connect and constitute a single component which gives rise $\beta_0^{r+0} = 1$. Furthermore, we can deduce that the C-C bond length of the benzene ring is about 1.40 Å, and the C-H bond length is around 1.08 Å, which are the real bond lengths in benzene molecule. Figure Figure 4.1 **c** shows that a 1-dimensional hole (1-cycle) is born when the filtration parameter r increase to 0.70 Å and dead when $r = 1.21$ Å. In Figures Figure 4.1 **b** and Figure 4.1 **c**, it can be seen that variants of 0-persistent 0-combinatorial Laplacian and 1

-combinatorial Laplacian matrices based on filtration give us the identical β_0^{r+0} and β_1^{r+0} information respectively.

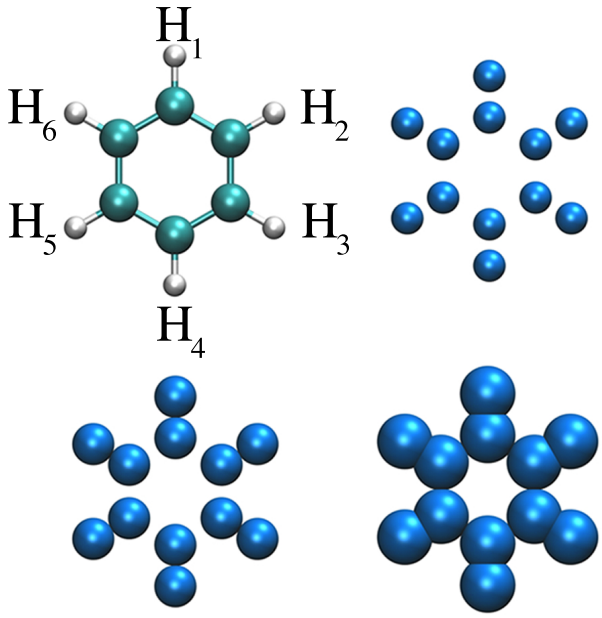

Figure 4.1: Benzene molecule and its topological changes during the filtration process.

The C-C bond length of benzene is 1.39 Å, and the C-H bond length is 1.09 Å. Due to the perfect hexagon structure of the benzene ring, we can calculate all of the distances between atoms. The shortest and longest distances between carbons and the hydrogen atoms are 1.09 Å and 3.87 Å. In Figure Figure 4.1**a**, a total of 10 changes of $(\tilde{\lambda}_2)_0^{r+0}$ values is observed at various radii. Table 4.1 lists all the distances between atoms and the values of radii when the changes of $(\tilde{\lambda}_2)_0^{r+0}$ occur. It can be seen that the distance between atoms approximately equals twice of the radius value when a jump of $(\tilde{\lambda}_2)_0^{r+0}$ occurs. Therefore, we can detect all the possible distances between atoms with the nonzero spectral information. Moreover, in Figure Figure 4.1 **b**, the values of the smallest nonzero eigenvalues of $\mathcal{L}_0^{r+0}$, $\hat{\mathcal{L}}_0^{r+0}$, and $\check{\mathcal{L}}_0^{r+0}$ change concurrently.

Table 4.1: Distances between atoms in the benzene molecule and the radii when the changes of $(\tilde{\lambda}_2)_0^{r+0}$ occur (Values increase from left to right).

Type	C_1-H_1	C_1-C_2	C_2-H_1	C_1-C_3	H_1-H_2	C_1-C_4	C_3-H_1	C_4-H_1	H_1-H_3	H_1-H_4
Distance (Å)	1.09	1.39	2.15	2.41	2.48	2.78	3.39	3.87	4.30	4.96
r (Å)	0.54	0.70	1.08	1.21	1.24	1.40	1.70	1.94	2.15	2.49

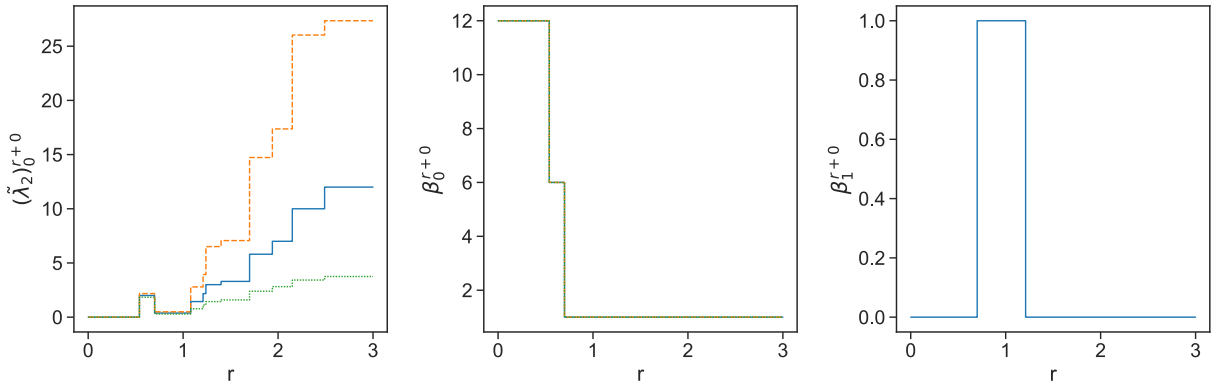

Figure 4.2: Persistent spectral analysis of the benzene molecule induced by filtration parameter r. Blue line, orange line, and green line represent $\mathcal{L}_0^{r+0}$, $\hat{\mathcal{L}}_0^{r+0}$, and $\check{\mathcal{L}}_0^{r+0}$ respectively. (a) Plot of the smallest non-zero eigenvalues with radius filtration under $\mathcal{L}_0^{r+0}$ (blue line), $\hat{\mathcal{L}}_0^{r+0}$ (red line), and $\check{\mathcal{L}}_0^{r+0}$ (green line). Total 10 jumps observed in this plot which represent 10 possible distances between atoms. (b) Plot of the number of zero eigenvalues (β_0^{r+0}) with radius filtration under $\mathcal{L}_0^{r+0}$, $\hat{\mathcal{L}}_0^{r+0}$, and $\check{\mathcal{L}}_0^{r+0}$ (three spectra are superimposed). When $r = 0.00$ Å, 12 atoms are disconnected with each other. After $r = 0.54$ Å, H atoms and their adjacent C atoms are connected with one another resulting in $\beta_0^{r+0} = 6$. With r keeps growing, all of the atoms are connected with one another and then $\beta_0^{r+0} = 1$. (c) Plot of the number of zero eigenvalues (β_1^{r+0}) with radius filtration under $\mathcal{L}_1^{r+0}$. When $r = 0.70$ Å, a 1-cycle created since all of the C atoms are connected and form a hexagon, resulting in $\beta_1^{r+0} = 1$. After the radius reached 1.21 Å, the hexagon disappears and $\beta_1^{r+0} = 0$.

4.1.2 Fullerene Analysis and Prediction

In 1985 Kroto et all discovered the first structure of C_{60} [139], which was confirmed by Kratschmer et al in 1990 [140]. Since then, the quantitative analysis of fullerene molecules has become an interesting research topic. The understanding of the fullerene structure-function relationship is important for nanoscience and nanotechnology. Fullerene molecules are only made of carbon atoms that have various topological shapes, such as the hollow spheres, ellipsoids, tubes, or rings. Due to the monotony of the atom type and the variety of geometric shapes, the minor heterogeneity of fullerene structures can be ignored.

The fullerene system offers a moderately large dataset with relatively simple structures. Therefore, it is suitable for validating new computational methods because every single change in the spectra is interpretable. The proposed persistent spectral theory, i.e., persistent spectral analysis, is applied to characterize fullerene structures and predict their stability.

All the structural data can be downloaded from CCL.NET Webpage. This dataset gives the coordinates of fullerene carbon atoms. In this section, we will analyze fullerene structures and predict the heat of formation energy.

4.1.2.1 Fullerene Structure Analysis

The smallest member of the fullerene family is C_{20} molecule with a dodecahedral cage structure. Note that 12 pentagons are required to form a closed fullerene structure. Following the Euler's formula, the number of vertices, edges, and faces on a polygon have the relationship $V - E + F = 2$. Therefore, the 20 carbon atoms in the dodecahedral cage form 30 bonds with the same bond length. The C_{20} is the only fullerene smaller than C_{60} that has the molecular symmetry of the full icosahedral point group I_h. C_{60} is a molecule that consists of 60 carbon atoms arranged as 12 pentagon rings and 20 hexagon rings. Unlike C_{20}, C_{60} has two types of bonds: $6:6$ bonds and $6:5$ bonds. The $6:6$ bonds are shorter than $6:5$ bonds, which can also be considered as "double bond" [141]. C_{60} is the most well-know fullerene with geometric symmetry I_h. Since C_{20} and C_{60} are highly symmetrical, they are ideal systems for illustrating the persistent spectral analysis.

Figure 4.3 (a) illustrates the radius filtration process built on C_{20}. As the radius increases, the solid balls corresponding to carbon atoms grow, and a sequence of $\mathcal{L}_0^{r+0}$ matrices can be defined through the overlap relations among the set of balls. At the initial state ($r = 0.00$ Å), all of the atoms are isolated from one another. Therefore, $\mathcal{L}_0^{r+0}$ is a zero matrix with dimension 20×20. Since the C_{20} molecule has the same bond length which can be denoted as $l(C_{20})$, once the radius of solid balls is greater than $l(C_{20})$, all of the

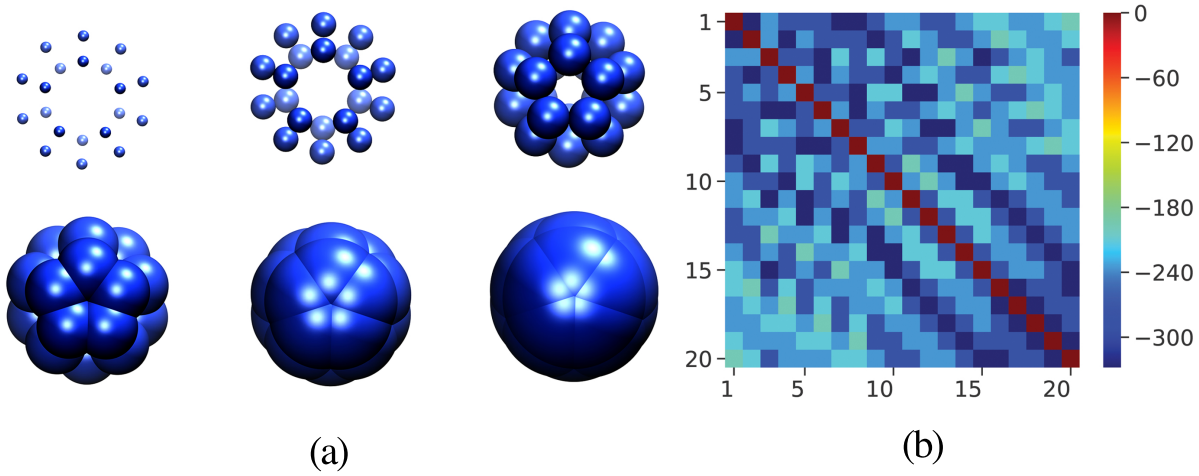

Figure 4.3: (a) Illustration of filtration built on fullerene C_{20}. Each carbon atom of C_{20} is plotted by its given coordinates, which are associated with an ever-increasing radius r. The solid balls centered at given coordinates keep growing along with the radius filtration parameter. (b) The accumulated $\mathcal{L}_0^{r+0}$ matrix for C_{20}. For clarity, the diagonal terms are set to 0.

balls are overlapped, which makes the system a singly connected component. Figure 4.3 (b) depicts the accumulated $\mathcal{L}_0^{r+0}$ for C_{20}. For C_{60}, the accumulated $\mathcal{L}_0^{r+0}$ is described in Figure 4.4 (a). Figure 4.4 (b)-(f) are the plots of $\mathcal{L}_0^{r+0}$ under different filtration r values. The blue cell located at the ith row and jth column means the balls centered at atom i and atom j connected with each other, i.e., a 1-simplex formed with its vertex to be i and j. When the radius filtration increases, more and more bluer cells are created. In Figure 4.4 (f), the color of cells, except the cells located in the diagonal, turns to blue, which means all of the carbon atoms are connected with one another at $r = 3.6$ Å. For clarity, we set the diagonal terms to 0.

In Figure 4.5, the blue solid line represents C_{20} properties and the dash orange line represents C_{60} properties. For Figure Figure 4.5 **a**, the blue line drops at $r = 0.72$ Å, which means the bond length of C_{20} is around 1.44 Å. The orange line drops at $r = 0.68$ Å and 0.72 Å, which means the "double bond" length of C_{60} is around 1.36 Å and the

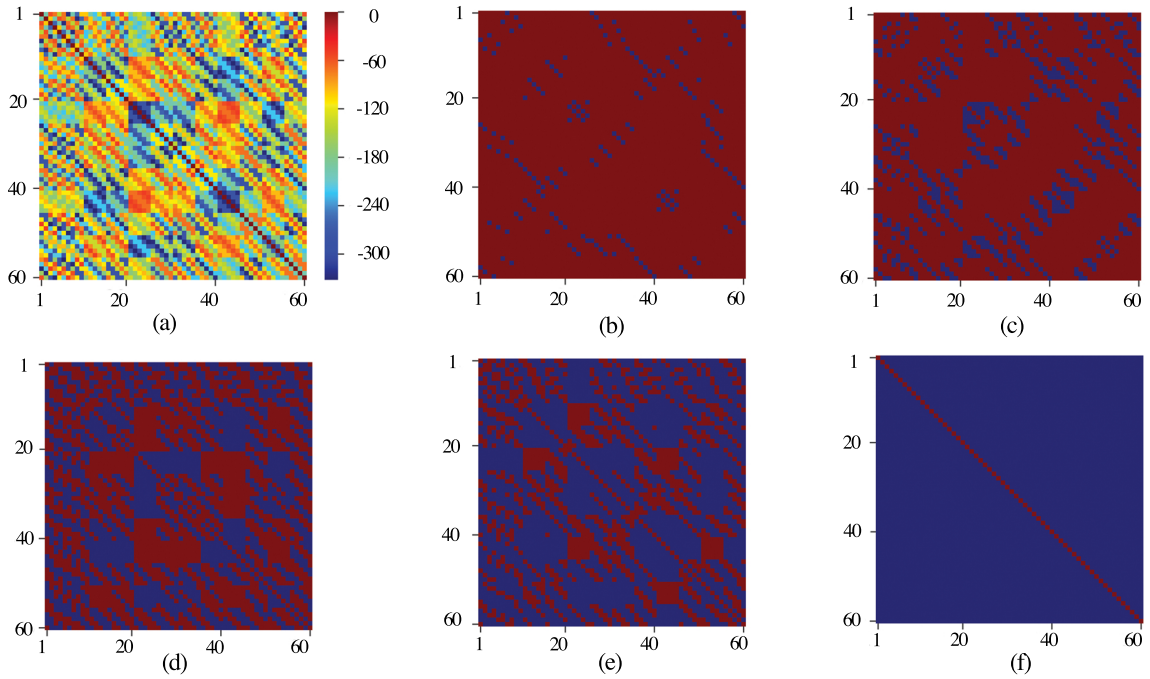

Figure 4.4: Illustration of persistent multiscale analysis of C_{60} in terms of 0-combinatorial Laplacian matrices (b)-(f) and their accumulated matrix (a) induced by filtration. As the value of filtration parameter r increases, high-dimensional simplicial complex forms and grows accordingly. (b), (c), (d), (e), and (d) demonstrate the 0-combinatorial Laplacian matrices (i.e., the connectivity among C_{60} atoms) at filtration $r = 1.0\,\text{Å}, 1.5\,\text{Å}, 2.5\,\text{Å}, 3.0\,\text{Å}$, and $3.6\,\text{Å}$, respectively. The blue cell located at the ith row and jth column represents the balls centered at atom i and atom j connected with each other. For clarity, the diagonal terms are set to 0 in all plots.

$6:5$ bond length is around $1.44\,\text{Å}$. Moreover, the total number of "double bond" is 30, yielding $\beta_0^{r+0} = 30$ when the radius of solid balls is over $0.68\,\text{Å}$. In conclusion, one can deduce the number of different types of bonds as well as the bond length information from the number of zero eigenvalues (i.e., β_0^{r+0}) under the radius filtration. Furthermore, the geometric information can also be derived from the plot of $(\tilde{\lambda}_2)_0^{r+0}$. Each jump in Figure Figure 4.5 **d** at a specific radius represents the change of geometric and topological structure. The smallest non-zero eigenvalue $(\tilde{\lambda}_2)_0^{r+0}$ of $\mathcal{L}_0^{r+0}$ matrices for C_{20} changes 5 times in Figure Figure 4.5 **d**, which means C_{20} has 5 different distances between carbon atoms. Furthermore, as $(\tilde{\lambda}_2)_0^{r+0}$ of C_{20} keeps increasing, the smallest vertex connectivity of the connected subgraph continues growing and the topological structure becomes steady.

As can be seen in the right-corner chart of Figure 4.3, the carbon atoms will finally grow to a solid object with a steady topological structure.

Figure Figure 4.5 **b** depicts the changes of Betti 1 value β_1^{r+0} (i.e., the number of zero eigenvalues for $\mathcal{L}_1^{r+0}$) under the filtration r. Since C_{20} has 12 pentagonal rings, β_1^{r+0} jumps to 11 when radius r equals to the half of the bond length of $l(C_{20})$. These eleven 1-cycles disappear at $r = 1.17$ Å. There are 12 pentagons and 20 hexagons in C_{60}, which results in $\beta_1^{r+0} = 12$ at $r = 0.72$ Å, $\beta_1^{r+0} = 31$ at $r = 1.17$ Å. All of the pentagons and hexagons disappear at $r = 1.22$ Å.

As the filtration process, even more structure information can be derived from the number of zero eigenvalues of $\mathcal{L}_2^{r+0}$ (i.e., β_2^{r+0}) in Figure Figure 4.5 **c**. For C_{20}, $\beta_2^{r+0} = 1$ when $r = 1.17$ Å, which corresponds to the void structure in the center of the dodecahedral cage. The void disappears at $r = 1.65$ Å since a solid structure is generated at this point. For fullerene C_{60}, 20 hexagonal cavities and a center void exist from 1.12 Å to 1.40 Å yielding $\beta_2^{r+0} = 21$. As the filtration goes, hexagonal cavities disappear which results β_2^{r+0} decrease to 1. The central void keeps alive until a solid block is formed at $r = 3.03$ Å. In a nutshell, we can deduce the number of different types of bonds, the bond length, and the topological invariants from the present persistent spectral analysis.

4.1.2.2 Fullerene stability prediction

Having shown that the detailed fullerene structural information can be extracted into the spectra of $\mathcal{L}_q^{r+0}$, we further illustrate that fullerene functions can be predicted from their structures by using our persistent spectral theory in this section. Similar structure-function analysis has been carried out by using other methods [136, 142, 143]. For small fullerene molecule series C_{20} to C_{60}, with the increase in the number of atoms, the ground-state heat of formation energies decrease [144, 1]. The left chart in Figure 4.6 describes this phenomenon. Similar patterns can also be found in the total energy (STO-3G/SCF at MM3) per atom and the average binding energy of C_{2n}. To analyze these patterns,

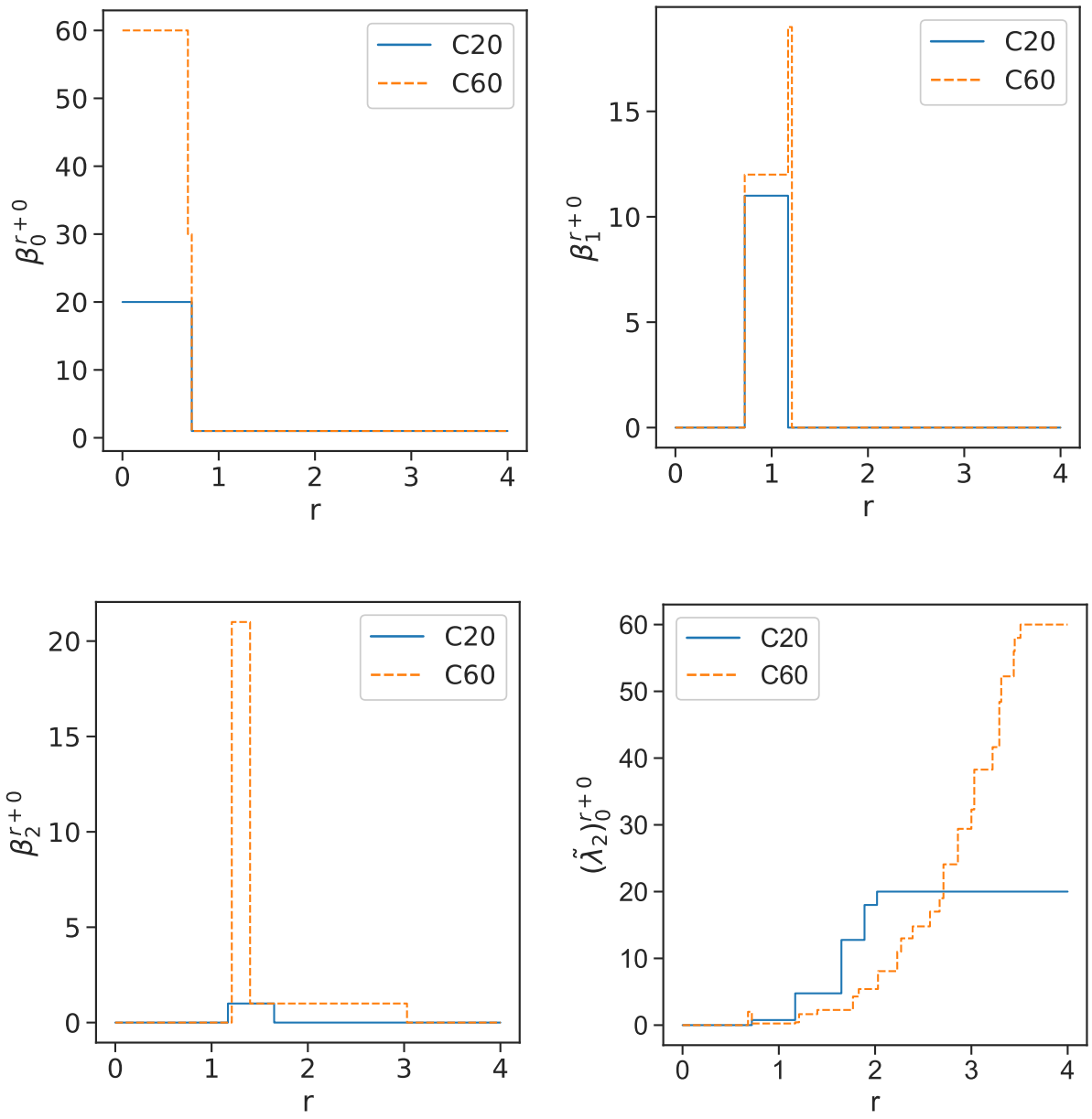

Figure 4.5: Illustration of persistent spectral analysis of C_{20} and C_{60} using the spectra of $\mathcal{L}_q^{r+0}$ ($q = 1, 2$ and 3). (a) The number of zero eigenvalues of $\mathcal{L}_0^{r+0}$, i.e., β_0^{r+0}, under radius filtration. (b) The number of zero eigenvalues of $\mathcal{L}_1^{r+0}$, i.e., β_1^{r+0} under radius filtration. (c) The number of zero eigenvalues of $\mathcal{L}_2^{r+0}$, i.e., β_2^{r+0} under radius filtration. (d) The smallest non-zero eigenvalue $(\tilde{\lambda}_2)_0^{r+0}$ under radius filtration. The radius grid spacing is 0.01 Å.

many theories have been proposed. Isolated pentagon rule assumes that the most stable fullerene molecules are those in which all the pentagons are isolated. Zhang et al. [1] stated that fullerene stability is related to the ratio between the number of pentagons and

the number of carbon atoms. Xia and Wei [142] proposed that the stability of fullerene depends on the average number of hexagons per atom. However, these theories all focused on the pentagon and hexagon information. More specifically, they use topological information to reveal the stability of fullerene. In contrast, we believe that the non-harmonic persistent spectra can also model the structure-function relationship of fullerenes. We hypothesize that the non-harmonic persistent spectra of $\mathcal{L}_0^{r+0}$ matrices are powerful enough to model the stability of fullerene molecules. To verify our hypothesis, we compute the summation, mean, maximal, standard deviation, variance of its eigenvalues, and $(\tilde{\lambda}_2)_0^{r+0}$ of the persistent spectra of $\mathcal{L}_0^{r+0}$ over various filtration radii r. We depict a plot with the horizontal axis represents radius r and the vertical axis represents the particular spectrum value, which is actually the same as Figure 4.5. Then we define the area under the plot of spectra with a negative sign as

$$A_\alpha = -\sum_{i=1} \Lambda_i^\alpha \delta r, \tag{4.1}$$

where δr is the radius grid spacing, in Figure 4.5, $\delta r = 0.01$ Å. Here, $\alpha =$ Sum, Avg, Max, Std, Var, Sec is the type index and thus, Λ_i^α represent the summation, mean, maximal, standard deviation, variance, and the smallest non-zero eigenvalue $(\tilde{\lambda}_2)_0^{r+0}$ of $\mathcal{L}_0^{r+0}$ at i-th radius step, respectively. The right chart in Figure 4.6 describes the area under the plot of spectra and closely resembles that of the heat of formation energy. We can see that generally the left chart and the middle chart show the same pattern. The integration of $(\tilde{\lambda}_2)_0^{r+0}$ decreases as the number of carbon atoms increases. However, the structural data we used might not be the same ground-state data as in Ref. [1], which results in C_{36} do not match the corresponding energy perfectly. Limited by the availability of the ground-state structural data, we are not able to analyze the full set of the fullerene family.

To quantitatively validate our model, we apply one of the simplest machine learning algorithms, linear least-squares method, to predict the heat of formation energy. The

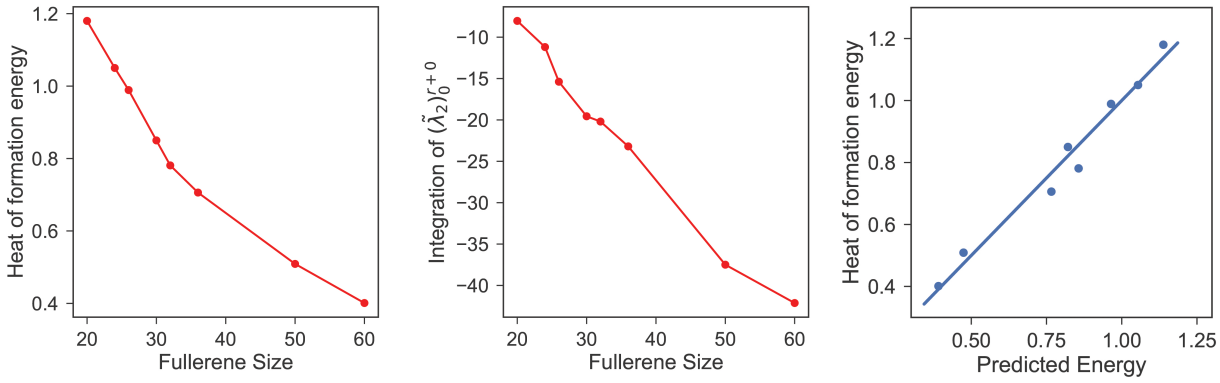

Figure 4.6: Persistent spectral analysis and prediction of fullerene heat formation energies. Left chart: the heat of formation energies of fullerenes obtained from quantum calculations [1]. Middle chart: PST model using the area under the plot of $(\tilde{\lambda}_2)_0^{r+0}$. Right chart: Correlation between the quantum calculation and the PST prediction. The highest correlation coefficient form the least-squares fitting is 0.986 with the type index of $\alpha = \text{Max}$.

Pearson correlation coefficient is defined as

$$C_c^{\alpha} = \frac{\sum_{i=1}^{N}(A_\alpha^i - \bar{A}_\alpha)(E_i - \bar{E})}{\left[\sum_{i=1}^{N}(A_\alpha^i - \bar{A}_\alpha)^2 \sum_{i=1}^{N}(E_i - \bar{E})^2\right]^{\frac{1}{2}}} \tag{4.2}$$

where A_α^i represents the theoretically predicted energy of the i-th fullerene molecule, E_i represents the heat of formation energy of the i-th fullerene molecule, and $\bar{A}_\alpha$ and $\bar{E}$ are the corresponding mean values. When $\alpha = \text{Max}$, the Pearson correlation coefficient is around 0.986. The right chart of Figure 4.6 plots the correlation between predicted energies and the heat of formation energy of the fullerene molecules computed from quantum mechanics [1]. These results agree very well.

Table 4.2: The heat of formation energy of fullerenes [1] and its corresponding predicted energies with $\alpha = \text{Max}$. The unit is EV/atom.

Fullerene type	C_{20}	C_{24}	C_{26}	C_{30}	C_{32}	C_{36}	C_{50}	C_{60}
Heat of formation energy	1.180	1.050	0.989	0.850	0.781	0.706	0.509	0.401
Predicted energy	1.138	1.050	0.964	0.821	0.857	0.766	0.474	0.391

The right chart of Figure 4.6 illustrates the fitting results under different type index α. Table 4.3 lists the correlation coefficient under different type index α. The highest corre-

lation coefficient is close to unity (0.986) obtained with $\alpha = \text{Max}$. The lowest correlation coefficient is 0.942 with $\alpha = \text{Sum}$. We can see that all the correlation coefficients are close to unity, which verifies our hypothesis that the non-harmonic spectra of $\mathcal{L}_0^{r+0}$ have the capacity of modeling the stability of fullerene molecules. Although we ignore the topological information (Betti numbers), our persistent spectral theory still works extremely well only with non-harmonic spectra, which means our persistent spectral theory is a powerful tool for quantitative data analysis and prediction.

Table 4.3: The correlation coefficients under different type index α.

Type index	Sum	Avg	Max	Std	Var	Sec
Correlation coefficient	0.942	0.985	0.986	0.969	0.977	0.981

4.1.3 Protein flexibility analysis

As clarified earlier, the number of zero eigenvalues of p-persistent q-Laplacian matrix (p-persistent qth Betti number) can also be derived from persistent homology. Persistent homology has been used to model fullerene stability [142]. In this section, we further illustrate the applicability of present persistent spectral theory by a case that non-harmonic persistent spectra offer a unique theoretical model whereas it may be difficult to come up with a suitable persistent homology model for this problem.

The protein flexibility is known to correlate with a wide variety of protein functions. It can be modeled by the beta factors or B-factors, which are also called Debye-Waller factors. B-factors are a measure of the atomic mean-square displacement or uncertainty in the X-ray scattering structure determination. Therefore, understanding the protein structure, flexibility, and function via the accurate protein B-factor prediction is a vital task in computational biophysics [145]. Over the past few years, quite many methods are developed to predict protein B-factors, such as GNM, [133], ANM [134], FRI, [146, 147] and MWCG [57, 145]. However, all of the aforementioned methods are based on a particular matrix derived from the graph network which is constructed using alpha

carbon as nodes and connections between nodes as edges. In this section, we apply our persistent spectral theory to create richer geometric information in B-factor prediction.

To illustrate our method, we consider protein 2Y7L whose total number of residues is $N = 319$. In this work, we employ the coarse-grained C_α representation of 2Y7L. Therefore, 319 particles are taken into consideration in protein 2Y7L. Similarly, like in the previous application of fullerene structure analysis, we treat each C_α atom as a 0-simplex at the initial setup and assign it a solid ball with a radius of r. By varying the filtration parameter r, we can obtain a family of $\mathcal{L}_0^{r+0}$. For each matrix $\mathcal{L}_0^{r+0}$, its corresponding ordered spectrum is given by

$$(\lambda_1)_0^{r+0}, (\lambda_2)_0^{r+0}, \cdots, (\lambda_N)_0^{r+0}.$$

Suppose the number of zero eigenvalues is m, then, we have $\beta_0^{r+0} = m$. Since $\mathcal{L}_0^{r+0}$ is symmetric, then eigenvectors of $\mathcal{L}_0^{r+0}$ corresponding to different eigenvalues must be orthogonal to each other. The Moore-Penrose inverse of $\mathcal{L}_0^{r+0}$ can be calculated by the non-harmonic spectra of $\mathcal{L}_0^{r+0}$:

$$(\mathcal{L}_0^{r+0})^{-1} = \sum_{k=m+1}^{N} \frac{1}{(\lambda_k)_0^{r+0}} [(u_k)_0^{r+0}((u_k)_0^{r+0})^T],$$

where T is the transpose and $(u_k)_0^{r+0}$ is the kth eigenvector of $\mathcal{L}_0^{r+0}$. The modeling of ith B-factor of 2Y7L at filtration parameter r can be expressed as

$$B_i^r = (\mathcal{L}_0^{r+0})_{ii}^{-1}, \forall i = 1, 2, \cdots, N,$$

and the final model of ith B-factor of 2Y7L is given by

$$B_i^{\text{PST}} = \sum_r w_r B_i^r + w_0, \forall i = 1, 2, \cdots, N,$$

where w_r and w_0 are fitting parameters which can be derived by linearly fitting B-factors from experimental data B^{Exp}. Consider the filtration radius from 2 to 12 with the grid spacing of 1, then totally 11 different $\mathcal{L}_0^{r+0}$ are created. By calculating all the non-harmonic spectra together with their eigenvectors, 11 Moore-Penrose inverse matrices $(\mathcal{L}_0^{r+0})^{-1}$ can

be constructed. Therefore, the predicted ith B-factor is

$$B_i^{\text{PST}} = \sum_{r=2}^{12} w_r B_i^r + w_0.$$

The specific values of w_r and w_0 can be found in Table A.16 and Table A.17 of Appendix Section A.2. Figure 4.7 (c) shows that the prediction B-factors are in an excellent agreement with the experimental B-factors of protein 2Y7L. The Pearson correlation coefficient is 0.925 [1].

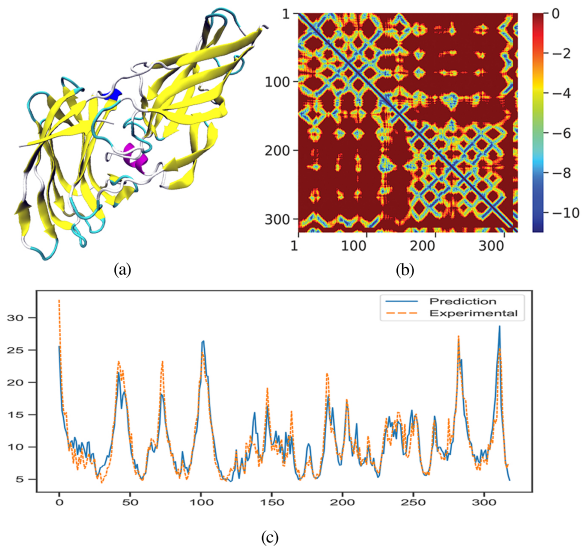

Figure 4.7: Illustration of persistent spectral prediction of protein B-factors. (a) Plot of the secondary structure of protein 2Y7L. (b) Accumulated persistent Laplacian matrix (For clarity, the diagonal terms are set to 0.). Note that the accumulated persistent Laplacian matrix maps out the detailed distance between each pair of residues. (c) Comparison of experimental B-factors and those predicted by PST for protein 2Y7L.

This example shows that our persistent spectral theory can be used beyond the persistent homology analysis. The number of zero eigenvalues of 0-persistent q-combinatorial Laplacian matrices fully recover the persistent barcode or persistent diagram of persistent homology. Additional spectral information from non-harmonic persistent spectra and persistent eigenvectors provides valuable information for data modeling, analysis, and prediction.

[1] We carry out feature scaling to make sure all B_i^r are on a similar scale.

4.1.4 Discussion and Conclusion

Spectral graph theory is a powerful tool for data analysis due to its ability to extract geometric and topological information. However, its performance can be quite limited for various reasons. One of them is that the current spectral graph theory does not provide a multiscale analysis. Motivated by persistent homology and multiscale graphs, we introduce persistent spectral theory as a unified paradigm to unveil both topological persistence and geometric shape from high-dimensional datasets.

For a point set $V \subset \mathbb{R}^n$ without additional structures, we construct a filtration using an $(n-1)$-sphere of a varying radius r centered at each point. A series of persistent combinatorial Laplacian matrices are induced by the filtration. It is noted that our harmonic persistent spectra (i.e., zero eigenvalues) fully recover the persistent barcode or persistent diagram of persistent homology. Specifically, the numbers of zero eigenvalues of persistent q-combinatorial Laplacian matrices are the q-dimensional persistent Betti numbers for the same filtration given filtration. However, additional valuable spectral information is generated from the non-harmonic persistent spectra. In this work, in addition to persistent Betti numbers and the smallest non-zero eigenvalues, five statistic values, namely, sum, mean, maximum, standard deviation, and variance, are also constructed for data analysis. We use a few simple two-dimensional (2D) and three-dimensional (3D) structures to carry out the proof of principle analysis of the persistent spectral theory. The detailed structural information can be incorporated into the persistent spectra of. For instant, for the benzene molecule, the approximate C-C bond and C-H bond length can be intuitively read from the plot of the 0-dimensional persistent Betti numbers. Moreover, persistent spectral theory also has the capacity to accurately predict the heat of formation energy of small fullerene molecules. We use the area under the plot of the persistent spectra to model fullerene stability and apply the linear least-squares method to fit our prediction with the heat of formation energy. The resulting correlation coefficient is close to 1, which shows that our persistent spectral theory has an excellent performance on

molecular data. Furthermore, we have applied our persistent spectral theory to the protein B-factor prediction. In this case, persistent homology does not give a straightforward model. This example shows that the additional non-harmonic persistent spectral information provides a powerful tool for dealing with molecular data.

It is pointed out that the proposed persistent spectral analysis can be paired with advanced machine learning algorithms, including various deep learning methods, for a wide variety of applications in data science. In particular, the further construction of element-specific persistent spectral theory and its application to protein-ligand binding affinity prediction and computer-aided drug design will be reported elsewhere.

4.2 Persistent Path Laplacian

Recent years witness the emergence of a variety of advanced mathematical tools in topological data analysis (TDA) [148]. As the main workhorse of TDA, persistent homology (PH) [100, 43, 122, 101] pioneered a new branch in algebraic topology, offering a powerful tool to decode the topological structures of data during filtration in terms of persistent Betti numbers. Persistent homology has had tremendous success in many areas of science and technology, such as biology [4], chemistry [5], drug discovery [6], 3D shape analysis [7], etc.

Inspired by the success of PH, other mathematical tools have been given due attention. One of them is de Rham-Hodge theory in differential geometry, which uses the differential forms to represent the cohomology of an oriented closed Riemannian manifold with boundary in terms of a topological Laplacian, namely Hodge Laplacian [8]. The de Rham-Hodge theory has been applied to computational biology [55], graphic [149], and robotics [150]. However, like homology, the de Rham-Hodge theory does not offer an in-depth analysis of data, which is a famous problem in spectral geometry [10]. To overcome this drawback, the evolutionary de Rham-Hodge theory [9] was introduced in terms of persistent Hodge Laplacian to offer a multiscale analysis of the de Rham-Hodge

theory. Defined on a family of evolutionary manifolds, the evolutionary de Rham-Hodge theory gives a new answer to, or at least reopens, the famous 55-years old question “can one hear the shape of a drum". [10] The persistent Hodge Laplacian captures both the topological persistence and the homotopic shape evolution of data during filtration.

Nevertheless, the evolutionary de Rham-Hodge theory is set up on Riemannian manifolds, which may be computationally demanding for large datasets. Hence, a similar multiscaled-based topological Laplacian, called persistent spectral graph (PSG) [11], was proposed by introducing a filtration to combinatorial graph Laplacians. PSG, aka persistent Laplacian (PL) [151], extends persistent homology to non-harmonic analysis of data, showing much advantage in sophisticated applications [152, 153]. Dealing with point cloud data instead of manifolds, PL encodes a point cloud to a family of simplicial complexes generated from filtration and analyzes both harmonic and non-harmonic spectra. It is worthy to notice that the harmonic spectra from the null spaces of PLs reveal the same topological persistence like that of persistent homology, whereas, the non-harmonic spectra of PLs capture the homotopic shape evolution of data during the filtration. Meanwhile, open-source software called HERMES [154] was developed for the simultaneous topological and geometric analysis of data. However, like persistent homology, PSG treats all data points equally. That is to say, each point does not carry any labeled information such as the type, mass, color, etc. Therefore, an extension of PSG, called persistent sheaf Laplacian (PSL), was proposed to generalize cellular sheaves [155, 156] for the multiscale analysis of point cloud data with attached labeled information [157]. PSL is also a topological Laplacian that carries topological information in its null space but tracks homotopic shape evolution during filtration. Another interesting development is the persistent Dirac Laplacian (PDL) by Ameneyro, Maroulas, and Siopsis [158]. PDL offers an efficient quantum computation of persistent Betti numbers across different scales. These new approaches have great potentials to deal with complex data in science and engineering.

It is noticed that the aforementioned homologies and topological Laplacians are in-

sensitive to asymmetry or directed relations, which limits their representational power in encoding structures that have directional information. For example, in gene regulation data, the directions of gene regulations are indicated by arrowheads or perpendicular edges in systems biology [159]. Therefore, a technique that can deal with directed graphs (digraphs) is of vital importance to inferring gene regulation relationships. Notably, the path homology [12] proposed by Grigor'yan, Lin, Muranov, and Yau provides a powerful tool to analyze datasets with asymmetric structures using the path complex. Particular cases of homologies of digraphs and their path cohomology were also discussed [12, 60]. The notion of path homology of digraphs has a richer mathematical structure than the earlier homology and Laplacian, opening new directions for both pure and applied mathematics. For example, path homology theory was extended to various objects such as quivers, multigraphs, digraphs pairs, cylinder, cone, hypergraphs, etc. [160, 161, 162] Path homology has drawn much attention from researchers in the TDA community. To encode richer information, Chowdhury and Mémoli extended path homology to a persistent framework on a directed network [13]. Wang, Ren, and Wu constructed a weighted path homology for weight digraphs and proved a persistent version of a Künneth-type formula for joins of weighted digraphs [163]. Recently, Dey, Li, and Wang have designed an efficient algorithm for 1-dimensional persistent path homology [164], which is useful in real applications.

Similar to persistent homology, persistent path homology cannot track the homotopic shape evolution of data during filtration. To overcome this limitation, we introduce path Laplacian as a new topological Laplacian to analyze the spectral geometry of data, in addition to its topology. Moreover, we introduce a filtration to path Laplacian to obtain a persistent path Laplacian (PPL), a new framework that captures both the topological persistence and shape evolution of directed graphs and networks. By varying the filtration parameter, one can construct a series of digraphs, which result in a family of persistent path Laplacian matrices. The harmonic spectra of the persistent path Laplacian recover

all the topological invariants of the digraphs, while the non-harmonic spectra provide additional geometric information, which can distinguish two systems when they are homotopy but geometrically different. PPL has potential applications in science, engineering, industry, and technology. This work is organized as follows: Section 2 reviews the necessary background on path homology. Section 3 describes path Laplacian and persistent path Laplacian. Detailed PPL matrix constructions are illustrated with various examples for the interested readers in Section 3 and Section 4.

4.2.1 Constructions of Persistent Path Laplacian for Tetra and Pyramid

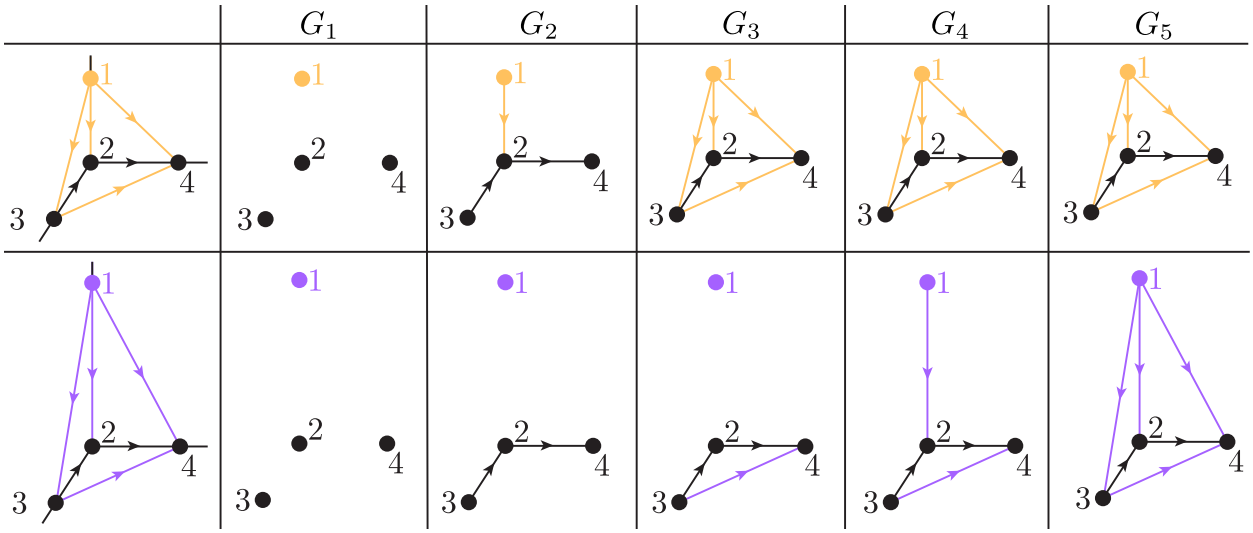

Figure 4.8: Illustration of filtration on a tetrahedron. Here, $1, 2, 3,$ and 4 represent four elementary 0-paths $e_1, e_2, e_3,$ and e_4. The top panel is a tetrahedron that has edge lengths $|e_{12}| = |e_{32}| = |e_{24}| = 1$ and $|e_{13}| = |e_{14}| = |e_{34}| = \sqrt{2}$. The bottom panel is a tetrahedron that has edge lengths $|e_{32}| = |e_{24}| = 1$, $|e_{34}| = \sqrt{2}$, $|e_{12}| = \sqrt{3}$, and $|e_{13}| = |e_{14}| = 2$.

One can get both abstract information (revealed by Betti numbers) and geometric information (revealed by non-harmonic spectra) from digraphs along filtration. For instance, Figure 4.8 illustrates the filtration on two tetrahedrons. The top panel is a tetrahedron (Tetra 1) with edge lengths $|e_{12}| = |e_{32}| = |e_{24}| = 1$, and $|e_{13}| = |e_{14}| = |e_{34}| = \sqrt{2}$. The bottom panel is another tetrahedron (Tetra 2) with edge lengths $|e_{12}| = \sqrt{3}$, $|e_{32}| = |e_{24}| = 1$, and $|e_{13}| = |e_{14}| = 2$, and $|e_{34}| = \sqrt{2}$. We say $G_1 = G^0, G_2 = G^1, G_3 = G^{\sqrt{2}}, G_4 = G^{\sqrt{3}},$

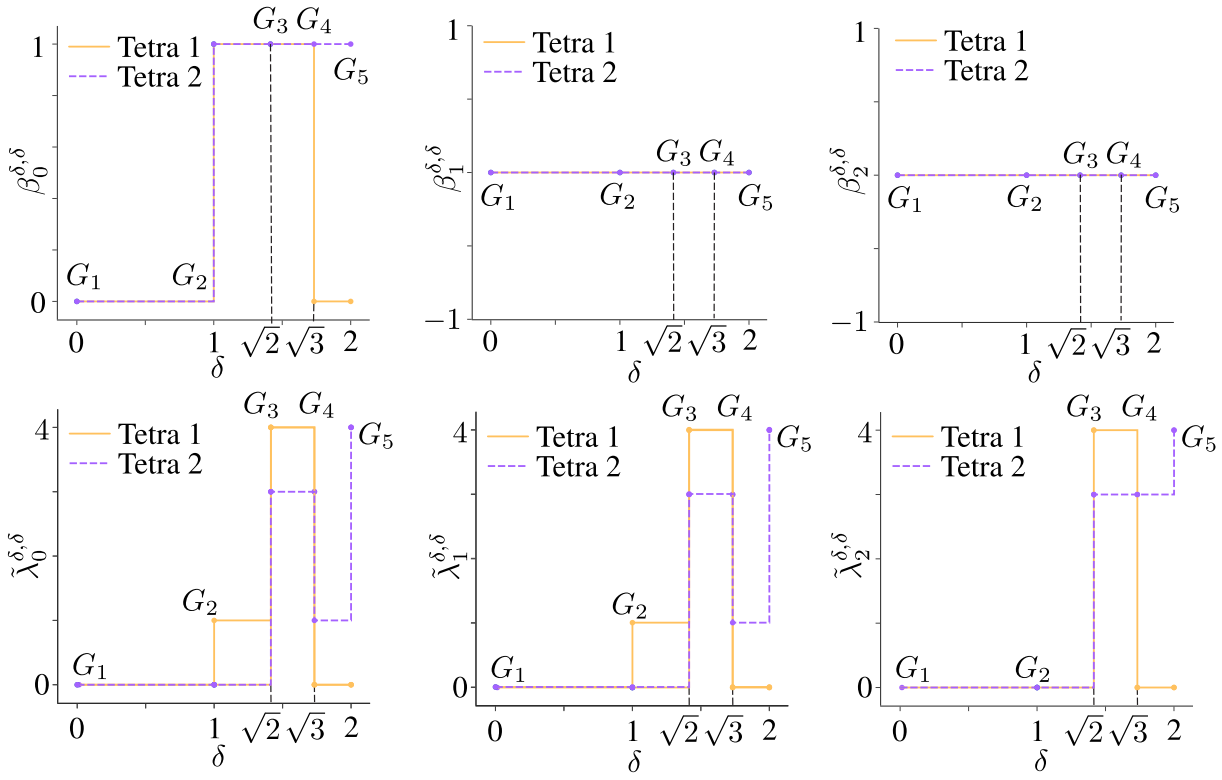

Figure 4.9: Comparison of Betti numbers and non-harmonic spectra of $L_n^{\delta,\delta}$ when $n = 0, 1,$ and 2 on tetrahedrons Tetra 1 and Tetra 2. Note that since $\beta_1^{\delta,\delta} = 0$ and $\beta_2^{\delta,\delta} = 0$ for Tetra 1 and Tetra 2, topological variants from persistent path homology cannot discriminate Tetra 1 and Tetra 2. However $\lambda_1^{\delta,\delta}$ and $\lambda_2^{\delta,\delta}$ show the differences between Tetra 1 and Tetra 2.

and $G_5 = G^{\sqrt{5}}$. Figure 4.9 shows the changes of $\beta_n^{\delta,\delta}$ and $\lambda_n^{\delta,\delta}$ of persistent n-th path Laplacian $L_n^{\delta,\delta}$ along filtration. It can be seen that by varying the filtration parameter δ from 0 to 1, the Betti 1 and Betti 2 are always 0. However, the smallest nonzero eigenvalue $\tilde{\lambda}_n^{\delta,\delta}$ of Tetra 1 and Tetra 2 have changes along filtration parameter δ. Additionally, when $n = 1, 2$, the $\tilde{\lambda}_n^{\delta,\delta}$ can distinguish Tetra 1 and Tetra 2, while $\beta_n^{\delta,\delta}$ cannot. This indicates that non-harmonic spectra of persistent path Laplacian can reveal more geometric information than the persistent Betti numbers in distinguishing similar topological structures. Notably, we remove all the isolated points from each digraph for the simplicity of calculation.

Moreover, a more complicated example is also illustrated in Figure 4.10 to describe

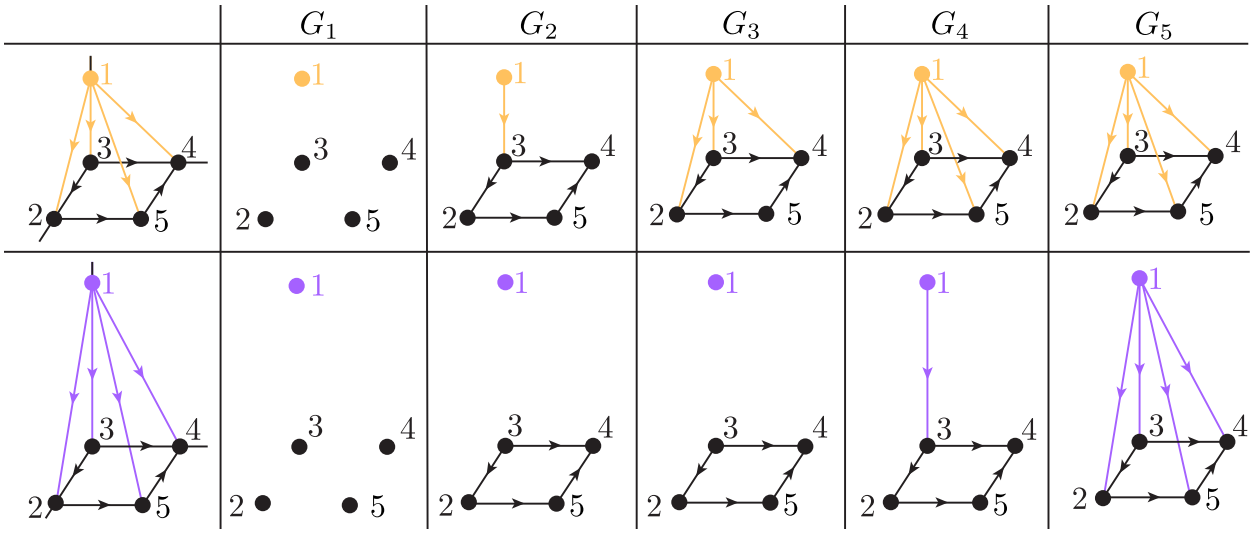

Figure 4.10: Illustration of filtration on a pyramid. Here, $1, 2, 3, 4$, and 5 represent five elementary 0-paths e_1, e_2, e_3, e_4, and e_5. The top panel is a pyramid that has edge lengths $|e_{13}| = |e_{25}| = |e_{32}| = |e_{34}| = |e_{54}| = 1$, $|e_{12}| = |e_{14}| = \sqrt{2}$, and $|e_{15}| = \sqrt{3}$. The bottom panel is a pyramid that has edge lengths $|e_{25}| = |e_{32}| = |e_{34}| = |e_{54}| = 1$, $|e_{12}| = |e_{14}| = 2$, and $|e_{15}| = \sqrt{5}$.

the filtration on two pyramids. The top panel is a pyramid (Pyra 1) with edge lengths $|e_{12}| = |e_{32}| = |e_{24}| = 1$, and $|e_{13}| = |e_{14}| = |e_{34}| = \sqrt{2}$. The bottom panel is a pyramid (Pyra 2) with edge lengths $|e_{12}| = \sqrt{3}$, $|e_{32},| = |e_{24}| = 1$, and $|e_{13}| = |e_{14}| = 2$, and $|e_{34}| = \sqrt{2}$. We say $G_1 = G^0, G_2 = G^1, G_3 = G^{\sqrt{2}}, G_4 = G^{\sqrt{3}}$, and $G_5 = G^{\sqrt{5}}$. Figure 4.11 depicts the changes of $\beta_n^{\delta,\delta}$ and $\lambda_n^{\delta,\delta}$ of persistent n-th path Laplacian $L_n^{\delta,\delta}$ for objects Pyra 1 and Pyra 2 along filtration. For Pyra 1 and Pyra 2, when $n = 0$ and $\delta = 1$, their corresponding digraphs form, which result in $\beta_0^{1,1} = 1$ and $\beta_1^{1,1} = 1$ for both Pyra 1 and Pyra 2. When $\delta = \sqrt{3}$, we have $\beta_1^{\sqrt{3},\sqrt{3}} = 0$ for Pyra 1 since the introducing of a new directed edges e_{15}. When $\delta = \sqrt{5}$, we have $\beta_1^{\sqrt{5},\sqrt{5}} = 0$ for Pyra 2 since the introducing of a new directed edges e_{15} kills the 1-cycle formed by e_{25}, e_{32}, e_{34}, and e_{54}. Furthermore, although Pyra 1 and Pyra 2 do not have exactly the same geometric structure, their share the same $\beta_2^{\delta,\delta}$ value from $\delta = 0$ to $\delta = \sqrt{5}$. However, Pyra 1 and Pyra 2 can be distinguished by the $\tilde{\lambda}_2^{\delta,\delta}$ along filtration. Therefore, we can see that similar to the PSG, one can use the non-harmonic spectra from the persistent path laplacian to reveal the intrinsic geometric information of a givens point-cloud dataset by varying the filtration parameters. In addition, the detailed

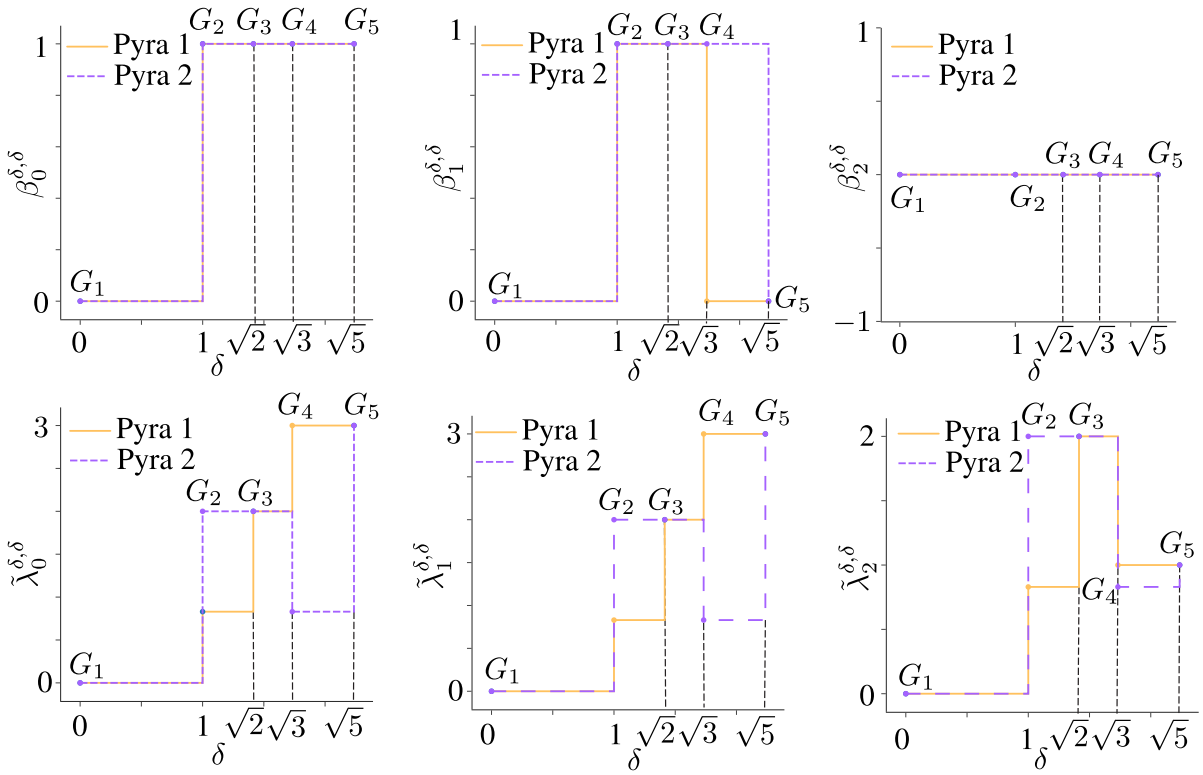

Figure 4.11: Comparison of Betti number and non-harmonic spectra of $L_n^{\delta,\delta}$ when $n = 0, 1$,c and 2 on pyramids Pyra 1 and Pyra 2. Note that since $\beta_2^{\delta,\delta} = 0$, it cannot distinguish Pyra 1 and Pyra 2. But $\lambda_2^{\delta,\delta}$ can tell the difference.

calculations of $L_n^{\delta,\delta}$ can be found in the Appendix.

4.2.2 Constructions of Persistent Path Laplacian for CB7

In this section, we apply the persistent path Laplacian to the analysis of the curcurbit[n]urils system. Cucurbiturils are macrocyclic molecules, which are made of glycoluril (=$C_6H_2N_4O_2$=) monomers linked by methylene bridges (-CH_2-). CBn is commonly used as an abbreviation of Cucurbiturils. Here, n is the number of glycoluril units. In this work, we consider CB7 as an example. The molecular formulas of CB7 is $C_{42}H_{14}N_{28}O_{14}$. The molecular structure of CB7 is obtained from the Supporting Information of Ref. [165].

Figure 4.12 illustrates how PPL is employed for a molecular system to extract its rich topological and geometric information. The first two charts of Figure 4.12**a** describe the

three-dimensional (3D) top view and side view of CB7. The green, blue, red, and gray colors represent C, N, O, and H atoms, respectively. The third chart of Figure 4.12**a** is a basic "Octagon-pentagon" unit that consists of two glycolurils. It can be seen that 7 glycolurils exist in CB7. The last chart of Figure 4.12**a** demonstrates the path direction assignment to pairs of atoms based on atomic electronegativity. The periodic table of electronegativity is given by the Pauling scale [166], in which the electronegativities of C, N, O, and H are 2.55, 3.04, 3.44, and 2.20, respectively. Then, we set the directions of edges following the order "H $\rightarrow$ C $\rightarrow$ N $\rightarrow$ O".

Figure 4.12**b** depicts the distance-based filtration of CB7. Here, structures $G_i (i = 1, 2, ..., 8)$ were obtained at the filtration radii of 0.200, 0.565, 0.710, 0.745, 0.800, 1.210, 1.315, and 1.800 Å, respectively. In our digraph notation, we denote these structures as $G_1 = G_0^{0.200}, G_2 = G_0^{0.565}, G_3 = G_0^{0.710}, G_4 = G_0^{0.745}$, $G_5 = G_0^{0.800}, G_6 = G_0^{1.210}, G_7 = G_0^{1.315}$, and $G_8 = G_0^{1.800}$. Note that, in the present formulation, all of the isolated points were removed from these digraphs.

Figure 4.12**c** illustrates the filtration-induced path complexes in the aforementioned $G_i (i = 1, 2, ..., 8)$. To clearly show the topological and geometric changes, only the path complexes in one "Octagon-pentagon" unit (or two glycolurils) are considered and depicted for each structure. For simplicity, only edges are presented. However, their path directions can be easily assigned based on their color map as shown in the last chart of Figure 4.12**a**.

Figure 4.12**d** depicts the PPL spectra of CB7. We can see that at the initial state (G_1) when $\delta = 0.200$ Å), total 98 atoms are isolated from one another. When radius $\delta = 0.565$ Å (G_2), C atoms on each pentagon are connected with their H atom neighborhoods. Therefore, four isolated components are formed in each glycoluril, which makes $\beta_0^{\delta,\delta} = 4 \times 7 = 28$. At G_3 ($r = 0.710$ Å), C atoms on each pentagon are connected with their N and O neighborhoods. At this stage, two more connected components are involved in one glycoluri structure, which makes $\beta_0^{\delta,\delta} = 6 \times 7 = 42$. Only one connected structure can

be formed if all of the atoms get connected with their neighborhood atoms. Therefore, $\beta_0^{\delta,\delta} = 1$ (see G_5 - G_8). Notably, the $\beta_2^{\delta,\delta}$ and $\tilde{\lambda}_2^{\delta,\delta}$ provide rich topological and geometric information when the filtration parameter δ increases.

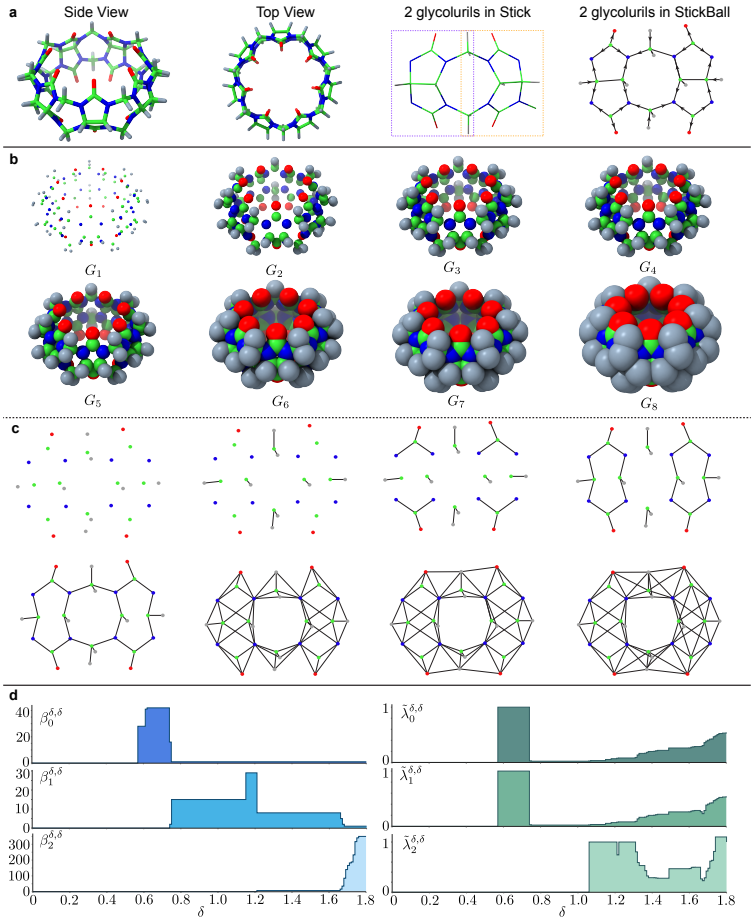

Figure 4.12: **a** The 3D structures of CB7, 2 glycolurils, and path direction assignment. Here, from left to right, the side view of CB7, top view of CB7, the structure of two glycoluril units (=$C_{10}H_4N_8O_4$=), and electronegativity-based path direction assignment are depicted as well. **b** Illustration of filtration-induced geometries $G_i (i = 1, 2, \ldots, 8)$ of CB7. Eight digraphs $G_1 = G_0^{0.200}, G_2 = G_0^{0.565}, G_3 = G_0^{0.710}, G_4 = G_0^{0.745}, G_5 = G_0^{0.800}, G_6 = G_0^{1.210}, G_7 = G_0^{1.315}, G_8 = G_0^{1.800}$ are constructed under filtration parameter δ. **c** Illustration of filtration-induced path complexes within two glycoluril units. Path directions can be inferred from their colors as shown in the last chart of **a**. **d** Betti numbers $\beta_n^{\delta,\delta}$ and non-harmonic spectra $\tilde{\lambda}_n^{\delta,\delta}$ of persistent path Laplacians ($L_n^{\delta,\delta}$ when $n = 0, 1$, and 2) for CB7.

This example shows that PPL can decode topological persistence and the shape evolution of a given molecular system with chemical- or biological-based directional assignment. Specifically, $\tilde{\lambda}_0^{\delta,\delta}$ can still offer geometric information when $\beta_0^{\delta,\delta}$ does not changes for

large radii. Therefore, PPL keeps revealing homotopic shape evolution when the topological invariant from persistent path homology does not change.

Additionally, unlike persistent Laplacian, high-order PPL operators provide rich topological information. For instance, when the filtration parameter δ increases to 1.68, $\beta_2^{\delta,\delta}$ from PPL dramatically goes up. Whereas, in persistent Laplacian, the value of Betti 2 is quite limited since the CB7 system can barely form 2-cycles at a similar filtration parameter using either Rips complex or alpha complex. This trait endows PPL with a better ability to characterize the geometry and topology of an object at large scales.

4.2.3 Discussion and Conclusion

Path homology, a rich mathematical concept introduced by Grigor'yan, Lin, Muranov, and Yau, has stimulated a variety of new developments in pure and applied mathematics, including much attention from the topological data analysis (TDA) community. Unlike original homology or persistent homology, path homology enables the treatment of directed graphs and networks. Persistent path homology bridges path homology with multiscale analysis, making it a powerful tool for practical applications. Nonetheless, these formulations are insensitive to homotopic shape evolution during filtration.

Topological Laplacians, including Hodge Laplacian, graph Laplacian, sheaf Laplacian, and Dirac Laplacian, are versatile mathematical tools that not only preserve all topological invariants but also describe geometric shapes. This work introduces a new topological Laplacian, namely persistent path Laplacian, as a new mathematical tool for the multi-scale analysis of directed graphs and networks. For a given data, the proposed persistent path Laplacian fully recovers the topological persistence of persistent homology in its harmonic spectra and meanwhile, captures homotopic shape evolution of the data during filtration in its non-harmonic spectra.

CHAPTER 5

HERMES: AN OPEN-SOURCE SOFTWARE FOR THE SPECTRAL ANALYSIS OF PERSISTENT LAPLACIANS

5.1 Introduction

As a branch of discrete mathematics, graph theory focuses on the relations among vertices or nodes (0-simplices), edges (1-simplices), faces (2-simplices), and their high-dimensional extensions. Benefiting from the capability of graph formulations that encode inter-dependencies among constituents of versatile data into simple representations, graph theory has been regarded as the mathematical scaffold in the study of various complex systems in biology, material science, physical infrastructure, and network science. However, traditional graphs only represent the pairwise relationships between entries. Therefore, hypergraphs, a generalization of graphs that describe the multi-way relationships of mathematical structures have been developed to capture the high-level complexity of data [167, 168]. Mathematically, graphs and hypergraphs are intrinsically related to the simplicial complexes, which have broader use in computational topology. Moreover, many other areas such as algebra, group theory, knot theory, spectral graph theory (SGT), algebraic topology (AT), and combinatorics are closely related to graph theory. Among them, the applications of SGT have been driven by various real-life problems in chemistry, physics, and life science in the past few decades [138, 169].

In its early days, spectral graph theory studied the properties of a graph by its graph Laplacian matrix and adjacency matrix. Later on, developments in spectral graph theory involved some geometric flavor. The explicit constructions of expander graphs rely on studying eigenvalues and isoperimetric properties of graphs. The discrete analog of Cheeger's inequality for graphs in Riemannian geometry is related to the study of manifolds [170]. Specifically, an eigenvalue of the Laplacian of a manifold is related to the

isoperimetric constant of the manifold, which motivates the study of graphs by employing manifolds. Benefiting from increasingly rich connections with differential geometry, spectral graph theory entered a new era [171]. One of the critical developments is the Laplacian on a compact Riemannian manifold in the context of the de Rham-Hodge theory [54, 55]. The harmonic part of the Hodge Laplacian spectrum contains the topological information, whereas the non-harmonic part of the Hodge Laplacian spectrum offers additional geometric information for shape analysis [56]. Indeed, the connectivity of a graph/topological space can be revealed from topological invariants. It is well-known that the number of eigenvalues in the harmonic spectra of qth-order persistent Laplacian represents the dimension of persistent q-cohomology of a graph [172, 53, 11], which builds the connection between spectral graph theory and algebraic topology.

Homology and cohomology are key concepts in the algebraic topology, which were developed to analyze and classify manifolds according to their cycles. Traditional homology is genuinely metric-independent, indicating that geometric information is barely considered [173]. Therefore, for practical computation, a new branch of algebraic topology named persistent homology (PH) [122, 124, 43] was implemented to create a sequence of topological spaces characterized by a filtration parameter, such as the radius of a ball or the level set of a real-valued function. As the most important realization of topological data analysis (TDA) [125, 128, 174], topological persistence has had great success in computational chemistry [5, 175] and biology [4, 44, 176, 177, 178]. For instance, the superior performance of using PH features of protein-drug complexes in the free energy prediction and ranking at D3R Grand Challenges, a worldwide competition series in computer-aided drug design [6], was a remarkable success for TDA. Additionally, a weighted persistent homology is proposed as a unified paradigm for the analysis of the biomolecular data system [179].

Recently, we introduced persistent spectral graph (PSG) theory to bridge persistent homology and spectral graph theory [11, 11]. The PSG theory extends the persistence no-

tion or multiscale analysis to algebraic graph theory. A family of spectral graphs induced by a filtration overcomes the difficulty of using traditional spectral graph theory in analyzing graph structures with a single geometry, giving rise to persistent spectral analysis (PSA). Additionally, the evolution of the null space dimension of the persistent Laplacian matrix (PLM) over the filtration offers the topological persistence. Therefore, PSG theory provides simultaneous TDA and PSA. Specifically, by varying a filtration parameter, a series of qth-order persistent Laplacians (or q-persistent Laplacian) provide persistent spectra. Notably, the persistent harmonic spectra of 0-eigenvalues span the null space of the q-th order persistent Laplacian and fully recover the persistent q-th Betti numbers or persistent barcodes [180] of the associated persistent homology. Specifically, the number of 0-eigenvalues of qth-order persistent Laplacian reveals the number of q-cocycles for a given point-cloud dataset. Moreover, the additional geometric shape information of the data will be unveiled in the non-harmonic spectra. For example, the spectral gap (the difference between the moduli of the first two smallest eigenvalues of a Laplacian) reveals the energy difference/density changes between the ground state and first excited state of a system/dataset. Additionally, the B-factor prediction performance can be significantly improved by using the non-harmonic spectra involved in the prediction model, as discussed in [11]. Recently, the theoretical properties and algorithms of PSGs have been further studied [151] and the application of PSG methods to drug discovery has been reported [181] . The de Rham-Hodge theory counterpart, called evolutionary de Rham-Hodge theory, has also been formulated [56].

Currently, many open-source packages have been developed for the applications of persistent homology, including Ripser [182], Dionysus [183], Gudhi [184], Perseus [123], DIPHA [185], Javaplex [186], CliqueTop [187], DioDe [188], Hera, Eirene, and "TDA" package in R [189]. These packages are able to construct a family of complexes with the point clouds data as input and calculate its corresponding Betti numbers, which are equivalent to the harmonic spectra of the persistent Laplacian. However, there is no soft-

ware package for simultaneous TDA and PSA. While we developed the theoretical part of the persistent spectral graph in 2019, we have not constructed efficient and robust software yet.

The objective of present work is to provide the first open-source package, dubbed highly efficient robust multidimensional evolutionary spectra (HERMES), for evaluating both the harmonic and non-harmonic spectra of persistent Laplacian matrices, which enable broad and convenient applications of the PSG method. In the present release, we consider an implementation in both alpha complexes [47] and Vietoris–Rips complexes. To verify the reliability of HERMES, 15 complicated 3D structures of proteins as well as two fullerene structures are used to calculate the spectra of qth-order persistent Laplacians for $q = 0, 1, 2$. Moreover, as a validation, the persistent harmonic spectra generated by HERMES are compared with those obtained from Gudhi and DioDe. Furthermore, with the use of the spectra of PLMs, molecular data abnormality detection is also discussed.

In a nutshell, HERMES provides a powerful tool in various applications such as drug discovery, protein flexibility analysis, and complex protein structures analysis. It can be potentially applied to various fields where persistent homology has had success.

5.2 Implementation

5.2.1 Construction of Alpha Shape

Recall that, given a set of points, the alpha shape with any α value is a subcomplex of Delaunay tessellation. Thus, to construct the filtration of alpha complexes, it is necessary to first compute the complete simplicial complex through the Delaunay tessellation formed by the set of points. A number of efficient implementations is available in existing software packages. Our implementation employs the Computational Geometry Algorithms Library (CGAL), an efficient and robust software package for many commonly used calculations. We then assign each simplex σ with an alpha value α_σ. Finally, the alpha shape

given at an α value α_0 is constructed by union of convex hulls of all the simplices σ satisfying $\alpha_\sigma \leq \alpha_0$, which naturally forms the nerve of balls centered at the given points truncated by the Voronoi regions, i.e., the corresponding alpha complex.

We illustrate our implementation with point sets P in 3D, as it is the most common use scenario. We also assume that all the points are in general positions, which means that no 4 points of P lie on the same plane and no 5 points of P lie on the same sphere. Given a simplex σ, which can be a point, an edge, a triangle or a tetrahedron, denote the open ball bounded by its minimal circumsphere as B_σ. The simplex σ is called *Gabriel* ([190]) if $B_\sigma \cap P = \emptyset$. Note that for vertices (0-simplices) the circumradius is considered 0. The above discussion can be directly adapted for 2D implementation by replacing circumsphere with circumcircle and omitting tetrahedra.

The filtration parameter α for every simplex σ can be defined as follows. If the simplex is Gabriel, the filtration value is the corresponding circumradius (for efficiency, we actually store its square) because the corresponding ball can be considered as an empty α-ball touching all its vertices. If the simplex is not Gabriel, the filtration value is the minimum of all the filtration values of the cofaces of σ that contain the points making the simplex non-Gabriel. When α value reaches that number, we will have an empty α-ball making the simplex α-exposed.

5.2.2 Implementation details for alpha shape

To ensure the valid calculation of the filtration parameter for non-Gabriel simplices, the filtration values are always computed from the highest dimension (tetrahedra) down to 0 (vertices). We initialize the filtration value for all the simplices to be positive infinity. For dimension k, we iterate through each k-simplex. If the current filtration value α_σ^2 is positive infinity, we assign the filtration value as the square of the corresponding circumradius. Then, we check every $(k-1)$-dimensional face τ in $\partial\sigma$. If the circumsphere of τ enclosed the other vertex of σ in the interior, it is not Gabriel, and does not correspond to

an empty α-ball. In this case, α_σ^2 is assigned to α_τ^2 if $\alpha_\sigma > \alpha_\tau$.

With this procedure, we ensure that α_σ for every simplex σ corresponding to the filtration value α is α-exposed to an empty α-ball. In other words, we ensure each simplex represented by its vertex index set $J \subseteq \{1, 2, ..., |P|\}$ is in the nerve of the R_i's, which are the intersections $R_i = V_i \cap B_i$ of Voronoi cells V_i's and balls B_i's around the points p_i's.

5.2.2.1 Boundary operator construction

With α_σ assigned, we sort the k-simplices with increasing filtration parameter value. This allows us to construct a single boundary operator B_q^∞ (the matrix representation of ∂_q^∞) for the entire filtration, which is that of the Delaunay tessellation. For any given α, we can read off the top left block of the full boundary matrix B_q^∞, i.e.,

$$\left(B_q^\alpha\right)_{ij} = \left(B_q^\infty\right)_{ij}, \quad \forall 1 \leq i \leq N_{q-1}^\alpha, 1 \leq j \leq N_q^\alpha, \tag{5.1}$$

where N_q^α is the number of q-simplices in the alpha complex with the filtration parameter α. Alternative, we can consider the $N_q^\alpha \times N_q^\infty$ projection matrix P_q^α from the Delaunay tessellation to the alpha complex, $\left(P_q^\alpha\right)_{ij} = \delta_{ij}$ (1 on the diagonal and 0 elsewhere), with which we have $B_q^\alpha = P_{q-1}^\alpha B_q^\infty (P_q^\alpha)^T$.

5.2.2.2 Persistent boundary operator

The construction of p-persistent boundary matrix $B_q^{\alpha,p}$ (the representation of operator $\eth_q^{\alpha,p}$) is more involved than reading off B_q^∞. We first construct the projection matrix $\mathbb{P}_q^{\alpha,p}$ from $C_q^{\alpha+p}$ to $\mathbb{C}_q^{\alpha,p}$. Then, the p-persistent boundary matrix can be assembled as $B_q^{\alpha,p} = P_{q-1}^\alpha B_q^\infty (\mathbb{P}_q^{\alpha,p})^T$.

To construct the projection matrix, we first note that it is the projection to the kernel of an operator that measures the difference between the boundary operator mapped onto $C_{q-1}^{\alpha+p}$ and the boundary restricted to C_{q-1}^α, $\text{Diff}_q^{\alpha,p} = (I_{q-1}^{\alpha+p} - R_{q-1}^{\alpha,p})^T B_q^{\alpha+p}$, where $R_q^{\alpha,p} =$

$P_q^{\alpha+p}(P_q^\alpha)^T P_q^\alpha (P_q^{\alpha+p})^T$ is the restriction from $C_q^{\alpha+p}$ to C_q^α and $I_q^{\alpha+p}$ is the identity matrix on $C_q^{\alpha+p}$.

Instead of storing a dense matrix, we propose to use a procedural representation involving the inverse of persistent Laplacians with gauge ([191]) to reduce the storage as well as speed up the computation. More specifically, we construct the projection matrix as follows

$$\mathbb{P}_q^{\alpha,p} = I_q^{\alpha+p} - (\tilde{\text{Diff}}_q^{\alpha,p})^T (\tilde{L}_{q-1}^{\alpha,p})^{-1} \tilde{\text{Diff}}_q^{\alpha,p}, \tag{5.2}$$

where $(\tilde{L}_{q-1}^{\alpha,p})^{-1}$ can be implemented through rank deficiency fixing in [191], and the restricted operator $\tilde{\text{Diff}}_q^{\alpha,p}$ is defined below. Note that this sparse linear equation solving approach is essentially the graph version of the harmonic extension described in Ref. [55].

The reason that the projection matrix can be defined this way is that starting from an arbitrary element $\omega_q \in C_q^{\alpha+p}$, we can modify it into $\omega_q - (\text{Diff}_q^{\alpha,p})^T f_{q-1} \in \mathbb{C}_q^{\alpha,p}$, where f_{q-1} is nonzero only in the difference complex $\text{Cl}(T_{\alpha+p} - T_\alpha)$, the closure of the difference between $T_{\alpha+p}$ and T_α. Denoting any chain f on the difference complex as $\tilde{f}$ and any operator B on it as $\tilde{B}^{\alpha,p}$, and the $\tilde{B}_q^{\alpha,p}(\tilde{B}_q^{\alpha,p})^T \tilde{f}_{q-1} = \tilde{B}_q^{\alpha,p}\tilde{\omega}_q$. Noticing that $\tilde{f}_{q-1}$ is determined up to a gauge transform $f_{q-1} - (\tilde{B}_{q-1}^{\alpha,p})^T \tilde{g}_{q-2}$ for some $(q-2)$-chain g_{q-2} in $\text{Cl}(T_{\alpha+p} - T_\alpha)$, we introduce the gauge fixing term $\tilde{B}_{q-1}^{\alpha,p} f_{q-1} = 0$, which leads us to the sparse linear system $\tilde{L}_{q-1}^{\alpha+p} \tilde{f}_{q-1} = \tilde{\text{Diff}}_q^{\alpha,p} \omega_q$ where the $\tilde{\text{Diff}}$ operator is the above operator projected to the difference complex. Note that fixing the rank deficiency of persistent Laplacians (in the difference complex) is computationally efficient as its kernel dimension is far smaller than that of the corresponding boundary or coboundary operators.

5.2.2.3 Persistent spectrum computation

The q-order p-persistent Laplacian operators can then be implemented by direct evaluation of $L_q^{\alpha,p} = B_{q+1}^{\alpha,p}(B_{q+1}^{\alpha,p})^T + (B_q^\alpha)^T B_q^\alpha$. Their spectra can be evaluated through any off-the-shelf sparse matrix eigensolver.

Thus, the dimension of the null space of $L_0^{\alpha,p}$ is the number of p-persistent connected components. The dimension of the null space of $L_1^{\alpha,p}$ is the number of p-persistent handles or tunnels. Similarly, the dimension of the null space of $L_2^{\alpha,p}$ is the number of p-persistent cavities.

5.2.3 Implementation Details for Rips Complex

The Vietoris–Rips complex at different filtration values is also considered in HERMES. Following the definition of the Vietoris–Rips complex, the implementation is straightforward. However, due to the large number of simplices, the calculation of non-harmonic spectra of PLMs $L_q^{t,p}$ can be resource-intensive. Therefore, we may set a maximum cutoff distance for the filtration r and an upper limit for persistent p for practical applications.

5.3 Validation

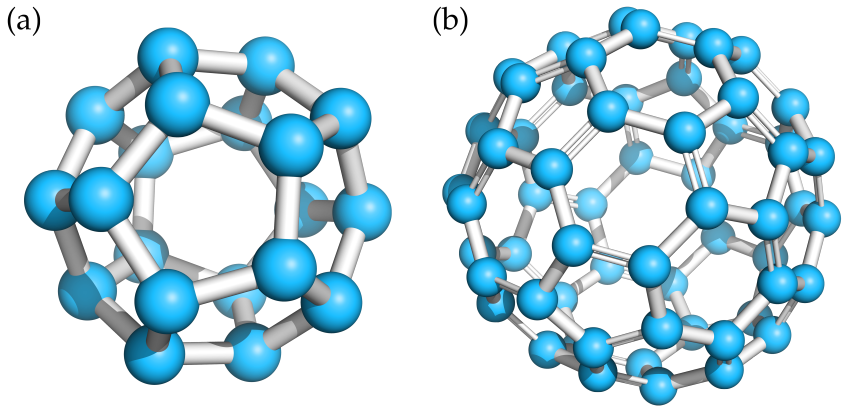

Figure 5.1: The 3D structures of C_{20} and C_{60}. (a) C_{20} molecule. A total of 12 pentagon rings can be found in C_{20}. (b) C_{60} molecule. 12 pentagon rings and 20 hexagon rings form the structure of C_{60} .

We construct the alpha complex at different filtration values from the finite cells of a Delaunay tessellation from the Computational Geometry Algorithms Library (CGAL). Moreover, the Vietoris–Rips complex at different filtration values is also constructed in HERMES. Gudhi and DioDe are two of the most frequently applied open-source libraries

that are able to compute Betti numbers (harmonic persistent spectra) based on CGAL, while Ripser is based on the blazing fast C++ Ripser package. As shown in [11], the 0-persistent qth Betti numbers $\beta_q^{\alpha,0}$ at filtration parameter t is the number of zero eigenvalues of qth-order 0-persistent Laplacian $\mathcal{L}_q^{t,0}$:

$$\beta_q^{t,0} = \dim(C_q^t) - \text{rank}(\mathcal{L}_q^{t,0}) = \dim \ker \mathcal{L}_q^{t,0}, \tag{5.3}$$

where $t = \alpha$ if we choose to construct the alpha complex, and $t = r$ if we choose to construct the Vietoris–Rips complex.

In fact, $\beta_q^{t,0}$ counts the number of q-cycles in the alpha complex K_t that persists in K_t. Although Gudhi and DioDe can calculate the number of zero eigenvalues, the non-harmonic persistent spectra also play an important role in applications as shown in our earlier work [11]. Therefore, we developed an open-source package HERMES, which not only tracks the topological changes from the persistent Betti numbers but also derives the geometric changes from the non-harmonic spectra of persistent Laplacians. In the following, we compare the Betti numbers $\beta_q^{t,p}$ that are calculated from HERMES with the Betti numbers that are derived from Gudhi and DioDe on a set of 2D and 3D points, aiming to validate the robustness and accuracy of HERMES.

5.3.1 Validation on Fullerene structures

In this section, we will validate the correctness of HERMES with simple systems such as C_{20} and C_{60} molecules with known persistent Betti numbers [4] for Rips complex. Moreover, the persistent Betti numbers for the alpha complex are also included in this section.

C_{20} molecule. The C_{20} molecule is the smallest member of the fullerene family, which has a dodecahedral cage structure as illustrated in Figure 5.1 (a). Both C_{20} and C_{60} have the molecular symmetry of the full icosahedral point group I_h. Figure 5.2 illustrates the persistent Betti numbers for Rips complex $\beta_0^{r,0.05}$, $\beta_1^{r,0.05}$, and $\beta_2^{r,0.05}$ (green curves) and the smallest non-zero eigenvalue $\lambda_0^{r,0.05}$, $\lambda_1^{r,0.05}$, and $\lambda_2^{r,0.05}$ (yellow curves) of C_20 that are

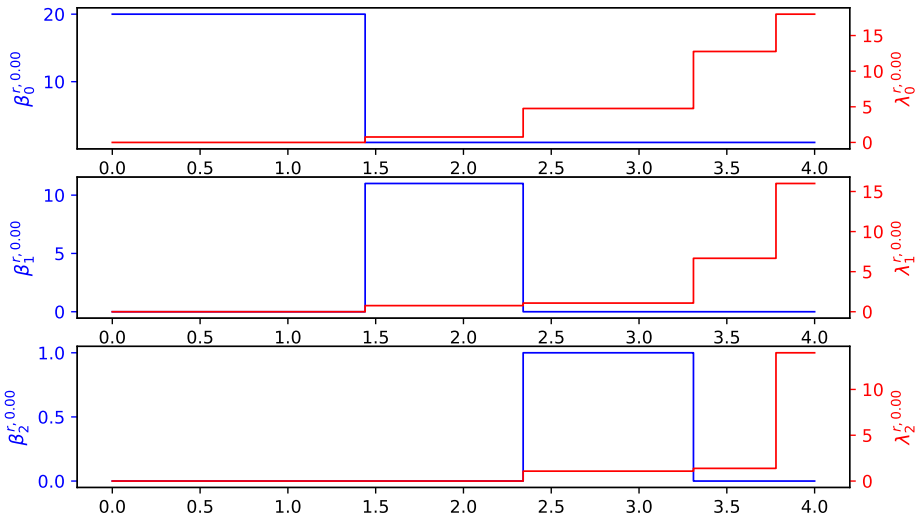

Figure 5.2: Illustration of the harmonic spectra (for Rips complex) $\beta_0^{r,0}$, $\beta_0^{r,0}$, and $\beta_2^{r,0}$ (green curves from top chart to bottom chart) and the smallest non-zero eigenvalue $\lambda_0^{r,0}$, $\lambda_1^{r,0}$, and $\lambda_2^{r,0}$ (yellow curves from top chart to bottom chart) of C_{20} molecule (the bottom left chart in Figure 5.6) at different filtration values α calculated from HERMES. Here, the x-axis represents the radius filtration value r (unit: Å), the left-y-axes represents the number of zero eigenvalues of $\mathcal{L}_0^{r,0}$, $\mathcal{L}_1^{r,0}$, and $\mathcal{L}_1^{r,0}$ from top to bottom, and the right-y-axes represents the first non-zero eigenvalue of $\mathcal{L}_0^{r,0}$, $\mathcal{L}_1^{r,0}$, and $\mathcal{L}_2^{r,0}$ from top to bottom.

computed from HERMES. Similarly, Figure 5.3 illustrates the persistent Betti numbers for the alpha complex $\beta_0^{\alpha,0.05}$, $\beta_1^{\alpha,0.05}$, and $\beta_2^{\alpha,0.05}$ (green curves) and the smallest non-zero eigenvalue the $\lambda_0^{\alpha,0.05}$, $\lambda_1^{\alpha,0.05}$, and $\lambda_2^{\alpha,0.05}$ (yellow curves) of C_20 that are computed from HERMES.

Note that although the Rips complex and the alpha complex have similar Betti-0 and Betti-1 patterns, their Betti-2 patterns differ from each other over the filtration range. Additionally, the non-harmonic spectra of the Rips complex and the alpha complex differ much from each other. Moreover, the non-harmonic spectra of the Rips complex appear to carry more information than those of the alpha complex.

C_{60} molecule. The C_{60} molecule is a well-known structure that is also called buckminsterfullerene. A total of 12 pentagon rings and 20 hexagon rings consist of C_{60}. Figure 5.1 (b) shows the 3D structure of C_{60}. Figure 5.4 and Figure 5.5 demonstrate the 0.05-

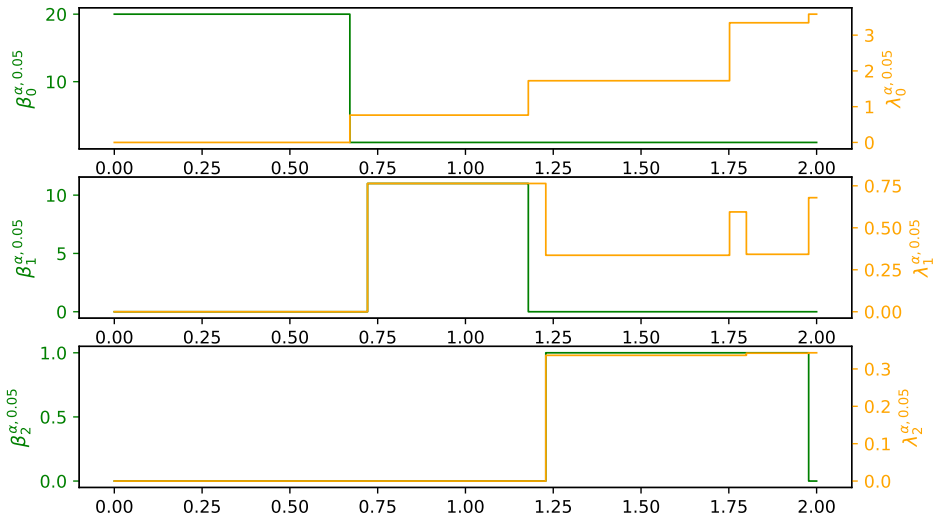

Figure 5.3: Illustration of the harmonic spectra (for alpha complex) $\beta_0^{\alpha,0.05}$, $\beta_0^{\alpha,0.05}$, and $\beta_2^{\alpha,0.05}$ (green curves from top chart to bottom chart) and the smallest non-zero eigenvalue $\lambda_0^{\alpha,0.05}$, $\lambda_1^{\alpha,0.05}$, and $\lambda_2^{\alpha,0.05}$ (yellow curves from top chart to bottom chart) of the C_{20} molecule (the bottom left chart in Figure 5.6) at different filtration value α calculated from HERMES. Here, the x-axis represents the radius filtration value α (unit: Å), the left-y-axes represents the number of zero eigenvalues of $\mathcal{L}_0^{\alpha,0.05}$, $\mathcal{L}_1^{\alpha,0.05}$, and $\mathcal{L}_1^{\alpha,0.05}$ from top to bottom, and the right-y-axes represents the first non-zero eigenvalue of $\mathcal{L}_0^{\alpha,0.05}$, $\mathcal{L}_1^{\alpha,0.05}$, and $\mathcal{L}_2^{\alpha,0.05}$ from top to bottom.

persistent Betti numbers for rips complex and alpha complex, respectively. Figure 5.2 - Figure 5.5 indicate the capacity of HERMES for the direct calculation of the persistent spectra of $\mathcal{L}_q^{r,p}$ and $\mathcal{L}_q^{\alpha,p}$ ($p > 0$).

5.3.2 Validation on proteins

In this section, we further validate HERMES using 15 proteins. Their Protein Data Bank (PDB) IDs of these proteins are 1CCR, 1NKO, 1O08, 1OPD, 1QTO, 1R7J, 1V70, 1W2L, 1WHI, 2CG7, 2FQ3, 2HQK, 2PKT, 2VIM, and 5CYT. The 3D structures of these 15 proteins can be downloaded from the PDB (https://www.rcsb.org/). Here, only the alpha carbon atoms are considered in our calculations. The harmonic spectra of HERMES are compared with the persistent Betti numbers of Gudhi and DioDe. Figure 5.6 illustrates

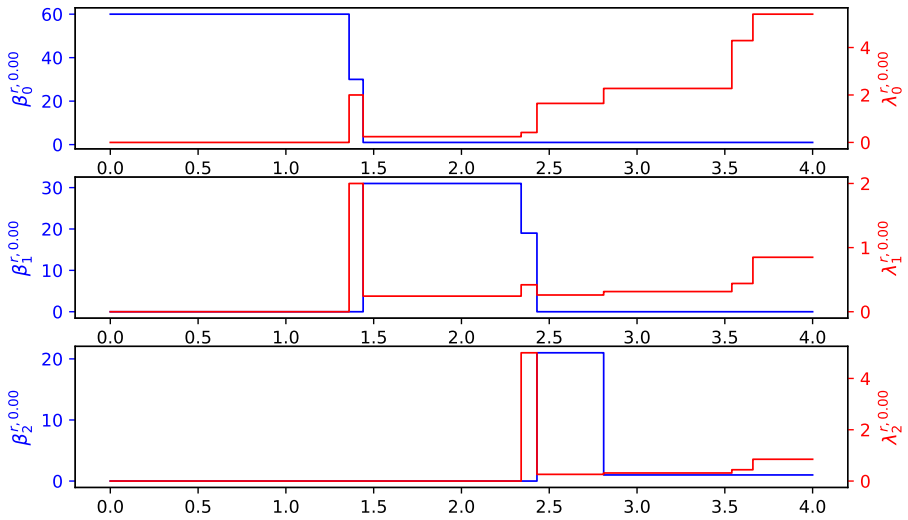

Figure 5.4: Illustration of the harmonic spectra $\beta_0^{r,0}$, $\beta_0^{r,0}$, and $\beta_2^{r,0}$ (blue curves from top chart to bottom chart) and the smallest non-zero eigenvalue $\lambda_0^{r,0}$, $\lambda_1^{r,0}$, and $\lambda_2^{r,0}$ (red curves from top chart to bottom chart) of the C_{60} molecule (the bottom left chart in Figure 5.6) at different filtration value α calculated from HERMES. Here, the x-axis represents the radius filtration value α (unit: Å), the left-y-axes represents the number of zero eigenvalues of $\mathcal{L}_0^{r,0}$, $\mathcal{L}_1^{r,0}$, and $\mathcal{L}_1^{r,0}$ from top to bottom, and the right-y-axes represents the first non-zero eigenvalue of $\mathcal{L}_0^{r,0}$, $\mathcal{L}_1^{r,0}$, and $\mathcal{L}_2^{r,0}$ from top to bottom.

the network structures of 15 proteins. For each protein, the color at atomic positions represents the normalized diagonal values of the accumulated 0th-order 0-persistent Laplacians: $\frac{1}{\max_i(\mathcal{L}_0^0)_{ii}}(\mathcal{L}_0^0)_{jj}$, with $\mathcal{L}_0^0 = \sum_\alpha \mathcal{L}_0^{\alpha,0}$. Here, the filtration α goes from $\sqrt{1.5}$ Å to $\sqrt{10}$ Å with the step size of 0.01 Å. Figure 5.7 depicts the persistent Betti numbers $\beta_q^{\alpha,0}$ (blue curve) of PDB ID 5CYT that are calculated from Gudhi, DioDe, and HERMES, together with the smallest non-zero eigenvalue $\lambda_q^{\alpha,0}$ (red curve) that are obtained only from HERMES.

It can be seen that all of these three packages return exactly the same persistent Betti numbers, suggesting that the calculation of our package HERMES is reliable. Additionally, the values of the smallest non-zero eigenvalues $\lambda_0^{\alpha,0}$ and $\lambda_1^{\alpha,0}$ increase around 1.86 Å, indicating the dramatic topological changes at this point. Similarly, with the increment of the α, the curve of $\lambda_2^{\alpha,0}$ also records the topological and geometric changes at a specific

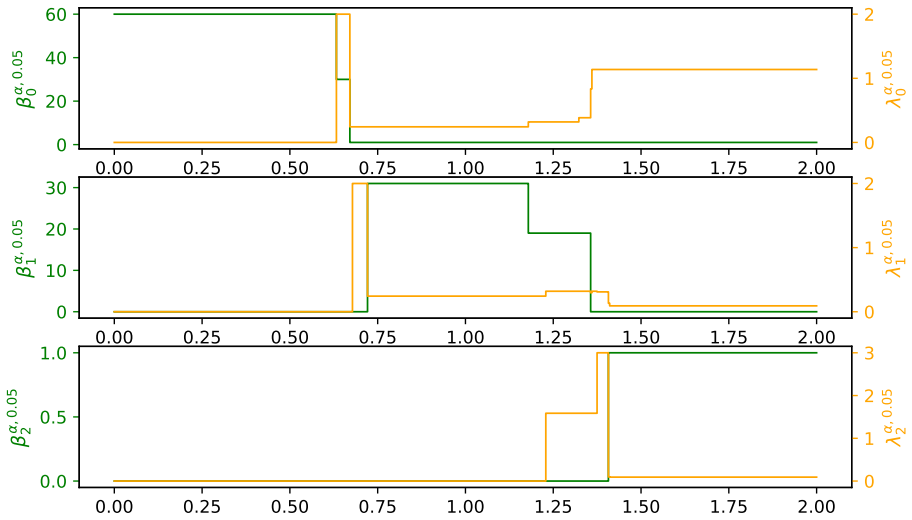

Figure 5.5: Illustration of the harmonic spectra $\beta_0^{\alpha,0.05}$, $\beta_0^{\alpha,0.05}$, and $\beta_2^{\alpha,0.05}$ (green curves from top chart to bottom chart) and the smallest non-zero eigenvalue $\lambda_0^{\alpha,0.05}$, $\lambda_1^{\alpha,0.05}$, and $\lambda_2^{\alpha,0.05}$ (yellow curves from top chart to bottom chart) of the C_{60} molecule (the bottom left chart in Figure 5.6) at different filtration value α calculated from HERMES. Here, the x-axis represents the radius filtration value α (unit: Å), the left-y-axes represents the number of zero eigenvalues of $\mathcal{L}_0^{\alpha,0.05}$, $\mathcal{L}_1^{\alpha,0.05}$, and $\mathcal{L}_1^{\alpha,0.05}$ from top to bottom, and the right-y-axes represents the first non-zero eigenvalue of $\mathcal{L}_0^{\alpha,0.05}$, $\mathcal{L}_1^{\alpha,0.05}$, and $\mathcal{L}_2^{\alpha,0.05}$ from top to bottom.

filtration value. The use of non-harmonic spectra for biophysical modeling was described in our earlier work [11].

To be noted, HERMES can also deal with the qth-order p-persistent Laplacians $\mathcal{L}_q^{\alpha,p}$. Figure 5.8 illustrates the persistent Betti numbers $\beta_0^{\alpha,0.5}$, $\beta_1^{\alpha,0.05}$, and $\beta_2^{\alpha,0.5}$ (green curves) and the smallest non-zero eigenvalue $\lambda_0^{\alpha,0.5}$, $\lambda_1^{\alpha,0.5}$, and $\lambda_2^{\alpha,0.5}$ (yellow curves) of 5CYT that are computed from HERMES, demonstrating the capacity of HERMES for the direct calculation of the persistent spectra of $\mathcal{L}_q^{\alpha,p}$ ($p > 0$). Compared with the middle chart of Figure 5.7, $\beta_1^{\alpha,0.5}$ in the middle chart of Figure 5.8 is always smaller than $\beta_1^{\alpha,0}$ at the same filtration α. Moreover, $\lambda_1^{\alpha,0.5}$ also goes up around 1.86 Å, which has the same behavior as $\lambda_1^{\alpha,0}$. Similar behaviors can be also observed from the bottom charts of Figure 5.7 and Figure 5.8.

Furthermore, HERMES can be used to detect the abnormality of a protein structure.

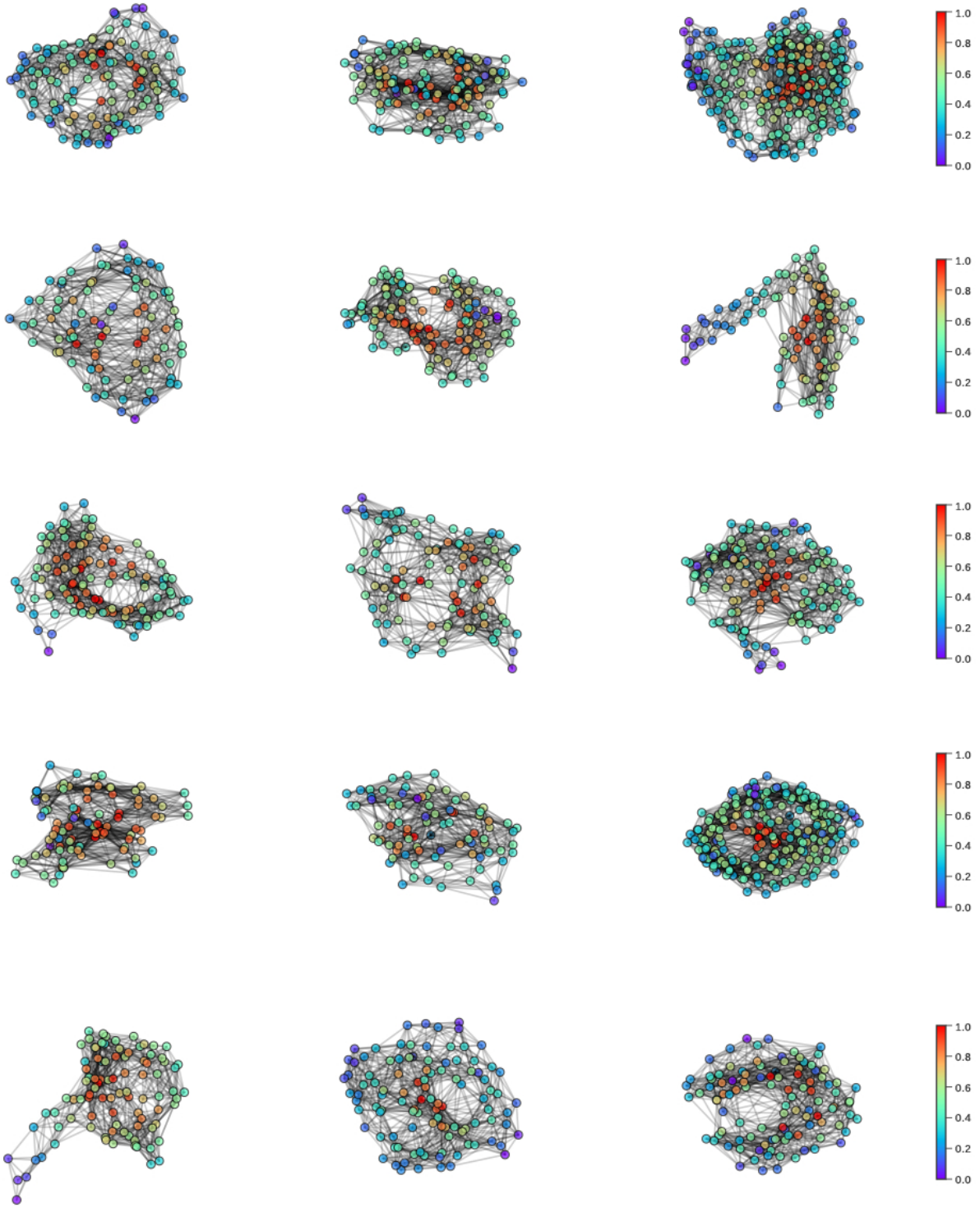

Figure 5.6: The alpha carbon network plots of 15 proteins: PDB IDs 1CCR, 1NKO, 1O08, 1OPD, 1QTO, 1R7J, 1V70, 1W2L, 1WHI, 2CG7, 2FQ3, 2HQK, 2PKT, 2VIM, and 5CYT from left to right and top to bottom. The color represents the normalized diagonal element of the accumulated Laplacian at each alpha carbon atom.

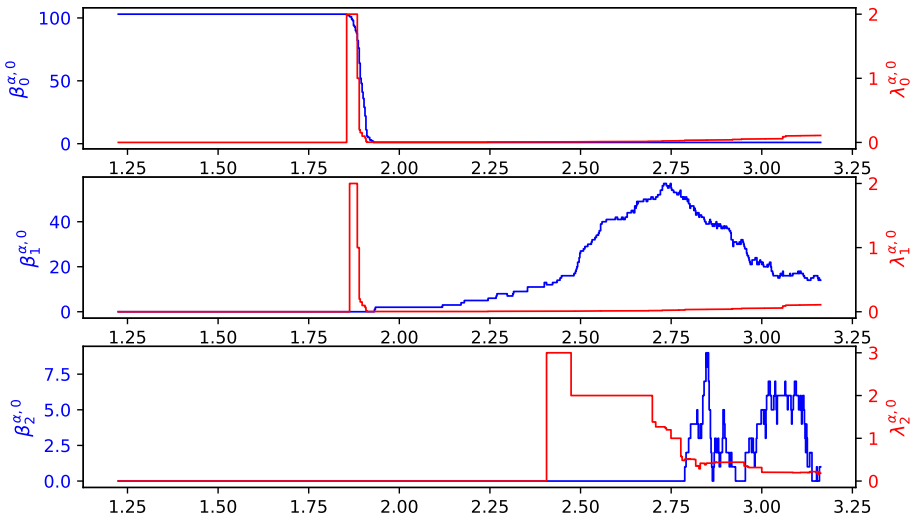

Figure 5.7: Illustration of the harmonic spectra $\beta_q^{\alpha,0}$ (blue curve) and the smallest non-zero eigenvalue $\lambda_q^{\alpha,0}$ (red curve) of PDB ID 5CYT (the bottom left chart in Figure 5.6) at different filtration values α when $q = 0, 1, 2$. The $\beta_q^{\alpha,0}$ are calculated from Gudhi, DioDe, and HERMES, and $\lambda_q^{\alpha,0}$ are obtained only from HERMES. Here, the x-axis represents the radius filtration value α (unit: Å), the left-y-axis represents the number of zero eigenvalues of $\mathcal{L}_q^{\alpha,0}$, and the right-y-axis represents the first non-zero eigenvalue of $\mathcal{L}_q^{\alpha,0}$. Note that the harmonic spectra from the three methods are indistinguishable.

Figure 5.9 (a) shows a 3D secondary structure of PDB 1O08, where the balls represent the alpha carbon atoms. The light blue, purple, and orange colors represent helix, sheet, and random coils of PDB ID 1O08. Figure 5.9 (b) depicts its harmonic spectra $\beta_q^{\alpha,0}$ (blue curve) and the smallest non-zero eigenvalue $\lambda_q^{\alpha,0}$ (red curve). Notably, two unusual onsets of $\beta_0^{\alpha,0}$ and $\beta_1^{\alpha,0}$ are detected when $\alpha << 1.9$ Å, indicating something is wrong with the structure data. Usually, the distance between the two alpha carbon atoms is around 3.8 Å. By examining the structure of PDB 1O08, we found that two pairs of alpha carbon atoms in PDB 1O08 have abnormal distances as marked with black frames. The distance of alpha carbon atoms in the upper box is 2.914 Å and that in the lower box is 2.996 Å, which are too short. The plots of the other proteins can be found in the Appendix. Similar structural defects were detected for PDB IDs 1V70, 2HQK, 2PKT, and 2VIM.

Although our package provides additional geometric information by calculating the

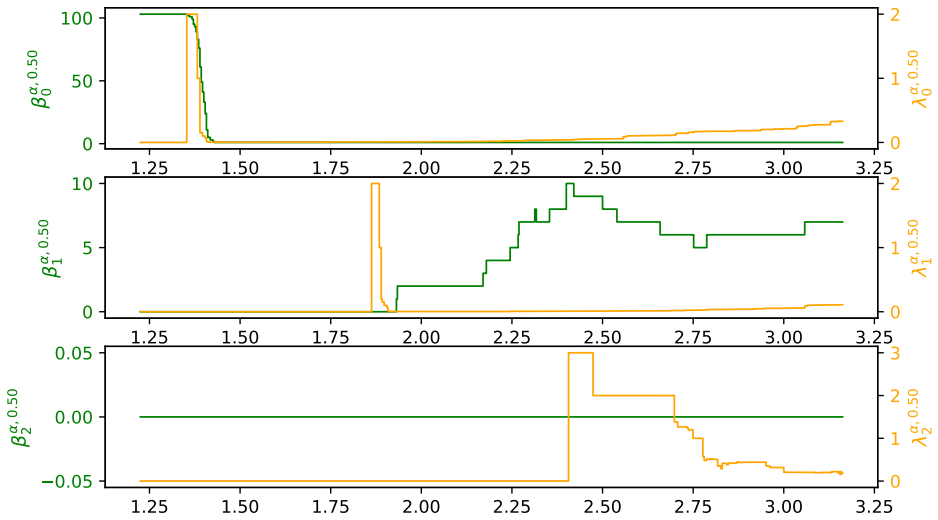

Figure 5.8: Illustration of the harmonic spectra $\beta_0^{\alpha,0.5}$, $\beta_0^{\alpha,0.5}$, and $\beta_2^{\alpha,0.5}$ (green curves from top chart to bottom chart) and the smallest non-zero eigenvalue $\lambda_0^{\alpha,0.5}$, $\lambda_1^{\alpha,0.5}$, and $\lambda_2^{\alpha,0.5}$ (yellow curves from top chart to bottom chart) of PDB ID 5CYT (the bottom left chart in Figure 5.6) at different filtration values α calculated from HERMES. Here, the x-axis represents the radius filtration value α (unit: Å), the left-y-axes represents the number of zero eigenvalues of $\mathcal{L}_0^{\alpha,0.5}$, $\mathcal{L}_1^{\alpha,0.5}$, and $\mathcal{L}_1^{\alpha,0.5}$ from top to bottom, and the right-y-axes represents the first non-zero eigenvalue of $\mathcal{L}_0^{\alpha,0.5}$, $\mathcal{L}_1^{\alpha,0.5}$, and $\mathcal{L}_2^{\alpha,0.5}$ from top to bottom.

non-harmonic spectra of qth-order persistent Laplacians, there are two limitations of HERMES. First, the construction of the Vietoris–Rips complex is the primary bottleneck in the calculation of non-harmonic spectra of persistent Laplacian matrices (PLMs). Additionally, the input format of HERMES is point cloud data. Other input formats, such as pairwise distances, point cloud with van der Waals radii, and volumetric density are not supported. These limitations will be addressed in our future implementation.

5.4 Discussion and Conclusion

While spectral graph theory has had tremendous success in data science to capture the geometric and topological information, it is limited by representing a graph structure at a given characteristic length scale, which hinders its practical application in data analysis. Motivated by the persistent (co)homology in dealing with a given initial data by

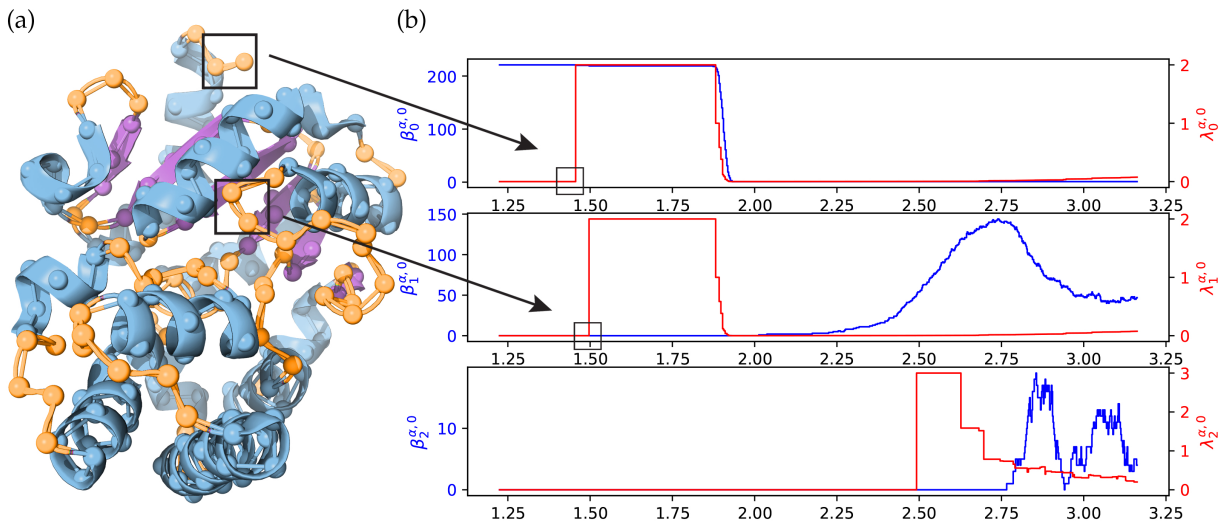

Figure 5.9: (a) The 3D secondary structure of PDB ID 1O08. The blue, purple, and orange colors represent helix, sheet, and random coils of PDB ID 1O08. The ball represents the alpha carbon of PDB ID 1O08. (b) Illustration of the harmonic spectra $\beta_q^{\alpha,0}$ (blue curve) and the smallest non-zero eigenvalue $\lambda_q^{\alpha,0}$ (red curve) of PDB ID 1O08 at different filtration values α when $q = 0, 1, 2$. The $\beta_q^{\alpha,0}$ are calculated from Gudhi, DioDe, and HERMES, and $\lambda_q^{\alpha,0}$ are calculated only from HERMES. Here, the x-axis represents the radius filtration value α (unit: Å), the left-y-axis represents for the number of zero eigenvalue of $\mathcal{L}_q^{\alpha,0}$, and the right-y-axis represents for the non-zero eigenvalues of $\mathcal{L}_q^{\alpha,0}$. Note that the harmonic spectra from three methods are indistinguishable.

constructing a family of simplicial complexes to track their topological invariants, and the multiscale graphs by creating a set of spectral graphs aiming to extract rich geometric information, we proposed persistent spectral graph (PSG) theory as a unified multiscale paradigm for simultaneous geometric and topological analysis [192]. PSG theory has stimulated mathematical analysis and algorithm development [151], as well as applications to drug discovery [181], and protein flexibility analysis [11].

To enable broad and convenient applications of the PSG method, we present an open-source software package called highly efficient robust multidimensional evolutionary spectra (HERMES). For a given point-cloud dataset, HERMES creates persistent Laplacian matrices (PLMs) at various topological dimensions via filtration. The spectrum of PLMs includes harmonic parts and non-harmonic parts. It turns out that the harmonic part spans the kernel spaces of PLMs and carries the full topological information of the

dataset. As a result, HERMES delivers the same topological data analysis (TDA) as does persistent homology. The non-harmonic part of PLMs provides valuable geometric analysis of the shape of data at various topological dimensions. The smallest non-zero eigenvalues are found to be very sensitive to data abnormality. In the present HERMES, both the alpha complex and the Vietoris–Rips complex are implemented. Due to the potentially large number of simplicies, the eigenvalue problem of persistent Laplacian for the Vietoris–Rips complex becomes memory-intensive for large systems. This difficulty may be overcome with approximate eigenvalue solvers. We will continue improving the efficiency of HERMES. HERMES has been extensively validated for its accuracy, robustness, and reliability by standard test datasets and a large number of complex protein structures, including comparison with Gudhi and DioDe.

CHAPTER 6

APPLICATIONS IN MATHEMATICAL MODELING OF VIROLOGY

6.1 Mutations on COVID-19 diagnostic targets

Severe acute respiratory syndrome coronavirus 2 (SARS-CoV-2), which was first reported in Wuhan in December 2019, is an unsegmented positive-sense single-stranded RNA virus that belongs to the β-coronavirus genus and coronaviridae family. Coronaviruses are some of the most sophisticated viruses with their genome size ranging from 26 to 32 kilobases in length. Caused by SARS-CoV-2, the coronavirus disease 2019 (COVID-19) pandemic outbreak has spread to more than 200 countries and territories with more than 15,012,731 infection cases and 619,150 fatalities worldwide by July 23, 2020 [193]. Additionally, travel restrictions, quarantines, and social distancing measures have essentially put the global economy on hold. Furthermore, we remain without efficacious testing, medications and vaccines for COVID-19. Undoubtedly, effective and widely available COVID-19 diagnostic testing, medications and vaccines would not only save lives, but would play a crucial role in a recovering worldwide economic[1].

There are three types of diagnostic tests for COVID-19, namely polymerase chain reaction (PCR) tests, antibody tests, and antigen tests. PCR tests detect the genetic material from the virus. Antibody tests, also called serological tests, examine the presence of antibodies produced from immune response to the virus infection. Antigen tests detect the presence of viral antigens, e.g., parts of the viral spike protein. PCR tests are relatively more accurate but take time to show the test result. The protein tests based on antibody or antigen can display test results in minutes but are relatively insensitive and subject to host immune response.

[1]This work is published on Nov 2020. No vaccines and medications available for COVID-19 at that time.

PCR diagnostic test reagents were designed based on early clinical specimens containing a full spectrum of SARS-CoV-2 [194], particularly the reference genome collected on January 5, 2020, in Wuhan (SARS-CoV-2, NC004718) [91]. Approved by the United States (US) Food and Drug Administration (FDA), the US Centers for Disease Control and Prevention (CDC) has detailed guidelines for COVID-19 diagnostic testing, called "CDC 2019-Novel Coronavirus (2019-nCoV) Real-Time RT-PCR Diagnostic Panel" (https://www.fda.gov/media/134922/download). The US CDC has designated two oligonucleotide primers from regions of the virus nucleocapsid (N) gene, i.e., N1 and N2, as probes for the specific detection of SARS-CoV-2. The panel has also selected an additional primer/probe set, the human RNase P gene (RP), as control samples. Many other diagnostic primers and probes based on RNA-dependent RNA polymerase (RdRP), envelope (E), and nucleocapsid (N) genes have been designed [195] and/or designated by the World Health Organization (WHO) as shown in Table S1 of the Supporting Material, which provides the details of 54 commonly used diagnostic primers and probes [196]. The diagnostic kits are often static over time, yet SARS-CoV-2 is undergoing fast mutations. Hence, it is reported that different primers and probes show nonuniform performance [197, 198, 199].

In this study, we genotype 31421 SARS-CoV-2 genome isolates in the globe and reveal numerous mutations on the COVID-19 diagnostic targets commonly used around the world, including those designated by the US CDC. We identify and analyze the SARS-CoV-2 mutation positions, frequencies, and encoded proteins in the global setting. These mutations may impact the diagnostic sensitivity and specialty, and therefore, they should be considered in designing new testing kits as the current effort in COVID-19 testing, prevention, and control. We propose diagnostic target selection and optimization based on nucleotide-based and gene-based mutation-frequency analysis.

6.1.1 Results and Analysis

Genotyping analysis We first genotype 31421 SARS-CoV-2 genome samples from the globe as of July 23, 2020. The genotyping results unravel 13402 single mutations among these virus isolates. Typically, a SARS-CoV-2 isolate can have eight co-mutations on average. A large number of mutations may occur on all of the SARS-CoV-2 genes and have broad effects on diagnostic kits, vaccines, and drug developments. Moreover, we cluster these mutations by k-means methods, resulting in globally at least six distinct subtypes of the SARS-CoV-2 genomes, from Cluster I to Cluster VI. Table 6.1 shows the mutation distribution clusters with sample counts (SC) and total single mutation counts (MC) in 20 countries.

Table 6.1: The mutation distribution clusters with sample counts (SC) and total single mutation counts (MC). The listed countries are United States (US), Canada (CA), Australia (AU), Germany (DE), France (FR), United Kingdom (UK), Italy (IT), Russia (RU), China (CN), Japan (JP), Korean (KR), India (IN), Iceland (IS), Brazil (BR), Spain (ES), Belgium (BE), Saudi Arabia (SA), Turkey (TR), Peru(PE), and Chile (CL).

	Cluster I		Cluster II		Cluster III		Cluster IV		Cluster V		Cluster VI	
Country	SC	MC	SC	MC	SC	MC	SC	MC	SC	MC	SC	MC
US	3252	24846	2013	14737	286	3686	2366	27012	562	3798	304	2706
CA	113	835	80	561	9	106	42	417	84	525	33	290
AU	173	1204	587	5048	75	1010	195	2127	165	885	132	1076
DE	69	504	25	121	5	58	26	209	27	144	43	366
FR	100	718	14	55	2	22	48	523	74	465	10	83
UK	295	2328	1927	12777	2171	27636	1623	16123	1890	11835	2919	25576
IT	1	8	8	104	33	561	24	308	57	283	24	192
RU	7	52	2	32	19	219	7	53	32	187	119	968
CN	3	22	287	1155	2	32	7	50	8	35	3	26
JP	18	134	243	1001	23	272	9	79	23	139	191	1676
KR	0	0	58	327	0	0	0	0	0	0	0	0
IN	29	212	268	3045	200	2703	399	4840	141	847	51	487
IS	66	446	103	595	30	345	10	89	152	924	59	525
ES	4	33	163	1198	3	33	37	365	170	1103	42	359
BR	3	26	7	51	78	1009	2	10	7	42	63	591
BE	56	411	85	400	66	783	115	1031	230	1381	141	1239
SA	16	110	9	61	0	0	14	126	17	133	1	7
TR	0	0	28	339	13	158	50	476	4	28	31	273
PE	2	12	5	36	10	124	5	48	9	58	2	17
CL	13	91	27	282	21	285	49	665	32	200	20	169

All of the countries are involved in six clusters except Korean (KR), Saudi Arabia (SA),

and Turkey (TR). Among them, China initially had samples only in clusters II and its sample distributions reached to other Clusters after March 2020. Cluster I, II, and IV dominate in the United States. Germany (DE) and France (FR) samples are mainly in Cluster I, IV, and VI. Italy (IT) samples are mainly in Clusters III, IV, V, and VI. Samples in Turkey (TR) are mainly in Cluster II, III, IV, and VI. Japan (JP) samples are dominated in Cluster II and VI, Korea (KR) samples belong to Cluster II only. Cluster II is common to all countries. Figure 6.1 depicts the distribution of six distinct clusters in the world. The light blue, dark blue, green, red, pink, and yellow represent Cluster I, Cluster II, Cluster III, Cluster IV, Cluster V, and Cluster VI, respectively. The color of the dominated Cluster decides the base color of each country. To be noted, although some countries have a lot of confirmed sequences, a very limited number of complete genome sequences are deposited in the GISAID, which causes the geographical bias in the Table 6.1.

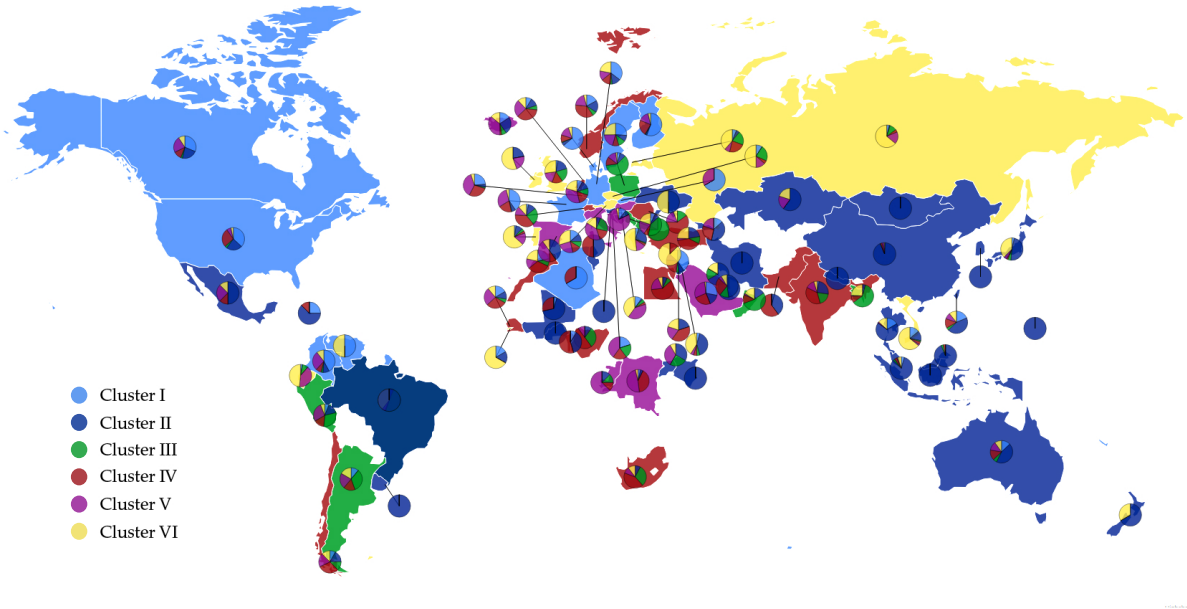

Figure 6.1: The scatter plot of six distinct clusters in the world in July 2020. The light blue, dark blue, green, red, pink, and yellow represent Cluster I, Cluster II, Cluster III, Cluster IV, Cluster V, and Cluster VI, respectively. The base color of each country is decided by the color of the dominated Cluster.

Mutations on Diagnostic Targets

Table 6.2: Summary of mutations on COVID-19 diagnostic primers and probes and their occurrence frequencies in clusters. Here, SC is the sample counts and MC is the mutation counts.

Primer	MC	SC	Cluster I	Cluster II	Cluster III	Cluster IV	Cluster V	Cluster VI
RX7038-N1 primer (Fw)[a]	15	79	5	14	12	28	14	6
RX7038-N1 primer (Rv)[a]	17	113	1	66	14	9	2	21
RX7038-N2 primer (Fw)[a]	7	60	3	10	24	21	1	1
RX7038-N2 primer (Rv)[a]	6	50	2	17	6	15	3	7
RX7038-N3 primer (Fw) [200]	13	287	4	224	13	26	14	6
RX7038-N3 primer (Rv) [200]	12	70	4	10	7	39	6	4
N1-U.S.-P [196]	15	856	4	782	20	31	15	4
N2-U.S.-P [196]	11	70	10	40	4	12	4	0
N3-U.S.-P [196]	16	84	5	27	15	21	10	6
N-Sarbeco-F[b] [195]	12	63	4	20	10	15	10	4
N-Sarbeco-P[b][195]	12	116	1	19	30	42	15	9
N-Sarbeco-R[b][195]	17	156	37	26	4	80	5	4
N-China-F [196]	23	26280	38	226	10873	139	17	14987
N-China-R [196]	17	217	5	15	17	157	8	15
N-China-P [196]	7	20	1	4	6	8	1	0
N-HK-F [196]	5	149	1	2	74	7	1	64
N-HK-R [196]	14	84	14	12	14	35	4	5
N-JP-F [196]	10	66	5	10	9	16	26	0
N-JP-P [196]	9	32	0	5	1	16	3	7
N-TL-F [196]	17	149	1	84	14	31	13	6
N-TL-R [196]	17	115	29	7	7	66	3	3
N-TL-P [196]	11	45	1	5	13	5	1	20
E-Sarbeco-F1[c]	5	23	0	0	10	9	2	2
E-Sarbeco-R2[c]	4	18	0	6	5	1	6	0
E-Sarbeco-P1[c]	9	48	1	29	6	9	3	0
nCoV-IP2-12669Fw[c]	3	50	0	17	12	11	0	10
nCoV-IP2-12759Rv[c]	11	739	123	244	77	168	127	0
nCoV-IP2-12696bProbe(+)[c]	8	17	2	4	1	6	4	0
nCoV-IP4-14059Fw[c]	3	9	0	0	7	2	0	0
nCoV-IP4-14146Rv[c]	11	38	7	7	9	9	1	5
nCoV-IP4-14084Probe(+)[c]	11	49	3	12	6	19	5	4
RdRP-SARSr-F2[d]	5	89	2	1	5	37	44	0
RdRP-SARSr-R1[d][195]	3	4	2	0	0	2	0	0
RdRP-SARSr-P2[d][195]	4	10	0	6	2	2	0	0
ORF1ab-China-F [196]	4	19	0	4	2	6	5	2
ORF1ab-China-R [196]	0	0	0	0	0	0	0	0
ORF1ab-China-P [196]	14	61	1	6	30	11	3	10
ORF1b-nsp14-HK-F [196]	6	12	2	1	6	3	0	0
ORF1b-nsp14-HK-R[196]	9	89	3	9	52	14	6	5
ORF1b-nsp14-HK-P[196]	6	37	2	1	9	13	0	12
SC2-F[e]	11	88	0	5	34	29	13	7
SC2-R[e]	0	0	0	0	0	0	0	0
NIID_WH-1_F501[201]	13	255	0	205	25	18	3	4
NIID_WH-1_R913[201]	14	128	1	94	9	18	4	2
NIID_WH-1_F509[201]	10	30	7	5	7	6	3	2
NIID_WH-1_R854[201]	9	261	63	25	33	117	5	18
NIID_WH-1_Seq[201] F519	19	130	8	89	17	11	3	2
NIID_WH-1_Seq R840[201]	12	66	6	9	21	8	3	19
WuhanCoV-spk1-f[201]	14	433	265	22	11	123	8	4
WuhanCoV-spk1-r[201]	4	10	0	2	3	1	2	2
NIID_WH-1_F24381[201]	20	494	275	30	16	153	13	7
NIID_WH-1_R24873[201]	5	15	1	4	3	7	0	0
NIID_WH-1_Seq_F24383[201]	21	503	275	30	22	153	13	10
NIID_WH-1_Seq_R24865[201]	6	17	2	4	5	6	0	0

Table 6.2 provides all mutations on various primers and probes and their occurring frequencies in various clusters, where SC is the sample counts and MC is the mutation counts. More detailed mutation information is given in Tables S4-S56 of the Supporting Material. We plot the mutation position and frequency for 54 primers and probes in this work in Figure 6.2 - Figure 6.6.

It is noted that N-China-F [196] is the mostly-used reagent among all primers/probes, but the primer target gene of SARS-CoV-2 has 15 mutations involving thousands of samples, which may account for low efficacy of certain COVID-19 diagnostic kits in China according to this website. Note that primers and probes typically have a small length of around 20 nucleotides.

Currently, most primers and probes used in the US target are the N gene [196]. However, Table 6.2 shows that a plurality of mutations has been found in all of the targets of the US CDC designated COVID-19 diagnostic primers. The targets of N gene primers and probes used in Japan, Thailand, and China, including Hong Kong, have undergone multiple mutations involving many clusters. Therefore, the N gene may not be an optimal target for diagnostic kits, and the current test kits targeting the N gene should be updated accordingly for testing accuracy.

It can be seen that so far, no mutation has been detected on ORF1ab-China-R and SC2-R, showing that they are two relatively reliable diagnostic primers. Notably, the targets of four E gene primers and probes have only six mutations. Also, no mutation has been found on the targets of ORF1ab-China-R and SC2-R. However, the target of nCoV-IP2-12759R recommended by Institute Pasteur, Paris has six mutations. Overall, targets of the envelope and RNA-dependent RNA polymerase based primers and probes have fewer mutations than the N gene. This observation leads to an assumption that the N gene is particularly prone to mutations.

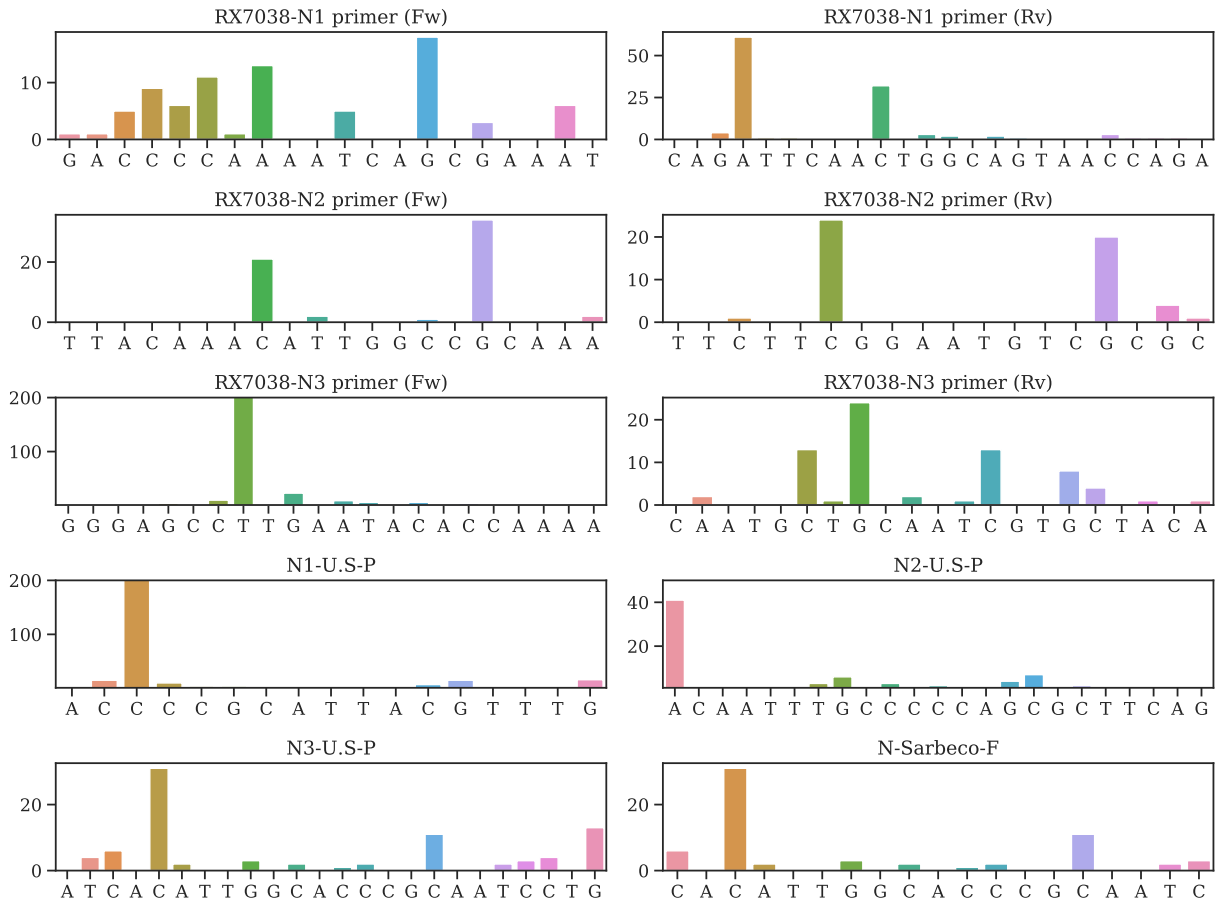

Figure 6.2: Illustration of mutation positions and frequencies on the primer and/or probes of RX7038-N1 primer (Fw), RX7038-N1 primer (Rv), RX7038-N2 primer (Fw), RX7038-N2 primer (Rv), RX7038-N3 primer (Fw), RX7038-N3 primer (Rv), N1-U.S.-P, N2-U.S.-P, N3-U.S.-P, N-Sarbeco-F.

6.1.2 Discussions

Mechanisms of mutation and mutation impact on diagnostics The accumulation of the frequency of virus mutations is due to natural selection, polymerase fidelity, cellular environment, features of recent epidemiology, random genetic drift, host immune responses, gene editing [202], replication mechanism, etc [203, 204]. SARS-CoV-2 has a higher fidelity in its transcription and replication process than other single-stranded RNA viruses because it has a proofreading mechanism regulated by NSP14 [205]. However, 13402 single mutations have been detected from 31421 SARS-CoV-2 genome isolates.

Due to technical constraints, genome sequencing is subject to errors. Some "muta-

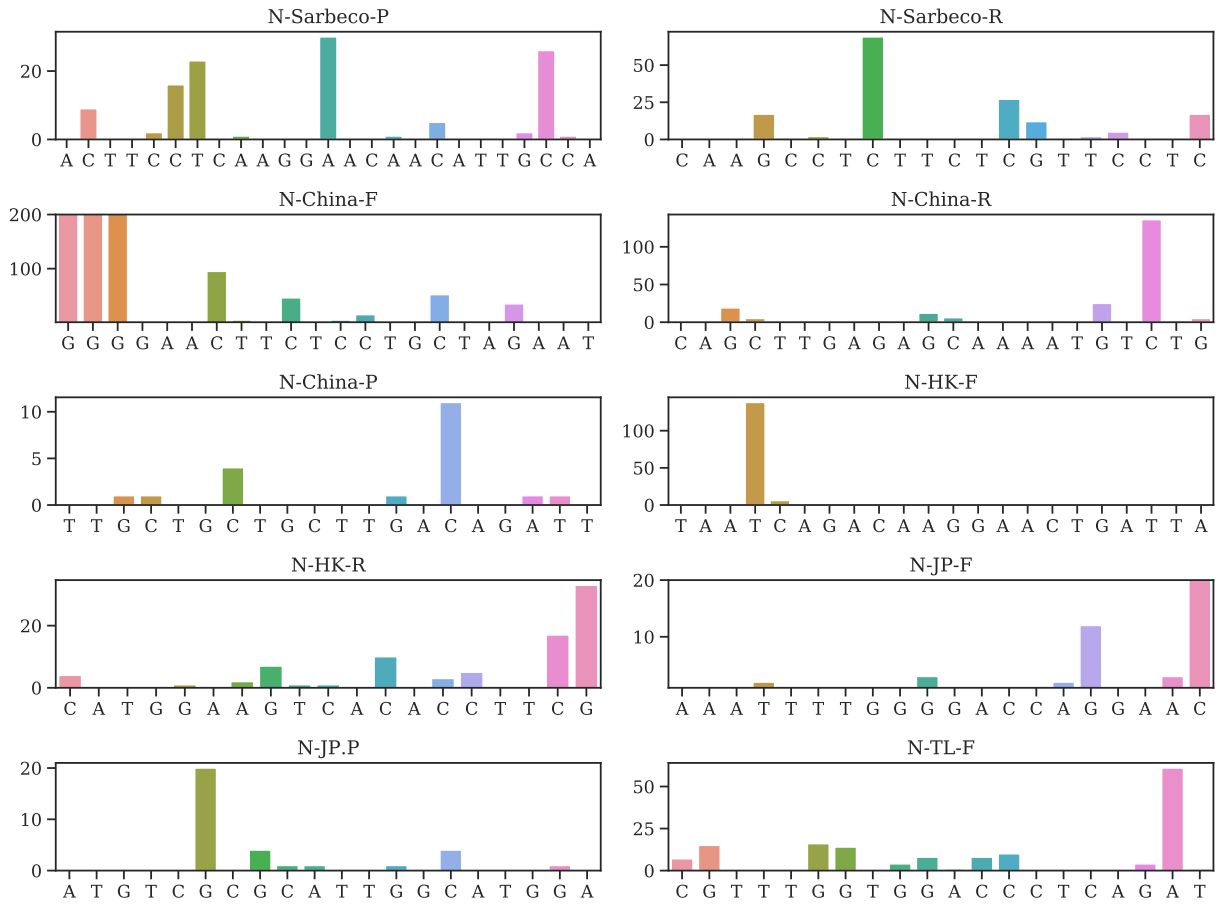

Figure 6.3: Illustration of mutation positions and frequencies on the primer and/or probes of N-Sarbeco-P, N-Sarbeco-R, N-China-F, N-China-R, N-China-P, N-HK-F, N-HK-R, N-JP-F, N-JP-P, N-TL-F.

tions" might result from sequencing errors, instead of actual mutations. Additionally, mRNA editing, such as APOBEC [202], in defending virus invasion in the human immune system can create fatal mutations. Both cases may lead to single-nucleotide polymorphisms (SNPs) without a descendant. We report that among all of 31421 genome isolates, 13402 individual mutations have at least one descendant.

It is well known that the sensitivity of diagnostic primers and probes depends on their target positions. Specifically, the beginning part of a primer or probe is not as important as its ending part. A high-frequency mutation on the right end of a primer or probe position of a target would possibly produce more false-negatives in diagnostics. Also, importantly,

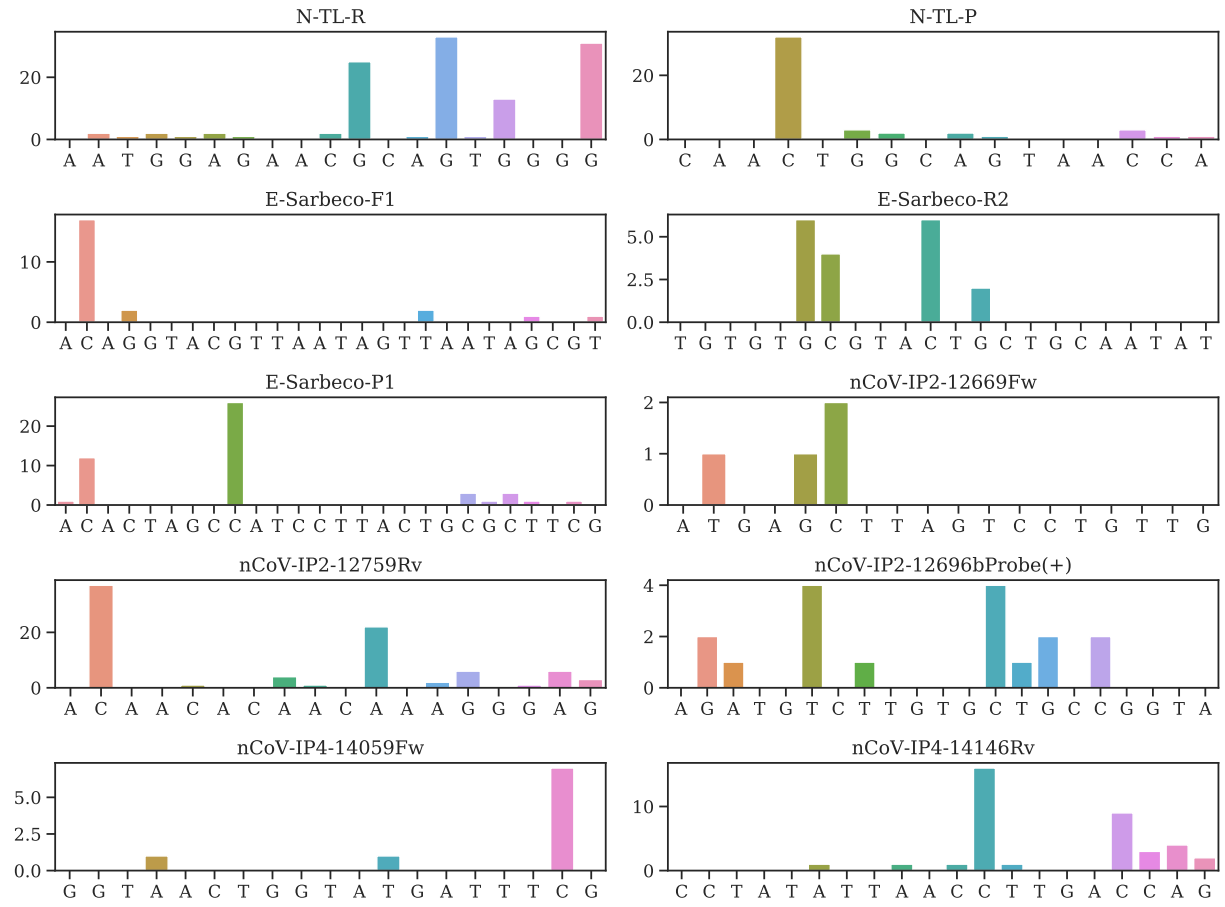

Figure 6.4: Illustration of mutation positions and frequencies on the primer and/or probes of N-TL-R, N-TL-P, E-Sarbeco-F1, E-Sarbeco-R2, E-Sarbeco-P1, nCoV-IP2-12669Fw, nCoV-IP2-12759Rv, nCoV-IP2-12696bProbe(+), nCoV-IP4-14059Fw, nCoV-IP4-14146Rv.

for primers involving significant mutations, polymerase chain reaction (PCR) annealing temperatures are estimated based on correctly matched sequences [206]. Annealing temperatures for primers and probes involving mutations of are given in Tables S4-S56 of the Supporting Material.

Nucleotide-based diagnostic target optimization Table 6.2 shows that the degree of mutations on various diagnostic targets vary dramatically. Therefore, it is of great importance to know how to select an optimal viral diagnostics target to avoid potential mutations. We discuss such a target optimization via both nucleotide-based analysis and gene-based mutation analysis.

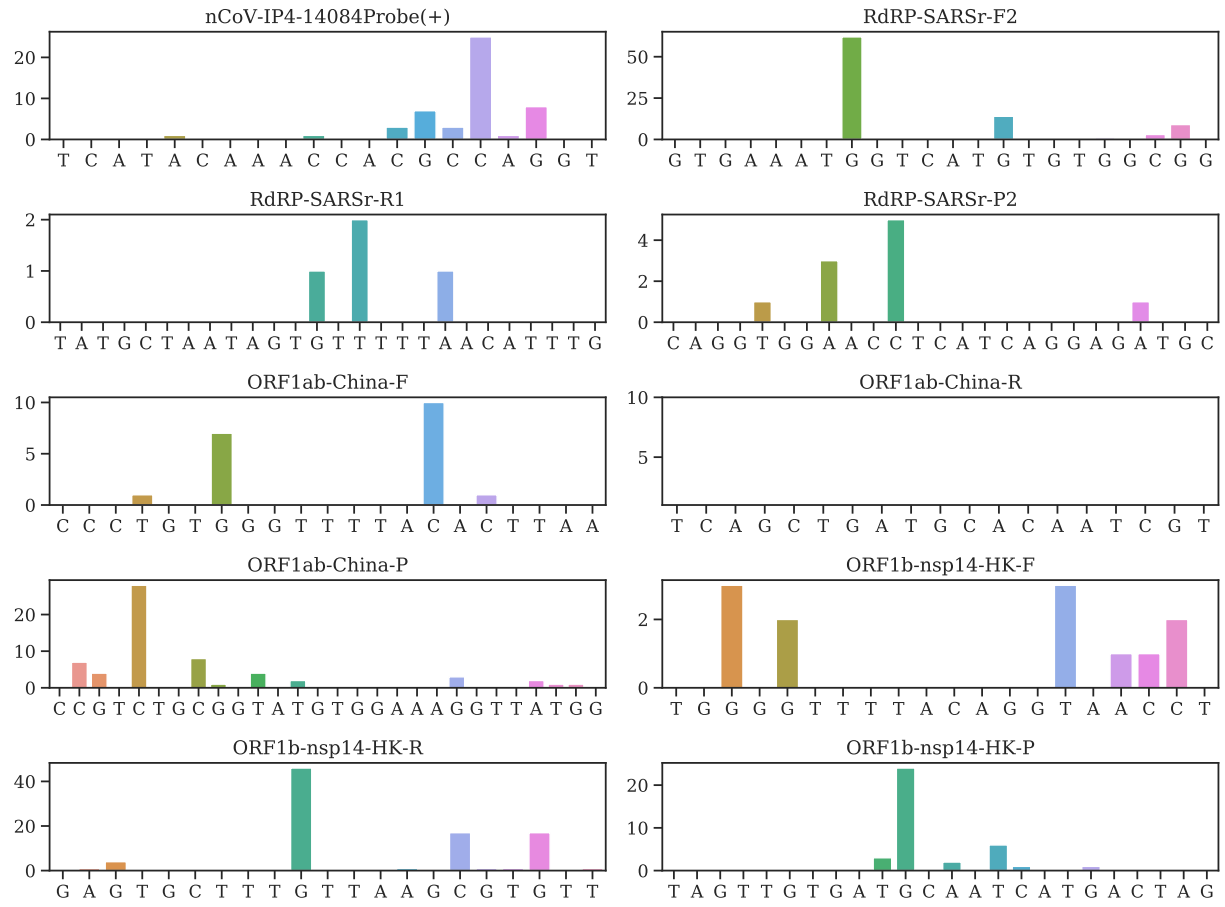

Figure 6.5: Illustration of mutation positions and frequencies on the primer and/or probes of nCoV-IP4-14084Probe(+), RdRP-SARSr-F2, RdRP-SARSr-R1, RdRP-SARSr-P2, ORF1ab-China-F, ORF1ab-China-R, ORF1ab-China-P, ORF1b-nsp14-HK-F, ORF1b-nsp14-HK-R, ORF1b-nsp14-HK-P.

Figure 6.7 illustrates the rates of 12 different types of mutations among 31421 SNP variants. It is interesting to note that 51.4% mutations on the SARS-CoV-2 are of C>T type, due to strong host cell mRNA editing knows as APOBEC cytidine deaminase [202]. Therefore, researchers should avoid cytosine bases as much as possible when designing the diagnostic test kits.

Gene-based diagnostic target optimization

To further understand how to design the most reliable SARS-CoV-2 diagnostic targets, we carry out gene-level mutation analysis. Figure 6.8 and Table 6.3 present the mutation ratio, i.e., the number of unique single-nucleotide polymorphisms (SNPs) over the

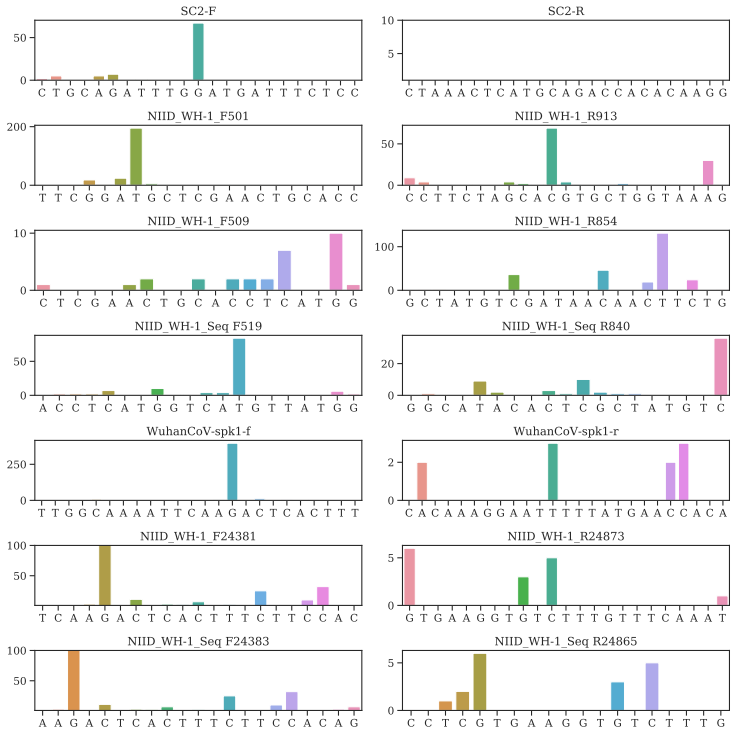

Figure 6.6: Illustration of mutation positions and frequencies on the primer and/or probes of SC2-F, SC2-R,NIID_WH-1_F501,NIID_WH-1_R913, NIID_WH-1_F509, NIID_WH-1_R85, NIID_WH-1_Seq F519, NIID_WH-1_Seq R840, WuhanCoV-spk1-f, WuhanCoV-spk1-r, NIID_WH-1_F24381, NIID_WH-1_R24873, NIID_WH-1_Seq F24383, NIID_WH-1_Seq R24865.

corresponding gene length, for each SARS-CoV-2 gene. A smaller mutation ratio for a given gene indicates a higher degree of conservativeness. Clearly, the ORF7b gene has the smallest mutation ratio of 0.155, while the ORF7a gene has the largest mutation ratio of 0.642. The N gene has the fourth-largest mutation rate of 0.558, which is very close to the largest ratio of 0.594 for the ORF3a gene and 0.559 for the ORF8 gene. Additionally, two ends of the SARS-CoV-2 genome, i.e., NSP1, NSP2, ORF10, N gene, ORF8, ORF7a, and ORF6, exception for ORF7b, have higher mutation ratios. Considering the mutation frequency, we introduce the mutation h-index, defined as the maximum value of h such that the given gene section has h single mutations that have each occurred at least h times. Normally, larger genes tend to have a higher h-index. Figure 6.8 shows that, with a moderate length, the N gene has the second-largest h-index of 44, which is close to the largest

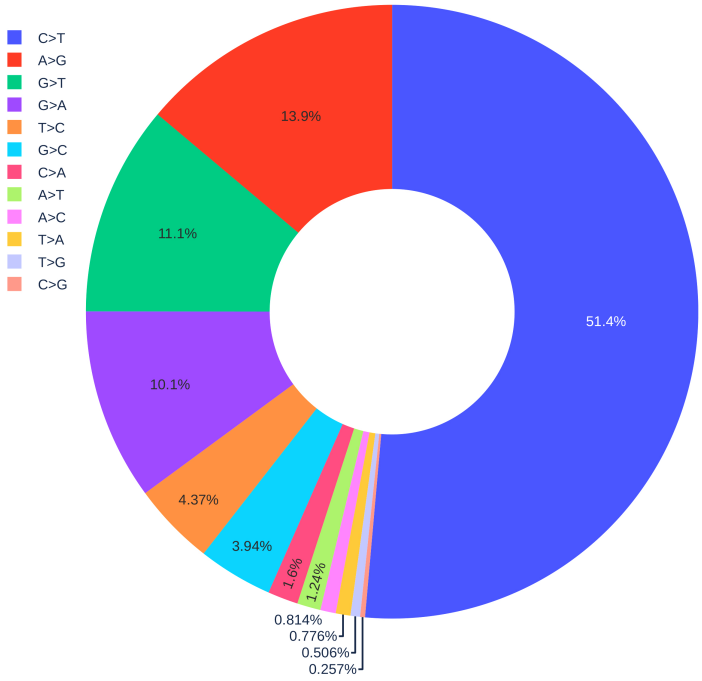

Figure 6.7: The pie chart of the distribution of 12 different types of mutations.

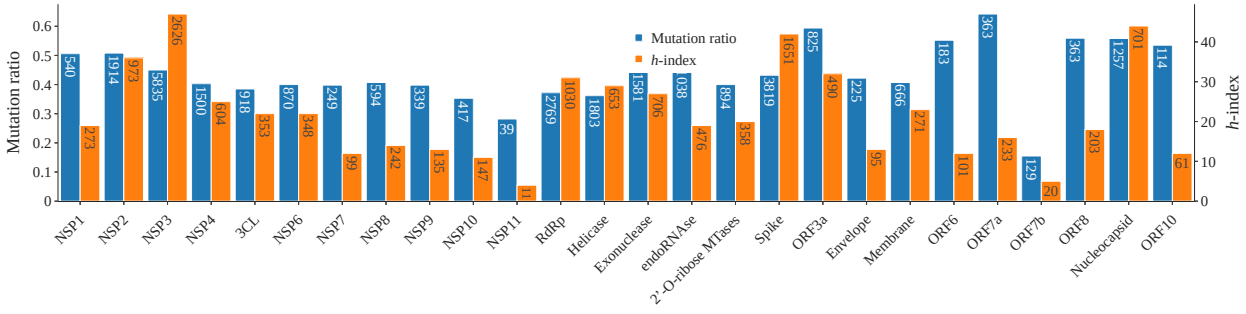

Figure 6.8: Illustration of SARS-CoV-2 mutation ratio and mutation h-index one various genes. For each gene, its length is given in the mutation ratio bar while the number of unique SNPs is given in the h-index bar.

h-index of 47 for NSP3. Therefore, selecting SARS-CoV-2 N gene primers and probes as diagnostic reagents for combating COVID-19 is not an optimal choice. Moreover, a few primers and probes used in Japan are designed on the spike and NSP2 gene. However, the high mutation ratio and h-index of spike and NSP2 gene indicate that these diagnostic reagents may not perform well. Furthermore, we design a website called Mutation Tracker to track the single mutations on 26 SARS-CoV-2 proteins, which will be an intuitive tool to inform other research on regions to be avoided in future diagnostic test development.

Table 6.3: Gene-specific statistics of SARS-CoV-2 single mutations on 26 proteins.

Gene type	Gene site	Gene length	Unique SNPs	mutation ratio	h-index
NSP1	266:805	540	273	0.506	19
NSP2	806:2719	1914	973	0.508	36
NSP3	2720:8554	5835	2626	0.450	47
NSP4	8555:10054	1500	604	0.403	25
NSP5(3CL)	10055:10972	918	353	0.385	22
NSP6	10973:11842	870	348	0.400	22
NSP7	11843:12091	249	99	0.398	12
NSP8	12092:12685	594	242	0.407	14
NSP9	12686:13024	339	135	0.398	13
NSP10	13025:13441	417	147	0.353	11
NSP11	13442:13480	39	11	0.282	4
RNA-dependent-polymerase	13442:16236	2796	1030	0.368	31
Helicase	16237:18039	1803	653	0.362	29
3′-to-5′ exonuclease	18040:19620	1581	706	0.447	27
endoRNAse	19621:20658	1038	476	0.459	19
2′-O-ribose methyltransferase	20659:21552	894	358	0.400	20
Spike protein	21563:25384	3819	1651	0.432	42
ORF3a protein	25393:26220	825	490	0.594	32
Envelope protein	26245:26472	225	95	0.422	13
Membrane glycoprotein	26523:27191	666	271	0.407	23
ORF6 protein	27202:27387	183	101	0.552	12
ORF7a protein	27394:27759	363	233	0.642	16
ORF7b protein	27756:27887	129	20	0.155	5
ORF8 protein	27894:28259	363	203	0.559	18
Nucleocapsid protein	28274:29533	1257	701	0.558	44
ORF10 protein	29558:29674	114	61	0.535	12

6.1.3 Conclusion

In summary, the targets of currently used COVID-19 diagnostic tests have numerous mutations that impact the diagnostic test accuracy in identifying COVID-19. There is a need for continued surveillance of viral evolution and diagnostic test performance, as the emergence of viral variants that are no longer detectable by certain diagnostics tests is a real possibility. A cocktail test kit is needed to mitigate mutations. We propose nucleotide-based and gene-based diagnostic target optimizations to design the most reliable diagnostic targets. We analyze a full list of SNPs for all 31421 genome isolates, including their positions and mutation types. This information, together with ranking of the degree of the conservativeness of SARS-CoV-2 genes or proteins given in Table 6.3, enables researchers to avoid non-conservative genes (or their proteins) and mutated nucleotide segments in designing COVID-19 diagnosis, vaccine, and drugs.

6.2 Mechanisms of SARS-CoV-2 evolution

The mechanism of mutagenesis is driven by various competitive processes [203, 204, 207, 208, 24], which can be categorized into 3 different scales with many factors as illustrated in Figure 6.9 **a**: 1) the molecular scale, 2) the organism scale, and 3) the population scale. From the molecular-scale perspective, the random shifts, replication errors, transcription errors, translation errors, viral proofreading, and viral recombination are the main driven sources. Moreover, the host gene editing induced by the adaptive immune response [24] and the recombination between the host and virus are the key-driven factors at the organism level. Furthermore, the natural selection popularized by Charles Darwin is a critical process, which favors mutations that have reproductive advantages for the virus to have adaptive traits in evolution. Such complicated mechanisms of viral mutagenesis make the comprehension of viral transmission and evolution a grand challenge.

Although there are 28,780 unique single mutations distributed evenly on the whole SARS-CoV-2 genome, the mutations on the S gene stand out among all 29 genes on SARS-

CoV-2 due to the mechanism of viral infection. Under assistant with host transmembrane protease, serine 2 (TMPRSS2), SARS-CoV-2 enters the host cell by interacting with its S protein and the host angiotensin-converting enzyme 2 (ACE2) [37] (See Figure 6.9 **b**). Later on, antibodies will be generated by the host immune system, aiming to eliminate the invading virus through direct neutralization or non-neutralizing binding [209, 210], which makes the S protein the main target for the current vaccines. Specifically, there is a short immunogenic fragment located on the S protein of SARS-CoV-2 that can facilitate the SARS-CoV-2 S protein binding with ACE2, which is called the receptor-binding domain (RBD) [211]. Studies have shown that the binding free energy (BFE) between the S RBD and the ACE2 is proportional to the infectivity [212, 213, 214, 37, 28]. Therefore, tracking and monitoring the RBD mutations and their corresponding BFE changes will expedite understanding the infectivity, transmission, and evolution of SARS-CoV-2, especially for the new SARS-CoV-2 variants, such as Alpha, Beta, Gamma, Delta, and Lambda, etc. [21]

The current prevailing variants Alpha, Beta, Gamma, Delta, Kappa, Theta, Lambda, and Mu carry at least one vital mutation at residues 452 and 501 on the S RBD [2]. Notably, in July 2020, we successfully predicted that residues 452 and 501 "have high chances to mutate into significantly more infectious COVID-19 strains" [41]. In the same work, we hypothesized that “natural selection favors those mutations that enhance the viral transmission" and provided the first evidence for infectivity-based natural selection. In other words, we revealed the mechanism of SARS-CoV-2 evolution and transmission based on very limited genome data in July 2020 [41]. Additionally, we predicted three categories of RBD mutations: 1) most likely (1149 mutations), 2) likely (1912 mutations), and 3) unlikely (625 mutations) [41]. Up to now, all of the RBD mutations we detected fall into our first category [102, 2]. Until now, all of the top 100 most observed RBD mutations have BFE change greater than the average BFE changes of -0.28kcal/mol (the average

[2]This work was published in 2020

BFE changes for all RBD mutations[215]). There are extremely low odds (i.e., $\frac{1}{1.27\times10^{30}}$) for 100 RBD mutations to accidentally have BFE changes simultaneously above the average value. This provides convincing evidence for our hypothesis that the transmission and evolution of new SARS-CoV-2 variants are governed by infectivity-based natural selection, despite all other competing mechanisms [41]. Our predictions rely on algebraic topology [100, 101, 4]-assisted deep learning [40, 41], but have been extensively validated [102, 99].

However, infectivity is not the only transmission pathway that governs viral evolution. Vaccine-resistant mutations or more precisely, antibody-resistant mutations, that can disrupt the protection of antibodies has become a viable mechanism for new variants to transmit among the vaccinated population since the vaccine was put on the market. In early January 2021, we have predicted that RBD mutations W353R, I401N, Y449D, Y449S, P491R, P491L, Q493P, etc., will weaken most antibody bindings to the S protein [102]. Later on, we have provided a list of most likely vaccine escape RBD mutations with high frequency, including S494P, Q493L, K417N, F490S, F486L, R403K, E484K, L452R, K417T, F490L, E484Q, and A475S [2]. Moreover, we have pointed out that Y449S and Y449H are two vaccine-resistant mutations, and “Y449S, S494P, K417N, F490S, L452R, E484K, K417T, E484Q, L452Q, and N501Y" are the top 10 mutations that will disrupt most antibodies with high-frequency [215]. As mentioned in Ref. [216], RBD mutations such as E484K/A, Y489H, Q493K, and N501Y found in late-stage evolved S variants “confer resistance to a common class of SARS-CoV-2 neutralizing antibodies", which suggests the viral evolution is also regulated by vaccine-resistant mutations.

6.2.1 Evolutionary trajectories of viral RBD single mutations

Studying the mechanisms of SARS-CoV-2 mutagenesis is beneficial to the understanding of viral transmission and evolution. The mainly driven force of viral evolution is regulated by natural selection, which is employed by two complementary transmission

pathways: 1) infectivity-based pathway and 2) vaccine-resistant pathway. We have discussed the infectivity-based pathways in Ref.[215] and [39]. This section focuses on the vaccine-resistant pathway and its impact on the transmission and evolution of SARS-CoV-2. To understand the mechanisms of vaccine-resistant mutations, we first analyze 1,983,328 complete SARS-CoV-2 genomes, and a total of 28,780 unique single mutations are decoded. Among them, there are 737 non-degenerate RBD mutations. The infectivity of SARS-CoV-2 is proportional to the BFE between the S RBD and ACE2 [212, 213, 214, 37, 28]. Therefore, the BFE change induced by a specific RBD mutation reveals whether the RBD mutation is an infectivity-strengthen or an infectivity-weaken mutation. Similarly, the BFE change between S RBD and antibody induced by a given mutation reveals whether this mutation will strengthen the binding between S and antibody or not.

Up to now, we have collected 130 antibody structures (see the Supporting Information S4), which includes Food and Drug Administration (FDA)-approved mAbs from Eli Lilly and Regeneron. For a specific RBD mutation, its antibody disruption count shows the number of antibodies that have antibody-S BFE changes smaller than -0.3 kcal/mol. The ACE2-S and antibody-S BFE changes induced by RBD mutations are predicted from our TopNetTree model [41], which is available at TopNetmAb. All of the predicted BFE changes induced by RBD mutations can be found at Mutation Analyzer. Figure 6.9 **c** illustrates the top 25 most observed RBD mutations. The height and color of each bar represent the ACE2-S BFE changes and frequency of each RBD mutation. The number at the top of each bar shows the antibody disruption count of each mutation. The detailed information can be viewed in Supplementary Information S4. It can be seen that 23 mutations have positive ACE2-S BFE changes, suggesting they are regulated by the infectivity-based transmission pathway.

Howbeit, 2 RBD mutations D427N and Y449S, have negative BFE changes. Notably, mutation Y449S has a significantly negative BFE change (-0.8112 kcal/mol) and a pretty large antibody disruption count (89), revealing a non-typical mechanism of mutagenesis.

Such a mutation with significantly negative ACE2-S BFE change together with a high antibody disruption count is called a vaccine-resistant or antibody-resistant mutation. Figure 6.9 **d** is the illustration of SARS-CoV-2 S protein (blue color) with human ACE2 (pink color), and the Y449 residue (purple color) is located on the random coil of the S protein. Among all of the vaccine-resistant mutations, Y449S has the highest frequency (1189). In addition, at residue 449, mutations Y449H, Y449N, Y449D are all vaccine-resistant mutations that have been observed in more than 20 SARS-CoV-2 genome isolates.

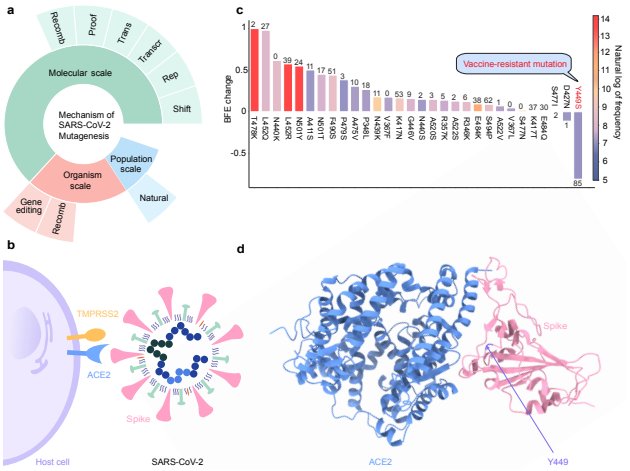

Figure 6.9: **a** The mechanism of mutagenesis. Nine mechanisms are grouped into three scales: 1) molecular-based mechanism (green color); 2) organism-based mechanism (red color); 3) population-based mechanism (blue color). The random shifts (Random), replication error (Rep), Transcription error (Transcr), viral proofreading (Proof), and recombination (Recomb) are the six molecular-based mechanisms. The gene editing and the host-virus recombination are the organism-based mechanism. In addition, the natural selection (Natural) is the population-based mechanism, which is the mainly driven source in the transmission of SARS-CoV-2. **b** A sketch of SARS-CoV-2 and its interaction with host cell. **c** Illustration of 25 single-site RBD mutations with top frequencies. The height of each bar shows the BFE change of each mutation, the color of each bar represents the natural log of frequency of each mutation, and the number at the top of each bar means the AI-predicted number of antibody and RBD complexes that may be significantly disrupted by a single site mutation. **d** Illustration of SARS-CoV-2 S protein with human ACE2. The blue chain represents the human ACE2, the pink chain represents the S protein, and the purple fragment on the S protein points out the two vaccine-resistant mutations Y449S/H.

To track the evolution trajectory of vaccine-resistant mutations, the BFE changes, log2 enrichment ratios [3], and log10 frequencies of RBD mutations are analyzed from April 30,

[3]Log2 enrichment ratio is collected from the experimental deep mutation enrichment data in Ref. [3]

2020, to August 23, 2021, in every 60 days, as illustrated in Figure 6.10. Here, the top 100 most observed RBD mutations are displayed. In Figure 6.10 **a**, red stars mark the vaccine-resistant mutations that have negative BFE changes. Although a few vaccine-resistant mutations S438F, I434K, Y505C, and Q506K were detected before November 2020, they had relatively low frequencies. However, since December 2020, such vaccine-resistant mutations were no longer in the top 100 most observed RBD mutation list, suggesting that in this period, the evolution of SARS-CoV-2 is mainly regulated by natural selection through the infectivity-based transmission pathway. Notably, in May 2021, two vaccine-resistant mutations Y449S and Y449H, came back to the top 100 most observed RBD mutation list. In addition, Y449S has a relatively high frequency. Such finding indicates that natural selection not only favors those mutations that enhance the transmission but also those mutations that can disrupt plenty of antibodies since SARS-CoV-2 vaccines started to provide protection among populations in early May. Similarly, patterns can be found in Figure 6.10 **b**, suggesting our AI-predicted BFE changes are highly consistent with the deep mutational enrichment ratio from experiments [3].

6.3 Mutational impacts on SARS-CoV-2 infectivity

Recently, the SARS-CoV-2 variants from the United Kingdom (UK), South Africa, and Brazil have received much attention for their increased infectivity, potentially high virulence, and possible threats to existing vaccines and antibody therapies. The question remains if there are other more infectious variants transmitted around the world. We carry out a large-scale study of 506,768 SARS-CoV-2 genome isolates from patients to identify many other rapidly growing mutations on the spike (S) protein receptor-binding domain (RBD). We reveal that essentially all 100 most observed mutations strengthen the binding between the RBD and the host angiotensin-converting enzyme 2 (ACE2), indicating the virus evolves toward more infectious variants. In particular, we discover new fast-growing RBD mutations N439K, S477N, S477R, and N501T that also enhance the RBD

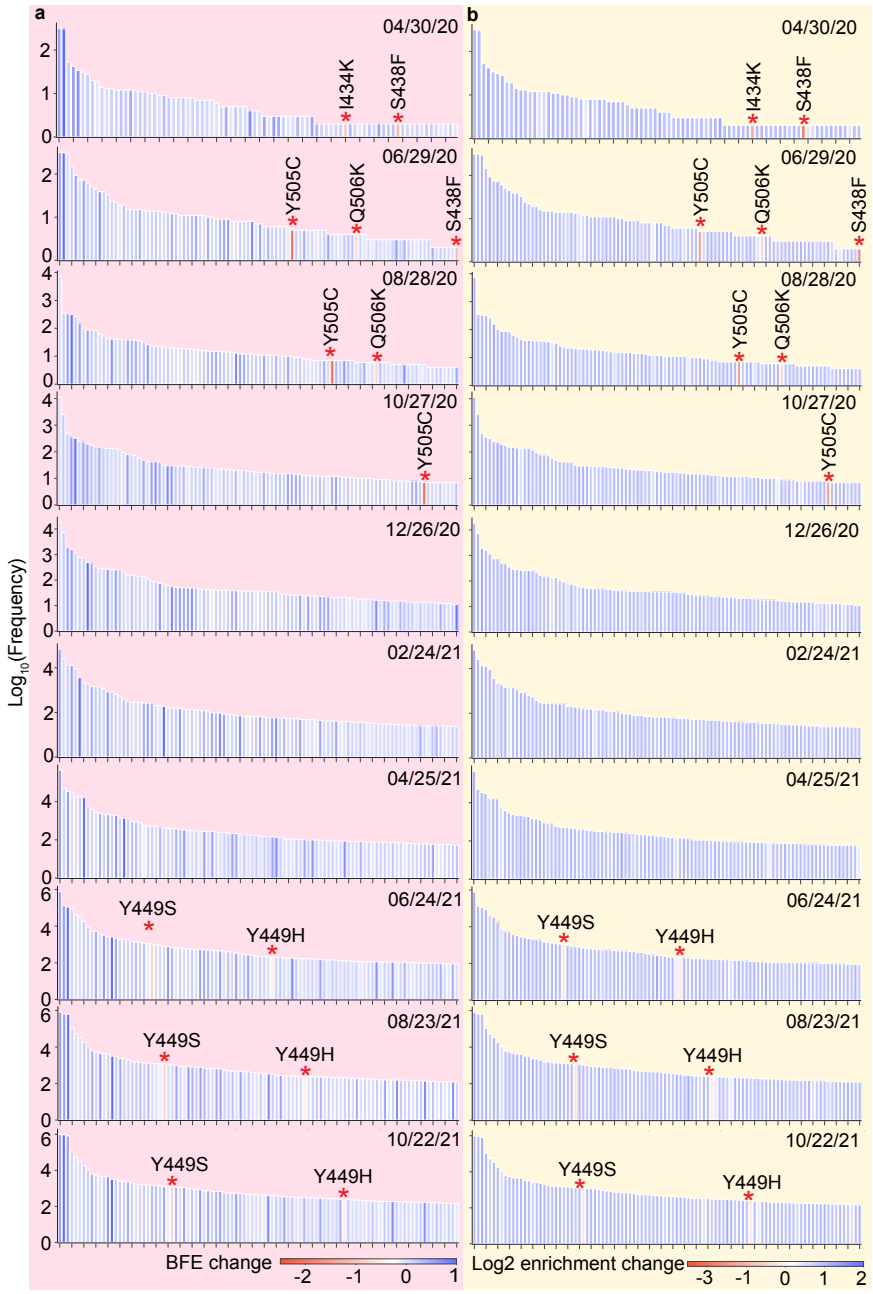

Figure 6.10: Most significant RBD mutations. **a** Time evolution of RBD mutations with its mutation-induced BFE changes per 60-day from April 30, 2020, to August 31, 2021. Here, only the top 100 most observed RBD mutations are displayed. The height and color of each bar represent the log frequency and ACE-S BFE change induced by a given RBD mutation. The red star marks the vaccine-resistant mutations with significantly negative BFE changes. **b** Time evolution of RBD mutations with its experimental mutation-induced log2 enrichment ratio changes per 60-day from April 30, 2020, to August 31, 2021. The height and color of each bar represent the log frequency and enrichment ratio change induced by a given RBD mutation. The red star marks vaccine-resistant mutations with significantly negative BFE changes.

and ACE2 binding. We further unveil that mutation N501Y involved in United Kingdom (UK), South Africa, and Brazil variants may moderately weaken the binding between the RBD and many known antibodies, while mutations E484K and K417N found in South Africa and Brazilian variants, L452R and E484Q found in India variants, can potentially disrupt the binding between the RBD and many known antibodies. Among these RBD mutations, L452R is also now known as part of the California variant B.1.427.

6.3.1 Impacts of S RBD single mutation on SARS-CoV-2 Infectivity

The RBD is located on the S1 domain of the S protein, which plays a vital role in binding with the human ACE2 to get entry into host cells. The mutations that are detected on the RBD may affect the binding process and lead to the BFE changes. In this section, we apply the TopNetTree model [217] to predict the mutation-induced BFE changes of RBD and ACE2. Figure 6.11 illustrates the predicted BFE changes for S protein and human ACE2 induced by single-site mutations on the RBD. Here, we consider 100 most observed mutations. The bar plot of the other mutations on S RBD can be found in the Supporting Information. In this figure, a total of 100 most observed mutations are displayed. Among them, 9 mutations induced negligible negative BFE changes, while the other 91 mutations are binding-strengthening mutations. Mutation T478K has the largest BFE change which is nearly 1 kcal/mol. It may have made the Mexico variant B.1.1.222 the most infectious observed variant.

To be noted, the residue T478 is not conservative among different species. The N501Y, S477N, L452R, N439K, and E484K mutations are the top mutations with significant frequencies. Among them, the N501Y and L452R mutations have a relatively high BFE change of 0.55 kcal/mol and 0.58kcal/mol. Moreover, the frequency and predicted BFE changes are both at a high level for mutations N501T, Y508H. Figure 6.12 illustrates the time evolution of 651 binding-strengthening (blue) and binding-weakening mutations (red) on the S protein RBD. Here, the y-axis reveals the natural log frequency of each mu-

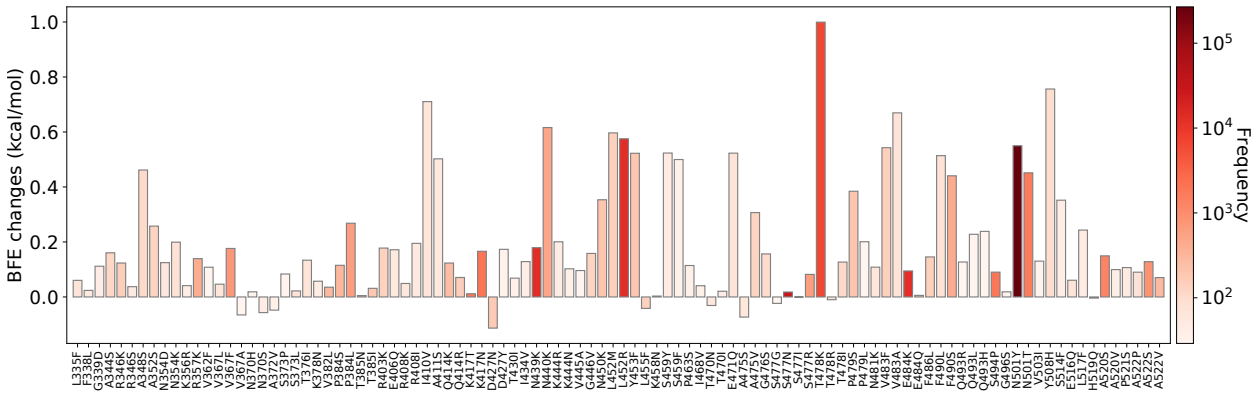

Figure 6.11: Illustration of SARS-CoV-2 mutation-induced BFE changes for the complexes of S protein and ACE2. Here, 100 most observed mutations on S RBD are illustrated.

tation. Based on the our previous findings in [41], at this stage, 651 out of 1149 RBD mutations that we predicted as "most likely" mutations have been observed, and none of the 1912 "likely" and 625 "unlikely" mutations are tracked on the S protein RBD, suggesting the reliability of our model for predicting the BFE changes of S protein RBD and ACE2. Among 651 mutations that are detected on RBD, mutations N501Y, S477N, L452R, N439K, and E484K have the highest frequency up to April 18, 2021.

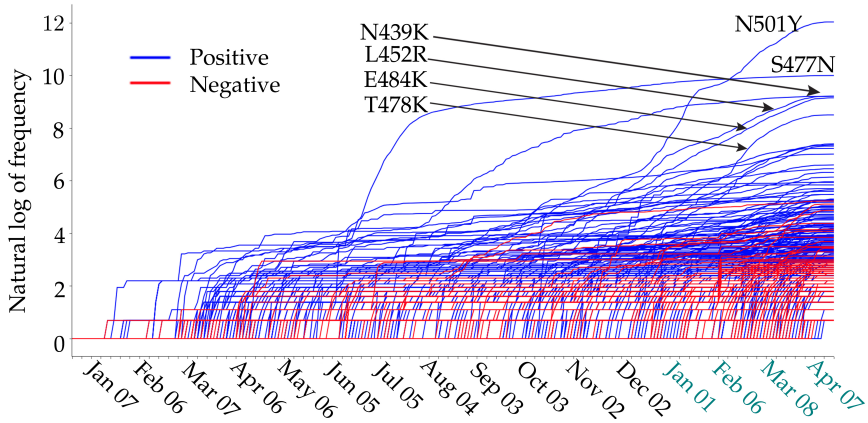

Figure 6.12: Illustration of the time evolution of 424 ACE2 binding-strengthening RBD mutations (blue) and 227 ACE2 binding-weakening RBD mutations (red) on the S protein RBD of SARS-CoV-2 from Jan 07, 2020 to April 18, 2021. The x-axis represents date and y-axis represents the natural log of frequency of each mutation.

It is important to track those mutations that have high frequency since the beginning of 2021. Table 6.4 gives such information for top 40 mutations in 2021. It can be seen that mutations N501Y, L452R, T478K, N501T, N550K, F490S, V483F, L452M, and A348S have

relatively high BFE changes of the binding of S protein and ACE2, suggesting that they may lead to more infectious variants.

Table 6.4: List of top 40 high-frequency (HF) mutations and their corresponding BFE changes (unit: kcal/mol) of the binding of S protein and ACE2. Here, count shows the frequency occurred in 2021.

Rank	HF mutation	Count	BFE change	Rank	HF mutation	Count	BFE change
Top 1	N501Y	168801	0.5499	Top 21	N450K	184	0.3535
Top 2	L452R	9843	0.5752	Top 22	E484Q	182	0.0057
Top 3	E484K	9350	0.0946	Top 23	P330S	182	0.0533
Top 4	S477N	9276	0.018	Top 24	A522V	179	0.0705
Top 5	N439K	6056	0.1792	Top 25	D427N	164	-0.1133
Top 6	T478K	4935	0.9994	Top 26	P479S	153	0.3844
Top 7	K417N	1634	0.1661	Top 27	V382L	151	0.0355
Top 8	K417T	1508	0.0116	Top 28	T385N	151	0.0049
Top 9	S494P	1483	0.0902	Top 29	Q414R	143	0.0708
Top 10	N501T	1295	0.4514	Top 30	R346K	135	0.1234
Top 11	A520S	819	0.1495	Top 31	T385I	127	0.0314
Top 12	A522S	621	0.1283	Top 32	R403K	121	0.1778
Top 13	V367F	536	0.1764	Top 33	L455F	99	-0.0415
Top 14	N440K	432	0.6161	Top 34	V483F	99	0.5428
Top 15	S477R	394	0.082	Top 35	A475V	96	0.3069
Top 16	P384L	389	0.2681	Top 36	G446V	86	0.1583
Top 17	R357K	373	0.1393	Top 37	L452M	83	0.5966
Top 18	F490S	363	0.4406	Top 38	A348S	82	0.4616
Top 19	P384S	263	0.1151	Top 39	T478I	81	0.1269
Top 20	Q414K	224	0.1234	Top 40	A352S	78	0.2576

Figure 6.13 shows the 3D structure of SARS-CoV-2 S protein RBD bound with ACE2. Here, we mark 13 mutations with either high frequency or high BFE changes. The blue and red colors represent the mutations that have positive and negative BFE changes, respectively. The darker the color is, the larger the absolute value of BFE changes is. While mutations occur everywhere on the spike protein, the ones that are most important to COVID-19 infectivity and the efficacy of antibodies and vaccines are located at the interface between the spike protein and ACE2 or antibodies.

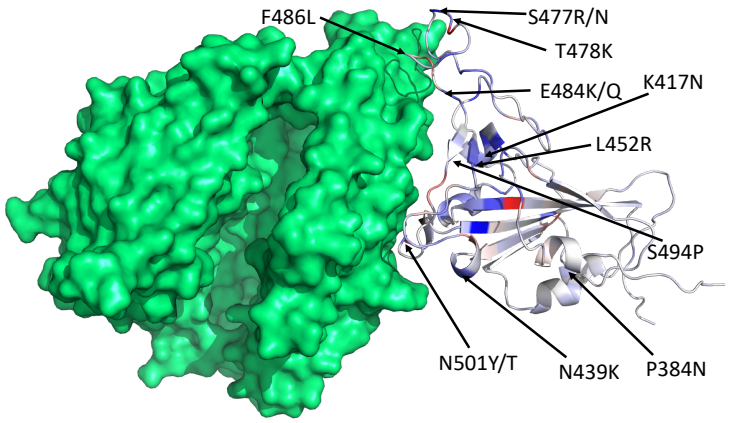

Figure 6.13: The 3D structure of SARS-CoV-2 S protein RBD bound with ACE2 (PDB ID: 6M0J). We choose blue and red colors to mark the binding-strengthening and binding-weakening mutations, respectively. Vaccine escape mutations described in Table 6.6 are labeled.

6.3.2 Impacts of S RBD co-mutations on SARS-CoV-2 Infectivity

To understand the molecular mechanisms of vaccine-escape mutations, we analyze single nucleotide polymorphisms (SNPs) of 1,489,884 complete SARS-CoV-2 genome sequences, resulting in 683 non-degenerate RBD mutations and their associated frequencies. A full set of mutation information is available on our interactive web page Mutation Tracker. The infectivity of each mutation is mainly determined by the mutation-induced BFE change to the binding complex of RBD and ACE2. To estimate the impact of each mutation on vaccines, we collect a library of 130 antibody structures (Supporting Information S2.1.2), including Food and Drug Administration (FDA)-approved mAbs from Eli Lilly and Regeneron. For a given RBD mutation, its number of antibody disruptions is given by the number of antibodies whose mutation-induced antibody-RBD BFE changes are smaller than -0.3kcal/mol (A list of names for antibodies that are disrupted by mutations can be found in the Supporting Information S2.1.1.). BFE changes following mutations are predicted by our deep learning model, TopNetTree [40]. We have created an interactive web page, Mutation Analyzer, to list all RBD mutations, their observed frequencies, their RBD-ACE2 BFE changes following mutations, their number of antibody disruptions, and various ranks. Figure 6.14 illustrates RBD mutations associated with prevailing SARS-

CoV-2 variants, time evolution trajectories of all RBD mutations, and the BFE changes of RBD-ACE2 and 130 RBD-antibodies induced by 75 significant mutations. A summary of our analysis is given in Table 6.5.

Table 6.5: Top 25 most observed S protein RBD mutations. Here, BFE change refers to the BFE change for the S protein and human ACE2 complex induced by a single-site S protein RBD mutation. A positive mutation-induced BFE change strengthens the binding between S protein and ACE2, which results in more infectious variants. Counts of antibody disruption represent the number of antibody and S protein complexes disrupted by a specific RBD mutation. Here, an antibody and S protein complex is to be disrupted if its binding affinity is reduced by more than 0.3 kcal/mol [2]. In addition, we calculate the antibody disruption ratio (%), which is the ratio of the number of disrupted antibody and S protein complexes over 130 known complexes. Ranks are computed from 683 observed RBD mutations.

Mutation	Worldwide		BFE change		Antibody disruption		
	Count	Rank	Change	Rank	Count	Ratio	Rank
N501Y	744354	1	0.5499	30	24	18.46	160
L452R	259345	2	0.5752	28	39	30.0	98
T478K	239619	3	0.9994	2	2	1.54	557
E484K	84167	4	0.0946	272	38	29.23	104
K417T	37748	5	0.0116	433	37	28.46	107
S477N	32673	6	0.0180	422	0	0.0	650
N439K	16154	7	0.1792	159	11	8.46	272
K417N	8399	8	0.1661	176	53	40.77	61
F490S	5617	9	0.4406	52	51	39.23	67
S494P	5119	10	0.0902	282	62	47.69	46
N440K	3379	11	0.6161	22	0	0.0	645
E484Q	3229	12	0.0057	442	30	23.08	130
L452Q	2858	13	0.9802	3	27	20.77	144
A520S	2727	14	0.1495	199	3	2.31	497
N501T	2054	15	0.4514	48	17	13.08	202
R357K	1973	16	0.1393	208	5	3.85	388
A522S	1959	17	0.1283	221	2	1.54	543
R346K	1686	18	0.1234	229	6	4.62	380
V367F	1395	19	0.1764	161	0	0.0	637
N440S	1361	20	0.1499	197	2	1.54	542
P384L	1155	21	0.2681	105	18	13.85	199
Y449S	1146	22	-0.8112	632	85	65.38	16
D427N	1106	23	-0.1133	558	1	0.77	589
R346S	1037	24	0.0374	386	20	15.38	182
A475V	891	25	0.3069	94	10	7.69	289

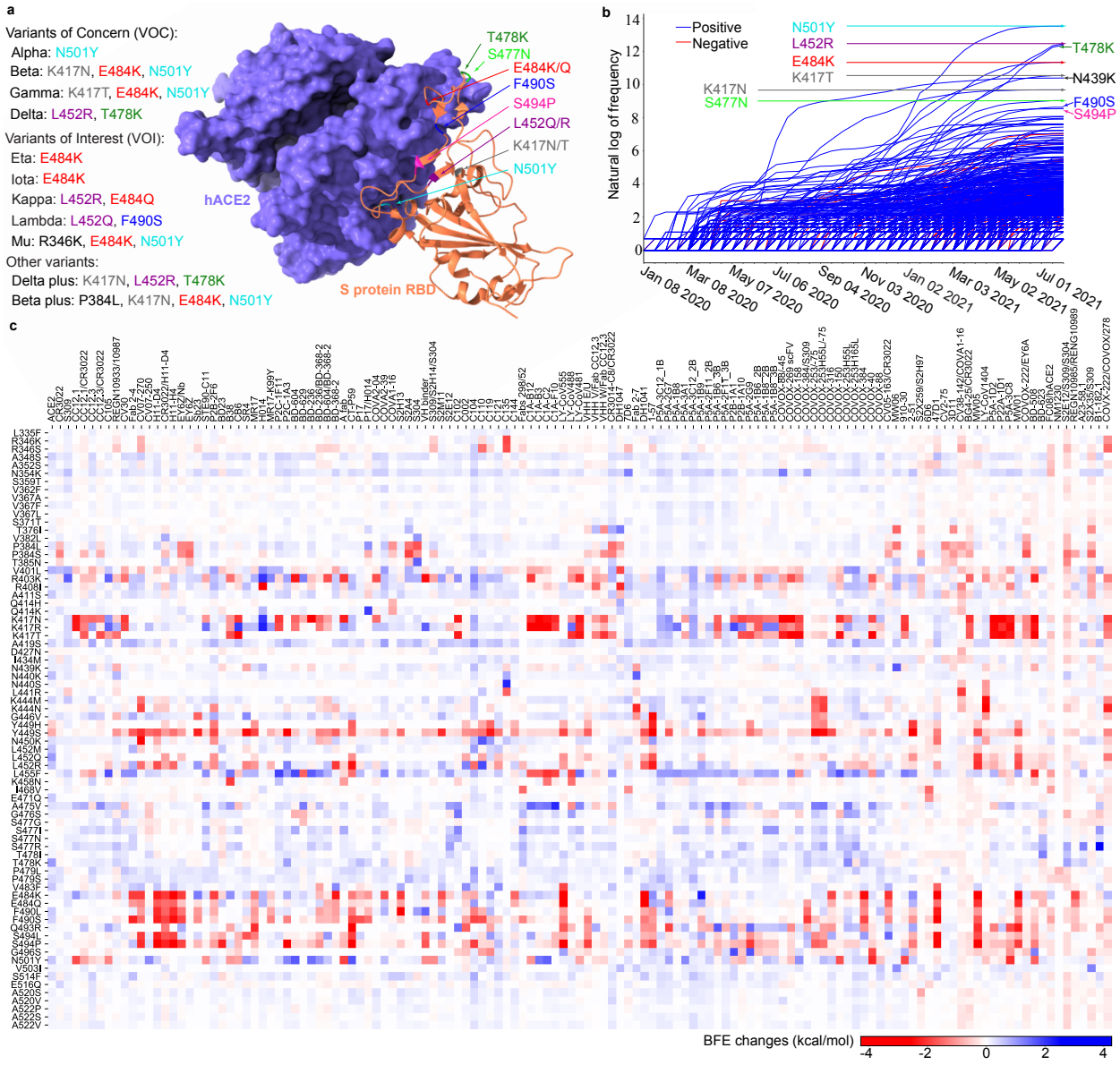

Figure 6.14: Most significant RBD mutations. **a** The 3D structure of SARS-CoV-2 S protein RBD and ACE2 complex (PDB ID: 6M0J). The RBD mutations in ten variants are marked with color. **b** Illustration of the time evolution of 455 ACE2 binding-strengthening RBD mutations (blue) and 228 ACE2 binding-weakening RBD mutations (red). The x-axis represents the date and the y-axis represents the natural log of frequency. There has been a surge in the number of infections since early 2021. **c** BFE changes of RBD complexes with ACE2 and 130 antibodies induced by 75 significant RBD mutations. A positive BFE change (blue) means the mutation strengthens the binding, while a negative BFE change (red) means the mutation weakens the binding. Most mutations, except for vaccine-resistant Y449H and Y449S, strengthen the RBD binding with ACE2. Y449S and K417N are highly disruptive to antibodies.

First, the 10 most observed or fast-growing RBD mutations are N501Y, L452R, T478K, E484K, K417T, S477N, N439K, K417N, F490S, and S494P, as shown in Table 6.5. Inclusively, these top mutations strengthen their BFEs and become more infectious, following the natural selection mechanism [41]. Figure 6.14**b** shows that the frequencies of the top three mutations increased dramatically since 2021 due to Alpha, Beta, Gamma, Delta, and other variants. Second, among the top 25 most observed RBD mutations, T478K, L452Q N440K, L452R, N501Y, N501T, F490S, A475V, and P384L are the 8 most infectious ones judged by their ability to strengthen the binding with ACE2, as shown in Figure 6.14**c**. The BFE changes of S protein and ACE2 for mutation T478K is nearly 1.00 kcal/mol, which strongly enhances the binding of the RBD-ACE2 complex [218]. Together with L452R (BFE change: 0.58kcal/mol), T478K makes Delta the most infectious variant in VOCs. Third, among the top 25 most observed RBD mutations, Y449S, S494P, K417N, F490S, L452R, E484K, K417T, E484Q, L452Q, and N501Y are the 10 most antibody disruptive ones, judged by their interactions with 130 antibodies shown in Figure 6.14**c**. It can be seen that mutations L452R, E484K, K417T, K417N, F490S, and S494P disrupt more than 30% of antibody-RBD complexes, while mutations E484K and K417T may disrupt nearly 30% antibody-RBD complexes, indicating their disruptive ability to the efficacy and reliability of antibody therapies and vaccines. The most dangerous mutations are the ones that are both infectivity-strengthening and antibody disruptive. Four RBD mutations, N501Y, L452R, F490S, and L452Q, appear in both lists and are key mutations in WHO's VOC and VOI lists. Among them, F490S and L452Q are the key RBD mutations in Lambda, making Lambda a more dangerous emerging variant than Delta. Note that high-frequency mutation S477N does not significantly weaken any antibody and RBD binding, and thus does not appear in any prevailing variants.

6.4 Mutational impacts on SARS-CoV-2 antibodies and vaccines

6.4.1 Impacts of S RBD single mutation on SARS-CoV-2 antibodies and vaccines

It is of paramount importance to track not only ACE2-binding-strengthening RBD mutations and FG mutations but also the antibody-binding-weakening RBD mutations. Our early work reported nearly 71% mutations on the S protein RBD will weaken the binding of S protein and antibodies, while 64.9% mutations on the RBD will strengthen the binding of S protein and ACE2, suggesting that these mutations may potentially enhance the infectivity of SARS-CoV-2 and make the existing antibodies less effective [217]. We call those mutations that weaken the binding of the S protein and most SARS-CoV-2 antibodies as antibody disrupting (AD) mutations [217]. Notably, most antibody disrupting mutations have negative BFE changes, suggesting that they will make the SARS-CoV-2 less infectious and thus, will not frequently occur due to natural selection. As a result, many of them may not be able to evade the existing vaccines in a population. Therefore, it is necessary to focus on the BFE changes of S protein and antibodies that are induced by 100 most observed mutations on S protein RBD.

In this work, we have collected a total of 106 antibodies. The detailed information of these 106 antibodies can be found in the Supporting Information. Figure 6.15 shows the BFE changes for the S protein and 106 antibody complexes together with ACE2 following 100 most observed mutations on the S protein RBD. The red color marks the mutation-induced negative BFE changes for the complexes of S protein and antibodies, which indicates that these mutations may weaken the binding and make the antibody less effective. Meanwhile, the green color represents the positive BFE changes induced by mutations, which suggests that these mutations may strengthen the binding of S protein and antibodies. From Figure 6.15, we can see that mutation E484K will disruptively weaken the binding of S protein with antibodies such as LY-CoV555 and DH1041, which are marked in dark red. Mutation S494P will disruptively weaken the binding of S protein with an-

tibodies such as H11-D4, H11-H4, and LY-CoV555. Mutation K417N will disruptively weaken the binding of S protein with a large number of antibodies. Moreover, mutation N501Y will moderately weaken the binding of S protein with antibodies such as CC12.1/CR3022, COVOX-88/-45, COVOX-88 etc.

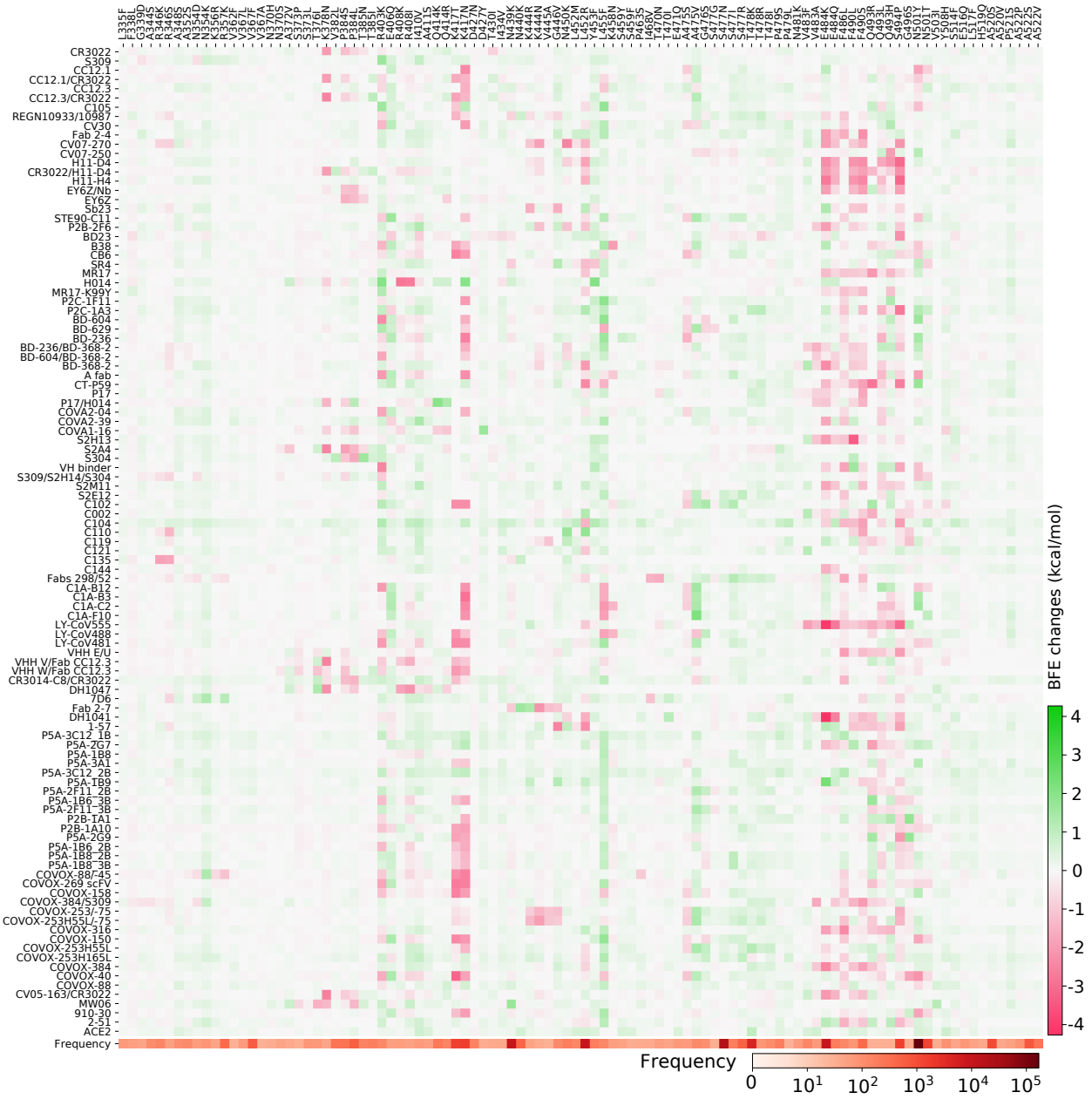

Figure 6.15: Illustration of SARS-CoV-2 S RBD 100 most observed mutations induced BFE changes for the complexes of S protein and 106 antibodies or ACE2. Here, red represents the negative changes that will weaken the binding, while green shows the positive changes that will strengthen the binding.

Considering the impact of the possible calculation error, we set -0.3 kcal/mol as the threshold of the binding of S protein and antibodies induced by AD mutations. Specifically, we say a mutation is an AD mutation to the binding complex of S protein and antibody if its BFE change for the complex is less than 0.3 kcal/mol.

We hypothesize that RBD mutations that can simultaneously strengthen the infectivity and disrupt the binding between the S protein and existing antibodies will pose imminent threats to the current crop of vaccines. We define a vaccine escape (VE) mutation as a high-frequency mutation that is an AD mutation for at least 24 (23%) different antibodies. We also define a vaccine-weakening (AW) mutation as a high-frequency mutation and AD mutation for 11 (10%) to 21 (20%) different antibodies.

Table 6.6: List of vaccine escape (VE) and vaccine weakening (VW) Their corresponding BFE changes (unit: kcal/mol) of the binding of S protein and ACE2 are provided as well. Here, the count shows the number of antibodies that will make a specific mutation to be an AD mutation.

VE Mutation	BFE change	Count	VW Mutation	BFE change	Count
S494P	0.0902	50	N501Y	0.5499	21
Q493L	0.2279	43	Q493R	0.1271	21
K417N	0.1661	43	R408I	0.1949	19
F490S	0.4406	42	Q493H	0.2385	18
F486L	0.1456	41	P384S	0.1151	18
R403K	0.1778	34	K378N	0.0573	16
E484K	0.0946	31	G496S	0.0187	15
L452R	0.5752	28	L455F	-0.0415	15
K417T	0.0116	28	I410V	0.7105	14
F490L	0.5139	25	R346S	0.0374	14
E484Q	0.0057	25	V483A	0.6695	13
A475S	-0.0732	24	K444N	0.1024	12
			N501T	0.4514	11
			P384L	0.2681	11

Table 6.6 lists vaccine-escape (VE) and vaccine-weakening (VW) RBD mutations together with their corresponding BFE changes (unit: kcal/mol) of the binding of S protein and ACE2. The count represents the number of antibodies that will make a specific mutation to be an AD mutation. We can see that VE mutations F490S, L452R, VW muta-

tions F490L, N501Y, V483A, and N501T have relatively high BFE changes of the binding of S protein and ACE2, suggesting that they are high-risk mutations. Moreover, L452R, N501Y, and N501T are also HF mutations, which should receive high attention.

6.4.2 Impacts of S RBD single mutation on SARS-CoV-2 antibodies and vaccines

The recent surge in COVID-19 infections is due to the occurrence of RBD co-mutations that combine two or more infectivity-strengthening mutations. The most dangerous future SARS-CoV-2 variants are highly likely to be RBD co-mutations that combine infectivity-strengthening mutation(s) with antibody disruptive mutation(s). A list of 1,139,244 RBD co-mutations that are decoded from 1,489,884 complete SARS-CoV-2 genome sequences can be found in Section S2.1.3 of the Supporting Information, and all of the non-degenerate RBD co-mutations with their frequencies, antibody disruption counts, total BFE changes, and the first detection dates and countries can be found in Section S2.1.4 of the Supporting Information.

Figure 6.16 illustrates the properties of S protein RBD 2, 3, and 4 co-mutations. The height of each bar shows the predicted total BFE change of each set of co-mutations on RBD, the color represents the natural log of frequency for each set of RBD co-mutations, and the number at the top of each bar is the AI-predicted number of antibody-RBD complexes that each set of RBD co-mutations may disrupt based on a total of 130 RBD and antibody complexes. Notably, for a specific set of co-mutations, the higher the number at the top of the bar is, the stronger ability to break through vaccines will be. From Figure 6.16, RBD 2 co-mutation set [L452R, T478K] (Delta variant) has the highest frequency (219,362) and the highest BFE change (1.575 kcal/mol). Moreover, the Delta variant would disrupt 40 antibody-RBD complexes, suggesting that Delta would not only enhance the infectivity but also be a vaccine breakthrough variant. Moreover, [L452Q, F490S] (Lambda) is another co-mutation with high frequency, high BFE changes (1.421 kcal/mol), and high antibody disruption count (59). In addition, Lambda is considered to be more dangerous

than Delta due to its higher antibody disruption count. Further, [R346K, E484K, N501Y] (Mu variant) has a BFE change of 0.768 kcal/mol and high antibody disruption count (60). It is not as infectious as Delta and Lambda, but has a similar ability as Lambda in escaping vaccines. Note that among all VOCs and VOIs, Beta has the highest ability to break through vaccines, but its infectivity is relatively low (BFE change: 0.656 kcal/mol).

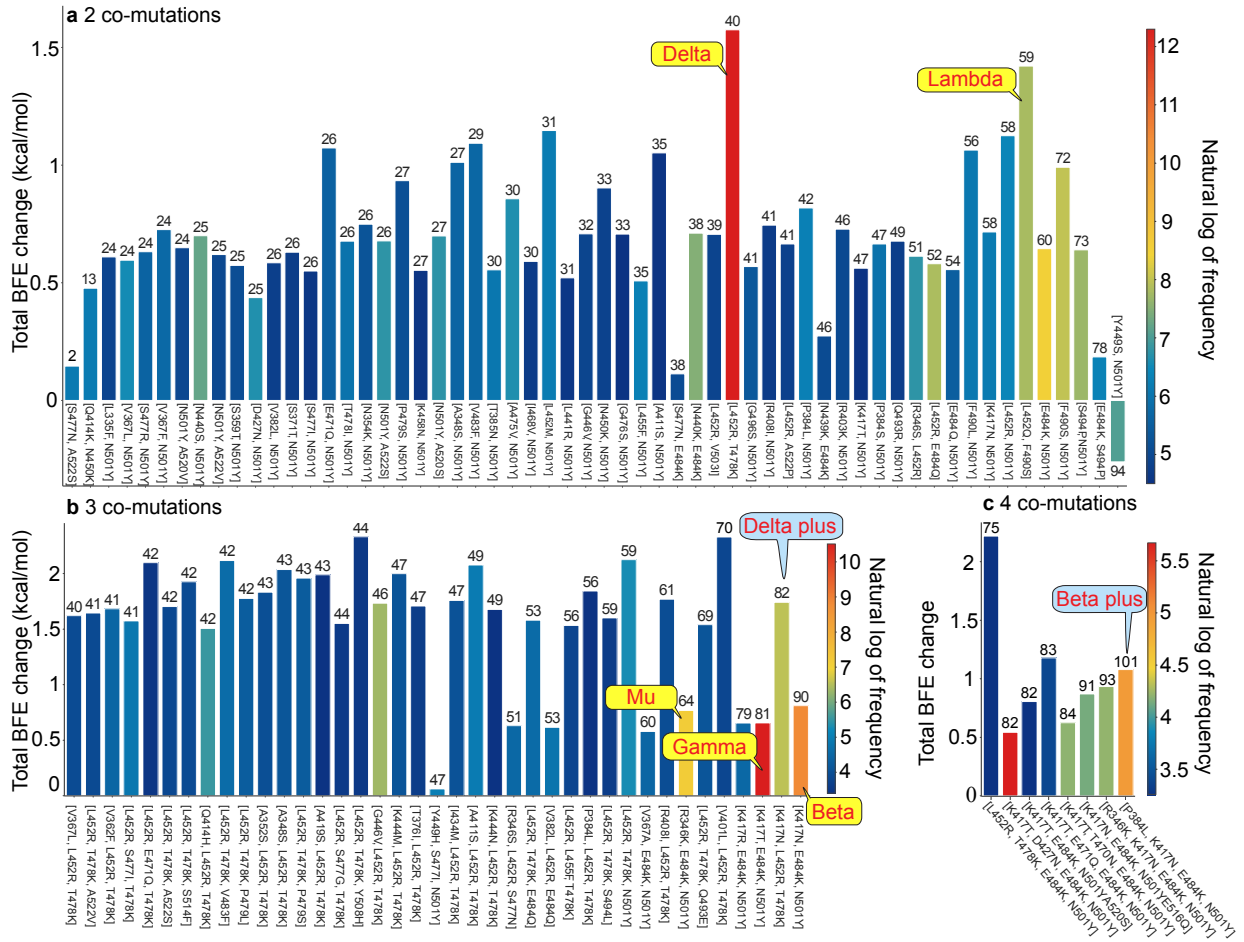

Figure 6.16: Properties of RBD co-mutations. **a** Illustration of RBD 2 co-mutations with a frequency greater than 90. **b** Illustration of RBD 3 co-mutations with a frequency greater than 30. **c** Illustration of RBD 2 co-mutations with a frequency greater than 20. Here, the x-axis lists RBD co-mutations and the y-axis represents the predicted total BFE change between S RBD and ACE2 of each set of RBD co-mutations. The number on the top of each bar is the AI-predicted number of antibody and RBD complexes that may be significantly disrupted by the set of RBD co-mutations, and the color of each bar represents the natural log of frequency for each set of RBD co-mutations. (Please check the interactive HTML files in the Supporting Information S2.2.4 for a better view of these plots.)

Furthermore, high-frequency 2 co-mutation sets [E484K, N501Y], [F490S, N501Y], and

[S494P, N501Y] are all considered to be the emerging variants that have the potential to escape vaccines. From Figure 6.16, three 3 co-mutation sets [R345K, E484K, N501Y] (Mu), [K417T, E484K, N501Y] (Gamma), and [K417N, E484K, N501Y] (Beta) draw our attention. They are all the prevailing three co-mutations with moderate BFE changes but very high antibody disruption count (more than 60). With a BFE change of 1.4 kcal/mol and antibody disruption count of 82, co-mutation set [K417N, L452R, T478K] (Delta plus) appears to be more dangerous than all of the current VOCs and VOIs.

For 4 co-mutations in Figure 6.16 **c**, [P384L, K417N, E484K, N501Y] (Beta plus) could penetrate all vaccines due to its highest antibody disruption count of 101. We would like to address that all of the co-mutations sets, except for [Y449S, N501Y] in Figure 6.16 have positive BFE changes, following natural selection. We anticipate that although co-mutation sets [V401L, L452R, T478K], [L452R, T478K, N501Y], [A411S, L452R, T478K], and [L452R, T478K, E484K, N501Y] have relatively low frequencies at this point, they may become dangerous variants soon due to their large BFE changes and antibody disruption counts.

It is important to understand the general trend of SARS-CoV-2 evolution. To this end, we carry out the statistical analysis of RBD co-mutations. Among 1,489,884 SARS-CoV-2 genome isolates, a total of 1,113 distinctive 2 co-mutations, 612 distinctive 3 co-mutations, and 217 distinctive 4 co-mutations are found. Figures 6.17 **a**, **b**, and **c** illustrate the 2D histograms of 2, 3, and 4 co-mutations, respectively. The x-axis is the number of antibody disruption counts, and the y-axis shows the total BFE change. Figure 6.17 **a** shows that there are 82 RBD 2 co-mutations that have BFE changes in the range of [0.600, 0.799] kcal/mol and will disruptive 40 to 49 antibodies. According to Figure 6.17 **b**, there are 170 unique 3 co-mutations that have large BFE changes of S protein and ACE2 in the range of [1.500, 1.999] kcal/mol. In Figure 6.17 **c**, it is seen that almost all of the 4 co-mutations on RBD have the BFE changes greater than 0.5 kcal/mol and weaken the binding of S protein with at least 60 antibodies. Figures 6.17**d**, **e**, and **f** are the histograms of total BFE

changes, natural log of frequencies, and antibody disruption counts for RBD 2, 3, and 4 co-mutations. It can be found that most of the 2, 3, and 4 RBD co-mutations have positive total BFE changes, and the larger number of RBD co-mutations is, the higher number of antibody disruption count will be. In summary, co-mutations with a larger number of antibody disruptive counts and high BFE changes will grow faster. We anticipate that when most of the population is vaccinated, vaccine-resistant mutations will become a more viable mechanism for viral evolution.

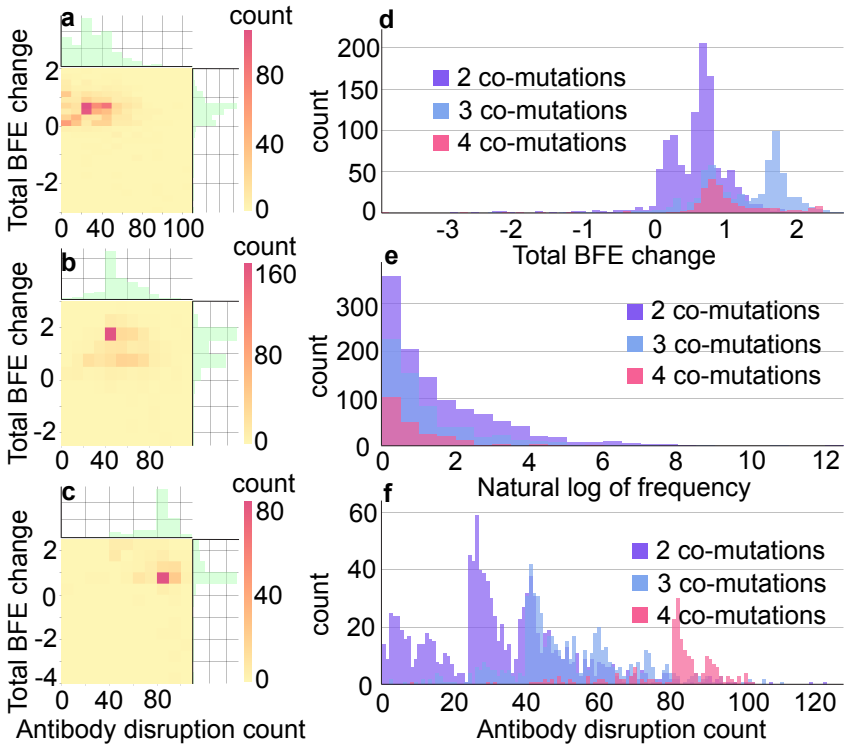

Figure 6.17: **a** 2D histograms of antibody disruption count and total BFE changes for RBD 2 co-mutations (unit: kcal/mol). **b** 2D histograms of antibody disruption count and total BFE changes (unit: kcal/mol) for RBD 3 co-mutations. **c** 2D histograms of antibody disruption count and total BFE changes (unit: kcal/mol) for RBD 4 co-mutations. **d** The histograms of total BFE changes (unit: kcal/mol) for RBD co-mutations. **e** The histograms of the natural log of frequency for RBD co-mutations. **f** The histograms of antibody disruption count for RBD co-mutations. In figures **a**, **b**, and **c**, the color bar represents the number of co-mutations that fall into the restriction of x-axis and y-axis. The reader is referred to the web version of these plots in the Supporting Information S2.2.2 and S2.2.3.

6.5 Validation

Here, we present a validation of our BFE change prediction for mutations on S protein RBD compared to the experimental deep mutational enrichment data [3]. Figure 6.18 presents a comparison between experimental deep mutational enrichment data and BFE change predictions on SARS-CoV-2 RBD binding to ACE2. In the heatmap of Figure 6.18, both BFE changes and enrichment ratios describe the affinity changes of the S protein RBD-ACE2 complex induced by mutations. It is obvious that the predicted BFE changes are highly correlated to the enrichment ratio data. Pearson correlation is 0.70. It should be noticed that the deep mutational scanning data from different labs might vary dramatically due to different experimental conditions. For example, the RBD deep mutational scanning data of the SARS-CoV-2 RBD binding to ACE2 reported by two teams [98, 3] have a relatively small Pearson correlation of 0.666.

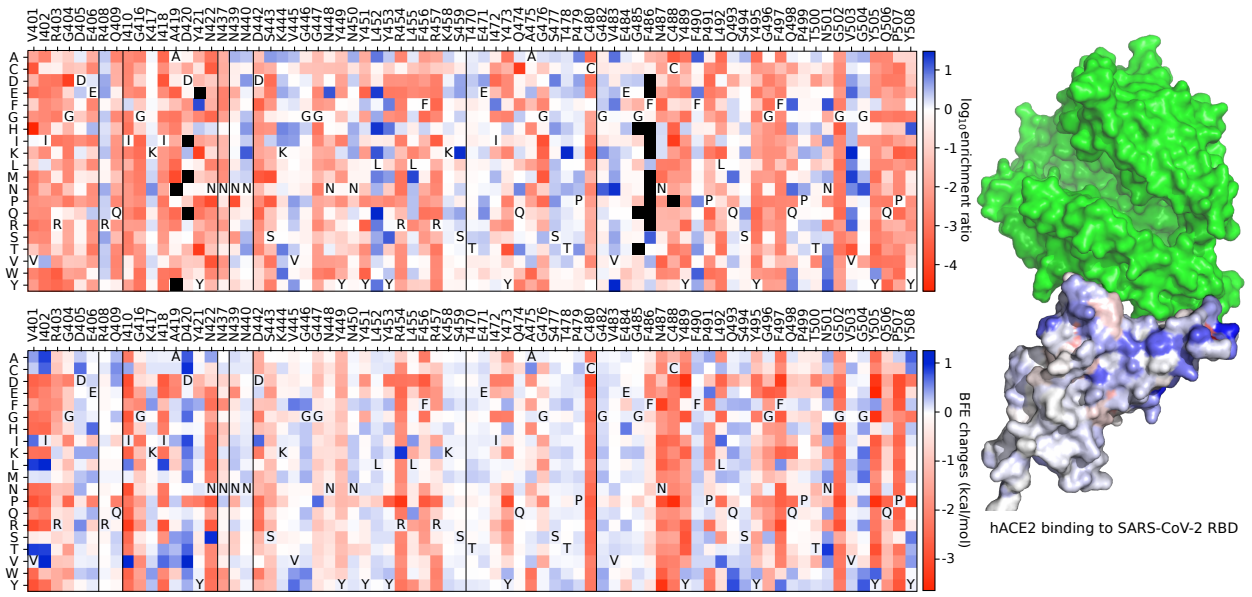

Figure 6.18: A comparison between experimental RBD deep mutation enrichment data and predicted BFE changes for SARS-CoV-2 RBD binding to ACE2 (6M0J) [3]. **T**op left: deep mutational scanning heatmap showing the average effect on the enrichment for single-site mutants of RBD when assayed by yeast display for binding to the S protein RBD [3]. **R**ight: RBD colored by average enrichment at each residue position bound to the S protein RBD. **B**ottom left: machine learning predicted BFE changes for single-site mutants of the S protein RBD.

The validation of our machine learning predictions for mutation-induced BFE changes compared to experimental data has been demonstrated in recently published papers [102, 99]. Firstly, we showed high correlations of experimental deep mutational enrichment data and predictions for the binding complex of SARS-CoV-2 S protein RBD and protein CTC-445.2 [102] and the binding complex of SARS-CoV-2 RBD and ACE2 [99]. In comparison with experimental data on the impacts of emerging variants on antibodies in clinical trials, our predictions achieve a Pearson correlation at 0.80 [99]. Considering the BFE changes induced by RBD mutations for ACE2 and RBD complex, predictions on mutations L452R and N501Y have a highly similar trend with experimental data [99]. Meanwhile, as we presented in [2], high-frequency mutations are all having positive BFE changes. Moreover, for multi-mutation tests, our BFE change predictions have the same pattern with experimental data of the impact of SARS-CoV-2 variants on major antibody therapeutic candidates, where the BFE changes are accumulative for co-mutations [99].

Recent studies on potency of mAb CT-P59 in vitro and in vivo against Delta variants[219] show that the neutralization of CT-P59 is reduced by L452R (13.22 ng/mL) and is retained against T478K (0.213 ng/mL). In our predictions [99], L452R induces a negative BFE change (-2.39 kcal/mol), and T478K produces a positive BFE change (0.36 kcal/mol). In Figure 3.2**b**, the fold changes for experimental and predicted values are presented. Additional, Figure 3.2**c** shows a comparison of the experimental pseudovirus infection changes and predicted BFE changes of ACE2 and S protein complex induced by mutations L452R and N501Y. The experimental data is obtained in a reference to D614G and reported in relative luciferase units [220]. It indicates that the binding of RBD and ACE2 dominates the infectivity of SARS-CoV-2. More details can be found in Section S6 of Supporting information.

6.6 Websites Designed

6.6.1 Mutation Tracker

Since the initial outbreak of the COVID-19, the raging pandemic caused by SARS-CoV-2 has lasted over two years. We do have many promising vaccines, but they might have side effects and their full side effects, particularly, long-term side effects, remain unknown. To make things worse, near 28734 unique mutations have been recorded for SARS-CoV-2 as shown by Mutation Tracker (See Figure 6.19). All of these reveal the sad reality that our current understanding of life science, virology, epidemiology, and medicine is severely limited.

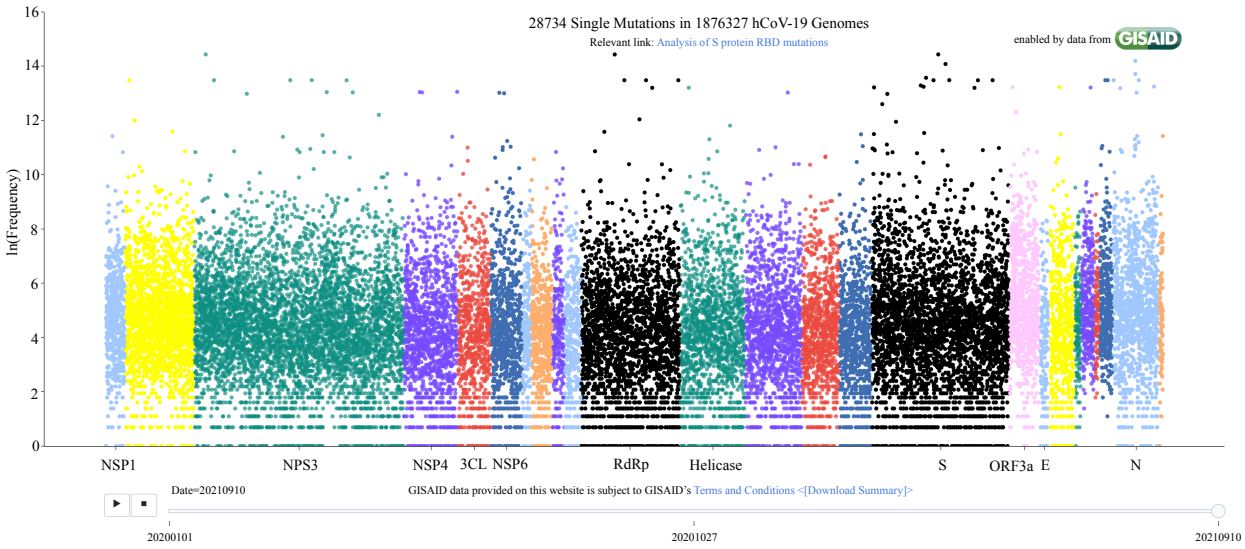

Figure 6.19: Illustration of SARS-CoV-2 mutations given by Mutation Tracker. Interactive version is available at Mutation Tracker.

6.6.2 Mutation Analyzer

The most observed SARS-CoV-2 RBD mutations are available at Mutation Analyzer (See Figure 6.20).

Analysis of observed S protein RBD mutations

Relevant link: Mutation Tracker for genome-wide analysis

Show 10 entries Search

Mutation	Worldwide observed		BFE change (kcal/mol)*		Antibody disruption		
	Counts	Rank	Value	Rank	Counts*	Ratio(%)*	Rank
N501Y	778190	1	0.5499	32	24	18.46	179
L452R	492276	2	0.5752	30	39	30.0	112
T478K	467835	3	0.9994	2	2	1.54	597
E484K	97264	4	0.0946	285	38	29.23	118
K417T	47315	5	0.0116	453	37	28.46	122
S477N	33170	6	0.0180	442	0	0.0	677
N439K	16505	7	0.1792	168	11	8.46	292
K417N	9415	8	0.1661	185	53	40.77	74
F490S	5971	9	0.4406	55	51	39.23	81
S494P	5263	10	0.0902	296	62	47.69	57

Showing 1 to 10 of 724 entries Previous 1 2 3 4 5 ... 73 Next

Figure 6.20: Illustration of the analysis of SARS-CoV-2 mutations given by interactive Mutation Analyzer that is available at Mutation Analyzer.

6.7 Discussion and Conclusion

Since the first COVID-19 case was reported in December 2019, this pandemic has led to four waves of infections, over 400 million reported cases globally, and near 6 million deaths. Despite the exciting progress in the developments of vaccines and monoclonal antibodies, their potential side effects, such as allergy reactions to COVID-19 vaccines, are not very clear. Additionally, the latest Omicron variant is able to evade current vaccines and compromise essentially all monoclonal antibodies. Although the Omicron variant may be less deadly than the original virus, there is no guarantee that future variants will be less virulent. Our present understanding of SARS-CoV-2 and COVID-19 is still quite poor.

Molecular modeling, simulation, and prediction of SARS-CoV-2 has contributed tremendously to the development of effective vaccines, drugs, and antibody therapies. Their role in combating COVID-19 is indispensable. For example, thank to an approach that integrates genotyping, biophysics, artificial intelligence, advanced mathematics, and experiment data, it is now well-understood that the SARS-CoV-2 evolution and transmission are

governed by natural selection [41]. This indicates the next SARS-CoV-2 variant will be increasingly more transmissible through high infectivity, robust vaccine breakthrough, and strong antibody resistance [221, 222]. This understanding cannot be achieved through individual experiments. Therefore, it is imperative to provide a literature review for the study of the molecular modeling, simulation, and prediction of SARS-CoV-2. Since the related literature is huge and varies in quality, we cannot collect all of the existing literature for the topic. However, we try to put forward a methodology-centered review in which we emphasize the methods used in various studies. To this end, we gather the existing theoretical and computational studies of SARS-CoV-2 concerning the aspects such as molecular modeling, biophysics, bioinformatics, cheminformatics, machine learning including deep learning, and mathematical approaches, aiming to provide a comprehensive, systematic, and indispensable component for the understanding of the molecular mechanism of SARS-CoV-2 and their interactions with host cells. Our review provides a methodology-centered description of the status of the molecular model, simulation, and prediction of SARS-CoV-2. We discuss both the traditional molecular theories, models, and methods and emergent machine learning algorithms and mathematical approaches.

Although various vaccines have been approved and in use, vaccine-breakthrough mutations have become a serious problem. Even with the promising news of new vaccines, COVID-19 as a global health crisis may still last for years before it is fully stopped globally. The research on SARS-CoV-2 will also last for many years. It will take researchers many more years to fully understand the molecular mechanism of coronaviruses, such as RNA proofreading, virus-host cell interactions, antibody-antigen interactions, protein-protein interactions, protein-drug interactions, viral regulation of host cell functions, and immune response. Even if we could control the transmission of SARS-CoV-2 in the future, newly emergent coronaviruses may still cause similar pandemic outbreaks. Therefore, the coronaviral studies will continue even after the current pandemic is fully under control.

Currently, epidemiologists, virologists, biologists, medical scientists, pharmacists, phar-

macologists, chemists, biophysicists, mathematicians, computer scientists, and many others are called to investigate various aspects of COVID-19 and SARS-CoV-2. This trend of a joint effort on COVID-19 investigations will continue beyond the present pandemic. The urgent need for the molecular mechanistic understanding of SARS-CoV-2 and COVID-19 will further stimulate the development of computational biophysical, artificial intelligence, and advanced mathematical methods. The theoretical, computational, and mathematical communities will benefit from this endeavor against the pandemic.

The year 2020 has witnessed the birth of human mRNA vaccines for the first time — a remarkable accomplishment in science and technology. Although there are more dark days ahead of us, humanity will prevail in a post-COVID-19 world. Science will emerge stronger against all pathogens and diseases in the future.

CHAPTER 7

DISSERTATION CONTRIBUTION

The main contributions of this dissertation are listed as follows:

- In Chapter 2, we propose two topological Laplacians: persistent Laplacians and persistent path Laplacians for the multiscale analysis of a given point-cloud dataset. The detailed construction process of persistent Laplacians and persistent path Laplacians are also included in Chapter 2. Notably, persistent Laplacians can extract rich topological and geometric information during filtration, and persistent path Laplacians are proposed to deal with asymmetric structures such as digraphs and networks.

- In Chapter 3, we set up a standard procedure to systematically decode nearly 30k unique single mutations from more than 2 million complete SARS-CoV-2 genome sequences in the GISAID database. In addition, we build a mathematical model called TopNetmAb, to detect the impact of single and co-mutations on the SARS-CoV-2 variants.

- In Chapter 4, we discuss applications of two new topological Laplacians in several systems, such as benzene, tetrahedron, pyramid, fullerene, curcurbit[n]urils systems, etc.

- In Chapter 5, we develop an open-source software package, called highly efficient robust multidimensional evolutionary spectra (HERMES), to enable broad applications of persistent Laplacians in science, engineering, and technology. To ensure the reliability and robustness of HERMES, we also validate the software with simple geometric shapes and complex datasets from three-dimensional (3D) protein structures.

- Chapter 6 shows our findings in the study of SARS-CoV-2, including the mechanisms of SARS-CoV-2 evolution, the mutational impacts on the infectivity, diagnostic targets, vaccines, and antibodies of SARS-CoV-2. Our standard procedures regarding date collection, pre-possessing, and model training integrate multiple techniques in computational biophysical, artificial intelligence, and advanced mathematics, which may facilitate the development of next-generation vaccines and antibody therapies against future SARS-CoV-2 variants.

The contents of this dissertation are mostly adopted from the following publications and preprints[1]:

- **Wang, R.**, Wei, G., Persistent Path Laplacian, arXiv, (2022)
- Gao, K.*, **Wang, R.***, Chen, J., Cheng, L., Frishcosy, J., Huzumi, Y., Qiu, Y., Schluckbier, T., Wei, X., and Wei, G., Methodology-centered review of molecular modeling, simulation, and prediction of SARS-CoV-2, *C*hemical Reviews, in press, (2022).
- **Wang, R.**, Chen, J., Hozumi, Y., Yin, C., and Wei, G., Emerging vaccine-breakthrough SARS-CoV-2 variants, *A*CS Infectious Diseases, 8(3), 546-556, (2022).
- Chen, J., **Wang, R.**, and Wei, G., Review of the mechanisms of SARS-CoV-2 evolution and transmission, (2021).
- **Wang, R.**, Chen, J., and Wei, G., Mechanisms of SARS-CoV-2 evolution revealing vaccine-resistant mutations in Europe and America, *T*he Journal of Physical Chemistry Letters, 12, 11850-11857, (2021)
- Chen, J., Gao, K., **Wang, R.**, and Wei, G., Revealing the threat of emerging SARS-CoV-2 mutations to antibody therapies, *J*ournal of Molecular Biology, 433(18), (2021)

[1](* co-first author)

- **Wang, R.**, Gao, K., Chen, J., and Wei, G., Vaccine-escape and fast-growing mutations in the United Kingdom, the United States, Singapore, Spain, South Africa, and other COVID-19-devastated countries, *G*enomics, 113(4), 2158-2170, (2021).

- Chen, J.*, Gao, K.*, **Wang, R.***, and Wei, G., Prediction and mitigation of mutation threats to COVID-19 vaccines and antibody therapies, *C*hemical Science, (2021).

- **Wang, R.**, Zhao, R., Ribando-Gros, Emily., Chen, J., Tong, Y., and Wei, G., HERMES: Persistent spectral graph software, *F*oundations of Data Science, 3(1), 67-97, (2021).

- **Wang, R.**, Hozumi, Y., Yin, C., and Wei, G., Decoding SARS-CoV-2 transmission, evolution and ramification on COVID-19 diagnosis, vaccine, and medicine, *J*ournal of Chemical Information and Modeling, 60, 5853-5865 (2020).

- **Wang, R.**, Duc D Nguyen and Wei, G., Persistent spectral graph, *I*nternational Journal for Numerical Methods in Biomedical Engineering, 36(9), e3376 (2020).

This work led to the following publications/preprints are not discussed in this dissertation[2]:

- Chen, J., **Wang, R.**, Gilby, N.B., and Wei, G., Omicron (B.1.1.529): Infectivity, vaccine breakthrough, and antibody resistance, *J*ournal of Chemical Information and Modeling, 62(2), 412-422, (2022).

- Gao, K., **Wang, R.**, Chen, J., Huang, F., and Wei, G., Perspectives on SARS-CoV-2 Main Protease Inhibitors, *J*ournal of Medicinal Chemistry, 64(23), 16922-16955, (2021).

- Jiang, J., **Wang, R.**, and Wei, G., GGL-Tox: Geometric graph learning for toxicity prediction, *J*ournal of Chemical Information and Modeling, 61(4), (2021).

[2](* co-first author)

- Hozumi, Y., **Wang, R.**, Yin, C., and Wei, G., UMAP-assisted K-means clustering of large-scale SARS-CoV-2 mutation datasets, *C*omputers in Biology and Medicine, 131, p.104264, (2021).

- Chen, J.*, Gao, K.*, **Wang, R.**, Duc Nguyen, and Wei, G., Review of COVID-19 antibody therapies, *A*nnual Review of Biophysics, 50, 1-30 (2021).

- **Wang, R.**, Chen, J., Gao, K., Hozumi, Y., Yin, C., and Wei, G., Analysis of SARS-CoV-2 mutations in the United States suggests presence of four substrains and novel variants, *C*ommunications Biology, 4,228 (2021).

- Chen, J., **Wang, R.**, and Wei, G., SARS-CoV-2 becoming more infectious as revealed by algebraic topology and deep learning. *C*ommunications in Information and Systems 21(1), 31-36 (2021).

- **Wang, R.**, Chen, J., Hozumi, Y., Yin, C., and Wei, G., Decoding Asymptomatic COVID-19 infection and transmission, *T*he Journal of Physical Chemistry Letters, 11, 10007-10015 (2020).

- Nguyen, D. D., Gao, K., Chen, J., **Wang, R.**, and Wei, G., Unveiling the molecular mechanism of SARS-CoV-2 main protease inhibition from 137 crystal structures using algebraic topology and deep learning, *C*hemical Sciences, 11, 12036 - 12046 (2020).

- **Wang, R.**, Hozumi, Y., Zheng, Y., Yin, C., and Wei, G., Host immune response driving SARS-CoV-2 evolution, *V*iruses, 12, 1095 (2020).

- **Wang, R.**, Hozumi, Y., Yin, C., Wei, G., Mutations on COVID-19 diagnostic targets, *G*enomics, 112, 5204-5213 (2020).

- Chen, J., **Wang, R.**, Wang, M., and Wei, G., Mutations strengthened SARS-CoV-2 infectivity, *J*ournal of Molecular Biology, 432, 5212-5226 (2020).

- Jiang, J., **Wang, R.**, Menglun Wang, Gao, K., Nguyen, D. D., and Wei, G., Boosting tree-assisted multitask deep learning for small scientific datasets. *J*ournal of Chemical Information and Modeling, 60 (3), 1235-1244 (2020).

APPENDICES

APPENDIX A

SUPPLEMENTARY MATERIALS IN PERSISTENT LAPLACIAN

A.1 Additional Laplacian matrices and their properties

In this section, we give a further description of additional boundary and Laplacian matrices and their properties involved in the filtration process in Figure 2.6.

Table A.1: $K_1 \to K_1$.

q	$q=0$	$q=1$	$q=2$
$\mathcal{B}_{q+1}^{1+0}$	/	/	/
$\mathcal{B}_q^1$	$\begin{matrix} 0 & 1 & 2 & 3 & 4 \\ [\,0 & 0 & 0 & 0 & 0\,] \end{matrix}$	/	/
$\mathcal{L}_q^{1+0}$	$\begin{bmatrix} 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 \end{bmatrix}$	/	/
β_q^{1+0}	5	/	/
$\dim(\mathcal{L}_q^{1+0})$	5	/	/
$\text{rank}(\mathcal{L}_q^{1+0})$	0	/	/
$\text{nullity}(\mathcal{L}_q^{1+0})$	5	/	/
$\text{Spectra}(\mathcal{L}_q^{1+0})$	$\{0,0,0,0,0\}$	/	/

Table A.2: $K_2 \to K_2$.

q	$q=0$	$q=1$	$q=2$
$\mathcal{B}_{q+1}^{2+0}$	$\begin{matrix} & 01 \\ 0 & \\ 1 & \\ 2 & \\ 3 & \\ 4 & \end{matrix}\begin{bmatrix} -1 \\ 1 \\ 0 \\ 0 \\ 0 \end{bmatrix}$	/	/
$\mathcal{B}_q^2$	$\begin{matrix} 0 & 1 & 2 & 3 & 4 \end{matrix}$ $[\ 0 \quad 0 \quad 0 \quad 0 \quad 0\ ]$	$\begin{matrix} & 01 \\ 0 & \\ 1 & \\ 2 & \\ 3 & \\ 4 & \end{matrix}\begin{bmatrix} -1 \\ 1 \\ 0 \\ 0 \\ 0 \end{bmatrix}$	/
$\mathcal{L}_q^{2+0}$	$\begin{bmatrix} 1 & -1 & 0 & 0 & 0 \\ -1 & 1 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 \end{bmatrix}$	[2]	/
β_q^{2+0}	4	0	/
$\dim(\mathcal{L}_q^{2+0})$	5	1	/
$\text{rank}(\mathcal{L}_q^{2+0})$	1	1	/
$\text{nullity}(\mathcal{L}_q^{2+0})$	4	0	/
$\text{Spectra}(\mathcal{L}_q^{2+0})$	$\{0,0,0,0,2\}$	2	/

Table A.3: $K_3 \to K_3$.

q	$q=0$	$q=1$	$q=2$
$\mathcal{B}_{q+1}^{3+0}$	$\begin{array}{c} \\ 0 \\ 1 \\ 2 \\ 3 \\ 4 \end{array}\begin{array}{c} \begin{array}{cccc} 01 & 12 & 23 & 03 \end{array} \\ \begin{bmatrix} -1 & 0 & 0 & -1 \\ 1 & -1 & 0 & 0 \\ 0 & 1 & -1 & 0 \\ 0 & 0 & 1 & 1 \\ 0 & 0 & 0 & 0 \end{bmatrix} \end{array}$	/	/
$\mathcal{B}_q^3$	$\begin{array}{c} \begin{array}{ccccc} 0 & 1 & 2 & 3 & 4 \end{array} \\ \begin{bmatrix} 0 & 0 & 0 & 0 & 0 \end{bmatrix} \end{array}$	$\begin{array}{c} \\ 0 \\ 1 \\ 2 \\ 3 \\ 4 \end{array}\begin{array}{c} \begin{array}{cccc} 01 & 12 & 23 & 03 \end{array} \\ \begin{bmatrix} -1 & 0 & 0 & -1 \\ 1 & -1 & 0 & 0 \\ 0 & 1 & -1 & 0 \\ 0 & 0 & 1 & 1 \\ 0 & 0 & 0 & 0 \end{bmatrix} \end{array}$	/
$\mathcal{L}_q^{3+0}$	$\begin{bmatrix} 2 & -1 & 0 & -1 & 0 \\ -1 & 2 & -1 & 0 & 0 \\ 0 & -1 & 2 & -1 & 0 \\ -1 & 0 & -1 & 2 & 0 \\ 0 & 0 & 0 & 0 & 0 \end{bmatrix}$	$\begin{bmatrix} 2 & -1 & 0 & 1 \\ -1 & 2 & -1 & 0 \\ 0 & -1 & 2 & 1 \\ 1 & 0 & 1 & 2 \end{bmatrix}$	/
β_q^{3+0}	2	1	/
$\dim(\mathcal{L}_q^{3+0})$	5	4	/
$\text{rank}(\mathcal{L}_q^{3+0})$	3	3	/
$\text{nullity}(\mathcal{L}_q^{3+0})$	2	1	/
$\text{Spectra}(\mathcal{L}_q^{3+0})$	$\{0,0,2,2,4\}$	$\{0,2,2,4\}$	/

Table A.4: $K_5 \to K_5$.

q	$q=0$	$q=1$	$q=2$
$\mathcal{B}_{q+1}^{5+0}$	$\begin{array}{c} \\ 0 \\ 1 \\ 2 \\ 3 \\ 4 \end{array}\begin{array}{c}\begin{array}{cccccc} 01 & 12 & 23 & 03 & 24 & 02 \end{array}\\ \begin{bmatrix} -1 & 0 & 0 & -1 & 0 & -1 \\ 1 & -1 & 0 & 0 & 0 & 0 \\ 0 & 1 & -1 & 0 & -1 & 1 \\ 0 & 0 & 1 & 1 & 0 & 0 \\ 0 & 0 & 0 & 0 & 1 & 0 \end{bmatrix}\end{array}$	$\begin{array}{c} \\ 01 \\ 12 \\ 23 \\ 03 \\ 24 \\ 02 \end{array}\begin{array}{c}\begin{array}{cc} 012 & 023 \end{array}\\ \begin{bmatrix} 1 & 0 \\ 1 & 0 \\ 0 & 1 \\ 0 & -1 \\ 0 & 0 \\ -1 & 1 \end{bmatrix}\end{array}$	$\begin{array}{c} \\ 012 \\ 023 \end{array}\begin{array}{c}\begin{array}{c} 0123 \end{array}\\ \begin{bmatrix} -1 \\ 1 \end{bmatrix}\end{array}$
$\mathcal{B}_q^5$	$\begin{array}{c}\begin{array}{ccccc} 0 & 1 & 2 & 3 & 4 \end{array}\\ \begin{bmatrix} 0 & 0 & 0 & 0 & 0 \end{bmatrix}\end{array}$	$\begin{array}{c} \\ 0 \\ 1 \\ 2 \\ 3 \\ 4 \end{array}\begin{array}{c}\begin{array}{cccccc} 01 & 12 & 23 & 03 & 24 & 02 \end{array}\\ \begin{bmatrix} -1 & 0 & 0 & -1 & 0 & -1 \\ 1 & -1 & 0 & 0 & 0 & 0 \\ 0 & 1 & -1 & 0 & -1 & 1 \\ 0 & 0 & 1 & 1 & 0 & 0 \\ 0 & 0 & 0 & 0 & 1 & 0 \end{bmatrix}\end{array}$	$\begin{array}{c} \\ 01 \\ 12 \\ 23 \\ 03 \\ 24 \\ 02 \end{array}\begin{array}{c}\begin{array}{cc} 012 & 023 \end{array}\\ \begin{bmatrix} 1 & 0 \\ 1 & 0 \\ 0 & 1 \\ 0 & -1 \\ 0 & 0 \\ -1 & 1 \end{bmatrix}\end{array}$
$\mathcal{L}_q^{5+0}$	$\begin{bmatrix} 3 & -1 & -1 & -1 & 0 \\ -1 & 2 & -1 & 0 & 0 \\ -1 & -1 & 4 & -1 & -1 \\ -1 & 0 & -1 & 2 & 0 \\ 0 & 0 & -1 & 0 & 1 \end{bmatrix}$	$\begin{bmatrix} 3 & 0 & 0 & 1 & 0 & 0 \\ 0 & 3 & -1 & 0 & -1 & 0 \\ 0 & -1 & 3 & 0 & 1 & 0 \\ 1 & 0 & 0 & 3 & 0 & 0 \\ 0 & -1 & 1 & 0 & 2 & -1 \\ 0 & 0 & 0 & 0 & -1 & 4 \end{bmatrix}$	$\begin{bmatrix} 4 & 0 \\ 0 & 4 \end{bmatrix}$
β_q^{5+0}	1	0	0
$\dim(\mathcal{L}_q^{5+0})$	5	6	2
$\operatorname{rank}(\mathcal{L}_q^{5+0})$	4	6	2
$\operatorname{nullity}(\mathcal{L}_q^{5+0})$	1	0	0
$\operatorname{Spectra}(\mathcal{L}_q^{5+0})$	$\{0,1,2,4,5\}$	$\{1,2,2,4,4,5\}$	$\{4,4\}$

Table A.5: $K_1 \to K_2$.

q	$q=0$	$q=1$	$q=2$
$\mathcal{B}_{q+1}^{1+1}$	$\begin{matrix} & 01 \\ 0 & -1 \\ 1 & 1 \\ 2 & 0 \\ 3 & 0 \\ 4 & 0 \end{matrix}$	/	/
$\mathcal{B}_q^1$	$\begin{matrix} 0 & 1 & 2 & 3 & 4 \\ 0 & 0 & 0 & 0 & 0 \end{matrix}$	/	/
$\mathcal{L}_q^{1+1}$	$\begin{bmatrix} 1 & -1 & 0 & 0 & 0 \\ -1 & 1 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 \end{bmatrix}$	/	/
β_q^{1+1}	4	/	/
$\dim(\mathcal{L}_q^{1+1})$	5	/	/
$\text{rank}(\mathcal{L}_q^{1+1})$	1	/	/
$\text{nullity}(\mathcal{L}_q^{1+1})$	4	/	/
$\text{Spectra}(\mathcal{L}_q^{1+1})$	$\{0,0,0,0,2\}$	/	/

Table A.6: $K_1 \to K_4$.

q	$q=0$	$q=1$	$q=2$
$\mathcal{B}_{q+1}^{1+3}$	$\begin{array}{c} \\ 0 \\ 1 \\ 2 \\ 3 \\ 4 \end{array} \begin{array}{c} \begin{array}{ccccc} 01 & 12 & 23 & 03 & 24 \end{array} \\ \begin{bmatrix} -1 & 0 & 0 & -1 & 0 \\ 1 & -1 & 0 & 0 & 0 \\ 0 & 1 & -1 & 0 & -1 \\ 0 & 0 & 1 & 1 & 0 \\ 0 & 0 & 0 & 0 & 1 \end{bmatrix} \end{array}$	/	/
$\mathcal{B}_q^1$	$\begin{array}{c} \begin{array}{ccccc} 0 & 1 & 2 & 3 & 4 \end{array} \\ \begin{bmatrix} 0 & 0 & 0 & 0 & 0 \end{bmatrix} \end{array}$	/	/
$\mathcal{L}_q^{1+3}$	$\begin{bmatrix} 2 & -1 & 0 & -1 & 0 \\ -1 & 2 & -1 & 0 & 0 \\ 0 & -1 & 3 & -1 & -1 \\ -1 & 0 & -1 & 2 & 0 \\ 0 & 0 & -1 & 0 & 1 \end{bmatrix}$	/	/
β_q^{1+3}	1	/	/
$\dim(\mathcal{L}_q^{1+3})$	5	/	/
$\text{rank}(\mathcal{L}_q^{1+3})$	4	/	/
$\text{nullity}(\mathcal{L}_q^{1+3})$	1	/	/
$\text{Spectra}(\mathcal{L}_q^{1+3})$	$\{0, 0.8299, 2, 2.6889, 4.4812\}$	/	/

Table A.7: $K_1 \to K_5$.

q	$q=0$	$q=1$	$q=2$
$\mathcal{B}_{q+1}^{1+4}$	$\begin{array}{c} \\ 0 \\ 1 \\ 2 \\ 3 \\ 4 \end{array} \begin{array}{c} \begin{array}{cccccc} 01 & 12 & 23 & 03 & 24 & 02 \end{array} \\ \left[\begin{array}{cccccc} -1 & 0 & 0 & -1 & 0 & -1 \\ 1 & -1 & 0 & 0 & 0 & 0 \\ 0 & 1 & -1 & 0 & -1 & 1 \\ 0 & 0 & 1 & 1 & 0 & 0 \\ 0 & 0 & 0 & 0 & 1 & 0 \end{array}\right] \end{array}$	/	/
$\mathcal{B}_q^1$	$\begin{array}{cccccc} 0 & 1 & 2 & 3 & 4 \\ \left[\begin{array}{ccccc} 0 & 0 & 0 & 0 & 0 \end{array}\right] \end{array}$	/	/
$\mathcal{L}_q^{1+4}$	$\left[\begin{array}{ccccc} 3 & -1 & -1 & -1 & 0 \\ -1 & 2 & -1 & 0 & 0 \\ -1 & -1 & 4 & -1 & -1 \\ -1 & 0 & -1 & 2 & 0 \\ 0 & 0 & -1 & 0 & 1 \end{array}\right]$	/	/
β_q^{1+4}	1	/	/
$\dim(\mathcal{L}_q^{1+4})$	5	/	/
$\text{rank}(\mathcal{L}_q^{1+4})$	4	/	/
$\text{nullity}(\mathcal{L}_q^{1+4})$	1	/	/
$\text{Spectra}(\mathcal{L}_q^{1+4})$	$\{0,1,2,4,5\}$	/	/

Table A.8: $K_1 \to K_6$.

q	$q=0$	$q=1$	$q=2$
$\mathcal{B}_{q+1}^{1+5}$	$\begin{array}{c} \\ 0 \\ 1 \\ 2 \\ 3 \\ 4 \end{array} \begin{array}{c} \begin{array}{ccccccc} 01 & 12 & 23 & 03 & 24 & 02 & 13 \end{array} \\ \left[\begin{array}{ccccccc} -1 & 0 & 0 & -1 & 0 & -1 & 0 \\ 1 & -1 & 0 & 0 & 0 & 0 & -1 \\ 0 & 1 & -1 & 0 & -1 & 1 & 0 \\ 0 & 0 & 1 & 1 & 0 & 0 & 1 \\ 0 & 0 & 0 & 0 & 1 & 0 & 0 \end{array}\right] \end{array}$	/	/
$\mathcal{B}_q^1$	$\begin{array}{c} \begin{array}{ccccc} 0 & 1 & 2 & 3 & 4 \end{array} \\ \left[\begin{array}{ccccc} 0 & 0 & 0 & 0 & 0 \end{array}\right] \end{array}$	/	/
$\mathcal{L}_q^{1+5}$	$\left[\begin{array}{ccccc} 3 & -1 & -1 & -1 & 0 \\ -1 & 3 & -1 & -1 & 0 \\ -1 & -1 & 4 & -1 & -1 \\ -1 & -1 & -1 & 3 & 0 \\ 0 & 0 & -1 & 0 & 1 \end{array}\right]$	/	/
β_q^{1+5}	1	/	/
$\dim(\mathcal{L}_q^{1+5})$	5	/	/
$\text{rank}(\mathcal{L}_q^{1+5})$	4	/	/
$\text{nullity}(\mathcal{L}_q^{1+5})$	1	/	/
$\text{Spectra}(\mathcal{L}_q^{1+5})$	$\{0,1,4,4,5\}$	/	/

Table A.9: $K_2 \to K_3$.

q	$q=0$	$q=1$	$q=2$
$\mathcal{B}_{q+1}^{2+1}$	$\begin{matrix} & 01 & 12 & 23 & 03 \\ 0 & -1 & 0 & 0 & -1 \\ 1 & 1 & -1 & 0 & 0 \\ 2 & 0 & 1 & -1 & 0 \\ 3 & 0 & 0 & 1 & 1 \\ 4 & 0 & 0 & 0 & 0 \end{matrix}$	/	/
$\mathcal{B}_q^2$	$\begin{matrix} 0 & 1 & 2 & 3 & 4 \\ 0 & 0 & 0 & 0 & 0 \end{matrix}$	$\begin{matrix} & 01 \\ 0 & -1 \\ 1 & 1 \\ 2 & 0 \\ 3 & 0 \\ 4 & 0 \end{matrix}$	/
$\mathcal{L}_q^{2+1}$	$\begin{bmatrix} 2 & -1 & 0 & -1 & 0 \\ -1 & 2 & -1 & 0 & 0 \\ 0 & -1 & 2 & -1 & 0 \\ -1 & 0 & -1 & 2 & 0 \\ 0 & 0 & 0 & 0 & 0 \end{bmatrix}$	[2]	/
β_q^{2+1}	2	0	/
$\dim(\mathcal{L}_q^{2+1})$	5	1	/
$\text{rank}(\mathcal{L}_q^{2+1})$	3	1	/
$\text{nullity}(\mathcal{L}_q^{2+1})$	2	0	/
$\text{Spectra}(\mathcal{L}_q^{2+1})$	$\{0,0,2,2,4\}$	2	/

Table A.10: $K_2 \to K_4$.

q	$q=0$	$q=1$	$q=2$
$\mathcal{B}_{q+1}^{2+2}$	$\begin{array}{c} \\ 0 \\ 1 \\ 2 \\ 3 \\ 4 \end{array}\begin{array}{c} \begin{array}{ccccc} 01 & 12 & 23 & 03 & 24 \end{array} \\ \begin{bmatrix} -1 & 0 & 0 & -1 & 0 \\ 1 & -1 & 0 & 0 & 0 \\ 0 & 1 & -1 & 0 & -1 \\ 0 & 0 & 1 & 1 & 0 \\ 0 & 0 & 0 & 0 & 1 \end{bmatrix} \end{array}$	/	/
$\mathcal{B}_q^2$	$\begin{array}{c} \begin{array}{ccccc} 0 & 1 & 2 & 3 & 4 \end{array} \\ \begin{bmatrix} 0 & 0 & 0 & 0 & 0 \end{bmatrix} \end{array}$	$\begin{array}{c} \\ 0 \\ 1 \\ 2 \\ 3 \\ 4 \end{array}\begin{array}{c} \begin{array}{c} 01 \end{array} \\ \begin{bmatrix} -1 \\ 1 \\ 0 \\ 0 \\ 0 \end{bmatrix} \end{array}$	/
$\mathcal{L}_q^{2+2}$	$\begin{bmatrix} 2 & -1 & 0 & -1 & 0 \\ -1 & 2 & -1 & 0 & 0 \\ 0 & -1 & 3 & -1 & -1 \\ -1 & 0 & -1 & 2 & 0 \\ 0 & 0 & -1 & 0 & 1 \end{bmatrix}$	[2]	/
β_q^{2+2}	1	0	/
$\dim(\mathcal{L}_q^{2+2})$	5	1	/
$\text{rank}(\mathcal{L}_q^{2+2})$	4	1	/
$\text{nullity}(\mathcal{L}_q^{2+2})$	1	0	/
$\text{Spectra}(\mathcal{L}_q^{2+2})$	$\{0, 0.8299, 2, 2.6889, 4.4812\}$	2	/

Table A.11: $K_2 \to K_5$.

q	$q=0$	$q=1$	$q=2$
$\mathcal{B}^{2+3}_{q+1}$	$\begin{array}{c} \\ 0 \\ 1 \\ 2 \\ 3 \\ 4 \end{array}\begin{array}{c}\begin{array}{cccccc} 01 & 12 & 23 & 03 & 24 & 02 \end{array}\\ \left[\begin{array}{cccccc} -1 & 0 & 0 & -1 & 0 & -1 \\ 1 & -1 & 0 & 0 & 0 & 0 \\ 0 & 1 & -1 & 0 & -1 & 1 \\ 0 & 0 & 1 & 1 & 0 & 0 \\ 0 & 0 & 0 & 0 & 1 & 0 \end{array}\right] \end{array}$	$\begin{array}{c} \\ 01 \end{array}\begin{array}{c}\begin{array}{cc} 012 & 023 \end{array}\\ \left[\begin{array}{cc} 1 & 0 \end{array}\right] \end{array}$	/
$\mathcal{B}^2_q$	$\begin{array}{c}\begin{array}{ccccc} 0 & 1 & 2 & 3 & 4 \end{array}\\ \left[\begin{array}{ccccc} 0 & 0 & 0 & 0 & 0 \end{array}\right] \end{array}$	$\begin{array}{c} \\ 0 \\ 1 \\ 2 \\ 3 \\ 4 \end{array}\begin{array}{c}\begin{array}{c} 01 \end{array}\\ \left[\begin{array}{c} -1 \\ 1 \\ 0 \\ 0 \\ 0 \end{array}\right] \end{array}$	/
$\mathcal{L}^{2+3}_q$	$\left[\begin{array}{ccccc} 3 & -1 & -1 & -1 & 0 \\ -1 & 3 & -1 & -1 & 0 \\ -1 & -1 & 4 & -1 & -1 \\ -1 & -1 & -1 & 3 & 0 \\ 0 & 0 & -1 & 0 & 1 \end{array}\right]$	[3]	/
β^{2+3}_q	1	0	/
$\dim(\mathcal{L}^{2+3}_q)$	5	1	/
$\text{rank}(\mathcal{L}^{2+3}_q)$	4	1	/
$\text{nullity}(\mathcal{L}^{2+3}_q)$	1	0	/
$\text{Spectra}(\mathcal{L}^{2+3}_q)$	$\{0,1,2,4,5\}$	3	/

Table A.12: $K_2 \to K_6$.

q	$q=0$	$q=1$	$q=2$
$\mathcal{B}_{q+1}^{2+4}$	$\begin{array}{c} \\ 0 \\ 1 \\ 2 \\ 3 \\ 4 \end{array}\begin{array}{c} \begin{array}{ccccccc} 01 & 12 & 23 & 03 & 24 & 02 & 13 \end{array} \\ \begin{bmatrix} -1 & 0 & 0 & -1 & 0 & -1 & 0 \\ 1 & -1 & 0 & 0 & 0 & 0 & -1 \\ 0 & 1 & -1 & 0 & -1 & 1 & 0 \\ 0 & 0 & 1 & 1 & 0 & 0 & 1 \\ 0 & 0 & 0 & 0 & 1 & 0 & 0 \end{bmatrix} \end{array}$	$\begin{array}{c} \\ 01 \end{array}\begin{array}{c} \begin{array}{cccc} 012 & 023 & 013 & 123 \end{array} \\ \begin{bmatrix} 1 & 0 & 1 & 0 \end{bmatrix} \end{array}$	/
$\mathcal{B}_q^2$	$\begin{array}{c} \\ \end{array}\begin{array}{c} \begin{array}{ccccc} 0 & 1 & 2 & 3 & 4 \end{array} \\ \begin{bmatrix} 0 & 0 & 0 & 0 & 0 \end{bmatrix} \end{array}$	$\begin{array}{c} \begin{array}{c} 01 \end{array} \\ \begin{bmatrix} -1 \\ 1 \\ 0 \\ 0 \\ 0 \end{bmatrix} \end{array}\begin{array}{c} \\ 0 \\ 1 \\ 2 \\ 3 \\ 4 \end{array}$	/
$\mathcal{L}_q^{2+4}$	$\begin{bmatrix} 3 & -1 & -1 & -1 & 0 \\ -1 & 2 & -1 & 0 & 0 \\ -1 & -1 & 4 & -1 & -1 \\ -1 & 0 & -1 & 2 & 0 \\ 0 & 0 & -1 & 0 & 1 \end{bmatrix}$	[4]	/
β_q^{2+4}	1	0	/
$\dim(\mathcal{L}_q^{2+4})$	5	1	/
$\text{rank}(\mathcal{L}_q^{2+4})$	4	1	/
$\text{nullity}(\mathcal{L}_q^{2+4})$	1	0	/
$\text{Spectra}(\mathcal{L}_q^{2+4})$	$\{0,1,4,4,5\}$	4	/

Table A.13: $K_3 \to K_5$.

q	$q=0$	$q=1$	$q=2$
$\mathcal{B}_{q+1}^{3+2}$	$\begin{array}{c} \\ 0 \\ 1 \\ 2 \\ 3 \\ 4 \end{array}\begin{array}{c}\begin{array}{cccccc} 01 & 12 & 23 & 03 & 24 & 02 \end{array}\\ \begin{bmatrix} -1 & 0 & 0 & -1 & 0 & -1 \\ 1 & -1 & 0 & 0 & 0 & 0 \\ 0 & 1 & -1 & 0 & -1 & 1 \\ 0 & 0 & 1 & 1 & 0 & 0 \\ 0 & 0 & 0 & 0 & 1 & 0 \end{bmatrix}\end{array}$	$\begin{array}{c} \\ 01 \\ 12 \\ 23 \\ 03 \end{array}\begin{array}{c}\begin{array}{cc} 012 & 023 \end{array}\\ \begin{bmatrix} 1 & 0 \\ 1 & 0 \\ 0 & 1 \\ 0 & -1 \end{bmatrix}\end{array}$	/
$\mathcal{B}_q^3$	$\begin{array}{c}\begin{array}{ccccc} 0 & 1 & 2 & 3 & 4 \end{array}\\ \begin{bmatrix} 0 & 0 & 0 & 0 & 0 \end{bmatrix}\end{array}$	$\begin{array}{c} \\ 0 \\ 1 \\ 2 \\ 3 \\ 4 \end{array}\begin{array}{c}\begin{array}{cccc} 01 & 12 & 23 & 03 \end{array}\\ \begin{bmatrix} -1 & 0 & 0 & -1 \\ 1 & -1 & 0 & 0 \\ 0 & 1 & -1 & 0 \\ 0 & 0 & 1 & 1 \\ 0 & 0 & 0 & 0 \end{bmatrix}\end{array}$	/
$\mathcal{L}_q^{3+2}$	$\begin{bmatrix} 3 & -1 & -1 & -1 & 0 \\ -1 & 2 & -1 & 0 & 0 \\ -1 & -1 & 4 & -1 & -1 \\ -1 & 0 & -1 & 2 & 0 \\ 0 & 0 & -1 & 0 & 1 \end{bmatrix}$	$\begin{bmatrix} 3 & 0 & 0 & 1 \\ 0 & 3 & -1 & 0 \\ 0 & -1 & 3 & 0 \\ 1 & 0 & 0 & 3 \end{bmatrix}$	/
β_q^{3+2}	1	0	/
$\dim(\mathcal{L}_q^{3+2})$	5	4	/
$\text{rank}(\mathcal{L}_q^{3+2})$	4	4	/
$\text{nullity}(\mathcal{L}_q^{3+2})$	1	0	/
$\text{Spectra}(\mathcal{L}_q^{3+2})$	$\{0,1,2,4,5\}$	$\{2,2,4,4\}$	/

Table A.14: $K_3 \to K_6$.

q	$q=0$	$q=1$	$q=2$
$\mathcal{B}_{q+1}^{3+3}$	$\begin{array}{c} \\ 0 \\ 1 \\ 2 \\ 3 \\ 4 \end{array}\begin{array}{c} \begin{array}{ccccccc} 01 & 12 & 23 & 03 & 24 & 02 & 13 \end{array} \\ \begin{bmatrix} -1 & 0 & 0 & -1 & 0 & -1 & 0 \\ 1 & -1 & 0 & 0 & 0 & 0 & -1 \\ 0 & 1 & -1 & 0 & -1 & 1 & 0 \\ 0 & 0 & 1 & 1 & 0 & 0 & 1 \\ 0 & 0 & 0 & 0 & 1 & 0 & 0 \end{bmatrix} \end{array}$	$\begin{array}{c} \\ 01 \\ 12 \\ 23 \\ 03 \end{array}\begin{array}{c} \begin{array}{cccc} 012 & 023 & 013 & 123 \end{array} \\ \begin{bmatrix} 1 & 0 & 1 & 0 \\ 1 & 0 & 0 & 1 \\ 0 & 1 & 0 & 1 \\ 0 & -1 & -1 & 0 \end{bmatrix} \end{array}$	/
$\mathcal{B}_q^3$	$\begin{array}{c} \begin{array}{ccccc} 0 & 1 & 2 & 3 & 4 \end{array} \\ \begin{bmatrix} 0 & 0 & 0 & 0 & 0 \end{bmatrix} \end{array}$	$\begin{array}{c} \\ 0 \\ 1 \\ 2 \\ 3 \\ 4 \end{array}\begin{array}{c} \begin{array}{cccc} 01 & 12 & 23 & 03 \end{array} \\ \begin{bmatrix} -1 & 0 & 0 & -1 \\ 1 & -1 & 0 & 0 \\ 0 & 1 & -1 & 0 \\ 0 & 0 & 1 & 1 \\ 0 & 0 & 0 & 0 \end{bmatrix} \end{array}$	/
$\mathcal{L}_q^{3+3}$	$\begin{bmatrix} 3 & -1 & -1 & -1 & 0 \\ -1 & 3 & -1 & -1 & 0 \\ -1 & -1 & 4 & -1 & -1 \\ -1 & -1 & -1 & 3 & 0 \\ 0 & 0 & -1 & 0 & 1 \end{bmatrix}$	$\begin{bmatrix} 4 & 0 & 0 & 0 \\ 0 & 4 & 0 & 0 \\ 0 & 0 & 4 & 0 \\ 0 & 0 & 0 & 4 \end{bmatrix}$	/
β_q^{3+3}	1	0	/
$\dim(\mathcal{L}_q^{3+3})$	5	4	/
$\text{rank}(\mathcal{L}_q^{3+3})$	4	4	/
$\text{nullity}(\mathcal{L}_q^{3+3})$	1	0	/
$\text{Spectra}(\mathcal{L}_q^{3+3})$	$\{0,1,4,4,5\}$	$\{4,4,4,4\}$	/

Table A.15: $K_4 \to K_6$.

q	$q=0$	$q=1$	$q=2$
$\mathcal{B}^{4+2}_{q+1}$	$\begin{array}{c} \\ 0 \\ 1 \\ 2 \\ 3 \\ 4 \end{array}\begin{array}{c} \begin{array}{ccccccc} 01 & 12 & 23 & 03 & 24 & 02 & 13 \end{array} \\ \begin{bmatrix} -1 & 0 & 0 & -1 & 0 & -1 & 0 \\ 1 & -1 & 0 & 0 & 0 & 0 & -1 \\ 0 & 1 & -1 & 0 & -1 & 1 & 0 \\ 0 & 0 & 1 & 1 & 0 & 0 & 1 \\ 0 & 0 & 0 & 0 & 1 & 0 & 0 \end{bmatrix} \end{array}$	$\begin{array}{c} \\ 01 \\ 12 \\ 23 \\ 03 \\ 24 \end{array}\begin{array}{c} \begin{array}{cccc} 012 & 023 & 013 & 123 \end{array} \\ \begin{bmatrix} 1 & 0 & 1 & 0 \\ 1 & 0 & 0 & 1 \\ 0 & 1 & 0 & 1 \\ 0 & -1 & -1 & 0 \\ 0 & 0 & 0 & 0 \end{bmatrix} \end{array}$	/
$\mathcal{B}^4_q$	$\begin{array}{c} \begin{array}{ccccc} 0 & 1 & 2 & 3 & 4 \end{array} \\ \begin{bmatrix} 0 & 0 & 0 & 0 & 0 \end{bmatrix} \end{array}$	$\begin{array}{c} \\ 0 \\ 1 \\ 2 \\ 3 \\ 4 \end{array}\begin{array}{c} \begin{array}{ccccc} 01 & 12 & 23 & 03 & 24 \end{array} \\ \begin{bmatrix} -1 & 0 & 0 & -1 & 0 \\ 1 & -1 & 0 & 0 & 0 \\ 0 & 1 & -1 & 0 & -1 \\ 0 & 0 & 1 & 1 & 0 \\ 0 & 0 & 0 & 0 & 1 \end{bmatrix} \end{array}$	/
$\mathcal{L}^{4+2}_q$	$\begin{bmatrix} 3 & -1 & -1 & -1 & 0 \\ -1 & 3 & -1 & -1 & 0 \\ -1 & -1 & 4 & -1 & -1 \\ -1 & -1 & -1 & 3 & 0 \\ 0 & 0 & -1 & 0 & 1 \end{bmatrix}$	$\begin{bmatrix} 4 & 0 & 0 & 0 & 0 \\ 0 & 4 & 0 & 0 & -1 \\ 0 & 0 & 4 & 0 & 1 \\ 0 & 0 & 0 & 4 & 0 \\ 0 & -1 & 1 & 0 & 2 \end{bmatrix}$	/
β^{4+2}_q	1	0	/
$\dim(\mathcal{L}^{4+2}_q)$	5	5	/
$\text{rank}(\mathcal{L}^{4+2}_q)$	4	5	/
$\text{nullity}(\mathcal{L}^{4+2}_q)$	1	0	/
$\text{Spectra}(\mathcal{L}^{4+2}_q)$	$\{0, 1, 4, 4, 5\}$	$\{1.2679, 4, 4, 4, 4.7321\}$	/

A.2 Parameters in the protein B-factor prediction

Table A.16: Fitting parameters from w_0 to w_5.

r	0	1	2	3	4	5
w_r	10.6102	0.2026	-0.0031	0.2169	0.3127	0.2815

Table A.17: Fitting parameters from w_6 to w_{11}.

r	6	7	8	9	10	11
w_r	-0.4623	1.0203	0.6110	-0.6872	-1.0695	4.4257

APPENDIX B

SUPPLEMENTARY MATERIALS IN PERSISTENT PATH LAPLACIAN

Table B.1 - Table B.14, we present the detailed matrix constructions, Betti numbers, and spectra for various digraphs as shown in Figure 4.10 top and bottom panels

Table B.1: Matrix construction of graph G_1 (with isolated points included) in the top panel of Figure 4.10.

n	$n=0$	$n=1$	$n=2$
Ω_n	span$\{e_1, e_2, e_3, e_4, e_5\}$	$\{0\}$	$\{0\}$
B_{n+1}	5×0 empty matrix	/	/
L_n	5×5 zero matrix	/	/
β_n	5	/	/
Spectra(L_n)	$\{0, 0, 0, 0, 0\}$	/	/

Table B.2: Matrix construction of graph G_1 (without isolated points) in the top panel of Figure 4.10.

n	$n=0$	$n=1$	$n=2$
Ω_n	$\{0\}$	$\{0\}$	$\{0\}$
B_{n+1}	/	/	/
L_n	/	/	/
β_n	/	/	/
Spectra(L_n)	/	/	/

Table B.3: Matrix construction of graph G_2 in the top panel of Figure 4.10.

n	$n=0$	$n=1$	$n=2$
Ω_n	span$\{e_1, e_2, e_3, e_4, e_5\}$	span$\{e_{13}, e_{25}, e_{32}, e_{34}, e_{45}\}$	$\{0\}$
B_{n+1}	$\begin{array}{c} \\ e_1 \\ e_2 \\ e_3 \\ e_4 \\ e_5 \end{array} \begin{array}{c} \begin{array}{ccccc} e_{13} & e_{25} & e_{32} & e_{34} & e_{45} \end{array} \\ \begin{pmatrix} -1 & 0 & 0 & 0 & 0 \\ 0 & -1 & 1 & 0 & 0 \\ 1 & 0 & -1 & -1 & 0 \\ 0 & 0 & 0 & 1 & 1 \\ 0 & 1 & 0 & 0 & -1 \end{pmatrix} \end{array}$	5×0 empty matrix	$\begin{pmatrix} \ & / & \ \end{pmatrix}$
L_n	$\begin{pmatrix} 1 & 0 & -1 & 0 & 0 \\ 0 & 2 & -1 & 0 & -1 \\ -1 & -1 & 3 & -1 & 0 \\ 0 & 0 & -1 & 2 & -1 \\ 0 & -1 & 0 & -1 & 2 \end{pmatrix}$	$\begin{pmatrix} 2 & 0 & -1 & -1 & 0 \\ 0 & 2 & -1 & 0 & -1 \\ -1 & -1 & 2 & 1 & 0 \\ -1 & 0 & 1 & 2 & 1 \\ 0 & -1 & 0 & 1 & 2 \end{pmatrix}$	$\begin{pmatrix} \ & / & \ \end{pmatrix}$
β_n	1	1	0
Spectra(L_n)	$\{0, 0.8299, 2, 2.6889, 4.4812\}$	$\{0, 0.8299, 2, 2.6889, 4.4812\}$	/

Table B.4: Matrix construction of graph G_3 in the top panel of Figure 4.10.

n	$n=0$	$n=1$	$n=2$
Ω_n	span$\{e_1, e_2, e_3, e_4, e_5\}$	span$\{e_{12}, e_{13}, e_{14}, e_{25}, e_{32}, e_{34}, e_{54}\}$	span$\{e_{132}, e_{134}\}$
B_{n+1}	$\begin{array}{c} \\ e_1 \\ e_2 \\ e_3 \\ e_4 \\ e_5 \end{array} \begin{array}{c} \begin{array}{ccccccc} e_{12} & e_{13} & e_{14} & e_{25} & e_{32} & e_{34} & e_{54} \end{array} \\ \begin{pmatrix} -1 & -1 & -1 & 0 & 0 & 0 & 0 \\ 1 & 0 & 0 & -1 & 1 & 0 & 0 \\ 0 & 1 & 0 & 0 & -1 & -1 & 0 \\ 0 & 0 & 1 & 0 & 0 & 1 & 1 \\ 0 & 0 & 0 & 1 & 0 & 0 & -1 \end{pmatrix} \end{array}$	$\begin{array}{c} \\ e_{12} \\ e_{13} \\ e_{14} \\ e_{25} \\ e_{32} \\ e_{34} \\ e_{54} \end{array} \begin{array}{c} \begin{array}{cc} e_{132} & e_{134} \end{array} \\ \begin{pmatrix} -1 & 0 \\ 1 & 1 \\ 0 & -1 \\ 0 & 0 \\ 1 & 0 \\ 0 & 1 \\ 0 & 0 \end{pmatrix} \end{array}$	2×0 empty matrix
L_n	$\begin{pmatrix} 3 & -1 & -1 & -1 & 0 \\ -1 & 3 & -1 & 0 & -1 \\ -1 & -1 & 3 & -1 & 0 \\ -1 & 0 & -1 & 3 & -1 \\ 0 & -1 & 0 & -1 & 2 \end{pmatrix}$	$\begin{pmatrix} 3 & 0 & 1 & -1 & 0 & 0 & 0 \\ 0 & 4 & 0 & 0 & 0 & 0 & 0 \\ 1 & 0 & 3 & 0 & 0 & 0 & 0 \\ -1 & 0 & 0 & 2 & -1 & 0 & -1 \\ 0 & 0 & 0 & -1 & 3 & 1 & 0 \\ 0 & 0 & 0 & 0 & 1 & 3 & 1 \\ 0 & 0 & 1 & -1 & 0 & 1 & 2 \end{pmatrix}$	$\begin{pmatrix} 3 & 1 \\ 1 & 3 \end{pmatrix}$
β_n	1	1	0
Spectra(L_n)	$\{0, 2, 3, 4, 5\}$	$\{0, 2, 2, 3, 4, 4, 5\}$	$\{2, 4\}$

Table B.5: Matrix construction of graph G_4 in the top panel of Figure 4.10.

n	$n=0$	$n=1$	$n=2$
Ω_n	span$\{e_1,e_2,e_3,e_4,e_5\}$	span$\{e_{12},e_{13},e_{14},e_{15},e_{25},e_{32},e_{34},e_{54}\}$	span$\{e_{125},e_{132},e_{134},e_{154}\}$
B_{n+1}	$\begin{array}{c} \\ e_1 \\ e_2 \\ e_3 \\ e_4 \\ e_5 \end{array}\begin{array}{c}\begin{array}{cccccccc} e_{12} & e_{13} & e_{14} & e_{15} & e_{25} & e_{32} & e_{34} & e_{54} \end{array}\\ \begin{pmatrix} -1 & -1 & -1 & -1 & 0 & 0 & 0 & 0 \\ 1 & 0 & 0 & 0 & -1 & 1 & 0 & 0 \\ 0 & 1 & 0 & 0 & 0 & -1 & -1 & 0 \\ 0 & 0 & 1 & 0 & 0 & 0 & 1 & 1 \\ 0 & 0 & 0 & 1 & 1 & 0 & 0 & -1 \end{pmatrix}\end{array}$	$\begin{array}{c} \\ e_{12} \\ e_{13} \\ e_{14} \\ e_{15} \\ e_{25} \\ e_{32} \\ e_{34} \\ e_{54} \end{array}\begin{array}{c}\begin{array}{cccc} e_{125} & e_{132} & e_{134} & e_{154} \end{array}\\ \begin{pmatrix} 1 & -1 & 0 & 0 \\ 0 & 1 & 1 & 0 \\ 0 & 0 & -1 & -1 \\ -1 & 0 & 0 & 1 \\ 1 & 0 & 0 & 0 \\ 0 & 1 & 0 & 0 \\ 0 & 0 & 1 & 0 \\ 0 & 0 & 0 & 1 \end{pmatrix}\end{array}$	4×0 empty matrix
L_n	$\begin{pmatrix} 4 & -1 & -1 & -1 & -1 \\ -1 & 3 & -1 & 0 & -1 \\ -1 & -1 & 3 & -1 & 0 \\ -1 & 0 & -1 & 3 & -1 \\ -1 & -1 & 0 & -1 & 3 \end{pmatrix}$	$\begin{pmatrix} 4 & 0 & 1 & 0 & 0 & 0 & 0 & 0 \\ 0 & 4 & 0 & 1 & 0 & 0 & 0 & 0 \\ 1 & 0 & 4 & 0 & 0 & 0 & 0 & 0 \\ 0 & 1 & 0 & 4 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 3 & -1 & 0 & -1 \\ 0 & 0 & 0 & 0 & -1 & 3 & 1 & 0 \\ 0 & 0 & 0 & 0 & 0 & 1 & 3 & 1 \\ 0 & 0 & 0 & 0 & -1 & 0 & 1 & 3 \end{pmatrix}$	$\begin{pmatrix} 3 & -1 & 0 & -1 \\ -1 & 3 & 1 & 0 \\ 0 & 1 & 3 & 1 \\ -1 & 0 & 1 & 3 \end{pmatrix}$
β_n	1	1	0
Spectra(L_n)	$\{0,3,3,5,5\}$	$\{1,3,3,3,3,5,5,5\}$	$\{1,3,3,5\}$

Table B.6: Matrix construction of graph G_5 in the top panel of Figure 4.10.

n	$n=0$	$n=1$	$n=2$
Ω_n	span$\{e_1,e_2,e_3,e_4,e_5\}$	span$\{e_{12},e_{13},e_{14},e_{15},e_{25},e_{32},e_{34},e_{54}\}$	span$\{e_{125},e_{132},e_{134},e_{154}\}$
B_{n+1}	$\begin{array}{c} \\ e_1 \\ e_2 \\ e_3 \\ e_4 \\ e_5 \end{array}\begin{array}{c}\begin{array}{cccccccc} e_{12} & e_{13} & e_{14} & e_{15} & e_{25} & e_{32} & e_{34} & e_{54} \end{array}\\ \begin{pmatrix} -1 & -1 & -1 & -1 & 0 & 0 & 0 & 0 \\ 1 & 0 & 0 & 0 & -1 & 1 & 0 & 0 \\ 0 & 1 & 0 & 0 & 0 & -1 & -1 & 0 \\ 0 & 0 & 1 & 0 & 0 & 0 & 1 & 1 \\ 0 & 0 & 0 & 1 & 1 & 0 & 0 & -1 \end{pmatrix}\end{array}$	$\begin{array}{c} \\ e_{12} \\ e_{13} \\ e_{14} \\ e_{15} \\ e_{25} \\ e_{32} \\ e_{34} \\ e_{54} \end{array}\begin{array}{c}\begin{array}{cccc} e_{125} & e_{132} & e_{134} & e_{154} \end{array}\\ \begin{pmatrix} 1 & -1 & 0 & 0 \\ 0 & 1 & 1 & 0 \\ 0 & 0 & -1 & -1 \\ -1 & 0 & 0 & 1 \\ 1 & 0 & 0 & 0 \\ 0 & 1 & 0 & 0 \\ 0 & 0 & 1 & 0 \\ 0 & 0 & 0 & 1 \end{pmatrix}\end{array}$	4×0 empty matrix
L_n	$\begin{pmatrix} 4 & -1 & -1 & -1 & -1 \\ -1 & 3 & -1 & 0 & -1 \\ -1 & -1 & 3 & -1 & 0 \\ -1 & 0 & -1 & 3 & -1 \\ -1 & -1 & 0 & -1 & 3 \end{pmatrix}$	$\begin{pmatrix} 4 & 0 & 1 & 0 & 0 & 0 & 0 & 0 \\ 0 & 4 & 0 & 1 & 0 & 0 & 0 & 0 \\ 1 & 0 & 4 & 0 & 0 & 0 & 0 & 0 \\ 0 & 1 & 0 & 4 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 3 & -1 & 0 & -1 \\ 0 & 0 & 0 & 0 & -1 & 3 & 1 & 0 \\ 0 & 0 & 0 & 0 & 0 & 1 & 3 & 1 \\ 0 & 0 & 0 & 0 & -1 & 0 & 1 & 3 \end{pmatrix}$	$\begin{pmatrix} 3 & -1 & 0 & -1 \\ -1 & 3 & 1 & 0 \\ 0 & 1 & 3 & 1 \\ -1 & 0 & 1 & 3 \end{pmatrix}$
β_n	1	0	0
Spectra(L_n)	$\{0,3,3,5,5\}$	$\{1,3,3,3,3,5,5,5\}$	$\{1,3,3,5\}$

Table B.7: Matrix construction of graph G_1 (with isolated points included) in the bottom panel of Figure 4.10.

n	$n=0$	$n=1$	$n=2$
Ω_n	span$\{e_1, e_2, e_3, e_4, e_5\}$	/	/
B_{n+1}	5×0 empty matrix	/	/
L_n	5×5 zero matrix	/	/
β_n	5	/	/
Spectra(L_n)	$\{0, 0, 0, 0, 0\}$	/	/

Table B.8: Matrix construction of graph G_1 (without isolated points) in the bottom panel of Figure 4.10.

n	$n=0$	$n=1$	$n=2$
Ω_n	$\{0\}$	$\{0\}$	$\{0\}$
B_{n+1}	/	/	/
L_n	/	/	/
β_n	/	/	/
Spectra(L_n)	/	/	/

Table B.9: Matrix construction of graph G_2 (with isolated points included) in the bottom panel of Figure 4.10.

n	$n=0$	$n=1$	$n=2$
Ω_n	span$\{e_1, e_2, e_3, e_4, e_5\}$	span$\{e_{25}, e_{32}, e_{34}, e_{54}\}$	$\{0\}$
B_{n+1}	$\begin{array}{c} \\ e_1 \\ e_2 \\ e_3 \\ e_4 \\ e_5 \end{array} \begin{array}{c} \begin{array}{cccc} e_{25} & e_{32} & e_{34} & e_{54} \end{array} \\ \begin{pmatrix} 0 & 0 & 0 & 0 \\ -1 & 1 & 0 & 0 \\ 0 & -1 & -1 & 0 \\ 0 & 0 & 1 & 1 \\ 1 & 0 & 0 & -1 \end{pmatrix} \end{array}$	4×0 empty matrix	$(\ /\ )$
L_n	$\begin{pmatrix} 0 & 0 & 0 & 0 & 0 \\ 0 & 2 & 0 & 0 & -2 \\ 0 & 0 & 1 & 1 & 0 \\ 0 & 0 & 1 & 2 & 1 \\ 0 & -2 & 0 & 1 & 3 \end{pmatrix}$	$\begin{pmatrix} 2 & 0 & 1 & -2 \\ 0 & 2 & -1 & 0 \\ 1 & -1 & 2 & -1 \\ -2 & 0 & -1 & 2 \end{pmatrix}$	$(\ /\ )$
β_n	2	1	0
Spectra(L_n)	$\{0, 0, 0.6571, 2.5293, 4.8136\}$	$\{0, 0.6571, 2.5293, 4.8136\}$	/

Table B.10: Matrix construction of graph G_2 (without isolated points) in the bottom panel of Figure 4.10.

n	$n=0$	$n=1$	$n=2$
Ω_n	span$\{e_2, e_3, e_4, e_5\}$	span$\{e_{25}, e_{32}, e_{34}, e_{54}\}$	$\{0\}$
B_{n+1}	$\begin{array}{c} \\ e_2 \\ e_3 \\ e_4 \\ e_5 \end{array} \begin{array}{c} \begin{array}{cccc} e_{25} & e_{32} & e_{34} & e_{54} \end{array} \\ \begin{pmatrix} -1 & 1 & 0 & 0 \\ 0 & -1 & -1 & 0 \\ 0 & 0 & 1 & 1 \\ 1 & 0 & 0 & -1 \end{pmatrix} \end{array}$	4×0 empty matrix	$(\ /\ )$
L_n	$\begin{pmatrix} 2 & -1 & 0 & -1 \\ -1 & 2 & -1 & 0 \\ 0 & -1 & 2 & -1 \\ -1 & 0 & -1 & 2 \end{pmatrix}$	$\begin{pmatrix} 2 & -1 & 0 & -1 \\ -1 & 2 & -1 & 0 \\ 0 & 1 & 2 & 1 \\ -1 & 0 & 1 & 2 \end{pmatrix}$	$(\ /\ )$
β_n	1	1	0
Spectra(L_n)	$\{0, 2, 2, 4\}$	$\{0, 2, 2, 4\}$	/

Table B.11: Matrix construction of graph G_3 (with isolated points included) in the bottom panel of Figure 4.10.

n	$n=0$	$n=1$	$n=2$
Ω_n	span$\{e_1, e_2, e_3, e_4, e_5\}$	span$\{e_{25}, e_{32}, e_{34}, e_{54}\}$	$\{0\}$
B_{n+1}	$\begin{array}{c} \\ e_1 \\ e_2 \\ e_3 \\ e_4 \\ e_5 \end{array} \begin{array}{c} \begin{array}{cccc} e_{25} & e_{32} & e_{34} & e_{54} \end{array} \\ \begin{pmatrix} 0 & 0 & 0 & 0 \\ -1 & 1 & 0 & 0 \\ 0 & -1 & -1 & 0 \\ 0 & 0 & 1 & 1 \\ 1 & 0 & 0 & -1 \end{pmatrix} \end{array}$	4×0 empty matrix	$(\ /\ )$
L_n	$\begin{pmatrix} 0 & 0 & 0 & 0 & 0 \\ 0 & 2 & 0 & 0 & -2 \\ 0 & 0 & 1 & 1 & 0 \\ 0 & 0 & 1 & 2 & 1 \\ 0 & -2 & 0 & 1 & 3 \end{pmatrix}$	$\begin{pmatrix} 2 & 0 & 1 & -2 \\ 0 & 2 & -1 & 0 \\ 1 & -1 & 2 & -1 \\ -2 & 0 & -1 & 2 \end{pmatrix}$	$(\ /\ )$
β_n	2	1	0
Spectra(L_n)	$\{0, 0, 0.6571, 2.5293, 4.8136\}$	$\{0, 0.6571, 2.5293, 4.8136\}$	/

Table B.12: Matrix construction of graph G_3 (without isolated points) in the bottom panel of Figure 4.10.

n	$n=0$	$n=1$	$n=2$
Ω_n	span$\{e_2, e_3, e_4, e_5\}$	span$\{e_{25}, e_{32}, e_{34}, e_{54}\}$	$\{0\}$
B_{n+1}	$\begin{array}{c} \\ e_2 \\ e_3 \\ e_4 \\ e_5 \end{array} \begin{array}{c} \begin{array}{cccc} e_{25} & e_{32} & e_{34} & e_{54} \end{array} \\ \begin{pmatrix} -1 & 1 & 0 & 0 \\ 0 & -1 & -1 & 0 \\ 0 & 0 & 1 & 1 \\ 1 & 0 & 0 & -1 \end{pmatrix} \end{array}$	4×0 empty matrix	$(\ /\ )$
L_n	$\begin{pmatrix} 2 & -1 & 0 & -1 \\ -1 & 2 & -1 & 0 \\ 0 & -1 & 2 & -1 \\ -1 & 0 & -1 & 2 \end{pmatrix}$	$\begin{pmatrix} 2 & -1 & 0 & -1 \\ -1 & 2 & -1 & 0 \\ 0 & 1 & 2 & 1 \\ -1 & 0 & 1 & 2 \end{pmatrix}$	$(\ /\ )$
β_n	1	1	0
Spectra(L_n)	$\{0, 2, 2, 4\}$	$\{0, 2, 2, 4\}$	/

Table B.13: Matrix construction of graph G_4 in the bottom panel of Figure 4.10.

n	$n=0$	$n=1$	$n=2$
Ω_n	$\text{span}\{e_1, e_2, e_3, e_4, e_5\}$	$\text{span}\{e_{13}, e_{25}, e_{32}, e_{34}, e_{45}\}$	$\{0\}$
B_{n+1}	$\begin{array}{c} \\ e_1 \\ e_2 \\ e_3 \\ e_4 \\ e_5 \end{array} \begin{array}{c} \begin{array}{ccccc} e_{13} & e_{25} & e_{32} & e_{34} & e_{45} \end{array} \\ \begin{pmatrix} -1 & 0 & 0 & 0 & 0 \\ 0 & -1 & 1 & 0 & 0 \\ 1 & 0 & -1 & -1 & 0 \\ 0 & 0 & 0 & 1 & 1 \\ 0 & 1 & 0 & 0 & -1 \end{pmatrix} \end{array}$	5×0 empty matrix	$(\ /\ )$
L_n	$\begin{pmatrix} 1 & 0 & -1 & 0 & 0 \\ 0 & 2 & -1 & 0 & -1 \\ -1 & -1 & 3 & -1 & 0 \\ 0 & 0 & -1 & 2 & -1 \\ 0 & -1 & 0 & -1 & 2 \end{pmatrix}$	$\begin{pmatrix} 2 & 0 & -1 & -1 & 0 \\ 0 & 2 & -1 & 0 & -1 \\ -1 & -1 & 2 & 1 & 0 \\ -1 & 0 & 1 & 2 & 1 \\ 0 & -1 & 0 & 1 & 2 \end{pmatrix}$	$(\ /\ )$
β_n	1	1	0
Spectra(L_n)	$\{0, 0.8299, 2, 2.6889, 4.4812\}$	$\{0, 0.8299, 2, 2.6889, 4.4812\}$	/

Table B.14: Matrix construction of graph G_5 in the bottom panel of Figure 4.10.

n	$n=0$	$n=1$	$n=2$
Ω_n	$\text{span}\{e_1, e_2, e_3, e_4, e_5\}$	$\text{span}\{e_{12}, e_{13}, e_{14}, e_{15}, e_{25}, e_{32}, e_{34}, e_{54}\}$	$\text{span}\{e_{125}, e_{132}, e_{134}, e_{154}\}$
B_{n+1}	$\begin{array}{c} \\ e_1 \\ e_2 \\ e_3 \\ e_4 \\ e_5 \end{array} \begin{array}{c} \begin{array}{cccccccc} e_{12} & e_{13} & e_{14} & e_{15} & e_{25} & e_{32} & e_{34} & e_{54} \end{array} \\ \begin{pmatrix} -1 & -1 & -1 & -1 & 0 & 0 & 0 & 0 \\ 1 & 0 & 0 & 0 & -1 & 1 & 0 & 0 \\ 0 & 1 & 0 & 0 & 0 & -1 & -1 & 0 \\ 0 & 0 & 1 & 0 & 0 & 0 & 1 & 1 \\ 0 & 0 & 0 & 1 & 1 & 0 & 0 & -1 \end{pmatrix} \end{array}$	$\begin{array}{c} \\ e_{12} \\ e_{13} \\ e_{14} \\ e_{15} \\ e_{25} \\ e_{32} \\ e_{34} \\ e_{54} \end{array} \begin{array}{c} \begin{array}{cccc} e_{125} & e_{132} & e_{134} & e_{154} \end{array} \\ \begin{pmatrix} 1 & -1 & 0 & 0 \\ 0 & 1 & 1 & 0 \\ 0 & 0 & -1 & -1 \\ -1 & 0 & 0 & 1 \\ 1 & 0 & 0 & 0 \\ 0 & 1 & 0 & 0 \\ 0 & 0 & 1 & 0 \\ 0 & 0 & 0 & 1 \end{pmatrix} \end{array}$	4×0 empty matrix
L_n	$\begin{pmatrix} 4 & -1 & -1 & -1 & -1 \\ -1 & 3 & -1 & 0 & -1 \\ -1 & -1 & 3 & -1 & 0 \\ -1 & 0 & -1 & 3 & -1 \\ -1 & -1 & 0 & -1 & 3 \end{pmatrix}$	$\begin{pmatrix} 4 & 0 & 1 & 0 & 0 & 0 & 0 & 0 \\ 0 & 4 & 0 & 1 & 0 & 0 & 0 & 0 \\ 1 & 0 & 4 & 0 & 0 & 0 & 0 & 0 \\ 0 & 1 & 0 & 4 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 3 & -1 & 0 & -1 \\ 0 & 0 & 0 & 0 & -1 & 3 & 1 & 0 \\ 0 & 0 & 0 & 0 & 0 & 1 & 3 & 1 \\ 0 & 0 & 0 & 0 & -1 & 0 & 1 & 3 \end{pmatrix}$	$\begin{pmatrix} 3 & -1 & 0 & -1 \\ -1 & 3 & 1 & 0 \\ 0 & 1 & 3 & 1 \\ -1 & 0 & 1 & 3 \end{pmatrix}$
β_n	1	0	0
Spectra(L_n)	$\{0, 3, 3, 5, 5\}$	$\{1, 3, 3, 3, 3, 5, 5, 5\}$	$\{1, 3, 3, 5\}$

BIBLIOGRAPHY

BIBLIOGRAPHY

[1] B. L. Zhang, C. H. Xu, C. Z. Wang, C. T. Chan, and K. M. Ho. Systematic study of structures and stabilities of fullerenes. *Physical Review B*, 46(11):7333–7336, 1992.

[2] Rui Wang, Jiahui Chen, Kaifu Gao, and Guo-Wei Wei. Vaccine-escape and fast-growing mutations in the united kingdom, the united states, singapore, spain, india, and other covid-19-devastated countries. *Genomics*, 113(4):2158–2170, 2021.

[3] Thomas W Linsky, Renan Vergara, Nuria Codina, Jorgen W Nelson, Matthew J Walker, Wen Su, Christopher O Barnes, Tien-Ying Hsiang, Katharina Esser-Nobis, Kevin Yu, et al. De novo design of potent and resilient hACE2 decoys to neutralize SARS-CoV-2. *Science*, 370(6521):1208–1214, 2020.

[4] Kelin Xia and Guo-Wei Wei. Persistent homology analysis of protein structure, flexibility, and folding. *International journal for numerical methods in biomedical engineering*, 30(8):814–844, 2014.

[5] Jacob Townsend, Cassie Putman Micucci, John H Hymel, Vasileios Maroulas, and Konstantinos D Vogiatzis. Representation of molecular structures with persistent homology for machine learning applications in chemistry. *Nature communications*, 11(1):1–9, 2020.

[6] Duc Duy Nguyen, Zixuan Cang, Kedi Wu, Menglun Wang, Yin Cao, and Guo-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.

[7] Primoz Skraba, Maks Ovsjanikov, Frederic Chazal, and Leonidas Guibas. Persistence-based segmentation of deformable shapes. In *2010 IEEE Computer Society Conference on Computer Vision and Pattern Recognition-Workshops*, pages 45–52. IEEE, 2010.

[8] Jozef Dodziuk. de Rham-Hodge theory for L2-cohomology of infinite coverings. *Topology*, 16(2):157–165, 1977.

[9] Jiahui Chen, Rundong Zhao, Yiying Tong, and Guo-Wei Wei. Evolutionary de rham-hodge method. *Discrete and continuous dynamical systems. Series B*, 26(7):3785, 2021.

[10] Mark Kac. Can one hear the shape of a drum? *The american mathematical monthly*, 73(4P2):1–23, 1966.

[11] Rui Wang, Duc Duy Nguyen, and Guo-Wei Wei. Persistent spectral graph. *International Journal for Numerical Methods in Biomedical Engineering*, page e3376, 2020.

[12] Alexander Grigor'yan, Yong Lin, Yuri Muranov, and Shing-Tung Yau. Homologies of path complexes and digraphs. *arXiv preprint arXiv:1207.2834*, 2012.

[13] Samir Chowdhury and Facundo Mémoli. Persistent path homology of directed networks. In *Proceedings of the Twenty-Ninth Annual ACM-SIAM Symposium on Discrete Algorithms*, pages 1152–1169. SIAM, 2018.

[14] Dafydd R Owen, Charlotte MN Allerton, Annaliesa S Anderson, Lisa Aschenbrenner, Melissa Avery, Simon Berritt, Britton Boras, Rhonda D Cardin, Anthony Carlo, Karen J Coffman, et al. An oral sars-cov-2 mpro inhibitor clinical candidate for the treatment of covid-19. *Science*, 374(6575):1586–1593, 2021.

[15] Kaifu Gao, Rui Wang, Jiahui Chen, Jetze J Tepe, Faqing Huang, and Guo-Wei Wei. Perspectives on sars-cov-2 main protease inhibitors. *Journal of medicinal chemistry*, 64(23):16922–16955, 2021.

[16] Matthew D Shin, Sourabh Shukla, Young Hun Chung, Veronique Beiss, Soo Khim Chan, Oscar A Ortega-Rivera, David M Wirth, Angela Chen, Markus Sack, Jonathan K Pokorski, et al. COVID-19 vaccine development and a potential nanomaterial path forward. *Nature Nanotechnology*, pages 1–10, 2020.

[17] Michael Day. COVID-19: four fifths of cases are asymptomatic, China figures indicate. *BMJ*, 369, 2020.

[18] Quan-Xin Long, Xiao-Jun Tang, Qiu-Lin Shi, Qin Li, Hai-Jun Deng, Jun Yuan, Jie-Li Hu, Wei Xu, Yong Zhang, Fa-Jin Lv, et al. Clinical and immunological assessment of asymptomatic SARS-CoV-2 infections. *Nature medicine*, 26(8):1200–1204, 2020.

[19] Rui Wang, Jiahui Chen, Yuta Hozumi, Changchuan Yin, and Guo-Wei Wei. Decoding asymptomatic COVID-19 infection and transmission. *The journal of physical chemistry letters*, 11(23):10007–10015, 2020.

[20] Stephen M Kissler, Christine Tedijanto, Edward Goldstein, Yonatan H Grad, and Marc Lipsitch. Projecting the transmission dynamics of SARS-CoV-2 through the postpandemic period. *Science*, 368(6493):860–868, 2020.

[21] Changchuan Yin. Genotyping coronavirus SARS-CoV-2: methods and implications. *Genomics*, 112(5):3588–3596, 2020.

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

[23] Rui Wang, Jiahui Chen, Kaifu Gao, Yuta Hozumi, Changchuan Yin, and Guo-Wei Wei. Analysis of SARS-CoV-2 mutations in the united states suggests presence of four substrains and novel variants. *Communications biology*, 4(1):1–14, 2021.

[24] Rui Wang, Yuta Hozumi, Yong-Hui Zheng, Changchuan Yin, and Guo-Wei Wei. Host immune response driving SARS-CoV-2 evolution. *Viruses*, 12(10):1095, 2020.

[25] 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*, 2020. PMID: 32530284.

[26] Nanshan Chen, Min Zhou, Xuan Dong, Jieming Qu, Fengyun Gong, Yang Han, Yang Qiu, Jingli Wang, Ying Liu, Yuan Wei, et al. Epidemiological and clinical characteristics of 99 cases of 2019 novel coronavirus pneumonia in Wuhan, China: a descriptive study. *The Lancet*, 395(10223):507–513, 2020.

[27] Roujian Lu, Xiang Zhao, Juan Li, Peihua Niu, Bo Yang, Honglong Wu, Wenling Wang, Hao Song, Baoying Huang, Na Zhu, et al. Genomic characterisation and epidemiology of 2019 novel coronavirus: implications for virus origins and receptor binding. *The Lancet*, 395(10224):565–574, 2020.

[28] Alexandra C Walls, Young-Jun Park, M Alejandra Tortorici, Abigail Wall, Andrew T McGuire, and David Veesler. Structure, function, and antigenicity of the SARS-CoV-2 spike glycoprotein. *Cell*, 181(2):281–292, 2020.

[29] Daniel Wrapp, Nianshuang Wang, Kizzmekia S Corbett, Jory A Goldsmith, Ching-Lin Hsieh, Olubukola Abiona, Barney S Graham, and Jason S McLellan. Cryo-EM structure of the 2019-nCoV spike in the prefusion conformation. *Science*, 367(6483):1260–1263, 2020.

[30] Christian Jean Michel, Claudine Mayer, Olivier Poch, and Julie Dawn Thompson. Characterization of accessory genes in coronavirus genomes. *Virology journal*, 17(1):1–13, 2020.

[31] Yosra A Helmy, Mohamed Fawzy, Ahmed Elaswad, Ahmed Sobieh, Scott P Kenney, and Awad A Shehata. The COVID-19 pandemic: a comprehensive review of taxonomy, genetics, epidemiology, diagnosis, treatment, and control. *Journal of Clinical Medicine*, 9(4):1225, 2020.

[32] Ahmad Abu Turab Naqvi, Kisa Fatima, Taj Mohammad, Urooj Fatima, Indrakant K Singh, Archana Singh, Shaikh Muhammad Atif, Gururao Hariprasad, Gulam Mustafa Hasan, and Md Imtaiyaz Hassan. Insights into SARS-CoV-2 genome, structure, evolution, pathogenesis and therapies: Structural genomics approach. *Biochimica et Biophysica Acta (BBA)-Molecular Basis of Disease*, 1866(10):165878, 2020.

[33] Jingfang Mu, Yaohui Fang, Qi Yang, Ting Shu, An Wang, Muhan Huang, Liang Jin, Fei Deng, Yang Qiu, and Xi Zhou. SARS-CoV-2 N protein antagonizes type I interferon signaling by suppressing phosphorylation and nuclear translocation of STAT1 and STAT2. *Cell discovery*, 6(1):1–4, 2020.

[34] Canrong Wu, Yang Liu, Yueying Yang, Peng Zhang, Wu Zhong, Yali Wang, Qiqi Wang, Yang Xu, Mingxue Li, Xingzhou Li, et al. Analysis of therapeutic targets for SARS-CoV-2 and discovery of potential drugs by computational methods. *Acta Pharmaceutica Sinica B*, 10(5):766–788, 2020.

[35] Dongwan Kim, Joo-Yeon Lee, Jeong-Sun Yang, Jun Won Kim, V Narry Kim, and Hyeshik Chang. The architecture of SARS-CoV-2 transcriptome. *Cell*, 181(4):914–921, 2020.

[36] Shutoku Matsuyama, Naganori Nao, Kazuya Shirato, Miyuki Kawase, Shinji Saito, Ikuyo Takayama, Noriyo Nagata, Tsuyoshi Sekizuka, Hiroshi Katoh, Fumihiro Kato, et al. Enhanced isolation of SARS-CoV-2 by TMPRSS2-expressing cells. *Proceedings of the National Academy of Sciences*, 117(13):7001–7003, 2020.

[37] Markus Hoffmann, Hannah Kleine-Weber, Simon Schroeder, Nadine Krüger, Tanja Herrler, Sandra Erichsen, Tobias S Schiergens, Georg Herrler, Nai-Huei Wu, Andreas Nitsche, et al. SARS-CoV-2 cell entry depends on ACE2 and TMPRSS2 and is blocked by a clinically proven protease inhibitor. *cell*, 181(2):271–280, 2020.

[38] Philip V'kovski, Annika Kratzel, Silvio Steiner, Hanspeter Stalder, and Volker Thiel. Coronavirus biology and replication: implications for SARS-CoV-2. *Nature Reviews Microbiology*, pages 1–16, 2020.

[39] Jiahui Chen, Kaifu Gao, Rui Wang, Duc Duy Nguyen, and Guo-Wei Wei. Review of covid-19 antibody therapies. *Annual review of biophysics*, 50:1–30, 2021.

[40] 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.

[41] 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.

[42] Peter Richardson, Ivan Griffin, Catherine Tucker, Dan Smith, Olly Oechsle, Anne Phelan, Michael Rawling, Edward Savory, and Justin Stebbing. Baricitinib as potential treatment for 2019-ncov acute respiratory disease. *Lancet (London, England)*, 395(10223):e30, 2020.

[43] Herbert Edelsbrunner and John Harer. Persistent homology-a survey. *Contemporary mathematics*, 453:257–282, 2008.

[44] Zixuan Cang and Guo-Wei Wei. Topologynet: Topology based deep convolutional and multi-task neural networks for biomolecular property predictions. *PLoS computational biology*, 13(7):e1005690, 2017.

[45] Daniel Hernández Serrano and Darío Sánchez Gómez. Centrality measures in simplicial complexes: applications of tda to network science. *arXiv preprint arXiv:1908.02967*, 2019.

[46] Slobodan Maletić and Milan Rajković. Consensus formation on a simplicial complex of opinions. *Physica A: Statistical Mechanics and its Applications*, 397(March):111–120, 2014.

[47] Herbert Edelsbrunner. Alpha shapes—a survey. *Tessellations in the Sciences*, 27:1–25, 2010.

[48] Georges Voronoi. Nouvelles applications des paramètres continus à la théorie des formes quadratiques. premier mémoire. sur quelques propriétés des formes quadratiques positives parfaites. *Journal für die reine und angewandte Mathematik*, 1908(133):97–102, 1908.

[49] Boris Delaunay et al. Sur la sphere vide. *Izv. Akad. Nauk SSSR, Otdelenie Matematicheskii i Estestvennyka Nauk*, 7(793-800):1–2, 1934.

[50] Franz Aurenhammer, Rolf Klein, and Der-Tsai Lee. *Voronoi diagrams and Delaunay triangulations*. World Scientific Publishing Company, 2013.

[51] Jude May. *Multivariate analysis*. Scientific e-Resources, 2018.

[52] Beno Eckmann. Harmonische funktionen und randwertaufgaben in einem komplex. *Commentarii Mathematici Helvetici*, 17(1):240–255, 1944.

[53] Daniel Hernández Serrano and Darío Sánchez Gómez. Higher order degree in simplicial complexes, multi combinatorial laplacian and applications of tda to complex networks. *arXiv preprint arXiv:1908.02583*, 2019.

[54] Franz W Kamber and Philippe Tondeur. de rham-hodge theory for riemannian foliations. *Mathematische Annalen*, 277(3):415–431, 1987.

[55] 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(8):1–38, 2020.

[56] Jiahui Chen, Rundong Zhao, Yiying Tong, and Guo-Wei Wei. Evolutionary de Rham-hodge method. *Discrete & Continuous Dynamical Systems-B*, 2020.

[57] Kelin Xia, Kristopher Opron, and Guo-Wei Wei. Multiscale gaussian network model (mgnm) and multiscale anisotropic network model (manm). *The Journal of chemical physics*, 143(20):11B616_1, 2015.

[58] Marcel Berger. *Geometry i*. Springer Science & Business Media, 2009.

[59] AA Grigor'yan, Yong Lin, Yu V Muranov, and Shing-Tung Yau. Path complexes and their homologies. *Journal of Mathematical Sciences*, 248(5):564–599, 2020.

[60] Alexander Grigor'yan, Yong Lin, Yuri Muranov, and Shing-Tung Yau. Cohomology of digraphs and (undirected) graphs. *Asian Journal of Mathematics*, 19(5):887–932, 2015.

[61] Gary Chartrand. *Introductory graph theory*. Courier Corporation, 1977.

[62] André Gomes and Daniel Miranda. Path cohomology of locally finite digraphs, hodge's theorem and the p-lazy random walk. *arXiv preprint arXiv:1906.04781*, 2019.

[63] Alexander Grigor'yan, Yong Lin, Yuri Muranov, and Shing-Tung Yau. Homotopy theory for digraphs. *arXiv preprint arXiv:1407.0234*, 2014.

[64] Danijela Horak and Jürgen Jost. Spectra of combinatorial laplace operators on simplicial complexes. *Advances in Mathematics*, 244:303–336, 2013.

[65] Martin Gollery. Bioinformatics: Sequence and genome analysis, david w. mount. cold spring harbor, ny: Cold spring harbor laboratory press, 2004, 692 pp. isbn 0-87969-712-1. *Clinical Chemistry*, 51(11):2219–2219, 2005.

[66] W John Wilbur and David J Lipman. Rapid similarity searches of nucleic acid and protein data banks. *Proceedings of the National Academy of Sciences*, 80(3):726–730, 1983.

[67] Stephen F Altschul, Warren Gish, Webb Miller, Eugene W Myers, and David J Lipman. Basic local alignment search tool. *Journal of molecular biology*, 215(3):403–410, 1990.

[68] Jian Ye, Scott McGinnis, and Thomas L Madden. Blast: improvements for better sequence analysis. *Nucleic acids research*, 34(suppl_2):W6–W9, 2006.

[69] David W Mount. Using the basic local alignment search tool (blast). *Cold Spring Harbor Protocols*, 2007(7):pdb–top17, 2007.

[70] Tao Zhang, Qunfu Wu, and Zhigang Zhang. Probable pangolin origin of sars-cov-2 associated with the covid-19 outbreak. *Current Biology*, 2020.

[71] Kangpeng Xiao, Junqiong Zhai, Yaoyu Feng, Niu Zhou, Xu Zhang, Jie-Jian Zou, Na Li, Yaqiong Guo, Xiaobing Li, Xuejuan Shen, et al. Isolation of sars-cov-2-related coronavirus from malayan pangolins. *Nature*, pages 1–4, 2020.

[72] Hongru Wang, Lenore Pipes, and Rasmus Nielsen. Synonymous mutations and the molecular evolution of sars-cov-2 origins. *Virus evolution*, 7(1):veaa098, 2021.

[73] Giuseppina La Rosa, Pamela Mancini, Giusy Bonanno Ferraro, Carolina Veneri, Marcello Iaconelli, Lucia Bonadonna, Luca Lucentini, and Elisabetta Suffredini. Sars-cov-2 has been circulating in northern italy since december 2019: Evidence from environmental monitoring. *Science of the total environment*, 750:141711, 2021.

[74] Ranjit Sah, Alfonso J Rodriguez-Morales, Runa Jha, Daniel KW Chu, Haogao Gu, Malik Peiris, Anup Bastola, Bibek Kumar Lal, Hemant Chanda Ojha, Ali A Rabaan, et al. Complete genome sequence of a 2019 novel coronavirus (sars-cov-2) strain isolated in nepal. *Microbiology resource announcements*, 9(11):e00169–20, 2020.

[75] Giuseppina La Rosa, Marcello Iaconelli, Pamela Mancini, Giusy Bonanno Ferraro, Carolina Veneri, Lucia Bonadonna, Luca Lucentini, and Elisabetta Suffredini. First detection of sars-cov-2 in untreated wastewaters in italy. *Science of The Total Environment*, 736:139652, 2020.

[76] Sandra Westhaus, Frank-Andreas Weber, Sabrina Schiwy, Volker Linnemann, Markus Brinkmann, Marek Widera, Carola Greve, Axel Janke, Henner Hollert, Thomas Wintgens, et al. Detection of sars-cov-2 in raw and treated wastewater in germany–suitability for covid-19 surveillance and potential transmission risks. *Science of The Total Environment*, 751:141750, 2021.

[77] Coronaviridae Study Group of the International et al. The species severe acute respiratory syndrome-related coronavirus: classifying 2019-ncov and naming it sars-cov-2. *Nature Microbiology*, 5(4):536, 2020.

[78] Desmond G Higgins and Paul M Sharp. Clustal: a package for performing multiple sequence alignment on a microcomputer. *Gene*, 73(1):237–244, 1988.

[79] Robert C Edgar. Muscle: a multiple sequence alignment method with reduced time and space complexity. *BMC bioinformatics*, 5(1):113, 2004.

[80] Kazutaka Katoh, George Asimenos, and Hiroyuki Toh. Multiple alignment of dna sequences with mafft. In *Bioinformatics for DNA sequence analysis*, pages 39–64. Springer, 2009.

[81] Kazutaka Katoh, Kazuharu Misawa, Kei-ichi Kuma, and Takashi Miyata. Mafft: a novel method for rapid multiple sequence alignment based on fast fourier transform. *Nucleic acids research*, 30(14):3059–3066, 2002.

[82] Julie D Thompson, Desmond G Higgins, and Toby J Gibson. Clustal w: improving the sensitivity of progressive multiple sequence alignment through sequence weighting, position-specific gap penalties and weight matrix choice. *Nucleic acids research*, 22(22):4673–4680, 1994.

[83] Mark A Larkin, Gordon Blackshields, Nigel P Brown, R Chenna, Paul A McGettigan, Hamish McWilliam, Franck Valentin, Iain M Wallace, Andreas Wilm, Rodrigo Lopez, et al. Clustal w and clustal x version 2.0. *bioinformatics*, 23(21):2947–2948, 2007.

[84] Naruya Saitou and Masatoshi Nei. The neighbor-joining method: a new method for reconstructing phylogenetic trees. *Molecular biology and evolution*, 4(4):406–425, 1987.

[85] Gordon Blackshields, Fabian Sievers, Weifeng Shi, Andreas Wilm, and Desmond G Higgins. Sequence embedding for fast construction of guide trees for multiple sequence alignment. *Algorithms for Molecular Biology*, 5(1):21, 2010.

[86] Johannes Söding. Protein homology detection by hmm–hmm comparison. *Bioinformatics*, 21(7):951–960, 2005.

[87] Michael Levandowsky and David Winter. Distance between sets. *Nature*, 234(5323):34–35, 1971.

[88] Thomas CoVer and Peter Hart. Nearest neighbor pattern classification. *IEEE transactions on information theory*, 13(1):21–27, 1967.

[89] Naomi S Altman. An introduction to kernel and nearest-neighbor nonparametric regression. *The American Statistician*, 46(3):175–185, 1992.

[90] Yuelong Shu and John McCauley. GISAID: Global initiative on sharing all influenza data–from vision to reality. *Eurosurveillance*, 22(13):30494, 2017.

[91] Fan Wu, Su Zhao, Bin Yu, Yan-Mei Chen, Wen Wang, Zhi-Gang Song, Yi Hu, Zhao-Wu Tao, Jun-Hua Tian, Yuan-Yuan Pei, et al. A new coronavirus associated with human respiratory disease in China. *Nature*, 579(7798):265–269, 2020.

[92] Sobin Kim and Ashish Misra. Snp genotyping: technologies and biomedical applications. *Annu. Rev. Biomed. Eng.*, 9:289–320, 2007.

[93] 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.

[94] 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.

[95] 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.

[96] Quanya Liu, Peng Chen, Bing Wang, Jun Zhang, and Jinyan Li. dbmpikt: a database of kinetic and thermodynamic mutant protein interactions. *Bmc Bioinformatics*, 19(1):1–7, 2018.

[97] Erik Procko. The sequence of human ace2 is suboptimal for binding the s spike protein of sars coronavirus 2. *BioRxiv*, 2020.

[98] Tyler N Starr, Allison J Greaney, Sarah K Hilton, Daniel Ellis, Katharine HD Crawford, Adam S Dingens, Mary Jane Navarro, John E Bowen, M Alejandra Tortorici, Alexandra C Walls, et al. Deep mutational scanning of SARS-CoV-2 receptor binding domain reveals constraints on folding and ACE2 binding. *Cell*, 182(5):1295–1310, 2020.

[99] Jiahui Chen, Kaifu Gao, Rui Wang, and Guo-Wei Wei. Revealing the threat of emerging sars-cov-2 mutations to antibody therapies. *Journal of molecular biology*, 433(18):167155, 2021.

[100] Gunnar Carlsson. Topology and data. *Bulletin of the American Mathematical Society*, 46(2):255–308, 2009.

[101] Herbert Edelsbrunner, David Letscher, and Afra Zomorodian. Topological persistence and simplification. In *Proceedings 41st Annual Symposium on Foundations of Computer Science*, pages 454–463. IEEE, 2000.

[102] Jiahui Chen, Kaifu Gao, Rui Wang, and Guo-Wei Wei. Prediction and mitigation of mutation threats to covid-19 vaccines and antibody therapies. *Chemical science*, 12(20):6929–6948, 2021.

[103] Delphine C Bas, David M Rogers, and Jan H Jensen. Very fast prediction and rationalization of pka values for protein–ligand complexes. *Proteins: Structure, Function, and Bioinformatics*, 73(3):765–783, 2008.

[104] Stephen F Altschul, Thomas L Madden, Alejandro A Schäffer, Jinghui Zhang, Zheng Zhang, Webb Miller, and David J Lipman. Gapped blast and psi-blast: a new generation of protein database search programs. *Nucleic acids research*, 25(17):3389–3402, 1997.

[105] Yuedong Yang, Rhys Heffernan, Kuldip Paliwal, James Lyons, Abdollah Dehzangi, Alok Sharma, Jihua Wang, Abdul Sattar, and Yaoqi Zhou. Spider2: A package to predict secondary structure, accessible surface area, and main-chain torsional angles by deep neural networks. In *Prediction of protein secondary structure*, pages 55–63. Springer, 2017.

[106] Beibei Liu, Bao Wang, Rundong Zhao, Yiying Tong, and Guo-Wei Wei. Eses: software for e ulerian solvent excluded surface, 2017.

[107] Todd J Dolinsky, Jens E Nielsen, J Andrew McCammon, and Nathan A Baker. Pdb2pqr: an automated pipeline for the setup of poisson–boltzmann electrostatics calculations. *Nucleic acids research*, 32(suppl_2):W665–W667, 2004.

[108] David A Case, Tom A Darden, Thomas E Cheatham, Carlos L Simmerling, Junmei Wang, Robert E Duke, Ray Luo, MRCW Crowley, Ross C Walker, Wei Zhang, et al. Amber 10. Technical report, University of California, 2008.

[109] Bernard R Brooks, Charles L Brooks III, Alexander D Mackerell Jr, Lennart Nilsson, Robert J Petrella, Benoît Roux, Youngdo Won, Georgios Archontis, Christian Bartels, Stefan Boresch, et al. Charmm: the biomolecular simulation program. *Journal of computational chemistry*, 30(10):1545–1614, 2009.

[110] Duan Chen, Zhan Chen, Changjun Chen, Weihua Geng, and Guo-Wei Wei. Mibpb: a software package for electrostatic analysis. *Journal of computational chemistry*, 32(4):756–770, 2011.

[111] Fabian Pedregosa, Gaël Varoquaux, Alexandre Gramfort, Vincent Michel, Bertrand Thirion, Olivier Grisel, Mathieu Blondel, Peter Prettenhofer, Ron Weiss, Vincent Dubourg, et al. Scikit-learn: Machine learning in python. *the Journal of machine Learning research*, 12:2825–2830, 2011.

[112] Nitish Srivastava, Geoffrey Hinton, Alex Krizhevsky, Ilya Sutskever, and Ruslan Salakhutdinov. Dropout: a simple way to prevent neural networks from overfitting. *The journal of machine learning research*, 15(1):1929–1958, 2014.

[113] Shlomo Hoory, Nathan Linial, and Avi Wigderson. Expander graphs and their applications. *Bulletin of the American Mathematical Society*, 43(4):439–561, 2006.

[114] Fan Chung. Laplacians and the cheeger inequality for directed graphs. *Annals of Combinatorics*, 9(1):1–19, 2005.

[115] Fan R. K. Chung. Spectral Graph Theory. *AMS*, 1997.

[116] Robert Grone, Russell Merris, and V S_ Sunder. The laplacian spectrum of a graph. *SIAM Journal on Matrix Analysis and Applications*, 11(2):218–238, 1990.

[117] Stephen J. Kirkland, Jason J. Molitierno, Michael Neumann, and Bryan L. Shader. On graphs with equal algebraic and vertex connectivity. *Linear Algebra and its Applications*, 341(1-3):45–56, 2002.

[118] Xiao-Dong Zhang. The laplacian eigenvalues of graphs: a survey. *arXiv preprint arXiv:1111.2897*, 2011.

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

[120] Timothy E Goldberg. Combinatorial laplacians of simplicial complexes. *Senior Thesis, Bard College*, 2002.

[121] Patrizio Frosini. Measuring shapes by size functions. In *Intelligent Robots and Computer Vision X: Algorithms and Techniques*, volume 1607, pages 122–133. International Society for Optics and Photonics, 1992.

[122] Afra Zomorodian and Gunnar Carlsson. Computing persistent homology. *Discrete & Computational Geometry*, 33(2):249–274, 2005.

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

[124] Gunnar Carlsson, Vin De Silva, and Dmitriy Morozov. Zigzag persistent homology and real-valued functions. In *Proceedings of the twenty-fifth annual symposium on Computational geometry*, pages 247–256. ACM, 2009.

[125] Vin De Silva and Robert Ghrist. Coverage in sensor networks via persistent homology. *Algebraic & Geometric Topology*, 7(1):339–358, 2007.

[126] 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.

[127] Peter Bubenik and Jonathan A Scott. Categorification of persistent homology. *Discrete & Computational Geometry*, 51(3):600–627, 2014.

[128] Tamal K Dey, Fengtao Fan, and Yusu Wang. Computing topological persistence for simplicial maps. In *Proceedings of the thirtieth annual symposium on Computational geometry*, page 345. ACM, 2014.

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

[130] Ramón García-Domenech, Jorge Gálvez, Jesus V. de Julián-Ortiz, and Lionello Pogliani. Some new trends in chemical graph theory. *Chemical Reviews*, 108(3):1127–1169, 2008.

[131] K. Balasubramanian. Applications of Combinatorics and Graph Theory to Spectroscopy and Quantum Chemistry. *Chemical Reviews*, 85(6):599–618, 1985.

[132] Ivan Gutman and Nenad Trinajstić. Graph theory and molecular orbitals. total φ-electron energy of alternant hydrocarbons. *Chemical Physics Letters*, 17(4):535–538, 1972.

[133] Ivet Bahar, Ali Rana Atilgan, and Burak Erman. Direct evaluation of thermal fluctuations in proteins using a single-parameter harmonic potential. *Folding and Design*, 2(3):173–181, 1997.

[134] A. R. Atilgan, S. R. Durell, R. L. Jernigan, M. C. Demirel, O. Keskin, and I. Bahar. Anisotropy of fluctuation dynamics of proteins with an elastic network model. *Biophysical Journal*, 80(1):505–515, 2001.

[135] Ivet Bahar, Ali Rana Atilgan, Melik C. Demirel, and Burak Erman. Vibrational dynamics of folded proteins: Significance of slow and fast motions in relation to function and stability. *Physical Review Letters*, 80(12):2733–2736, 1998.

[136] Kristopher Opron, Kelin Xia, and Guo Wei Wei. Communication: Capturing protein multiscale thermal fluctuations, 2015.

[137] David Bramer and Guo-Wei Wei. Multiscale weighted colored graphs for protein flexibility and rigidity analysis. *The Journal of chemical physics*, 148(5):054103, 2018.

[138] Duc Nguyen and Guo-Wei Wei. Agl-score: Algebraic graph learning score for protein-ligand binding scoring, ranking, docking, and screening. *Journal of Chemical Information and Modeling*, 2019.

[139] H.W. Kroto, J.R. Heath, S.C. O'Brien, R.F. Curl, and R E Smalley. C₆₀: Buckminsterfullerene. *Nature*, 318(14):162–163, 1985.

[140] W. Krätschmer, Lowell D. Lamb, K. Fostiropoulos, and Donald R. Huffman. Solid C60: a new form of carbon. *Nature*, 347(6291):354–358, 1990.

[141] B C Yadav and Ritesh Kumar. Structure , properties and applications of fullerenes. *International Journal of Nanotechnology and Applications ISSN*, 0973(1):15–24, 2008.

[142] Kelin Xia, Xin Feng, Yiying Tong, and Guo Wei Wei. Persistent homology for the quantitative prediction of fullerene stability. *Journal of computational chemistry*, 36(6):408–422, 2015.

[143] Kelin Xia and Guo-Wei Wei. Persistent homology analysis of protein structure, flexibility, and folding. *International Journal for Numerical Methods in Biomedical Engineering*, (June):814–844, 2014.

[144] B. L. Zhang, C. Z. Wang, K. M. Ho, C. H. Xu, and C. T. Chan. The geometry of small fullerene cages: C20 to C70. *The Journal of Chemical Physics*, 97(7):5007–5011, 1992.

[145] David Bramer and Guo-Wei Wei. Blind prediction of protein b-factor and flexibility. *The Journal of chemical physics*, 149(13):134107, 2018.

[146] Kristopher Opron, Kelin Xia, and Guo-Wei Wei. Fast and anisotropic flexibility-rigidity index for protein flexibility and fluctuation analysis. *The Journal of chemical physics*, 140(23):06B617_1, 2014.

[147] Kelin Xia, Kristopher Opron, and Guo-Wei Wei. Multiscale multiphysics and multidomain models—flexibility and rigidity. *The Journal of chemical physics*, 139(19):11B614_1, 2013.

[148] Jelena Grbic, Jie Wu, Kelin Xia, and Guo-Wei Wei. Aspects of topological approaches for data science. *Foundations of Data Science*, 2022.

[149] Yiying Tong, Santiago Lombeyda, Anil N Hirani, and Mathieu Desbrun. Discrete multiscale vector field decomposition. *ACM transactions on graphics (TOG)*, 22(3):445–452, 2003.

[150] Yoshihiko Mochizuki and Atsushi Imiya. Spatial reasoning for robot navigation using the helmholtz-hodge decomposition of omnidirectional optical flow. In *2009 24th International Conference Image and Vision Computing New Zealand*, pages 1–6. IEEE, 2009.

[151] Facundo Mémoli, Zhengchao Wan, and Yusu Wang. Persistent Laplacians: properties, algorithms and implications. *42nd Conference on Very Important Topics, Digital Object Identifier: 10.4230/LIPIcs.CVIT.2016.23*, 2020.

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

[153] Jiahui Chen, Yuchi Qiu, Rui Wang, and Guo-Wei Wei. Persistent laplacian projected omicron ba. 4 and ba. 5 to become new dominating variants. *arXiv preprint arXiv:2205.00532*, 2022.

[154] 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.

[155] Allen Dudley Shepard. *A cellular description of the derived category of a stratified space*. PhD thesis, Brown University, 1985.

[156] Jakob Hansen and Robert Ghrist. Toward a spectral theory of cellular sheaves. *Journal of Applied and Computational Topology*, 3(4):315–358, 2019.

[157] Xiaoqi Wei and Guo-Wei Wei. Persistent sheaf laplacians. *arXiv preprint arXiv:2112.10906*, 2021.

[158] Bernardo Ameneyro, Vasileios Maroulas, and George Siopsis. Quantum persistent homology. *arXiv preprint arXiv:2202.12965*, 2022.

[159] Terri A Long, Siobhan M Brady, and Philip N Benfey. Systems approaches to identifying gene regulatory networks in plants. *Annual review of cell and developmental biology*, 24:81–103, 2008.

[160] Alexander Grigor'yan, Yuri Muranov, Vladimir Vershinin, and Shing-Tung Yau. Path homology theory of multigraphs and quivers. In *Forum mathematicum*, volume 30, pages 1319–1337. De Gruyter, 2018.

[161] Alexander Grigor'yan, Rolando Jimenez, Yuri Muranov, and Shing-Tung Yau. On the path homology theory of digraphs and eilenberg–steenrod axioms. *Homology, Homotopy and Applications*, 20(2):179–205, 2018.

[162] Alexander Grigor'yan, Rolando Jimenez, Yuri Muranov, and Shing-Tung Yau. Homology of path complexes and hypergraphs. *Topology and its Applications*, 267:106877, 2019.

[163] Yong Lin, Shiquan Ren, Chong Wang, and Jie Wu. Weighted path homology of weighted digraphs and persistence. *arXiv preprint arXiv:1910.09891*, 2019.

[164] Tamal K Dey, Tianqi Li, and Yusu Wang. An efficient algorithm for 1-dimensional (persistent) path homology. *arXiv preprint arXiv:2001.09549*, 2020.

[165] Kaifu Gao, Jian Yin, Niel M Henriksen, Andrew T Fenley, and Michael K Gilson. Binding enthalpy calculations for a neutral host–guest pair yield widely divergent salt effects across water models. *Journal of chemical theory and computation*, 11(10):4555–4564, 2015.

[166] Linus Pauling. The nature of the chemical bond. iv. the energy of single bonds and the relative electronegativity of atoms. *Journal of the American Chemical Society*, 54(9):3570–3582, 1932.

[167] Sinan G Aksoy, Cliff Joslyn, Carlos Ortiz Marrero, Brenda Praggastis, and Emilie Purvine. Hypernetwork science via high-order hypergraph walks. *EPJ Data Science*, 9(1):16, 2020.

[168] Stephane Bressan, Jingyan Li, Shiquan Ren, and Jie Wu. The embedded homology of hypergraphs and applications. *arXiv preprint arXiv:1610.00890*, 2016.

[169] Daniel A Spielman. Spectral graph theory and its applications. In *48th Annual IEEE Symposium on Foundations of Computer Science (FOCS'07)*, pages 29–38. IEEE, 2007.

[170] Jeff Cheeger. A lower bound for the smallest eigenvalue of the Laplacian. In *Proceedings of the Princeton conference in honor of Professor S. Bochner*, pages 195–199, 1969.

[171] Fan RK Chung and Fan Chung Graham. *Spectral graph theory*, volume 92. American Mathematical Soc., 1997.

[172] Joel Friedman. Computing betti numbers via combinatorial laplacians. *Algorithmica*, 21(4):331–346, 1998.

[173] Tomasz Kaczynski, Konstantin Mischaikow, and Marian Mrozek. *Computational homology*, volume 157. Springer Science & Business Media, 2006.

[174] Peter Bubenik, Peter T Kim, et al. A statistical approach to persistent homology. *Homology, homotopy and Applications*, 9(2):337–362, 2007.

[175] Yongjin Lee, Senja D Barthel, Paweł Dłotko, S Mohamad Moosavi, Kathryn Hess, and Berend Smit. Quantifying similarity of pore-geometry in nanoporous materials. *Nature communications*, 8(1):1–8, 2017.

[176] Vasileios Maroulas, Cassie Putman Micucci, and Farzana Nasrin. Bayesian Topological Learning for Classifying the Structure of Biological networks. *arXiv preprint arXiv:2009.11974*, 2020.

[177] Maria-Veronica Ciocanel, Riley Juenemann, Adriana T Dawes, and Scott A McKinley. Topological data analysis approaches to uncovering the timing of ring structure onset in filamentous networks. *Bulletin of Mathematical Biology*, 83(3):1–25, 2021.

[178] Ioannis Sgouralis, Andreas Nebenfuhr, and Vasileios Maroulas. A bayesian topological framework for the identification and reconstruction of subcellular motion. *SIAM Journal on Imaging Sciences*, 10(2):871–899, 2017.

[179] Zhenyu Meng, D Vijay Anand, Yunpeng Lu, Jie Wu, and Kelin Xia. Weighted persistent homology for biomolecular data analysis. *Scientific reports*, 10(1):1–15, 2020.

[180] Gunnar Carlsson, Afra Zomorodian, Anne Collins, and Leonidas J Guibas. Persistence barcodes for shapes. *International Journal of Shape Modeling*, 11(02):149–187, 2005.

[181] Zhenyu Meng and Kelin Xia. Persistent spectral based machine learning (perspect ml) for drug design. *arXiv:2002.00582*, 2020.

[182] Ulrich Bauer. Ripser: a lean C++ code for the computation of Vietoris–Rips persistence barcodes. *Software available at https://github. com/Ripser/ripser*, 436, 2017.

[183] Dmitriy Morozov. Dionysus Software, 2012.

[184] GUDHI Project. GUDHI User and reference manual, 2015.

[185] Ulrich Bauer, Michael Kerber, and Jan Reininghaus. Dipha (a distributed persistent homology algorithm). *Software available at https://github. com/DIPHA/dipha*, 2014.

[186] Henry Adams, Andrew Tausz, and Mikael Vejdemo-Johansson. JavaPlex: A research software package for persistent (co) homology. In *International Congress on Mathematical Software*, pages 129–136. Springer, 2014.

[187] Chad Giusti, Eva Pastalkova, Carina Curto, and Vladimir Itskov. Clique topology reveals intrinsic geometric structure in neural correlations. *Proceedings of the National Academy of Sciences*, 112(44):13455–13460, 2015.

[188] Dmitriy Morozov and Primoz Skraba. DioDe Software.

[189] Brittany T Fasy, Jisu Kim, Fabrizio Lecci, Clement Maria, David L Millman, and Maintainer Jisu Kim. Package 'TDA', 2019.

[190] Michael Kerber and Herbert Edelsbrunner. The medusa of spatial sorting: 3D kinetic alpha complexes and implementation. *arXiv preprint arXiv:1209.5434*, 2012.

[191] Rundong Zhao, Mathieu Desbrun, Guo-Wei Wei, and Yiying Tong. 3D hodge decompositions of edge-and face-based vector fields. *ACM Transactions on Graphics (TOG)*, 38(6):1–13, 2019.

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

[193] WHO. Coronavirus disease 2019 (COVID-19) situation report – 172. *Coronavirus Disease (COVID-2019) Situation Reports*, 2020.

[194] Jasper Fuk-Woo Chan, Cyril Chik-Yan Yip, Kelvin Kai-Wang To, Tommy Hing-Cheung Tang, Sally Cheuk-Ying Wong, Kit-Hang Leung, Agnes Yim-Fong Fung, Anthony Chin-Ki Ng, Zijiao Zou, Hoi-Wah Tsoi, et al. Improved molecular diagnosis of covid-19 by the novel, highly sensitive and specific covid-19-rdrp/hel real-time reverse transcription-pcr assay validated in vitro and with clinical specimens. *Journal of clinical microbiology*, 58(5):e00310–20, 2020.

[195] Victor M Corman, Olfert Landt, Marco Kaiser, Richard Molenkamp, Adam Meijer, Daniel KW Chu, Tobias Bleicker, Sebastian Brünink, Julia Schneider, Marie Luisa Schmidt, et al. Detection of 2019 novel coronavirus (2019-nCoV) by real-time RT-PCR. *Eurosurveillance*, 25(3):2000045, 2020.

[196] Buddhisha Udugama, Pranav Kadhiresan, Hannah N Kozlowski, Ayden Malekjahani, Matthew Osborne, Vanessa YC Li, Hongmin Chen, Samira Mubareka, Jonathan Gubbay, and Warren CW Chan. Diagnosing COVID-19: The disease and tools for ddtection. *ACS nano*, 2020.

[197] Yujin Jung, Gun-Soo Park, Jun Hye Moon, Keunbon Ku, Seung-Hwa Beak, Chang-Seop Lee, Seil Kim, Edmond Changkyun Park, Daeui Park, Jong-Hwan Lee, et al. Comparative analysis of primer–probe sets for rt-qpcr of covid-19 causative virus (sars-cov-2). *ACS infectious diseases*, 6(9):2513–2523, 2020.

[198] Susanne Pfefferle, Svenja Reucher, Dominic Nörz, and Marc Lütgehetmann. Evaluation of a quantitative rt-pcr assay for the detection of the emerging coronavirus SARS-CoV-2 using a high throughput system. *Eurosurveillance*, 25(9):2000152, 2020.

[199] Chantal BF Vogels, Anderson F Brito, Anne Louise Wyllie, Joseph R Fauver, Isabel M Ott, Chaney C Kalinich, Mary E Petrone, Marie-Louise Landry, Ellen F Foxman, and Nathan D Grubaugh. Analytical sensitivity and efficiency comparisons of SARS-CoV-2 qrt-pcr assays. *medRxiv*, 2020.

[200] Arun K Nalla, Amanda M Casto, Meei-Li W Huang, Garrett A Perchetti, Reigran Sampoleo, Lasata Shrestha, Yulun Wei, Haiying Zhu, Keith R Jerome, and Alexander L Greninger. Comparative performance of SARS-CoV-2 detection assays using seven different primer/probe sets and one assay kit. *Journal of Clinical Microbiology*, 2020.

[201] Kazuya Shirato, Naganori Nao, Harutaka Katano, Ikuyo Takayama, Shinji Saito, Fumihiro Kato, Hiroshi Katoh, Masafumi Sakata, Yuichiro Nakatsu, Yoshio Mori, et al. Development of genetic diagnostic methods for novel coronavirus 2019 (ncov-2019) in japan. *Japanese journal of infectious diseases*, pages JJID–2020, 2020.

[202] Kate N Bishop, Rebecca K Holmes, Ann M Sheehy, and Michael H Malim. APOBEC-mediated editing of viral RNA. *Science*, 305(5684):645–645, 2004.

[203] Rafael Sanjuán and Pilar Domingo-Calap. Mechanisms of viral mutation. *Cellular and Molecular Life Sciences*, 73(23):4433–4448, 2016.

[204] Nathan D Grubaugh, William P Hanage, and Angela L Rasmussen. Making sense of mutation: what D614G means for the COVID-19 pandemic remains unclear. *Cell*, 182(4):794–795, 2020.

[205] Marion Sevajol, Lorenzo Subissi, Etienne Decroly, Bruno Canard, and Isabelle Imbert. Insights into RNA synthesis, capping, and proofreading mechanisms of SARS-coronavirus. *Virus Research*, 194:90–99, 2014.

[206] Hatim T Allawi and John SantaLucia. Thermodynamics and nmr of internal g.t mismatches in dna. *Biochemistry*, 36(34):10581–10594, 1997.

[207] Tugba G Kucukkal, Marharyta Petukh, Lin Li, and Emil Alexov. Structural and physico-chemical effects of disease and non-disease nsSNPs on proteins. *Current Opinion in Structural Biology*, 32:18–24, 2015.

[208] Peng Yue, Zhaolong Li, and John Moult. Loss of protein structure stability as a major causative factor in monogenic disease. *Journal of molecular biology*, 353(2):459–473, 2005.

[209] Jiahui Chen, Kaifu Gao, Rui Wang, Duc Duy Nguyen, and Guo-Wei Wei. Review of COVID-19 antibody therapies. *Annual Review of Biophysics*, 50:1–30, 2020.

[210] Peter Chen, Ajay Nirula, Barry Heller, Robert L Gottlieb, Joseph Boscia, Jason Morris, Gregory Huhn, Jose Cardona, Bharat Mocherla, Valentina Stosor, et al. SARS-CoV-2 neutralizing antibody LY-CoV555 in outpatients with COVID-19. *New England Journal of Medicine*, 384(3):229–237, 2021.

[211] Wanbo Tai, Lei He, Xiujuan Zhang, Jing Pu, Denis Voronin, Shibo Jiang, Yusen Zhou, and Lanying Du. Characterization of the receptor-binding domain (RBD) of 2019 novel coronavirus: implication for development of RBD protein as a viral attachment inhibitor and vaccine. *Cellular & molecular immunology*, 17(6):613–620, 2020.

[212] Wendong Li, Zhengli Shi, Meng Yu, Wuze Ren, Craig Smith, Jonathan H Epstein, Hanzhong Wang, Gary Crameri, Zhihong Hu, Huajun Zhang, et al. Bats are natural reservoirs of SARS-like coronaviruses. *Science*, 310(5748):676–679, 2005.

[213] Xiu-Xia Qu, Pei Hao, Xi-Jun Song, Si-Ming Jiang, Yan-Xia Liu, Pei-Gang Wang, Xi Rao, Huai-Dong Song, Sheng-Yue Wang, Yu Zuo, et al. Identification of two critical amino acid residues of the severe acute respiratory syndrome coronavirus spike protein for its variation in zoonotic tropism transition via a double substitution strategy. *Journal of Biological Chemistry*, 280(33):29588–29595, 2005.

[214] Huai-Dong Song, Chang-Chun Tu, Guo-Wei Zhang, Sheng-Yue Wang, Kui Zheng, Lian-Cheng Lei, Qiu-Xia Chen, Yu-Wei Gao, Hui-Qiong Zhou, Hua Xiang, et al. Cross-host evolution of severe acute respiratory syndrome coronavirus in palm civet and human. *Proceedings of the National Academy of Sciences*, 102(7):2430–2435, 2005.

[215] Rui Wang, Jiahui Chen, Yuta Hozumi, Changchuan Yin, and Guo-Wei Wei. Emerging vaccine-breakthrough SARS-CoV-2 variants. *arXiv preprint arXiv:2103.08023*, 2021.

[216] Sarah A Clark, Lars E Clark, Junhua Pan, Adrian Coscia, Lindsay GA McKay, Sundaresh Shankar, Rebecca I Johnson, Vesna Brusic, Manish C Choudhary, James Regan, et al. SARS-CoV-2 evolution in an immunocompromised host reveals shared neutralization escape mechanisms. *Cell*, 184(10):2605–2617, 2021.

[217] Jiahui Chen, Kaifu Gao, Rui Wang, and Guowei Wei. Prediction and mitigation of mutation threats to COVID-19 vaccines and antibody therapies. *arXiv preprint arXiv:2010.06357*, 2020.

[218] Sarah Cherian, Varsha Potdar, Santosh Jadhav, Pragya Yadav, Nivedita Gupta, Mousumi Das, Partha Rakshit, Sujeet Singh, Priya Abraham, Samiran Panda, et al. SARS-CoV-2 Spike Mutations, L452R, T478K, E484Q and P681R, in the Second Wave of COVID-19 in Maharashtra, India. *Microorganisms*, 9(7):1542, 2021.

[219] Soo-Young Lee, Dong-Kyun Ryu, Hanmi Noh, Jongin Kim, Ji-Min Seo, Cheolmin Kim, Carel van Baalen, Aloys SL Tijsma, Hyo-Young Chung, Min-Ho Lee, et al. Therapeutic efficacy of CT-p59 against P. 1 variant of SARS-CoV-2. *bioRxiv*, 2021.

[220] Xianding Deng, Miguel A Garcia-Knight, Mir M Khalid, Venice Servellita, Candace Wang, Mary Kate Morris, Alicia Sotomayor-González, Dustin R Glasner, Kevin R Reyes, Amelia S Gliwa, et al. Transmission, infectivity, and antibody neutralization of an emerging SARS-CoV-2 variant in California carrying a L452R spike protein mutation. *MedRxiv*, 2021.

[221] 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.

[222] Rui Wang, Jiahui Chen, Yuta Hozumi, Changchuan Yin, and Guo-Wei Wei. Emerging vaccine-breakthrough SARS-CoV-2 variants. *ACS Infectious Diseases*, 2022.