

Lattice Paths for Persistent Diagrams with Application to COVID-19 Virus Spike Proteins

Moo K. Chung¹ and Hernando Ombao²

¹University of Wisconsin, Madison, USA

²King Abdullah University of Science and Technology, Thuwal, Saudi Arabia
mkchung@wisc.edu, hernando.ombao@kaust.edu.sa

Abstract. Topological data analysis, including persistent homology, has undergone significant development in recent years. However, one outstanding challenge is to build a coherent statistical inference procedure on persistent diagrams. The paired dependent data structure, which are the births and deaths in persistent diagrams, adds complexity to statistical inference. In this paper, we present a new lattice path representation for persistent diagrams. A new exact statistical inference procedure is developed for lattice paths via combinatorial enumerations. The proposed lattice path method is applied to study the topological characterization of the protein structures of the COVID-19 virus. We demonstrate that there are topological changes during the conformational change of spike proteins, a necessary step in infecting host cells.

1 Introduction

Despite its rigorous mathematical foundation developed since its introduction 20 years ago [11], persistent homology still suffers from numerous statistical and computational problems. It has not yet become a standard tool in medical imaging. This approach has been applied to a wide variety of data including brain networks [9], protein structures [12], and RNA viruses [5]. However, most of these methods only serve as exploratory tools that provide descriptive summary statistics rather than formal inference. The main difficulty is due to the heterogeneous nature of topological features, which do not have a one-to-one correspondence across persistent diagrams. Motivated by these challenges, we propose a more principled topological inference procedure through lattice paths.

Lattice paths are widely studied algebraic objects in combinatorics and may have potential applications in persistent homology [2,16,18]. Previously, they have been used to differentiate 0th and 1st Betti numbers in functional brain networks [9]. However, the method is limited to a specific type of filtration that produces only up to the 1st Betti numbers. Here, we propose a procedure that is generalizable to arbitrary persistent diagrams. The main advantage of the lattice path approach is that it provides probabilistic statements about the similarity of two persistent diagrams. This is often needed to produce some baseline quantitative measure, such as the p -value, commonly used in biomedical research [7,9]. Existing methods for computing p -values usually rely on approximate resampling techniques: jackknife [8], bootstrap [1] and the permutation test [9].

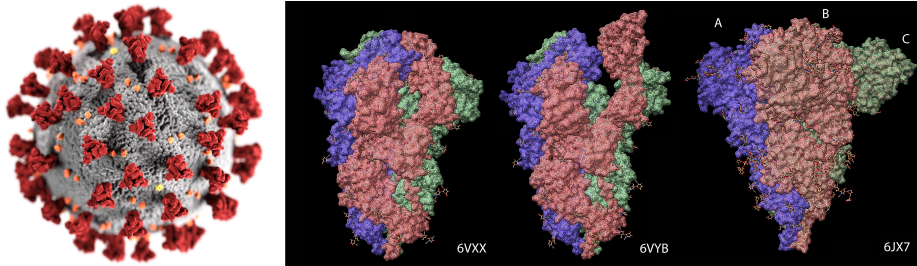


Fig. 1: Left: Schematic representation of COVID-19 virus with spike proteins (red). Right: Spike proteins of the three different corona viruses. The spike proteins consist of three similarly shaped interwinding substructures identified as A (blue), B (red) and C (green) domains.

Our approach here is different because it is analytic and thus compute the exact probability .

The main contributions of this study are the following: (1) a new data representation via Dyck and lattice paths; (2) the first analytic approach for computing probabilities without resampling; (3) the first topological study on the shape of COVID-19 virus spikes proteins. The proposed lattice path method was used to differentiate the conformational changes of the COVID-19 virus spike proteins (Figure 1). This demonstration is particularly relevant due to the potential for advancing vaccine development [19,4].

2 Preliminary: simplicial homology for protein structures

High dimensional objects, such as proteins and molecules, can be modeled as a point cloud data V consisting of p number of points (atoms) indexed as $V = \{1, 2, \dots, p\}$. Suppose that the distance ρ_{ij} between points i and j satisfies the metric properties. For proteins, we can simply use the Euclidean distance between atoms in a molecule. Then $\mathcal{X} = (V, \rho)$, $\rho = (\rho_{ij})$ is a metric space where we can build a filtration necessary for persistent homology. If we connect points following some criterion on the distance, they will form a simplicial complex which will follow the topological structure of the molecule [10,15,21]. Note that the k -simplex is the convex hull of $k+1$ points in V . A simplicial complex is a finite collection of simplices such as points (0-simplex), lines (1-simplex), triangles (2-simplex) and higher dimensional counterparts. In particular, the *Rips complex* \mathcal{X}_ϵ is a simplicial complex, whose k -simplices are formed by $(k+1)$ points which are pairwise within distance ϵ [13]. While a graph has at most 1-simplices, the Rips complex has at most $(p-1)$ -simplices. The Rips complex induces a hierarchical nesting structure called the Rips filtration $\mathcal{X}_{\epsilon_0} \subset \mathcal{X}_{\epsilon_1} \subset \mathcal{X}_{\epsilon_2} \subset \dots$ for filtration values $0 = \epsilon_0 < \epsilon_1 < \epsilon_2 < \dots$. The filtration is quantified through k -cycles where 0-cycles are the connected components, 1-cycles are loops while 2-cycles are a 3-simplices (tetrahedron) without interior. The Betti numbers β_k , count the number of independent k -cycles. During the Rips filtration, the i -th

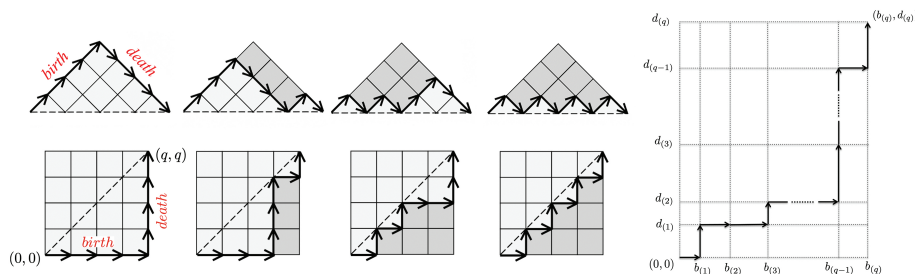


Fig. 2: Top: 4 different Dyck paths out of 14 possible paths for $q = 4$. Bottom: corresponding lattice paths. Right: Weighted lattice path equivalent to a persistent diagram.

k -cycles are born at filtration value b_i and die at d_i . The collection of all the paired filtration values $\{(b_1, d_1), \dots, (b_q, d_q)\}$ displayed as 1D intervals is called the *barcode* and displayed as scatter points in 2D plane is called the *persistent diagram*. Since $b_i < d_i$, the scatter points in the persistent diagram are traditionally displayed above the line $y = x$ line by taking births in the x -axis and deaths in the y -axis.

3 Methods

Dyck paths. The first step in the proposed *lattice path method* is to sort the set of all the birth and death values in the filtration as an order statistic $c : c_{(1)} < c_{(2)} < \dots < c_{(2q)}$, where $c_{(i)}$ is one of the birth or death values. The subscript (i) denotes the i -th smallest value. We will simply call such sequence as the *birth-death process*. Every possible valid sequence of birth and death values can be viewed as forming a probability space, where each valid sequence is likely to happen with equal probability. During the filtration, the sequence of birth and death occurs somewhat randomly but still maintains a specific pairing structure.

There exists a one-to-one relation between the ordering information and Dyck paths if we identify births with \nearrow and deaths with \searrow [2,16]. If we trace the arrows, we obtain the Dyck path (Figure 2) [18]. A valid Dyck path always starts at $y = 0$ and ends at $y = 0$. At any moment during the filtration, a Dyck path cannot go below $y = 0$. The total number of Dyck paths is called the Catalan number $\kappa_p = \frac{1}{q+1} \binom{2q}{q}$. The first few Catalan numbers are $\kappa_1 = 1, \kappa_2 = 2, \kappa_3 = 5$ and $\kappa_4 = 14$. More rapid changes in the direction of Dyck paths imply more fleeting fluctuations indicative of smaller topological signals. Less fluctuations indicate larger persistence and thus larger topological structures. The first path in Figure 2 has larger persistence while the last path has smaller persistence.

Lattice paths. If we rotate the Dyck paths clockwise at 45° and flip vertically, we obtain equivalent *monotone lattice paths* consisting of a sequence of \rightarrow (uparrow) and \uparrow (downarrow). Figure 2 displays corresponding monotone *lattice paths* between $(0,0)$ and (q,q) in \mathbb{Z}^2 where the path does not pass above the diagonal

line $y = x$ [18]. During the filtration, there cannot be more deaths than births and thus the path must lie below the diagonal line. The total area below Dyck paths can be used to quantify the Dyck paths [6]. Since

$$\text{Area below Dyck path} = \frac{q^2}{2} - \text{total area of boxes below lattice path,}$$

we will simply use lattice paths for quantification. If we tabulate how the area of boxes change over the x -coordinate in a lattice path, it is monotone. In the first path in Figure 2, the number of boxes below the first and the last lattice paths are $(0, 0, 0, 0)$ and $(0, 1, 2, 3)$. The area below the path is related to persistence. A barcode with smaller persistences (last path in Figure 2) will have more boxes (dark gray boxes) while longer persistences will have fewer boxes (first path in Figure 2). Given the sequence of heights of piled-up boxes, we can recover the corresponding lattice path by tracing the outline of boxes. We can further recover the original pairing information about births and deaths.

Remarks. In the situation where the birth values are identical at $-\infty$, persistent diagrams line up vertically as $(0, d_{(i)})$ and hence we simply augment them as $((i-1)\delta, d_{(i)})$. For instance, the Rips filtration for 0-cycles line up vertically.

Weighted lattice paths. The lattice and Dyck path representations only encode the ordering information about how births and deaths are paired, and do not encode the actual filtration values. This is remedied by adaptively weighting the length of arrows in lattice paths. We sort the set of birth values b_i and death values d_i as the order statistics:

$$b_{(1)} < b_{(2)} < \dots < b_{(q-1)} < b_{(q)}, \quad d_{(1)} < d_{(2)} < \dots < d_{(q-1)} < d_{(q)}.$$

We start at origin $(0, 0)$. When we encounter a birth $b_{(i)}$, we take the horizontal step to $b_{(i)}$. When we encounter a death $d_{(i)}$, take the vertical step to $d_{(i)}$ (Figure 2). The weighted lattice paths contains the same topological information as the original persistent diagram. Using the weighted lattice paths, we map a birth-death process to a monotone function ϕ :

Theorem 1. *There exists a one-to-one map from a birth-death process to a monotone function ϕ with $\phi(0) = 0$ and $\phi(1) = q$.*

Proof. We explicitly construct such a function ϕ . Consider the sequence of heights (or death values) as we traverse the weighed lattice path: $h : h_1 \leq h_2 \leq \dots \leq h_q$. The heights may not strictly increase (Figure 2-right). If births occurs r times sequentially in the birth-death process, the heights h will have r repeated identical death values $d_{(i)}$ as a subsequence. To make the subsequence strictly increasing, we simply add $\delta(0, 1, 2, \dots, r-1)(d_{(i+1)} - d_{(i)})$ to the repetition for $\delta \leq \frac{1}{r}$ (Figure 3 for example). Denote the transformed sequence as $h' = (h'_1, \dots, h'_q)$. $\phi(t)$ is given as a step function

$$\phi(t) = \begin{cases} 0 & \text{if } t \in [0, \frac{h'_1}{q}) \\ j & \text{if } t \in [\frac{h'_j}{q}, \frac{h'_{j+1}}{q}) \text{ for } j = 1, \dots, q-1 \\ q & \text{if } t \in [\frac{h'_q}{q}, 1] \end{cases} .$$

From $\phi(t)$, the original sequence h and the original birth-death process can be recovered exactly. Such map from a birth-death process to ϕ is one-to-one. \square

Remarks. Note $\|h - h'\|_2 \rightarrow 0$ as $\delta \rightarrow 0$. So by making δ as small as possible, we can construct a strictly monotone h' to be arbitrarily close to h . The normalized step function $\phi(t)/q$ can be viewed as an *empirical cumulative distribution* and many statistical tools for analyzing distributions can be readily applied. Figure 4-bottom displays the lattices paths and the normalized step functions of 1-cycles corresponding to the spike proteins used in the study. The proposed *lattice path method* uses $\phi(t)/q$ as a new topological functional descriptor – which is one of the main contributions in this paper.

Exact topological inference. The task now is to develop a method for testing the *topological equivalence* of two birth-death processes:

$$C^1 : c_1^1 < c_2^1 < \dots < c_{q_1}^1, \quad C^2 : c_1^2 < c_2^2 < \dots < c_{q_2}^2,$$

where each element is either a birth or death. Let ϕ_1 and ϕ_2 be the step functions corresponding to C^1 and C^2 . The topological distance

$$D(\phi_1, \phi_2) = \sup_{t \in [0,1]} \left| \frac{\phi_1(t)}{q_1} - \frac{\phi_2(t)}{q_2} \right|$$

will be used as the test statistic for testing the equivalence of C^1 and C^2 . The normalizing denominators q_1 and q_2 ensures that the value of step functions are in $[0, 1]$. The statistic $D(\phi_1, \phi_2)$ is the upper bound of area difference under $\phi_1(t)/q_1$ and $\phi_2(t)/q_2$:

$$\int_0^1 \left| \frac{\phi_1(t)}{q_1} - \frac{\phi_2(t)}{q_2} \right| dt \leq D(\phi_1, \phi_2).$$

Theorem 2. *Under the null hypothesis of the equivalence of C^1 and C^2 ,*

$$P(D(\phi^1, \phi^2) \geq d) = 1 - \frac{A_{q_1, q_2}}{\binom{q_1 + q_2}{q_1}},$$

where $A_{u,v}$ satisfies $A_{u,v} = A_{u-1,v} + A_{u,v-1}$ with the boundary condition $A_{q_1, 0} = A_{0, q_2} = 1$ within the band $|u/q_1 - v/q_2| < d$.

Proof. The statement can be proved similarly as the combinatorial construction of the Kolmogorov-Smirnov test that also uses the lattice path enumeration [3,14,9]. First, we combine monotonically increasing sequences C^1 and C^2 and

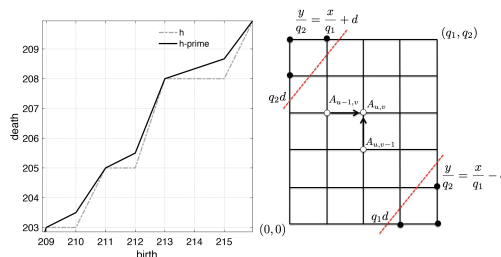


Fig. 3: Left: The heights of boxes below lattice paths h (dotted line) is made into strictly increasing to h' (solid line). Right: The problem of lattice path enumeration between $(0, 0)$ and (q_1, q_2) with the constraint $|x/q_1 - y/q_2| < d$.

sort them into a bigger monotone sequence of size $q_1 + q_2$. Then, we represent the combined sequence as the sequence of \rightarrow and \uparrow respectively depending on if they are coming from C^1 or C^2 . Under the null, there is no preference and they equally likely come from C^1 or C^2 . If we follow the sequence of arrows, it forms a monotone lattice path from $(0, 0)$ to (q_1, q_2) . In total, there are $\binom{q_1+q_2}{q_1}$ possible equally likely lattice paths that forms the sample space. This is equivalent to the total number of permutations in a two-sample test.

From Theorem 1, the values of $\phi_1(t)$ and $\phi_2(t)$ are integers from 0 to q . Thus, $(\phi_1(t), \phi_2(t))$ are the (x, y) -coordinates in lattice paths. Note

$$\sup_t \left| \frac{\phi_1(t)}{q_1} - \frac{\phi_2(t)}{q_2} \right| < d$$

if and only if $|\phi_1(t)/q_1 - \phi_2(t)/q_2| < d$ for all t . Such points are all the lattice points satisfying within the band bounded by two lines $y/q_2 = x/q_1 \pm d$ (dotted red lines in Figure 3). Thus, the probability of the event $D < d$ is equal the number of valid paths over the every possible paths:

$$P(D \geq d) = 1 - P(D < d) = 1 - \frac{A_{q_1, q_2}}{\binom{q_1+q_2}{q_1}}, \quad (1)$$

where $A_{u, v}$ is the total number of valid paths from $(0, 0)$ to (u, v) . Since there are only two paths leading to (u, v) , we write

$$A_{u, v} = A_{u-1, v} + A_{u, v-1}.$$

If there is no constraint, $A_{u, v} = \binom{u+v}{u}$. However, due to the constraint, $A_{u, v}$ has to be iteratively computed with the boundary condition $A_{u, 0} = A_{0, v} = 1$ for all u and v . \square

For numerical implementation, we start from the bottom left corner $(0, 0)$ and move toward the top right corner (q_1, q_2) in solving the recursion numerically:

$$A_{0, 0} = 0 \rightarrow A_{1, 0} = A_{0, 1} = 1, A_{1, 1} = 2 \rightarrow \dots \rightarrow A_{q_1, q_2} = A_{q_1-1, q_2} + A_{q_1, q_2-1}.$$

Computing A_{q_1, q_2} iteratively requires at most $q_1 \cdot q_2$ operations while the permutation test will cause a computational bottleneck for large q_1 and q_2 . Thus, the proposed *lattice path method* computes the exact p -value substantially faster than the permutation test. Since most protein molecules consist of thousands of atoms, q_1 and q_2 should be sufficiently large to apply the asymptotic:

Theorem 3. $\lim_{q_1, q_2 \rightarrow \infty} P\left(\sqrt{\frac{q_1 q_2}{q_1 + q_2}} D \geq d\right) = 2 \sum_{j=1}^{\infty} (-1)^{j-1} e^{-2j^2 d^2}.$

The result follows by treating ϕ_1/q_1 and ϕ_2/q_2 as the empirical cumulative distribution functions and using the asymptotic expansion of the KS-test [14,17]. Subsequently, the p -value under null is given by

$$p\text{-value} = 2e^{-d_0^2} - 2e^{-8d_0^2} + 2e^{-18d_0^2} \dots,$$

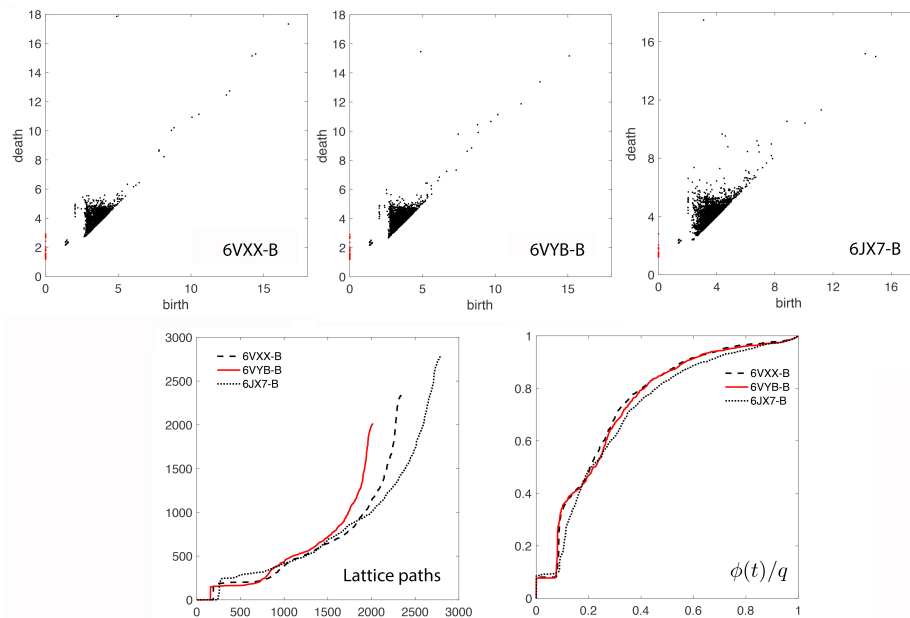


Fig. 4: Top: persistent diagrams of three different spike proteins. The red dots are 0-cycles and the black dots are 1-cycles. The units are in Å (angstrom). Bottom: the corresponding lattice paths and normalized step functions $\phi(t)/q$.

where d_o is the observed value of $\sqrt{\frac{q_1 q_2}{q_1 + q_2}} D$. For any large observed value $d_o \geq 2$, the second term is in the order of 10^{-14} and insignificant. Even for a small value of d_o , the expansion converges quickly and 5 terms should be more than sufficient to use Theorem 3.

4 Experiment: spike proteins of COVID-19 virus

The proposed lattice path method is used to study the topological structure of the severe acute respiratory syndrome coronavirus 2 (SARS-Cov-2), which is often called COVID-19. Since the start of the global pandemic (approximately December 2019), COVID-19 has already caused 3.85 million deaths in the world as of June 2021. The COVID-19 virus is specific member of a much broader coronavirus family, which all have spike proteins surrounding the spherically shaped virus similar to the sun's corona. The glycoprotein spikes that bind with receptors on cells and consequently cause severe infection. The atomic structure of spike proteins can be determined through the cryogenic electron microscopy (cryo-EM) [19,4]. Figure 1-left illustrates spike proteins (colored red) that surrounds the spherically shaped virus. Each spike consists of three similarly shaped protein molecules with rotational symmetry often identified as A, B and C domains. The spike proteins have two distinct conformations identified as *open* and

closed states, where the domain’s opening is necessary for interfacing with the host cell’s surface for membrane fusion and viral entry (Figure 1-right). Indeed, most current vaccine efforts focus on preventing the open state from interfacing with the host cell. Hence, this line of research is of prime importance in vaccine development and therapeutics [4].

In this study, we analyzed the spikes of three different coronaviruses identified as 6VXX, 6VYB [19] and 6JX7 [20]. The 6VXX and 6VYB are the closed and open states of SARS-Cov-2 from human while 6JX7 is feline coronavirus (Figure 1). All the domains of 6VXX have exactly 7604 atoms and are expected to be topologically identical. Applying the lattice path method to 1-cycles, we tested the topological equivalence of the B-domain and the A- and C-domains within 6VXX. The normalized step functions are almost identical and the observed topological distances are 0.0090 and 0.0936, which give the p -value of 1.00 each. As expected, the method concludes that they are topologically equivalent. Figure 4-bottom displays the lattice paths and the normalized step functions for domain B of 6VXX. The plots for other domains are visually almost inseparable and hence not shown. The closed domain B of 6VXX is also compared against the open domain B of 6VYB. The open state has a significantly reduced number of atoms at 6865 due to the conformational change that may change the topology as well. The observed topological distance is 0.2 and the test shows the extremely small p -values of 8.1123×10^{-38} confirming the topological change. The persistent diagrams of both closed and open states are almost identical in smaller birth and death values below 6 Å (angstrom) (Figure 4-top). The major difference is in the scatter points with larger birth and death values. The lattice path method confirms that the local topological structures are almost identical while the global topological structures are different.

The domain B of 6VXX is also compared against the domain B of feline coronavirus 6JX7 consisting of 9768 atoms. Since 6JX7 is not from human, it is expected that they are different. The topological distance is 0.9194 and p -values is 0.00×10^{-38} confirming that the topological nature of spikes are different. This shows the biggest difference among all the comparisons done in this study.

5 Conclusions

In this paper, we proposed a new representation of persistent diagrams using the Dyck and lattice paths. The novel representation enables us to perform the statistical inference combinatorially through the proposed lattice path method by enumerating every possible valid lattice paths analytically. The lattice path method is subsequently used to analyze the coronavirus spike proteins. The normalized step function $\phi(t)/q$ for all the spike proteins show fairly stable consistent global monotone pattern but with localized differences. We demonstrated the lattice path method has the ability to statistically discriminate between the conformational changes of the spike protein that is needed in the transmission of the virus. We hope that the our new representation enables scientists in their effort to automatically identify the different types and states of coronaviruses in a more principled manner.

Acknowledgement

The schematic illustration of COVID-19 virus is provided by Alissa Eckert and Dan Higgins of Disease Control and Prevention (CDC), US. The proteins 6VXX and 6VYB are provided by Alexander Walls of University of Washington. The protein 6JX7 is provided by Tzu-Jing Yang of National Taiwan University. Figure 2-left is modified from an image in Wikipedia. This study is supported by NIH R01 EB022856 and R01 EB028753, NSF MDS-2010778, and CRG from the King Abdullah University of Science and Technology (KAUST).

References

1. M. Ahmed, B. T. Fasy, and C. Wenk. Local persistent homology based distance between maps. In *Proceedings of the 22nd ACM SIGSPATIAL International Conference on Advances in Geographic Information Systems*, pages 43–52, 2014.
2. L.J. Billera, S.P. Holmes, and K. Vogtmann. Geometry of the space of phylogenetic trees. *Advances in Applied Mathematics*, 27:733–767, 2001.
3. W. Böhm and K. Hornik. A Kolmogorov-Smirnov test for r samples. *Institute for Statistics and Mathematics*, Research Report Series:Report 105, 2010.
4. Y. Cai, J. Zhang, T. Xiao, H. Peng, S.M. Sterling, R.M. Walsh, S. Rawson, S. Rits-Volloch, and B. Chen. Distinct conformational states of SARS-CoV-2 spike protein. *Science*, 369:1586–1592, 2020.
5. J. M. Chan, G. Carlsson, and R. Rabadan. Topology of viral evolution. *Proceedings of the National Academy of Sciences*, 110(46):18566–18571, 2013.
6. R. Chapman. Moments of dyck paths. *Discrete mathematics*, 204:113–117, 1999.
7. Frédéric Chazal, Brittany Terese Fasy, Fabrizio Lecci, Alessandro Rinaldo, Aarti Singh, and Larry Wasserman. On the bootstrap for persistence diagrams and landscapes. *arXiv preprint arXiv:1311.0376*, 2013.
8. M.K. Chung, J.L. Hanson, J. Ye, R.J. Davidson, and S.D. Pollak. Persistent homology in sparse regression and its application to brain morphometry. *IEEE Transactions on Medical Imaging*, 34:1928–1939, 2015.
9. M.K. Chung, H. Lee, A. DiChristofano, H. Ombao, and V. Solo. Exact topological inference of the resting-state brain networks in twins. *Network Neuroscience*, 3:674–694, 2019.
10. H. Edelsbrunner and J. Harer. *Computational topology: An introduction*. American Mathematical Society, 2010.
11. H. Edelsbrunner, D. Letscher, and A. Zomorodian. Topological persistence and simplification. In *Foundations of Computer Science, 2000. Proceedings. 41st Annual Symposium on*, pages 454–463. IEEE, 2000.
12. M. Gameiro, Y. Hiraoka, S. Izumi, M. Kramar, K. Mischaikow, and V. Nanda. A topological measurement of protein compressibility. *Japan Journal of Industrial and Applied Mathematics*, 32:1–17, 2015.
13. R. Ghrist. Barcodes: The persistent topology of data. *Bulletin of the American Mathematical Society*, 45:61–75, 2008.
14. J. D. Gibbons and S. Chakraborti. *Nonparametric Statistical Inference*. Chapman & Hall/CRC Press, 2011.
15. J.C. Hart. Computational topology for shape modeling. In *Proceedings of the International Conference on Shape Modeling and Applications*, pages 36–43, 1999.

16. R. Simion. Noncrossing partitions. *Discrete Mathematics*, 217(1-3):367–409, 2000.
17. N.V. Smirnov. Estimate of deviation between empirical distribution functions in two independent samples. *Bulletin of Moscow University*, 2:3–16, 1939.
18. R.P. Stanley. *Enumerative combinatorics. Vol. 2*. Cambridge Studies in Advanced Mathematics, 1999.
19. A.C. Walls, Y.-J. Park, M.A. Tortorici, A. Wall, A.T. McGuire, and D. Vesler. Structure, function, and antigenicity of the SARS-CoV-2 spike glycoprotein. *Cell*, 181:281–292, 2020.
20. T.-J. Yang, Y.-C. Chang, T.-P. Ko, P. Draczkowski, Y.-C. Chien, Y.-C. Chang, K.-P. Wu, K.-H. Khoo, H.-W. Chang, and S.-T. D. Hsu. Cryo-EM analysis of a feline coronavirus spike protein reveals a unique structure and camouflaging glycans. *Proceedings of the National Academy of Sciences*, 117:1438–1446, 2020.
21. A.J. Zomorodian. *Topology for computing*, volume 16 of *Cambridge Monographs on Applied and Computational Mathematics*. Cambridge University Press, Cambridge, 2009.