Topology in Biology

Ann Sizemore Blevins and Danielle S. Bassett

Contents

Abstract

Fueled by increasing computing power and ever-growing datasets, novel methods for complex systems analyses seem to emerge daily. With this whirlwind flows excitement and opportunity as new methods allow us to view our system of interest with a fresh perspective. One set of approaches that has offered particularly deep insights into complex systems is that of applied topology, also known as the field of topological data analysis (TDA). Topological data analysis specializes in representing and analyzing the entirety of the system and therefore can often remove the need for thresholds and filtering when selecting data, as is commonly done when studying biological systems. In this chapter we will briefly introduce the topological perspective that makes the above possible, and then we will dive into three methods within TDA through a biological lens. First, we cover persistent homology, which detects evolving topological cavities within data. Next, we join data to graphs in order to construct sheaves or structures that can help us understand systems with linear relations between adjacent nodes. We

A. S. Blevins (✉) · D. S. Bassett
Department of Bioengineering, University of Pennsylvania, Philadelphia, PA, USA
e-mail: dsb@seas.upenn.edu

© Springer Nature Switzerland AG 2020

B. Sriraman (ed.), *Handbook of the Mathematics of the Arts and Sciences*,
https://doi.org/10.1007/978-3-319-70658-0_87-1

then describe path signatures, which allow us to detect lead-lag behavior in time series. While detailing each of these three main concepts, we note how we can view each as a perspective shift from an analysis we might already perform to a similar analysis using the topological language. As we discuss, this reframing into the language of topology can provide extensive additional insight into our system of interest, and at present, much of this added information remains ripe for interpretation and application.

Keywords

Applied topology · Computational biology · Computational neuroscience

Introduction

When studying biology, we often encounter big data extracted from complex biological systems. A complex biological system is a system comprised of many interacting units which collectively serve a greater biological purpose. Advances in single cell sequencing, multimodal imaging of the human brain and body, and world-wide genome collections comprise a small subset of the technological innovations that allow us to probe complex biological systems at a new level of detail. As we are introduced to each new dataset – perhaps high-dimensional and tough to grasp, perhaps microscopic and difficult to directly observe – we often find ourselves asking a deceptively basic question: what does my system look like?

In order to answer this question, we can take advantage of a plethora of techniques within easy reach. Statistics and machine learning provide us with unsupervised methods such as principal component analysis (Lever et al., 2017; Pearson, 1901), independent component analysis (Comon, 1994; Yao et al., 2012), canonical correlation analysis (Jordan, 1875; Naylor et al., 2010), non-negative matrix factorization (Devarajan, 2008; Lee and Seung, 2001), and t-distributed stochastic neighbor embedding (Kobak and Berens, 2019; van der Maaten and Hinton, 2008), which allow us to uncover latent structure and classes within the data. Additionally supervised methods such as regression, decision trees (Koch et al., 2013; Quinlan, 1987), support vector machines (Boser et al., 1992; Yang, 2004), and neural networks (Cartwright, 2008; Kleene, 1951) provide predictions of new observations based on user-defined labels. Each field offers unique methods that yield particular insights into the system under study, but, as with any method, are also tailgated by their inescapable parfait of assumptions, limitations, advantages, and disadvantages. To gain a full understanding of the complex biological system under study, it is becoming increasingly important for today's scientist to equip themselves with a medley of analysis techniques in order to appropriately and comprehensively investigate their system.

Here we invite the reader to add topological methods to their technique tool-belt. Topology offers a framework that focuses more on relations between elements and how those relations and elements together combine into the whole system and less on precise calculated values of per-element statistics. As such, topological strategies offer methodological advantages such as robustness to noise and freedom from predefined coordinates (Carlsson, 2009; Cohen-Steiner et al., 2007; Ghrist, 2008). Furthermore, topology as a field of mathematics frequently promotes visualizations for understanding new concepts, and as a result, topological methods for data analysis can yield visual insight into the global organization of data. Moreover, mathematicians have poured a great effort into making current topological data analysis tools publicly available (Bauer et al., 2017; Henselman and Ghrist, 2016; Maria et al., 2014; Morozov et al., 2012; Tausz et al., 2014), offering intuitive explanations of the underlying mathematics (Edelsbrunner and Morozov, 2012; Munch, 2017; Otter et al., 2017; Patania et al., 2017), and continuing to demonstrate novel concepts for data analysis (Fasy and Wang, 2016; Purvine et al., 2018; Robins et al., 2004). Currently, much of the work in developing novel topological data analysis techniques continues, as scientists on both the data analysis and mathematics sides of the community repeatedly find that topological ideas naturally emerge within the analysis of complex systems.

In this chapter we will introduce topological thinking for biological data analysis. We will first discuss the field of topology and how topological thinking connects local and global scales of a system. Next, we will present three different methodologies or frameworks for topological data analysis; persistent homology, sheaf cohomology, and path signatures. For each framework we describe current methods in other fields, how we can translate a system and associated question into the topological language, the topological method itself, previous uses for biological research, and the leading edge of method development. Finally, we forecast future developments at the intersection between topological data analysis and biology, and we invite the reader to imagine how topology may shape their own system of interest.

What and Why Topology?

Let us begin with two examples from biology: leaf vasculature and neural firing patterns. Interestingly, while having a tree-like vasculature network optimizes efficient transportation of consistent loads of water and nutrients within leaves, more evolutionarily mature plants show looping within their networks (Katifori and Magnasco, 2012; Melville, 1969; Roth-Nebelsick et al., 2001). Why might a looping architecture be advantageous to a species? If an insect damages the leaf or if the leaf becomes otherwise punctured, the redundant pathways formed from loops in the vasculature will provide alternative routes for nutrient transportation (Fig. 1, left). Additionally, under varying loads, it has been shown that a looped architecture is more optimal for transport than tree structures (Katifori et al., 2010). Within a time-evolving system such as a population of connected neurons, a feedback loop in the structure can induce repeating patterns of activity after only one stimulus event (Fig. 1, right) (Ju et al., 2018). We often find such stereotyped repeating firing patterns in songbirds, in which unique patterns of repeated firing correspond to song

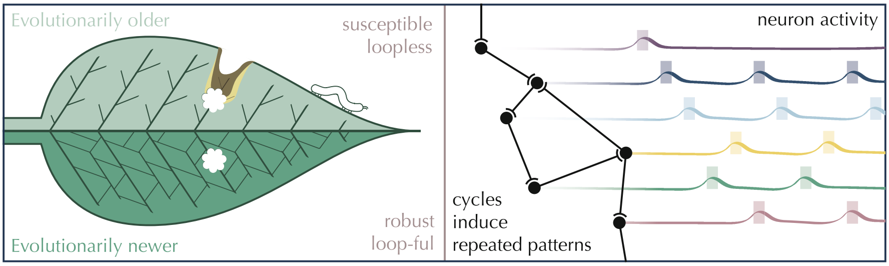

Fig. 1 **Naturally topology. (Left)** Earlier plant species show branching vasculature structures, leaving their distribution system vulnerable to pointed attack. On the other hand, more recently evolved plants often contain many nested loops in their vasculature, which promotes robustness. **(Right)** Network of connected neurons. When the top neuron fires only once, the feedback loop within the network structure induces a recurring, multi-neuron firing pattern

syllables (Mackevicius et al., 2019; Okubo et al., 2015), as well as many rhythmic movements such as walking (Brown, 1914; Marder and Bucher, 2001).

In the above systems, the difference between looping and non-looping organizations is formally a difference in their topology. Note, though the term "topology" often refers to network structure, we use "topology" to mean the features studied in the topology subfield of mathematics. Briefly, topology as a subject in mathematics concerns itself with how local relations assemble into the global structure of an object. Thinking topologically about a system comprises a middle ground between strictly local and explicitly global perspectives. For example, if we imagine our complex system as a Lego set or tangram puzzle, local analyses focus on individual blocks and their particular neighborhoods. Global statistics, on the other hand, describe the end product – the solved puzzle or the constructed object in our examples. Topology, however, is more akin to the instruction manual which helps us understand how the small, basic pieces fit together in order to form the completed object.

Although topology has been formally studied as a mathematical discipline for over a century, only recently has this mathematical language begun to be used in areas of science that are more applied. Beginning with Edelsbrunner (1995), Edelsbrunner et al. (2000), Dey et al. (1999), and Carlsson (2009) published near the turn of the millennium, topological thinking found its way into big data analysis with the implementation of two distinct tools: (i) Mapper (Singh et al., 2007), an algorithm used to visualize and analyze high-dimensional data based on overlapping clustering in the original space, and (ii) persistent homology (Edelsbrunner et al., 2000; Zomorodian and Carlsson, 2005), which detects topological voids within data (see the next section for further details). Early work employed these topological tools to discover a new subset of breast cancer (Singh et al., 2007), detect the Klein bottle on which natural images live (Carlsson et al., 2008), and identify the cyclicity in the evoked neural firing patterns of the macaque visual cortex (Singh et al., 2008). Since these early works, topological methods have flourished as the topological perspective has granted new insights to both neural (Bendich et al., 2016; Curto, 2017; Giusti et al., 2016, in press, 2015; Stolz, 2014) and biological systems (Braslavsky and Stavans, 2018; Edelsbrunner and Koehl, 2017; Meng et al., 2020; Nanda and Sazdanović, 2014; Rabadán and Blumberg, 2019; Schlick and Olson, 1992; Thom, 1969).

Topological methods provide particular advantages for big data analysis, and these advantages have fueled the field's rapid expansion. First, topological methods are often coordinate-free, meaning generally that we care less about specific measurement values and we care more about the relationships between objects. So, for example, if collected datapoints formed a circular pattern, topology would care less about the exact span, say, of the dataset, and more about the fact that the data globally organized into a circle. Second, topology describes data in a conceptually and visually more intuitive way than many standard methods from the field of statistics. Finally, topological methods report intrinsic characteristics of systems in a manner that is robust to many standard sources of error in biological data. For more discussion on these points, we invite the reader to peruse (Carlsson, 2009; Ghrist, 2008; Munch, 2017; Otter et al., 2017).

Still, topology and methods derived therefrom can feel quite foreign due to the perceived high start-up cost of learning new definitions, new mathematics, and additional software implementations. In this work, we will show that oftentimes we can find an entry point into topological methods by considering an abstraction that we are familiar with or an analysis that we frequently perform, which has a natural counterpart or analogous approach in topology. From this entry point, we will smoothly translate into the algebraic topological language and show how, once we can employ this topological perspective, we gain access to powerful mathematics existing behind the new terminology. We then illustrate how these new mathematical tools can offer more information about the biological system under study and suggest that much of this information is ripe for creative questions and novel hypotheses. For additional intuition, we collated a list of tutorials for relevant software, which we share at https://github.com/asizemore/TDA_tutorials.

Finding Topological Cavities: Persistent Homology

We can understand many biological systems as a collection of interacting units that together have some emergent behavior. For example, we understand the brain as a collection of regions that structurally connect but also organize into large functional modules in order to support a range of behaviors (Fig. 2a, left) (Baum et al., 2017; Power et al., 2011). In a similar manner, many collagen fiber types (e.g., type V (Shoulders and Raines, 2009)) require multiple chain types to form the fully functional coil structure (Fig. 2a, middle) (Bella, 2016; Gopalakrishnan et al., 2004). In social systems, disease spreading not only relies on the individuals themselves but also takes advantage of groups of individuals interacting with one another (Fig. 2a, right) (Iacopini et al., 2019; Petri and Barrat, 2018). Systems such as these are often amenable to graph-theoretic analyses, in which we ask how the graph (or network) representing our system is organized. To answer this question, we often draw from the fields of network science and graph theory, which provide methods to quantify the role of individual nodes in the network as well as the global architecture of the network.

Network science has enabled us to both quantify and visualize complex systems, which in turn has improved our understanding of biological systems (Gosak et al., 2018; Sporns, 2010) and diseases (Menche et al., 2015; Zanzoni et al., 2009). By nature, representing a system as a graph implicitly assumes that interactions are explicitly dyadic. However many biological systems fundamentally violate this assumption. For example, protein complexes form interactions between multiple subunits; populations of neurons can act collectively; and communicating cells can operate in unison. If we shift our perspective from graph theory to algebraic topology, instead of a graph representation with nodes and edges as dyadic building blocks, we model the system as a *simplicial complex* with nodes and multi-node blocks called *simplices* (Fig. 2b). We can still use pairwise interactions if they are appropriate, but we can just as readily describe a larger relation between many nodes and have all these polyadic relations fit into the same framework. The reframing of our system as a simplicial complex offers us an expanded view of the system and, as we will see, also supports an expanded repertoire of possible analysis techniques. Most prominently, the simplicial complex lets us detect topological cavities within the system, such as those illustrated within the leaf vasculature and neuronal connection pattern in Fig. 1.

Diving into the mathematics, a simplicial complex is composed of nodes and simplices (Fig. 2b). A k-simplex is a collection of $k + 1$ nodes that all-to-all interact and together form a cohesive unit. Then by definition any subset of nodes in a simplex also forms a simplex (Fig. 2b, left). Note that we can have a collection of nodes that are all-to-all connected, but their connections could be exclusively pairwise (dyadic), or we can have larger simplices that connect the nodes based on node sets that form a cohesive unit (polyadic). We refer the interested reader to Kozlov (2007) and Edelsbrunner and Harer (2010) for formal definitions. In biology we see simplices as groups of nodes that work together; a protein complex, a collection of co-firing neurons, or co-functioning and structurally connected brain regions could form a simplex. The mathematical power of the simplicial complex comes from knowing that any subset of nodes in a simplex also forms a smaller simplex, called a *face*. This fact implies that we understand how the 2-simplices fit into the skeleton of 1-simplices (edges), how any 3-simplices fit into the known organization of 2-simplices (filled triangles), and so on. Visually we imagine a 2-simplex as having a footprint of 1-simplices and similarly a 3-simplex as having a shell of 2-simplices that marks its place in the complex.

If we keep track of which k-simplices *are* footprints of $(k + 1)$-simplices, then we also know which k-simplices *are not*; this distinction allows us to find which loops of k-simplices are not footprints of $(k + 1)$-simplices. Said another way, we can use the simplicial complex to find where (topological) cavities reside within the complex (Fig. 2b, right). This technique is called *homology*. Homology intakes a simplicial complex and returns the cavities (of each dimension) within the complex

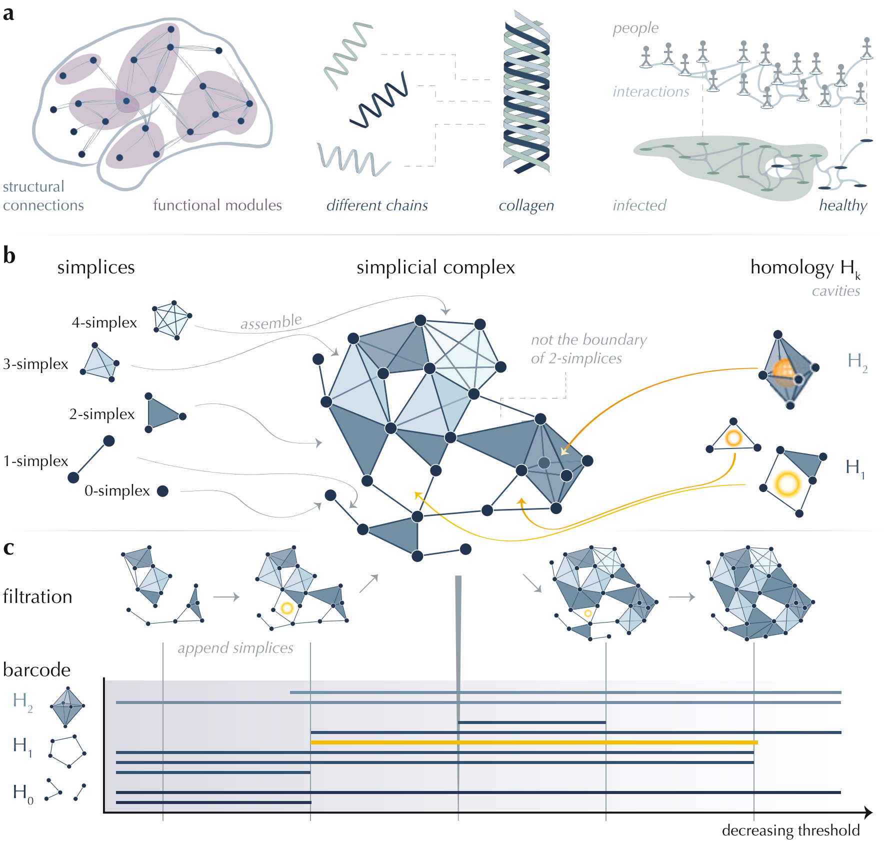

Fig. 2 Persistent homology detects evolving cavities within weighted or growing systems. **(a)** Examples of beyond pairwise interactions in biological systems. (Left) Nodes within the structural brain network are related by functionally determined modules. (Middle) Three types of collagen chains wind together into one type of fully functional collagen fiber. (Right) Higher-order interactions in social systems can influence the spread of disease. **(b)** Building blocks called k-simplices (left) *assemble* into the simplicial complex (middle) on which we can use homology H_k to detect *cavities* in each dimension (right). Homology operates by finding loops that are not boundaries of higher-dimensional simplices, exemplified by the triangle of edges that is *not the boundary of 2-simplices*. In this simplicial complex, homology detects five cavities of dimension 1 (examples highlighted with yellow and orange circles) and one of dimension 2 (orange globe). **(c)** Example filtered simplicial complex (top) in which we *append simplicies* at each step. Often we attach new simplices based on a decreasing threshold of connectivity strength within the dataset so that simplices connecting most similar nodes are added first and simplices defining weak connections are added later. We record the persistent homology in the barcode (bottom) which draws one line for each persistent cavity. We highlight in yellow a particular persistent cavity across the filtered simplicial complex as an example

(see Fig. 2b, right for examples of cavities in dimensions 1 and 2). Homology in dimension k, denoted by H_k, reports cavities formed from k-simplices, with the special case of dimension 0 in which H_0 documents connected components in the simplicial complex.

Furthermore, we can extend homology to a growing or weighted simplicial complex using *persistent homology*, which records the evolution of cavities along a growing sequence of simplicial complexes (Fig. 2c, top) called a *filtration*. We can create a filtration from weighted networks (Giusti et al., 2015; Petri et al., 2013), point clouds (Carlsson et al., 2008), growing systems (Blevins and Bassett, 2020; Rieck et al., 2017; Sizemore et al., 2018), and more. In such growing or weighted systems, topological cavities may form, evolve, and collapse based on the ordering, weights, or connections within the system. For example, the persistent cavity marked with a yellow circle in Fig. 2c first emerges at the second step, persists through the addition of more simplices, and finally gets filled in at the fifth step. Shown in Fig. 2c, the *barcode plot* below records these persistent cavities by assigning one bar per persistent cavity spanning from the first appearance of the cavity (called the birth) to the point in which the cavity becomes tessellated (called the death). Generally we interpret the longest living persistent cavities as those most descriptive of the global organization, since if a cavity persists throughout many simplex additions without being filled, then it is intuitive that the persistent cavity is an essential feature of system structure. In contrast, we often interpret short-lived persistent cavities as topological noise, although interesting ideas in Bubenik et al. (2020) have recently challenged this assumption. Fundamentally, the barcode reveals the strength and longevity of incompressible topological features (cavities) in our system. Said another way, persistent homology decomposes a weighted or growing system into independent persistent cavities, ignoring topologically redundant information.

The ability of persistent homology to detect cavities within complex systems has fueled the wide application of this method, especially in biological systems. For example, persistent homology of functional brain networks can be used to distinguish between placebo and drug states Petri et al. (2014), and persistent homology of structural brain networks reveals topological cavities that exist across adult humans (Sizemore et al., 2018). At a much smaller scale, the authors in (Gameiro et al., 2015) found that topological cavities within crystal structures predict the compressibility of a protein. Additionally we expect that finding long-persisting gaps within a growing epidemic may uncover mechanisms of inherent resistance (Fig. 2c, left). Analyses such as those listed have provided a wealth of new insights into the system structure, as the topological perspective provides information complementary to but not overlapping with that which is provided by network science. For example, edge-weighted network models classified based on persistent homology features do not necessarily share similar graph statistics (Sizemore et al., 2017).

Persistent homology has also recently been extended to the multiparameter (multiple filtration functions) case (Betancourt et al., 2018; Chambers and Letscher, 2018; Corbet et al., 2019; Scaramuccia et al., 2020) and to time-evolving systems

(Gasparovic et al., 2019; Kramár et al., 2016; Munch, 2013; Sanderson et al., 2017). For example, in Yoo et al. (2016) the authors use persistence vineyards to identify coherent functional states of the brain's dynamic functional connectivity estimated from an individual as they engage in a gaming task or as they simply rest. One can also use the persistent homology outputs as features that can then be used by standard statistical inference and machine learning algorithms (Adams et al., 2017; Chen et al., 2019; Kališnik, 2019; Monod et al., 2019; Piangerelli et al., 2018; Pun et al., 2018; Stolz, 2018), as well as deep learning (Brüel-Gabrielsson et al., 2019; Carlsson and Gabrielsson, 2018). The authors in Piangerelli et al. (2018) used similar ideas to classify healthy volunteers and epilepsy patients based on topological summaries of individual's EEG recordings. Moving from large-scale neural recordings to cellular recordings, recent work demonstrates the utility of topological methods in studying the activity of hippocampal neuronal populations, which displays a ring structure during wake and sleep that relates to place-tuning and head-direction (Rubin et al., 2019). Additional extensions include recovering circular coordinates for data (De Silva et al., 2011; Perea, 2018; Wang et al., 2011), topology for medical imaging (Damiano and McGuirl, 2018; Qaiser et al., 2016), understanding collective motion (Bhaskar, 2019; Topaz et al., 2015), and more. Despite these applications, the connection between homology and biological systems remains far from complete. The leading edge of persistent homology applications asks more specifically how cavities (in any dimension) directly relate to biological functions and what behaviors are evoked or supported by topological cavities within the system.

Data Systems and Solutions: Sheaves

While on the topic of systems with units and relations, let us next dig a little deeper into what we generally call "connections" between units. Imagine we observe three cells, each signaling to the next cell as in Fig. 3a. In this scenario, our system is constructed such that the second cell always signals at twice the rate of the first and such that the third cell signals at half the rate of the second. We then administer two different drugs in order to modulate signaling. Perhaps drug *A* globally slows down signaling by slowing all transcription, resulting in halved signal release rates for all cells (Fig. 3a, middle). In contrast, drug *B* causes a loss of certain receptors, which will alter the linear signaling relations between cells (Fig. 3a, bottom). If we can only observe the effect of each drug on the third cell (glowing green reporter molecules), we might conclude that both drugs function similarly. If we instead can observe the signaling output of each cell, we might conclude that both drugs directly affect the third cell since we observe a halved signaling rate in both conditions, even though drug *B* actually works only on the relationship between the signaling rates of the first and second cells. Importantly, in this scenario, we could unintentionally misinterpret the drug mechanism by considering only individual signal production rates and disregarding the relationships between cells. Similarly, consider a collection of neuronal populations as depicted in Fig. 3b. We can often observe the states of all populations, such as their average frequency or firing rate, but the relations between the populations remain difficult to directly infer. If instead we could learn not only which populations affect which others but also the more precise linear relation between populations, we could, for example, use the learned relations to predict the effect of stimulation in living systems.

The setup just described is nothing new to investigators working in computer science and applied graph theory. Indeed, finite element solvers (Arnold et al., 2010) could find the space of solutions to our puzzle, or we could frame the entire model as a connection graph (Singer and Wu, 2012) or matrix-weighted graph (Trinh et al., 2018). A disadvantage of these approaches is that the encoding and overall manipulation of the system may no longer be intuitive, nor does it necessarily maintain a strong connection between the system representation and the original biological system. Alternatively, if we consider the system to be a set of units (nodes) that locally abide by consistency rules (defined across edges), then we can encode the system as a *sheaf*, a well-studied construct in topology. In its general form, a *sheaf* is an incredibly versatile modeling object; nevertheless, for this exposition we will think mainly of a sheaf as a graph that has observable information in the form of vector spaces attached to nodes, and whose edges specify how values at neighboring nodes should relate to one another. In addition to maintaining our intuition about the system, the sheaf *language* brings with it a myriad of powerful mathematical operations.

Diving into the mathematics, we will break sheaves down into three basic parts (although since we are using topology, all parts are of course intimately related): the data as vector spaces at nodes and edges (Fig. 3d), maps between data following graph edges (Fig. 3e), and agreement (or lack thereof) across edges (Fig. 3f). First, each node n_i of the underlying graph (Fig. 3c, left) can pick any value (Fig. 3c, middle) from its attached vector space V_i (Fig. 3c, right). This value may be a differential activity level that comes from vector space $\mathbb{R}$, coordinates from $\mathbb{R}^2$, or a simple on-off indicator. Edges also have vector spaces attached (e.g., $V_{1,2}$ in Fig. 3c) that will help us relate happenings at connected nodes to one another. Vector spaces at nodes and edges are called *stalks* (Fig. 3d). Next, if two nodes n_i and n_j connect via an edge $e_{i,j}$, then we define linear maps $f_{v_i,e_{i,j}}$ and $f_{v_j,e_{i,j}}$ called *restriction maps* that detail how our choice of values (filled boxes ■) at nodes n_i and n_j map to values (open boxes □) within the edge vector space $V_{i,j}$ (Fig. 3e). Looking at our setup in Fig. 3e, we have the value ■ at node 1 mapping via $f_{v_1,e_{1,2}}$ to □, a value in $V_{1,2}$. Similarly, node 2 has value ■ which maps to □ in $V_{1,2}$. Now we have a sheaf $\mathcal{F}$, which is the entire object containing the vector spaces (on nodes and edges), and linear maps from node vector spaces to edge vector spaces following the connections in a graph (This definition is intended for an introductory level exposition to assist relatively naive readers. Please see Curry (2014), Robinson et al. (2014, 2017) for more formal definitions.).

Returning to our example sheaf in Fig. 3e, notice that when we choose values at each vertex and then map them to values on the edges, the two values (now in the edge vector space) may be different, as seen across edge $e_{1,2}$, or may be the

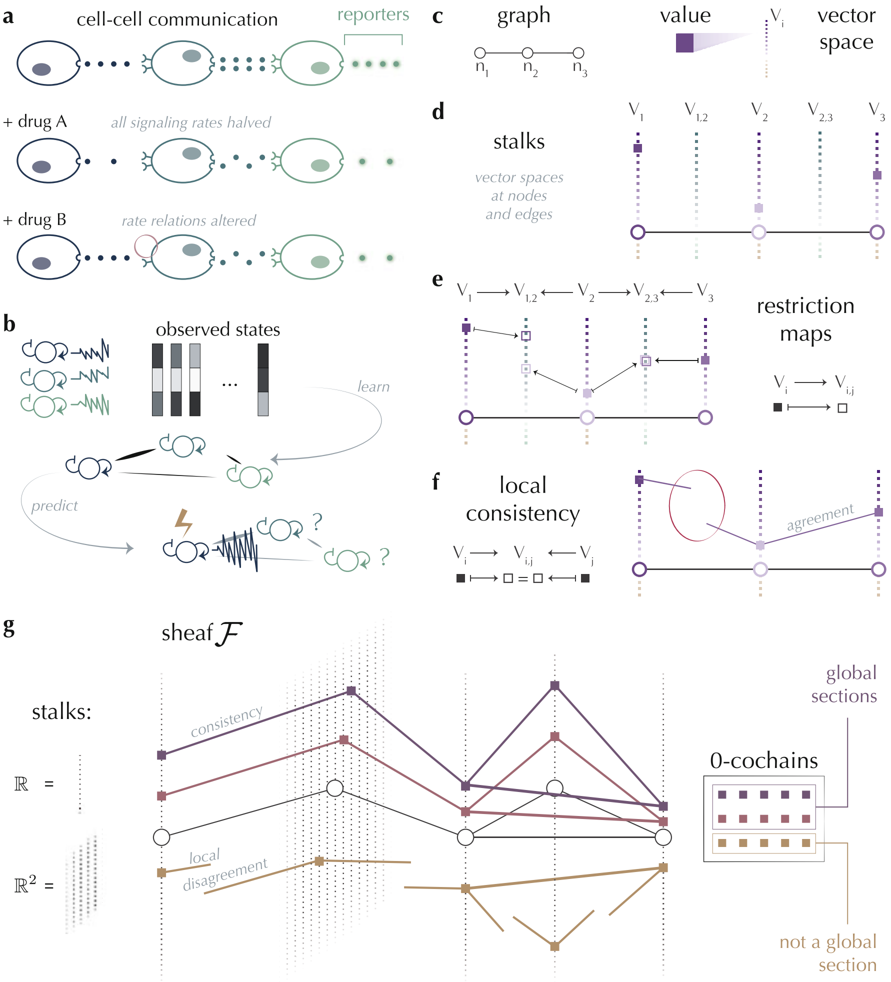

Fig. 3 **Sheaves encode linear relations between nodes with data.** Examples of systems with linear relations between units. **(a)** Paracrine signaling between cells (from left to right) in which we can observe the system via the reporter molecules depicted in glowing green. With the addition of drug A, *all signaling rates are halved*, while with the addition of drug, B *rate relations are altered*. **(b)** Recording states across time from active neuronal populations (top) allows us to *learn* relations between these populations (middle) and consequently *predict* the affect of stimulation (bottom). **(c)** Our illustrations for the underlying graph (left), chosen node values (middle), and vector spaces (right). **(d)** In a sheaf on a graph, stalks are *vector spaces at nodes and edges*, denoted by V_i and $V_{i,j}$ for node n_i and edge $e_{i,j}$, respectively. **(e)** Graph edges determine linear maps called restriction maps from node vector spaces to edge vector spaces, sending the node values (filled squares) to an edge value (outlined squares). **(f)** When both node values map to the same edge value (right pair of nodes), we say there is local consistency or *agreement* across that edge. **(g)** An example sheaf $\mathcal{F}$ with node stalks either $\mathbb{R}$ or $\mathbb{R}^2$. The purple and rose 0-chains are also linearly related global sections (the purple is twice the rose), while the gold 0-chain is not a global section

same, as seen across edge $e_{2,3}$. More formally we have that in the first case, the difference across edge $e_{1,2}$ is $f_{v_1,e_{1,2}}(■) - f_{v_2,e_{1,2}}(□) \neq 0$, while the difference across edge $e_{2,3}$ is $f_{v_2,e_{2,3}}(□) - f_{v_3,e_{2,3}}(■) = 0$. When we choose values □ and ■ on connected nodes n_2 and n_3, respectively, such that the difference across the edge $(f_{v_2,e_{2,3}}(□) - f_{v_3,e_{2,3}}(■))$ is 0, we say there is *local consistency* or agreement across the edge. In Fig. 3f we visually illustrate the agreement, or lack thereof, across edges either by having node values (filled boxes) connect via lines when they agree across the edge (right) or by showing stubs emerging from filled boxes if the values do not agree (left).

Although what we just described was a very small scenario, we can also build larger, more complicated sheaves, as shown in Fig. 3g. When we choose one value (or vector) for each node, this is called a *0-cochain*. Just as in the smaller case of Fig. 3e, f, values of a 0-cochain may or may not agree across edges. In our large sheaf, we can of course check for local agreement across edges individually, but when *all* node values agree across *all* edges in our sheaf, then our 0-cochain is called a *global section*, which we think of as a solution to the sheaf. In applications, we might think of cochains as multi-node signals or system states, and then satisfying local consistency constraints across all edges, or finding global sections, would be akin to finding particular states that offer maximal agreement with our model. Thinking through the cell-cell communication example in Fig. 3a, we first have some idea of how the non-perturbed system functions; cells (nodes) respond to signals from neighboring nodes in a linear fashion, so that the signaling rate of the middle cell is twice that of the first, and the signaling rate of the right cell is half that of the second (To be very particular, we say that the intercellular rate of change of signaling molecules from cell a to cell b is the cochain value at cell a, since signaling molecules can be both created and destroyed. Then stalks above both nodes (cells) and edges (cell signaling pairs) are $\mathbb{R}$.). Our initial setup of cell signal output respects the relations so we have a global section. After the addition of drug A, we still have a global section, as signaling rates respect the original linear relationships. However, after the addition of drug B, our cochain is no longer a global section of the sheaf with the originally defined linear relationships, which alerts us to the possibility of different mechanisms of action between drugs A and B.

If we have modeled a system as a sheaf, how can we detect global sections? Impressively, we can use homological ideas similar to those described in the previous section to compute all linearly independent global sections of a given sheaf. Instead of using boundary relations as in homology, in this particular sheaf scenario, we consider *co*-boundary relations, or the edges in which nodes participate (recall boundary relations move downward from edges to their participating nodes). Using these upward co-boundary relations between dimensions (here from nodes to edges), we receive the space of global sections as the zeroth cohomology group H^0 (see Hansen 2019a; Joslyn et al. 2014; Robinson et al. 2014 for more detailed descriptions). Said another way, given a system of nodes and linear relations, we can use sheaf theory to calculate all linearly independent solutions. Importantly for biological systems, we can also perform the reverse analysis: that is, given a set of measurements of a system, we can learn the underlying sheaf that generates the observed data, under the assumption that each measurement records an approximate global section (Hansen and Ghrist, 2019c). Or furthermore, we can use the sheaf Laplacian, similar in spirit to the graph Laplacian, to understand dynamics on a sheaf (Hansen and Ghrist, 2019d). All of these possibilities are equally accessible when the underlying object is a simplicial complex instead of a graph, or when the data come from larger vector spaces or even simple categories.

In practice, the above analyses are topics of current research (see Hansen and Ghrist 2019b,c) and mark the leading edge of applied sheaf theory. To the best of our knowledge, sheaves have rarely been used to formalize a biological system, although they clearly have the requisite features to do so. In Vepstas (2019), for example, the author describes how sheaves can model biochemical reaction systems. Expanding further, given enough gene expression data, we could reconstruct a sheaf defining how genes interact under particular conditions and then test if after changing conditions we see a different global section of the same sheaf (indicating changing conditions did not effect relationships between genes, only expression) or if the sheaf altogether changes (echoing the example in Fig. 3a, in which both the values and relationships between nodes change). Vector spaces above nodes could be of dimension N, so we could also use sheaves to naturally encode and describe relations between cells with multiple recordings (e.g., multiplexed imaging in which we tag multiple macromolecules). Additionally, sheaves just as elegantly encode data over simplicial complexes and posets instead of graphs, as well as have the ability to incorporate categorical, ordinal, or other types of data beyond vector spaces. Specifically, sheaves have been incredibly useful in modeling sensor networks in which each node records features, and in which relations between nodes impose a consistency constraint between corresponding sensors (Robinson et al., 2017). This modeling as a sheaf allows us to understand how (and if) sensors connected in this way can reach a consensus. In biology this may be similar to how we integrate information from multiple senses in different areas of the body, or how we understand an image by neuron populations capturing and agreeing upon different visual features.

All of the above examples begin with an emphasis on global sections. But beyond using the intuitions of global sections (solutions to the sheaf), fundamentally larger questions remain. As we discussed earlier, we can compute a global section by considering one layer of an entire mathematical construction on the sheaf, specifically dimension 0 (hence the 0-cochains) of the sheaf cohomology. This approach is analogous to computing the homology of a simplicial complex but only interpreting cavities in dimension 1. How can we interpret higher dimensional cohomology groups in sheaves that describe biological data? Cohomology in dimension 1 is intuitively described as obstructions to global sections, although the precise biological interpretation remains an interesting and open question. To summarize, sheaves not only encode complex interactions and solve computationally difficult problems in one elegant mathematical framework, but they also allow for *even more* higher-level descriptions of the system than many current methods, making sheaves an untapped framework with incredible potential.

Lead-Lag Relationships: Path Signatures

Often in complex biological systems, a signal arrives, then some time passes, and finally the system responds. This sequence from signal to response exemplifies a lead-lag relationship, a time-dependent association found extensively throughout biological processes. For example, neuronal responses to stimuli, cellular reactions to an adjusting firmness of environment, and an organism's capacity to learn, all involve lead-lag relationships at multiple spatial and temporal scales. Perhaps more importantly, lead-lag relationships that have been detected from data can suggest causal mechanisms of the observed biological function.

Despite their relevance for biological systems and their utility in probing the causes of biological phenomena, lead-lag relationships remain difficult to detect in practice. When handed a collection of time series traces describing the activity of many units (e.g., the activity of many brain regions as shown in Fig. 4a), we can apply a few different methods. Perhaps the most straightforward approach is to calculate the cross-correlation between signal pairs. However, this method only offers an accurate assessment when signals are periodic, that is, when the signal loops back onto itself at a fixed and regular time interval. Unfortunately, many

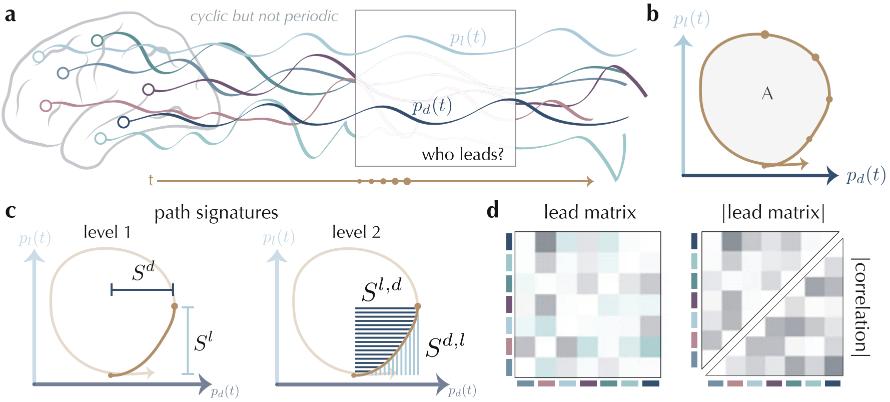

Fig. 4 Path signatures uncover leading and lagging relationships in time series data. (a) Example brain activity fluctuations as *cyclic but not periodic* paths. Considering the dark and light blue traces $p_d(t)$ and $p_l(t)$, respectively, which region leads and by how much? **(b)** Plotting the progression of time along the $p_d(t)$, $p_l(t)$ axes (dark blue and light blue, respectively) results in a large oriented area, A. **(c)** Depiction of path signature definitions for levels 1 and 2, during the highlighted portion of the path. **(d)** Computed lead matrix for the time-series in panel **(a)** along with the absolute lead matrix and absolute correlation matrix for comparison. All diagonal values set to zero

biological signals tend to exhibit cyclicity rather than periodicity; a cyclical signal is one that repeats, but the repeating sections do not need to have a regular duration. The existence of cyclicity precludes the use of the cross-correlation approach, and we must therefore return to the drawing board (we refer the reader to Section 2.1 of Giusti and Lee (2018) for further discussion of cyclicity and cross-correlation). In addition to the break from periodicity, the lag-time between the leading and lagging signal may not remain constant across time, as perhaps the system globally speeds up or slows down.

With these constraints in mind, we can instead concoct a recipe of integration and Stokes's theorem in order to detect lead-lag relationships between cyclic signals. Specifically, if we plot two cyclic signals such as the light blue $p_l(t)$ ∿ and dark blue $p_d(t)$ ∿ against each other and allow time to progress starting from $t = t_0$, we will trace out a closed loop as shown in Fig. 4b. Importantly, since the dark blue strongly leads the light blue signal over time, the so-called oriented area A within this closed loop will be large (and positive), while perfectly anticorrelated or un-correlated pairs would result in a near zero oriented area. We can calculate the oriented area A using a specific application of Stoke's theorem (or Green's theorem). We know that the area $A = \iint 1dA$ and by Stoke's theorem we have $\iint 1dA = \oint xdy = \oint ydx = \frac{1}{2}\oint xdy - ydx$, where x and y are the position along these axes. For us, if our path starts at $(p_d(t_0), p_l(t_0))$, then our movement along the x, y directions would be $p_d(t) - p_d(t_0)$ and $p_l(t) - p_l(t_0)$. Incorporating all of the above together we see that we can write the oriented area as

$$A = \frac{1}{2}\Big(\int_{t_0}^{t_f} (p_d(t) - p_d(t_0))p_l'(t)dt - \int_{t_0}^{t_f} (p_l(t) - p_l(t_0))p_d'(t)dt\Big),$$

and we assume for the sake of exposition that paths begin at time $t_0 = 0$ and progress until $t_f = 1$. Note that regardless of the system speed, the same curve and consequently the same oriented area will emerge. Additionally the signals do not need to be periodic in order to detect the lead-lag relationships via this method. This approach is so clean that we might naively think our story is complete. However, the evaluations of the two integrals used in the oriented area equation are only two elements of the *path signature* – an infinitely long sequence of numbers that uniquely describes the entire time-evolving system up to reparameterization and translation. While an infinitely long sequence of numbers seems not so helpful for finite datasets, we will see that even using a few of these descriptors and the mathematics behind them can be quite powerful.

If we reframe our collection of time series $(p_1(t), p_2(t), \ldots, p_n(t))$ as parts of a multidimensional path $\Gamma(t) = (p_1(t), p_2(t), \ldots, p_n(t))$, then what we are trying to measure is how the components of Γ influence each other. Intuitively, if we imagine considering the activity of brain regions over time, $\Gamma(t)$ would amount to the path that the whole brain takes over time, but since the brain activity is defined by activity on regions, we can look also at the activity of each region $p_i(t)$. On the other hand, $\Gamma(t)$ itself is a path in $\mathbb{R}^n$, and therefore descriptors of Γ will include descriptors of the individual evolution and interactions of the path components.

Specifically, the *path signature* is an infinite sequence of numbers that uniquely describes Γ up to translation and reparameterization, so that all the information about Γ exists within the path signature. We can calculate each element of the path signature from iterated integrals, each sequentially involving different combinations of the component paths. Specifically, the first-level path signature terms are $S^i(t_f) = \int_{t_0}^{t_f} p_i'(t)dt = p_i(t_f) - p_i(t_0)$ for each $i = 1, \ldots, n$ and quantify the distance moved from $t = t_0$ to $t = t_f$ in the i^{th} component (Fig. 4c, left, for light blue and dark blue components within the highlighted time segment). The second level (Fig. 4c, right) records how component i and component j interact and is defined as

$$S^{i,j}(t_f) = \int_{t_0}^{t_f} S^i(t)p_j'(t)dt = \int_{t_0}^{t_f} (p_i(t) - p_i(0))p_j'(t)dt\,.$$

Graphically, $S^{i,j}$ records the area between the top left of the $[t_0, t_f]$ box and the path $(p_i(t), p_j(t))$, while $S^{j,i}$ records the bottom right area (Fig. 4c, right). Notice the second level path signature terms exactly match those needed to calculate the oriented area above, and indeed we can now rewrite A as $\frac{1}{2}(S^{i,j} - S^{j,i})$. So in fact we have been using parts of the path signature all along. Repeating the oriented area computation for all pairs of brain regions constructs the *lead matrix* (Fig. 4d, left), which we can use for downstream investigative analyses. For demonstrative purposes, in Fig. 4d we also compare the oriented area to the absolute value of the correlation between each pair of signals shown in Fig. 4a. In this example experiment, the lead matrix finds the lead-lag relationship between the light and dark blue traces as one of the strongest in the set, while the correlation matrix suggests many alternative path pairs that show stronger relations.

The above pipeline has so far been implemented in only a handful of biological studies. In Baryshnikov and Schlafly (2016) the authors applied this lead-lag analysis to resting state fMRI data from the Human Connectome Project and found sequences of brain regions activated in cyclic order. Additionally the authors of Zimmerman et al. (2018) used path signatures to unfurl differences between the lead-lag relationships in the brains of tinnitus patients and healthy controls. Discussed at length in Lyons et al. (2002, 2007), Friz and Victoir (2010), one can use path signatures to understand (multidimensional) stochastic processes, also known in that literature as rough paths. Path signatures also provide ample features for machine learning techniques (see Chevyrev and Kormilitzin (2016); Lyons (2014) for introductions and examples). What more could lead-lag analyses reveal in biological data, from genes involved in drug response to neurons responsible for bird songs (Okubo et al., 2015)? Yet again we gain a powerful (and computationally fast) tool by using only *part* of a larger algebraic structure: here, the entire path signature. If indeed level two is so insightful, how do we interpret level three or higher? Additionally, new mathematical explorations describe how one could use more involved constructions of path signatures to understand families of paths such as multiple biological time series (Giusti and Lee, 2018). As with sheaves and persistent homology, it is the author's opinion that the potential of path signature analyses, given its rapid computation and few assumptions, has yet to be fully exercised in the applied realm.

Where Are We Going?

Extracting insights from the reams of data now being acquired from complex biological systems requires tools from multiple disciplines applied in creative ways. And more likely than not, our next challenge – be it analyses of whole-brain imaging, high-throughput single-cell RNAseq, or population dynamics – will also require interdisciplinary communication, a new collection of techniques, and creativity. In such a world, the importance of an ever-broadening range of analytical tools and understanding of such tools only increases. It is unlikely that any one method or line of thinking will complete an analysis, but instead combining tools from topology, statistics, and other fields can only strengthen our ability to derive biological understanding and pinpoint biological mechanisms.

In the specific case of topology, we often need only to reframe our current object or system under study as a similar construction in the topological language to access the wealth of supported mathematical methods. As discussed in this work, instead of a graph we construct a simplicial complex, solving linear relations between elements we can see as finding the global section of a sheaf, and we can find lead-lag relationships with level two path signature terms. After reframing the problem in the language of topology, we can confidently proceed with an incredible number of analyses supported and developed throughout centuries of mathematics.

Oftentimes we ask, "Does method A outperform method B in terms of accuracy or speed?" With the increase in readily available computing power, the speed matters less. On some occasions, taking the topological approach may be the fastest and most accurate, but on other occasions it may fall behind state-of-the-art algorithms. However, we emphasize that this competition is not the point. The end goal of topology is not, for example, to compute second level path signature terms, but instead to understand the *whole* signature sequence. The entry points we describe above only brush the surface of what these topological methods naturally return. For example, although we currently interpret dimension 0 or dimension 1 homology best (in most cases), algebraic topology offers a holistic understanding of the structure by both detecting evolving cavities and their algebraic relations in *all* dimensions and providing intuitive assembly instructions for the system. What more does the entire, holistic structure tell us about the underlying biology? How can we creatively harness the information from higher dimensions or algebraic relations in path signatures, persistent homology, and sheaf cohomology? And how do we interpret the additional information in our specific system? We eagerly look forward to answers to some of these questions as topological data analyses continue to reach into the careful study of biological systems.

Citation Diversity Statement

Recent work in other fields has identified a bias in citations such that papers from women and other minorities are undercited relative to the number of such papers in the field (Caplar et al., 2017; Chakravartty et al., 2018; Dion et al., 2018; Dworkin et al., 2020; Maliniak et al., 2013; Thiem et al., 2018). Unfortunately, the fields where this citation bias has been studied do not include applied mathematics, and therefore we do not have a way to determine whether the percentages found in our paper are reflective of the field or biased in some way. Nevertheless, we sought to proactively consider choosing references that reflect the diversity of the field in thought, form of contribution, gender, and other factors. Our current paper contains references that are (first author)-(last author) approximately 60% man-man, 12% man-woman, 11% woman-man, and 14% woman-woman, after excluding self-citations. We look forward to future work that could help us to better understand how to support equitable practices in applied mathematics.

Acknowledgments We are extremely grateful to Lia Papadopolous for early development of this manuscript. We also graciously thank Darrick Lee, Chad Giusti, and Jakob Hansen for helpful discussions. Finally we appreciate the iterative feedback and insightful comments that Jason Kim, Erin Teich, Christopher Lynn, Jennifer Stiso, Ari Kahn, Harang Ju, Eli Cornblath, Likhitha Kolla, Brennan Klein, and Leonardo Torres provided throughout manuscript preparation. This work was supported by the Paul G. Allen Family Foundation; the John D. and Catherine T. MacArthur foundation; the Army Research Office (Bassett-W911NF-14-1-0679, Grafton-W911NF-16-1-0474, DCIST-W911NF-17-2-0181); the National Institute of Mental Health (2-R01-DC-009209-11, R01-MH112847, R01-MH107235, R21-M MH-106799); the National Institute of Child Health and Human Development (1R01HD086888-01); the National Institute of Neurological Disorders and Stroke (R01 NS099348); and the National Science Foundation (BCS-1441502, BCS-1430087, NSF PHY-1554488, and BCS-1631550). The content is solely the responsibility of the authors and does not necessarily represent the official views of any of the funding agencies.

References

Adams H, Emerson T, Kirby M, Neville R, Peterson C, Shipman P, Chepushtanova S, Hanson E, Motta F, Ziegelmeier L (2017) Persistence images: a stable vector representation of persistent homology. J Mach Learn Res 18(1):218–252

Arnold D, Falk R, Winther R (2010) Finite element exterior calculus: from Hodge theory to numerical stability. Bull Am Math Soc 47(2):281–354

Baryshnikov Y, Schlafly E (2016) Cyclicity in multivariate time series and applications to functional MRI data. In: 2016 IEEE 55th conference on decision and control (CDC). IEEE, pp 1625–1630

Bauer U, Kerber M, Reininghaus J, Wagner H (2017) Phat–persistent homology algorithms toolbox. J Symb Comput 78:76–90

Baum GL, Ciric R, Roalf DR, Betzel RF, Moore TM, Shinohara RT, Kahn AE, Vandekar SN, Rupert PE, Quarmley M et al (2017) Modular segregation of structural brain networks supports the development of executive function in youth. Curr Biol 27(11):1561–1572

Bella J (2016) Collagen structure: new tricks from a very old dog. Biochem J 473(8):1001–1025

Bendich P, Marron JS, Miller E, Pieloch A, Skwerer S (2016) Persistent homology analysis of brain artery trees. Ann Appl Stat 10(1):198

Betancourt C, Chalifour M, Neville R, Pietrosanu M, Tsuruga M, Darcy I, Heo G (2018) Pseudo-multidimensional persistence and its applications. In: Research in computational topology. Springer, pp 179–202

Bhaskar D, Manhart A, Milzman J, Nardini JT, Storey KM, Topaz CM, Ziegelmeier L (2019) Analyzing collective motion with machine learning and topology. Chaos: An Interdisciplinary J Nonlinear Sci 29(12):123125

Blevins AS, Bassett DS (2020) Reorderability of node-filtered order complexes. https://journals.aps.org/pre/abstract/10.1103/PhysRevE.101.052311

Boser BE, Guyon IM, Vapnik VN (1992) A training algorithm for optimal margin classifiers. In: Proceedings of the fifth annual workshop on Computational learning theory, pp 144–152

Braslavsky I, Stavans J (2018) On a non-trivial application of algebraic topology to molecular biology. Available at SSRN 3188351

Brown TG (1914) On the nature of the fundamental activity of the nervous centres; together with an analysis of the conditioning of rhythmic activity in progression, and a theory of the evolution of function in the nervous system. J Physiol 48(1):18–46

Brüel-Gabrielsson R, Nelson BJ, Dwaraknath A, Skraba P, Guibas LJ, Carlsson G (2019) A topology layer for machine learning. arXiv preprint arXiv:1905.12200

Bubenik P, Hull M, Patel D, Whittle B (2020) Persistent homology detects curvature. Inverse Problems. IOP Publishing 36(2):025008

Caplar N, Tacchella S, Birrer S (2017) Quantitative evaluation of gender bias in astronomical publications from citation counts. Nat Astron 1(6):0141

Carlsson G (2009) Topology and data. Bull Am Math Soc 46(2):255–308

Carlsson G, Gabrielsson RB (2018) Topological approaches to deep learning. arXiv preprint arXiv:1811.01122

Carlsson G, Ishkhanov T, De Silva V, Zomorodian A (2008) On the local behavior of spaces of natural images. Int J Comput Vis 76(1):1–12

Cartwright HM (2008) Artificial neural networks in biology and chemistry – the evolution of a new analytical tool. In: Artificial neural networks, pp 1–13. Springer

Chakravartty P, Kuo R, Grubbs V, McIlwain C (2018) # communicationsowhite. J Commun 68(2):254–266

Chambers EW, Letscher D (2018) Persistent homology over directed acyclic graphs. In: Research in computational topology. Springer, pp 11–32

Chen C, Ni X, Bai Q, Wang Y (2019) A topological regularizer for classifiers via persistent homology. In: The 22nd international conference on artificial intelligence and statistics, pp 2573–2582

Chevyrev I, Kormilitzin A (2016) A primer on the signature method in machine learning. arXiv preprint arXiv:1603.03788

Cohen-Steiner D, Edelsbrunner H, Harer J (2007) Stability of persistence diagrams. Discret Comput Geom 37(1):103–120

Comon P (1994) Independent component analysis, a new concept? Signal Process 36(3):287–314

Corbet R, Fugacci U, Kerber M, Landi C, Wang B (2019) A kernel for multi-parameter persistent homology. Comput Graph X:100005

Curry JM (2014) Sheaves, cosheaves and applications. Ph.D. thesis, The University of Pennsylvania

Curto C (2017) What can topology tell us about the neural code? Bull Am Math Soc 54(1):63–78

Damiano DB, McGuirl MR (2018) A topological analysis of targeted in-111 uptake in spect images of murine tumors. J Math Biol 76(6):1559–1587

De Silva V, Morozov D, Vejdemo-Johansson M (2011) Persistent cohomology and circular coordinates. Discret Comput Geom 45(4):737–759

Devarajan K (2008) Nonnegative matrix factorization: an analytical and interpretive tool in computational biology. PLoS Comput Biol 4(7):1–12

Dey TK, Edelsbrunner H, Guha S (1999) Computational topology. Contemp Math 223:109–144

Dion ML, Sumner JL, Mitchell SM (2018) Gendered citation patterns across political science and social science methodology fields. Polit Anal 26(3):312–327

Dworkin JD, Linn KA, Teich EG, Zurn P, Shinohara RT, Bassett DS (2020) The extent and drivers of gender imbalance in neuroscience reference lists. Nature Neuroscience. https://doi.org/10.1038/s41593-020-0658-y

Edelsbrunner H (1995) The union of balls and its dual shape. Discret Comput Geom 13(3–4):415–440

Edelsbrunner H, Harer J (2010) Computational topology: an Introduction. American Mathematical Society, Providence

Edelsbrunner H, Koehl P (2017) Handbook of discrete and computational geometry. Chapman and Hall/CRC

Edelsbrunner H, Letscher D, Zomorodian A (2000) Topological persistence and simplification. In: Proceedings 41st annual symposium on foundations of computer science. IEEE, pp 454–463

Edelsbrunner H, Morozov D (2012) Persistent homology: theory and practice. Technical report, Lawrence Berkeley National Lab (LBNL), Berkeley

Fasy BT, Wang B (2016) Exploring persistent local homology in topological data analysis. In: 2016 IEEE international conference on acoustics, speech and signal processing (ICASSP). IEEE, pp 6430–6434

Friz PK, Victoir NB (2010) Multidimensional stochastic processes as rough paths: theory and applications, vol 120. Cambridge University Press, Cambridge

Gameiro M, Hiraoka Y, Izumi S, Kramar M, Mischaikow K, Nanda V (2015) A topological measurement of protein compressibility. Jpn J Ind Appl Math 32(1):1–17

Gasparovic E, Gommel M, Purvine E, Sazdanovic R, Wang B, Wang Y, Ziegelmeier L (2019) Local versus global distances for zigzag persistence modules. arXiv preprint arXiv:1903.08298

Ghrist R (2008) Barcodes: the persistent topology of data. Bull Am Math Soc 45(1):61–75

Giusti C, Ghrist R, Bassett DS (2016, in press) Two's company, three (or more) is a simplex: algebraic-topological tools for understanding higher-order structure in neural data. J Complex Netw 41:1–14

Giusti C, Lee D (2018) Path space cochains and population time series analysis. arXiv preprint arXiv:1811.03558

Giusti C, Pastalkova E, Curto C, Itskov V (2015) Clique topology reveals intrinsic geometric structure in neural correlations. Proc Natl Acad Sci 112(44):13455–13460

Gopalakrishnan B, Wei-Man Wang, Greenspan DS (2004) Biosynthetic processing of the pro-$\alpha 1$ (v) pro-$\alpha 2$ (v) pro-$\alpha 3$ (v) procollagen heterotrimer. J Biol Chem 279(29):30904–30912

Gosak M, Markovič R, Dolenšek J, Rupnik MS, Marhl M, Stožer A, Perc M (2018) Network science of biological systems at different scales: a review. Phys Life Rev 24:118–135

Hansen J (2019a) A gentle introduction to sheaves on graphs. Available at http://www.jakobhansen.org/publications/gentleintroduction.pdf

Hansen J, Ghrist R (2019b) Toward a spectral theory of cellular sheaves. Springer, J Appl Comput Topol 3(4):315–358

Hansen J, Ghrist R (2019c) Learning sheaf laplacians from smooth signals. In: ICASSP 2019–2019 IEEE international conference on acoustics, speech and signal processing. IEEE, pp 5446–5450

Hansen J, Ghrist R (2019d) Distributed Optimization with Sheaf Homological Constraints. 57th Annual Allerton Conference on Communication, Control, and Computing (Allerton). IEEE, pp 565–571

Henselman G, Ghrist R (2016) Matroid filtrations and computational persistent homology. arXiv preprint arXiv:1606.00199

Iacopini I, Petri G, Barrat A, Latora V (2019) Simplicial models of social contagion. Nat Commun 10(1):2485

Jordan C (1875) Essai sur la géométrie à n dimensions. Bulletin de la Société Mathématique de France 3:103–174

Joslyn CA, Hogan EA, Robinson M (2014) Towards a topological framework for integrating semantic information sources. STIDS. 93–96

Ju H, Kim JZ, Bassett DS (2018) Network topology of neural systems supporting avalanche dynamics predicts stimulus propagation and recovery. bioRxiv. Cold Spring Harbor Laboratory 504761

Kališnik S (2019) Tropical coordinates on the space of persistence barcodes. Found Comput Math 19(1):101–129

Katifori E, Magnasco MO (2012) Quantifying loopy network architectures. PLoS One 7(6):e37994

Katifori E, Szöllősi GJ, Magnasco MO (2010) Damage and fluctuations induce loops in optimal transport networks. Phys Rev Lett 104(4):048704

Kleene SC (1951) Representation of events in nerve nets and finite automata. Technical report, Rand Project Air Force, Santa Monica

Kobak D, Berens P (2019) The art of using t-SNE for single-cell transcriptomics. Nat Commun 10(1):1–14

Koch Y, Wolf T, Sorger PK, Eils R, Brors B (2013) Decision-tree based model analysis for efficient identification of parameter relations leading to different signaling states. PLoS One 8(12):1–10

Kozlov D (2007) Combinatorial algebraic topology, vol 21. Springer Science & Business Media. Berlin, Germany

Kramár M, Levanger R, Tithof J, Suri B, Xu M, Paul M, Schatz MF, Mischaikow K (2016) Analysis of kolmogorov flow and rayleigh–bénard convection using persistent homology. Physica D: Nonlinear Phenomena 334:82–98

Lee DD, Seung HS (2001) Algorithms for non-negative matrix factorization. In: Advances in neural information processing systems, pp 556–562

Lever J, Krzywinski M, Altman N (2017) Points of significance: principal component analysis. Nature Methods. Nat Pub Group 14(7):641–642

Lyons T (2014) Rough paths, signatures and the modelling of functions on streams. arXiv preprint arXiv:1405.4537

Lyons T, Qian Z, Qian Z et al (2002) System control and rough paths. Oxford University Press, Oxford

Lyons TJ, Caruana M, Lévy T (2007) Differential equations driven by rough paths. Springer, Berlin

van der Maaten L, Hinton G (2008) Visualizing data using t-SNE. J Mach Learn Res 9(2):2579–2605

Mackevicius EL, Bahle AH, Williams AH, Gu S, Denisenko NI, Goldman MS, Fee MS (2019) Unsupervised discovery of temporal sequences in high-dimensional datasets, with applications to neuroscience. Elife 8:e38471

Maliniak D, Powers R, Walter BF (2013) The gender citation gap in international relations. Int Organ 67(4):889–922

Marder E, Bucher D (2001) Central pattern generators and the control of rhythmic movements. Curr Biol 11(23):R986–R996

Maria C, Boissonnat J-D, Glisse M, Yvinec M (2014) The GUDHI library: simplicial complexes and persistent homology. In: International congress on mathematical software. Springer, pp 167–174

Melville R (1969) Leaf venation patterns and the origin of the angiosperms. Nature 224(5215):121

Menche J, Sharma A, Kitsak M, Ghiassian SD, Vidal M, Loscalzo J, Barabási A-L (2015) Uncovering disease-disease relationships through the incomplete interactome. Science 347(6224):1257601

Meng Z, Vijay Anand D, Lu Y, Wu J, Xia K (2020) Weighted persistent homology for biomolecular data analysis. Scientific reports. Nat Pub Group 10(1):1–15

Monod A, Kališnik S, Patinõ Galindo JA, Crawford L (2019) Tropical sufficient statistics for persistent homology. SIAM J Appl Algebr Geom 3(2):337–371

Morozov D (2012) Dionysus library for computing persistent homology. Software available at http://www.mrzv.org/software/dionysus2

Munch E (2013) Applications of persistent homology to time varying systems. Ph.D. thesis

Munch E (2017) A user's guide to topological data analysis. J Learn Anal 4(2):47–61

Nanda V, Sazdanović R (2014) Simplicial models and topological inference in biological systems. In: Discrete and topological models in molecular biology. Springer, pp 109–141

Naylor MG, Lin X, Weiss ST, Raby BA, Lange C (2010) Using canonical correlation analysis to discover genetic regulatory variants. PLoS One 5(5):1–6

Okubo TS, Mackevicius EL, Payne HL, Lynch GF, Fee MS (2015) Growth and splitting of neural sequences in songbird vocal development. Nature 528(7582):352

Otter N, Porter MA, Tillmann U, Grindrod P, Harrington HA (2017) A roadmap for the computation of persistent homology. EPJ Data Sci 6(1):17

Patania A, Vaccarino F, Petri G (2017) Topological analysis of data. EPJ Data Sci 6(1):7

Pearson K (1901) LIII. On lines and planes of closest fit to systems of points in space. Lond Edinb Dublin Philos Mag J Sci 2(11):559–572

Perea JA (2018) Multiscale projective coordinates via persistent cohomology of sparse filtrations. Discret Comput Geom 59(1):175–225

Petri G, Expert P, Turkheimer F, Carhart-Harris R, Nutt D, Hellyer PJ, Vaccarino F (2014) Homological scaffolds of brain functional networks. J R Soc Interface 11(101):20140873

Petri G, Barrat A (2018) Simplicial activity driven model. Phys Rev Lett 121(22):228301

Petri G, Scolamiero M, Donato I, Vaccarino F (2013) Topological strata of weighted complex networks. PLoS One 8(6):e66506

Piangerelli M, Rucco M, Tesei L, Merelli E (2018) Topological classifier for detecting the emergence of epileptic seizures. BMC Res Notes 11(1):392

Power JD, Cohen AL, Nelson SM, Wig GS, Barnes KA, Church JA, Vogel AC, Laumann TO, Miezin FM, Schlaggar BL, Petersen SE (2011) Functional network organization of the human brain. Neuron 72(4):665–678

Pun CS, Xia K, Lee SX (2018) Persistent-homology-based machine learning and its applications–a survey. arXiv preprint arXiv:1811.00252

Purvine E, Aksoy S, Joslyn C, Nowak K, Praggastis B, Robinson M (2018) A topological approach to representational data models. In: International conference on human interface and the management of information, pp 90–109. Springer

Qaiser T, Sirinukunwattana K, Nakane K, Tsang Y-W, Epstein D, Rajpoot N (2016) Persistent homology for fast tumor segmentation in whole slide histology images. Proc Comput Sci 90:119–124

Quinlan JR (1987) Simplifying decision trees. International journal of man-machine studies, Elsevier 27(3):221–234

Rabadán R, Blumberg AJ (2019) Topological data analysis for genomics and evolution: topology in biology. Cambridge University Press, Cambridge

Rieck B, Fugacci U, Lukasczyk J, Leitte H (2017) Clique community persistence: A topological visual analysis approach for complex networks. IEEE Trans Vis Comput Graph 24(1):822–831

Robins V, Abernethy J, Rooney N, Bradley E (2004) Topology and intelligent data analysis. Intell Data Anal 8(5):505–515

Robinson M (2014) Topological signal processing. Springer, Berlin, Germany

Robinson M (2017) Sheaves are the canonical data structure for sensor integration. Inf Fusion 36:208–224

Anita Roth-Nebelsick, Uhl D, Mosbrugger V, Kerp H (2001) Evolution and function of leaf venation architecture: a review. Ann Bot 87(5):553–566

Rubin A, Sheintuch L, Brande-Eilat N, Pinchasof O, Rechavi Y, Geva N, Ziv Y (2019) Revealing neural correlates of behavior without behavioral measurements. Nat Commun 10(1):1–14

Sanderson N, Shugerman E, Molnar S, Meiss JD, Bradley E (2017) Computational topology techniques for characterizing time-series data. In: International symposium on intelligent data analysis. Springer, pp 284–296

Scaramuccia S, Iuricich F, Leila De Floriani, Landi C (2020) Computing multiparameter persistent homology through a discrete morse-based approach. Comput Geo, Elsevier 89:101623

Schlick T, Olson WK (1992) Trefoil knotting revealed by molecular dynamics simulations of supercoiled dna. Science 257(5073):1110–1115

Shoulders MD, Raines RT (2009) Collagen structure and stability. Annu Rev Biochem 78:929–958

Singer A, Wu H-T (2012) Vector diffusion maps and the connection laplacian. Commun Pure Appl Math 65(8):1067–1144

Singh G, Mémoli F, Carlsson GE (2007) Topological methods for the analysis of high dimensional data sets and 3D object recognition. In: SPBG, pp 91–100

Singh G, Memoli F, Ishkhanov T, Sapiro G, Carlsson G, Ringach DL (2008) Topological analysis of population activity in visual cortex. J Vis 8(8):11–11

Sizemore A, Giusti C, Bassett DS (2017) Classification of weighted networks through mesoscale homological features. J Complex Netw, Oxford University Press 5(2):245–273

Sizemore AE, Giusti C, Kahn A, Vettel JM, Betzel RF, Bassett DS (2018) Cliques and cavities in the human connectome. J Comput Neurosci, Springer 44(1):115–145

Sizemore AE, Karuza EA, Giusti C, Bassett DS (2018) Knowledge gaps in the early growth of semantic feature networks. Nat Hum Behav 2(9):682–692

Sporns O (2010) Networks of the brain. MIT Press, Cambridge

Stolz B (2014) Computational topology in neuroscience. Master's thesis, University of Oxford

Stolz BJ, Emerson T, Nahkuri S, Porter MA, Harrington HA (2018) Topological data analysis of task-based FMRI data from experiments on schizophrenia. arXiv preprint arXiv:1809.08504

Tausz A, Vejdemo-Johansson M, Adams H (2014) JavaPlex: a research software package for persistent (co)homology. In: Hong H, Yap C (eds) Proceedings of ICMS 2014. Lecture Notes in Computer Science, vol 8592, pp 129–136. Software available at http://appliedtopology.github.io/javaplex/

Thiem Y, Sealey KF, Ferrer AE, Trott AM, Kennison R (2018) Just ideas? The status and future of publication ethics in philosophy: a white paper. Technical report

Thom R (1969) Topological models in biology. Topology 8(3):313–335

Topaz CM, Ziegelmeier L, Halverson T (2015) Topological data analysis of biological aggregation models. PLoS One 10(5):e0126383

Trinh MH, Van Nguyen C, Lim Y-H, Ahn H-S (2018) Matrix-weighted consensus and its applications. Automatica 89:415–419

Vepstas L (2019) Sheaves: a topological approach to big data. arXiv preprint arXiv:1901.01341

Wang B, Summa B, Pascucci V, Vejdemo-Johansson M (2011) Branching and circular features in high dimensional data. IEEE Trans Vis Comput Graph 17(12):1902–1911

Yang ZR (2004) Biological applications of support vector machines. Brief Bioinform 5(4):328–338

Yao F, Coquery J, Lê Cao K-A (2012) Independent principal component analysis for biologically meaningful dimension reduction of large biological data sets. BMC bioinform 13(1):24

Yoo J, Kim EY, Ahn YM, Ye JC (2016) Topological persistence vineyard for dynamic functional brain connectivity during resting and gaming stages. J Neurosci Methods 267:1–13

Zanzoni A, Soler-López M, Aloy P (2009) A network medicine approach to human disease. FEBS Lett 583(11):1759–1765

Zimmerman BJ, Abraham I, Schmidt SA, Baryshnikov Y, Husain FT (2018) Dissociating tinnitus patients from healthy controls using resting-state cyclicity analysis and clustering. Netw Neurosci 3(1):67–89

Zomorodian A, Carlsson G (2005) Computing persistent homology. Discret Comput Geom 33(2):249–274