Topological Data Analysis of Task-Based fMRI Data from Experiments on Schizophrenia

Bernadette J. Stolz

Mathematical Institute, University of Oxford, Oxford, UK

E-mail: stolz@maths.ox.ac.uk

Tegan Emerson

United States Naval Research Laboratory, Washington, D.C., USA

E-mail: tegan.emerson@nrl.navy.mil

Satu Nahkuri

F. Hoffmann-La Roche AG, Basel, Switzerland

E-mail: satu.nahkuri@roche.com

Mason A. Porter[1,2,3]

[1]Mathematical Institute, University of Oxford, Oxford, UK
[2]Department of Mathematics, University of California, Los Angeles, Los Angeles, USA
[3]CABDyN Complexity Centre, University of Oxford, Oxford, UK

E-mail: mason@math.ucla.edu

Heather A. Harrington

Mathematical Institute, University of Oxford, Oxford, UK

E-mail: harrington@maths.ox.ac.uk

Abstract. We use methods from computational algebraic topology to study functional brain networks, in which nodes represent brain regions and weighted edges represent similarity of fMRI time series from each region. With these tools, which allow one to characterize topological invariants such as loops in high-dimensional data, we are able to gain understanding into low-dimensional structures in networks in a way that complements traditional approaches based on pairwise interactions. In the present paper, we analyze networks constructed from task-based fMRI data from schizophrenia patients, healthy controls, and healthy siblings of schizophrenia patients using persistent homology, which allows us to explore the persistence of topological structures such as loops at different scales in the networks. We use persistence landscapes, persistence images, and Betti curves to create output summaries from our persistent-homology calculations, and we study the persistence landscapes and images using k-means clustering and community detection. Based on our analysis of persistence landscapes, we find that the members of the sibling cohort have topological

features (specifically, their 1-dimensional loops) that are distinct from the other two cohorts. From the persistence images, we are able to distinguish all three subject groups and to determine the brain regions in the loops (with four or more edges) that allow us to make these distinctions.

Keywords: persistent homology, networks, fMRI, persistence landscapes, persistence images, functional networks, functional brain networks Submitted to: *Phys. Biol.*

1. Introduction

Schizophrenia is a chronic psychiatric disorder that affects more than 21 million people worldwide [1]. Up to 80% of the risk factors appear to be genetic, although it has proven difficult to identify the specific genes that are involved in the disease [2]. The disease usually commences in early adulthood, and symptoms range from hallucinations and avolition to cognitive deficits (such as impaired working memory) [1, 3]. The cause of the cognitive deficits is thought to originate from compromised functional integration between neural subsystems [3–6]. There can be significant differences in the properties of time series from imaging measurements of healthy versus schizophrenic individuals. Different studies have found seemingly contradictory results when comparing functional magnetic resonance imaging (fMRI) time series from two distinct brain regions in a schizophrenia patient and a healthy control. The majority of studies have concluded that schizophrenia patients have less-similar time series across different brain regions [7]. Zalesky *et al.* [8] suggested that such reduced similarity may arise from an altered coupling between brain regions and local decoherence within brain regions in schizophrenia patients. However, some studies have observed that schizophrenia patients have more-similar series than controls across brain regions. For a detailed discussion of these seemingly contradictory findings, see [9]. In some cases, methodological steps in fMRI analyses seem to yield increases in these similarities, but abnormal neurodevelopment or drug treatment may play a role in increasing them in other cases [9].

One approach to study the human brain is to construct a (possibly time-dependent) neuronal network based on experimental data and then analyze the network's structure and dynamics to gain insights into its properties [10–17]. One can form a so-called *functional network* [10, 11, 18–20], in which each node represents a brain region, and the edges between them are weighted based on some measure of the similarity between their fMRI time series. In Fig. 1, we show a pipeline of how to construct a functional network from fMRI time-series data. Note that when interpreting fMRI studies, it is very important to consider the cautionary notes in [21].

Studies of functional networks of schizophrenia patients have revealed that such networks differ significantly from the functional networks of healthy controls [6, 7, 22–26].

For example, schizophrenia patients can have rather different community structure from controls [24, 27]. In one paper, Alexander-Bloch *et al.* [24] observed that a small subset of brain regions lead to significant differences in the community assignments in schizophrenia patients, whereas the communities for healthy subjects appear to be consistent with each other. Moreover, the maximum modularity of functional networks appears to be smaller for schizophrenia patients than in healthy controls [24, 28]. Two recent papers, Flanagan *et al.* [27] and Towlson *et al.* [29], compared the network structures of schizophrenia patients and healthy controls under the effects of different drugs and a placebo.

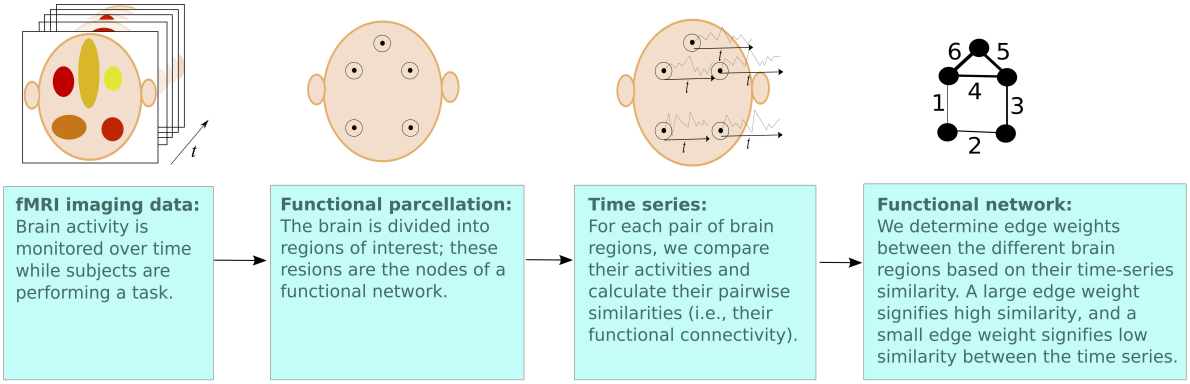


Figure 1: Pipeline to construct functional networks from imaging data (e.g., fMRI data).

An increasingly popular approach for the analysis of functional networks is to use ideas from computational algebraic topology, as they allow one to characterize topological invariants (such as connectedness and loops) in high-dimensional data [30–35]. In contrast to standard methods of network analysis [36], employing computational topology allows one to explicitly go beyond pairwise connections; this is helpful for gaining global understanding of low-dimensional structures in networks. Although one can also use frameworks such as hypergraphs [37] to study higher-order network structures (see, e.g., [38]), such a formalism does not by itself give direct information about the shape or scale of mesoscale features in networks. By contrast, *persistent homology* (PH), the most prominent approach in topological data analysis, allows one to explore the persistence of features, such as connectedness or loops, in data sets [39,40]. Persistent homology has led to interesting insights in a variety of fields (for examples, see [41–45]); and it has been used increasingly in neuronal networks, leading to several promising insights [20, 46–55].

In this paper, we construct functional networks using fMRI data from schizophrenia patients, healthy controls, and siblings of schizophrenia patients. We create a nested sequence of networks in which we add edges, one by one, to the networks in order from largest edge weights to smallest. (In the unlikely case of two edges having the exact same weight, we add both edges simultaneously in one step.) We then construct a weight rank clique filtration (WRCF) [48] by determining cliques and tracking their changes in every step of the network sequence. We then compute PH and Betti numbers [34, 56]

of the WRCF and examine the results by applying tools from statistics and machine learning, respectively, on the persistence landscapes and persistence images that result from our computation of PH. We compare our findings from these two approaches and also study Betti numbers using Betti curves. We focus on loops (with four or more edges)‡ in the networks in our nested sequence, rather than on connected components, because one can also study the latter using more conventional approaches, such as by examining the spectrum of the combinatorial graph Laplacian [36, 37].

Our paper proceeds as follows. We introduce the data set and the mathematical methods in an intuitive way in Section 2, present our findings in Section 3, and discuss our comparisons in the context of current biological research in Section 4. We give some additional details about a few results in Appendix A.

2. Methods

2.1. Data set: fMRI data of schizophrenia patients, siblings of schizophrenia patients, and healthy controls

We use a data set that consists of time series from blood oxygen level dependent (BOLD) functional magnetic resonance imaging (fMRI) data collected from 281 subjects (54 schizophrenia patients, 50 healthy siblings of schizophrenia patients, and 177 healthy controls) with 120 time steps (where the length of 1 time step corresponds to $\Delta t = 2\,\mathrm{s}$). The brain regions were determined according to the Montreal Neurological Institute template [57]. Prior to obtaining the time series, the fMRI data were corrected for head motion, and they were normalized and smoothed with a Gaussian filter. The voxel-wise signal intensities were normalized to the whole-brain global mean. The data set was acquired by Bertolino, Blasi, and their collaborators as part of a larger fMRI data set over a period of approximately 10 years. Subsets of the data set have been studied previously [58–60]; these previous studies of the data did not include the data for siblings.

The experimentalists obtained fMRI images while subjects were performing a block paradigm of a so-called 'n-back task'. During an n-back task, subjects are presented with a sequence of numbers. In each step m of the sequence, they are first shown a number and then asked to recall the number from sequence step $m - n$. For example, during a 2-back task, subjects are shown a sequence $\{\dots, x_{i-1}, x_i, x_{i+1}, x_{i+2}, \dots\}$ and are asked to recall number x_{i-1} while being shown number x_{i+1}, recall number x_i while being shown number x_{i+2}, and so on. For the present data set, the stimuli consisted of alternating blocks of 30 seconds each of 0-back tasks and 2-back tasks.

We preprocess the data in the following way. For each subject and time step, we calculate the mean signal for white-matter brain regions, the mean signal for regions

‡ We use the term 'loop' to refer to at least four edges in a network that are connected in a way that forms a cycle. Conventionally, loops (other than self-loops) in undirected graphs must have at least 3 edges, and loops in directed graphs must have at least 2 edges. In our paper, we adapt this terminology to represent the topological features that we detect in our simplicial complexes.

that consist of cerebrospinal fluid, and the mean of the global signal. In addition to these mean values, we also use the squares and cubes of the global signal means, as well as head-motion parameters (3 translation and 3 rotation parameters), to construct 11×120 subject-specific design matrices. We then perform linear regression for each time series using MATLAB's command for the Moore–Penrose pseudoinverse § PINV(); we exclude brain regions without grey matter from our calculations. We then use the residuals from the regression as our time series for the 120 brain regions that we list in Tables A1–A5.

2.2. Functional connectivity

We obtain functional networks from the fMRI time series for each subject by using the 120 distinct brain regions (see Tables A1–A5) as the nodes of the networks and calculating Pearson correlations‖ (without a time lag) between the nodes' time series as a measure of pairwise functional connectivity. The values of the pairwise functional connectivity give the edge weights between the brain regions in the functional networks.

In all but one of our analyses, we consider four contiguous time regimes of 30 time points each, yielding four functional networks per subject. (The exception is Subsection 3.4, in which we use each subject's full time series, which consists of 120 time points, to construct a single functional network for each subject.) Although the four time regimes correspond temporally to one 0-back and one 2-back task each, our separation into time regimes is motivated by an interest in potential developments in the dynamics over time, rather than in relating the fMRI response to the task. We summarize each functional network in an adjacency matrix $A = A(\text{subject}, \text{time regime})$, whose entry A_{ij} is given by the edge weight between node i and node j. We apply a statistical threshold, described in [63], to the weighted adjacency matrices without modifying the remaining edge weights. To obtain the thresholded adjacency matrices, we estimate p-values for the correlations using the MATLAB function CORRCOEF and retain only those entries whose p-value is less than 0.05. Using this type of thresholding, we retain at most 44% of all edges in the network and on average 20–30% of the edges. We then separate each adjacency matrix into a positive and a negative part, $A = A^+ + A^-$, and study only the positive A^+ part of the adjacency matrix¶. In Fig. 2, we show a diagram of the steps that we perform to construct our functional networks. (In the one case (see Subsection 3.4) in which we study one functional network per subject instead of four, we skip the step in which we split time series; we perform all other steps in the same way.) In our computations in which we consider the four time regimes separately (i.e., except for Subsection 3.4), we treat all subjects and all time regimes together as one

§ As some of the matrices are ill-conditioned, there are variations in the resulting networks across different runs of the preprocessing code. However, in our observations, the matrices differ by only up to 0.2% of entries after two runs of preprocessing.

‖ There are numerous ways to measure functional connectivity [11, 61, 62]. For a discussion in the context of schizophrenia research, see [7].

¶ By using this approach, we avoid the interpretation of negative correlations between time series.

data set.

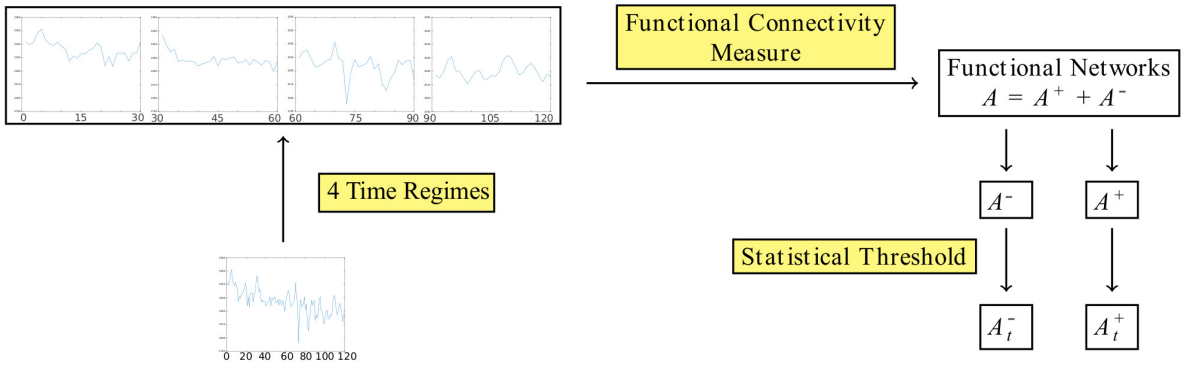


Figure 2: Steps that we perform on the preprocessed time series of each brain region to construct a functional network for each subject during each of four time regimes. We study the positive parts of the resulting networks using persistent homology.

2.3. Persistent homology

Persistent homology (PH) is a technique from topological data analysis, which aims to understand the 'shape' of data [39]. PH is based on the topological concept of *homology*, which is used to study the shape of objects, disregarding changes from stretching and bending.

We motivate our use of PH for brain networks by considering different types of cheese and how they differ in their homology. Calculating homology allows one to differentiate between the shape of a stereotypical Swiss cheese (of the Emmental sort) with holes and the shape of a mozzarella cheese by providing information about the presence or absence of holes in the cheeses. (See Fig. 3 for examples of the aforementioned cheeses.) One can thereby consider the space surrounding the holes; these are the so-called *loops*. However, homology does not give information about the geometry of the cheeses; for example, it does not 'see' that the Swiss cheese is a cube or that the mozzarella is a sphere (unless it happens to be hollow), as it only detects the differences in the number of holes.

We now give a brief intuitive introduction to a few concepts behind homology and PH for network data. For more mathematical introductions, see [20, 31, 32, 34, 39, 56].

2.3.1. Simplicial complexes To study the characteristics of topological spaces [64], such as the Swiss cheese or the mozzarella, we consider small, simplified pieces ('morsels'), on which we can perform computations more easily. When reassembled, the morsels carry the same overall topological information as the original space. We begin building these morsels (i.e., 'spaces', to be more formal) using a discrete set of points, which we call 'nodes'. We then add 'edges' to connect pairs of nodes; 'triangles', which consist of three nodes, three edges, and a face; 'tetrahedra'; and so on. Formally, these elements

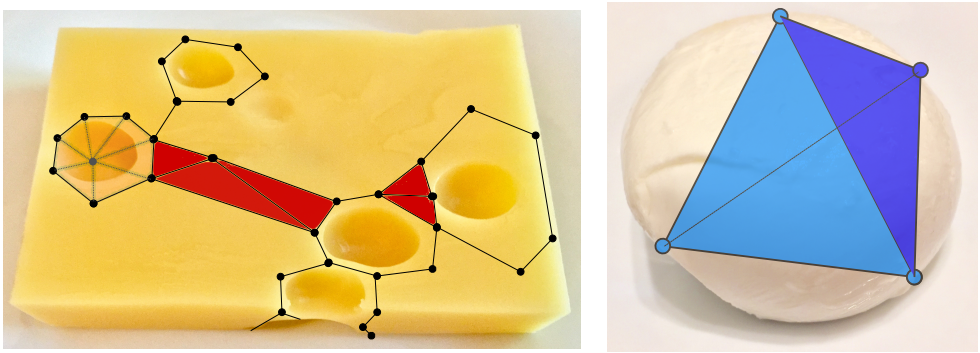
(a) Swiss (Emmental) Cheese **(b)** Mozzarella Cheese

Figure 3: An example of two topologically-different objects. Homology detects the topological differences by counting the number of holes in the cheeses.

are called k-simplices, where k indicates the dimension of the simplex. A point is a 0-simplex, an edge is a 1-simplex, a triangle is a 2-simplex, and the tetrahedron is a 3-simplex.

We can combine different simplices to capture different aspects of a topological space. For example, to capture the holes in the Emmental cheese, we glue together a collection of triangles and edges around the holes, enclosing the same number of holes as in the original cheese. Note that we can only capture the holes that are enclosed inside the cheese (using the triangles), as one can deform the visible holes on the surface into a smooth surface of the cheese. For demonstrative purposes, we therefore assume that the Emmental cheese in Fig. 3 is a cross section of a larger cheese that encloses the holes that are visible in the image.

One can combine simplices to obtain a *simplicial complex* Σ, and we take the *dimension* of Σ to be the dimension of its highest-dimensional simplex. We show examples of simplicial complexes in Fig. 4, where we again note that we are assuming that the Emmental cheese is only a cross section of a larger hunk of cheese.



(a) Swiss (Emmental) Cheese **(b)** Mozzarella Cheese

Figure 4: Simplicial complexes approximate topological spaces and capture their properties.

2.3.2. Homology and Betti numbers Homology assigns a family of vector spaces (called *homology groups* in more general settings) to a simplicial complex. For a given dimension, the vector spaces capture the topological features in that dimension. For example, for dimension 0, homology gives a vector space whose elements are connected components; for dimension 1, homology gives a vector space that has loops as its elements. The dimensions of these vector spaces are called *Betti numbers*, where β_D denotes the Betti number for dimension D. One can interpret the first three Betti numbers (β_0, β_1, and β_2), respectively, as representing the number of connected components, the number of 1-dimensional holes (i.e., loops), and the number of 2-dimensional holes (as found in the Emmental cheese) in a simplicial complex.

2.3.3. Weight rank clique filtration (WRCF) Similarly to being able to distinguish between two types of cheese, we are interested in whether we can use homology (and specifically persistent homology) to distinguish between functional networks of schizophrenia patients, siblings of schizophrenia patients, and healthy controls.

In a network, we take a loop to consist of a sequence of four or more nodes and edges that begins and ends at the same node. If two loops surround the same hole and can be deformed into one another in the space without tearing open one of the loops, then one counts the loops only once, and we construe them to be different *representatives* (also called *generators*) of a loop.

To obtain simplicial complexes from a weighted network, we construct a so-called *filtration*. A filtration is a sequence of embedded simplicial complexes that starts with the empty complex:

$$\emptyset = \Sigma_0 \subseteq \Sigma_1 \subseteq \Sigma_2 \subseteq \cdots \subseteq \Sigma_{\max} = \Sigma\,.$$

One can obtain a filtration from data in various ways. When given data in the form of a weighted network, the easiest way is to filter by weights [65]. In the first filtration step, one includes all nodes and the edge(s) with the largest weight in the simplicial complex. In the second step of the filtration, one adds the edge(s) with the second-largest weight to the simplicial complex from step one, and so on. In this way, one obtains a sequence of embedded simplicial complexes that fulfills the properties of a filtration. To construct a *weight rank clique filtration* (WRCF) [48], one performs one additional step: Whenever three edges in a simplicial complex of a filtration form a triangle, one fills in the associated face, and one construes the triangle as a 2-simplex. Similarly, when four nodes are all connected pairwise by edges, the nodes form a (filled) tetrahedron (i.e., a 3-simplex). We use the WRCF to analyze our weighted networks. The WRCF has been applied to weighted neuronal networks in several previous studies, including [20, 48, 50, 66].

We can use homology to study topological features, such as loops, in every step of a filtration and determine the amount of persistence of a feature with respect to the filtration [39]. We say that a topological feature h in a given dimension is *born* at filtration step m if the homology group of Σ_m is the first homology group of a simplicial

complex in the filtration to include the feature. Similarly, we say that a topological feature *dies* at filtration step n if it is present in the homology group of Σ_{n-1} but not in the homology group of Σ_n. The lifetime of a feature in a filtration is defined as the *persistence* p. That is,

$$p = n - m\,. \tag{1}$$

If a feature persists until the last filtration step, we say that it has *infinite persistence*. Persistence was first used as a measure to rank topological features by their lifetime in a filtration in [30].

Ideally, one performs a WRCF on a fully connected functional network. However, because of the high computational cost, this is often impossible in practice. We avoid this issue by thresholding our weighted networks before analyzing them.

2.4. Representations of persistent homology

There are multiple ways to represent the output of persistent homology (PH) calculations and to visualize the persistence of topological features and their location within a filtration. The most common representations are barcodes and persistence diagrams. In recent years, a desire to leverage the output of PH computations for machine-learning and data-mining tasks has resulted in the development of alternative representations to both barcode and persistence diagrams [39]. Two of these alternative representations are persistence landscapes [67, 68] and persistence images [69]. In the following subsubsections we describe barcodes, persistence diagrams, persistence landscapes and persistence images.

2.4.1. Barcodes A common representation of the output of PH calculations is a *barcode* [32]. See Fig. 5 for an example. A D-dimensional barcode is a plot of time intervals $[\text{birth}, \text{death})$ that indicate the birth and death of topological features of dimension D. The horizontal axis represents the filtration steps, and each D-dimensional topological feature in the filtration is represented by a bar that starts at the filtration step at which the feature is born and ends at the filtration step at which it dies. In a 0-dimensional barcode, each bar corresponds to a connected component, and the length of a bar indicates how long a particular component is disconnected from other components in a simplicial complex. Similarly, in a 1-dimensional barcode, each bar corresponds to a loop in a simplicial complex.

2.4.2. Persistence diagrams As an alternative to a barcode, one can use a *persistence diagram* (PD) [70], which is a planar representation of a barcode that conveys the same information: each $[\text{birth}, \text{death})$ interval in the barcode is mapped to birth–death coordinates, where the horizontal coordinate of a point represents the birth time of a feature in the filtration and its vertical coordinate represents the death time of that feature. Alternatively, one can use a birth–persistence coordinate system, which is particularly useful when examining persistence images. Points that are farther

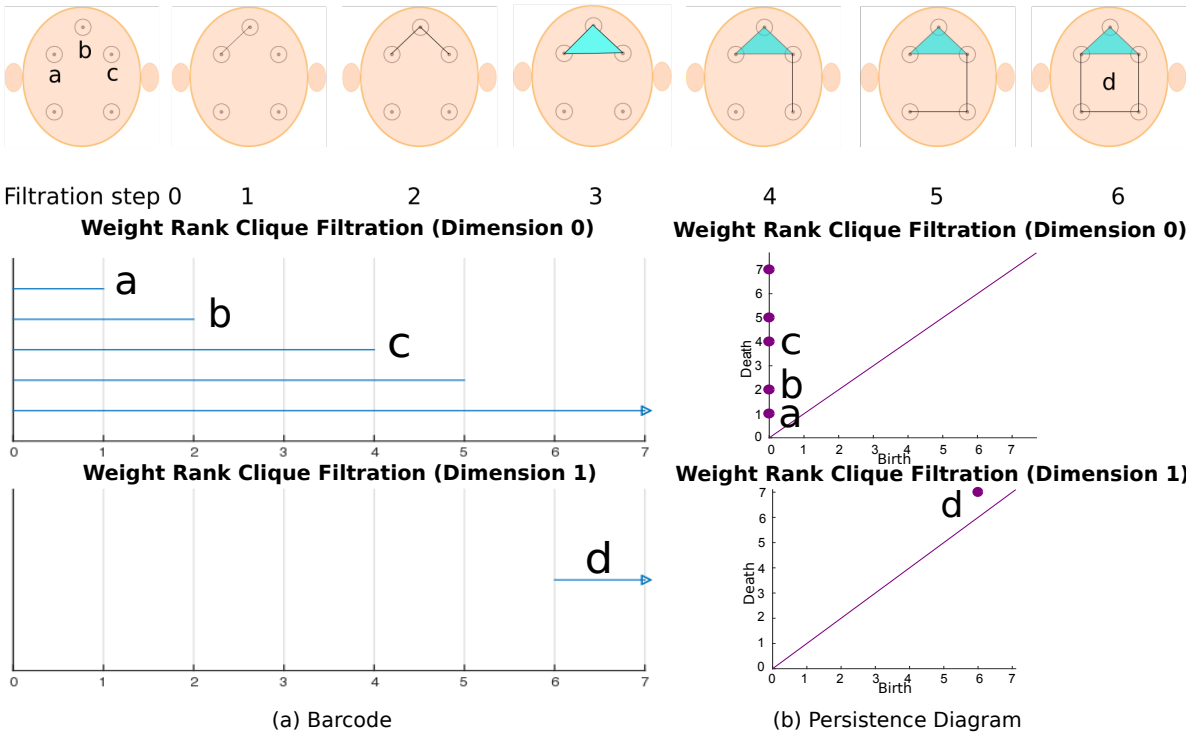


Figure 5: Example of a weight rank clique filtration (WRCF) of a neuronal network and the corresponding (a) barcodes and (b) persistence diagrams (PDs) in dimension 0 and 1. The neuronal network consists of different brain regions (indicated by circles), which we interpret as the nodes (indicated by dots) of a network, and weighted edges between the nodes. To construct the filtration, we add the nodes in step 0, followed by the edge with the largest weight in step 1, the edge with the second-largest weight in step 2, and so on. As soon as three nodes are all connected pairwise by edges, we cover the resulting region with a triangle. When four nodes are all connected pairwise, we fill in a tetrahedron. In a 0-dimensional barcode, each connected component is represented by a bar starting when the component is born and ending when it dies (e.g., when two components combine with each other). In a 1-dimensional barcode, each bar represents a loop, which consists of 4 or more edges and starts and ends at the same node. In persistence diagrams, on represents topological features by points rather than by bars. The distance of a point to the diagonal (the purple line) indicates the persistence of the corresponding feature in the filtration.

away from the diagonal identity line represent more-persistent topological features in a filtration. We show an example of a persistence diagram in Fig. 5. As with barcodes, one can treat persistence diagrams as mathematical objects, and one can endow the space of persistence diagrams with a distance.

2.4.3. Persistence landscapes A *persistence landscape* (PL) [67, 68] is a sequence of piecewise-linear functions that one can use to visualize and analyze the information in a barcode or persistence diagram. Instead of using a bar and its length to represent a feature and its persistence, one now interprets each topological feature as a peak, whose height is determined by the feature's persistence and whose location corresponds to the

feature's location in the filtration. In contrast to a barcode or persistence diagram, a PL has three dimensions. As in a barcode, the horizontal axis represents the filtration step. The other two dimensions of a PL are the persistence of a feature and the different layers of the PL.

To create a PL from a barcode, one first defines peak functions for each bar. For a given [birth, death) interval in a barcode, one constructs the function

$$f_{[\text{birth,death})}(x) = \begin{cases} 0\,, & \text{if } x \notin (\text{birth}, \text{death})\,, \\ x - \text{birth}\,, & \text{if } x \in \left(\text{birth}, \frac{\text{birth}+\text{death}}{2}\right]\,, \\ -x + \text{death}\,, & \text{if } x \in \left(\frac{\text{birth}+\text{death}}{2}, \text{death}\right)\,. \end{cases} \tag{2}$$

One then collapses the collection of peak functions onto the horizontal axis of the barcode. For a barcode that consists of the collection $\{[\text{birth}, \text{death})_l\}_{l=1}^{t}$ of intervals, the qth layer (with $q \geq 0$) of the PL (i.e., the qth *PL*) is the following set of functions:

$$\begin{aligned} &\lambda_q : \mathbb{R} \to \mathbb{R}\,, \\ &\lambda_q(x) = q\text{th-largest value of } \left\{f_{[\text{birth,death})_l}(x)\right\}_{l=1}^{t}\,. \end{aligned} \tag{3}$$

If the qth-largest value does not exist, $\lambda_q(x) = 0$. The 0th layer of a PL consists of the maximum function values among the collection of functions evaluated across the filtration. Similarly, the 1st layer of the PL consists of the second-largest values of the collection of functions evaluated across a filtration. Other layers are defined analogously. The *persistence landscape* λ of a barcode $\{[\text{birth}, \text{death})_l\}_{l=1}^{t}$ is defined as the sequence $\{\lambda_q\}$ of the functions λ_q. We illustrate the pipeline from a barcode to a PL in Fig. 6.

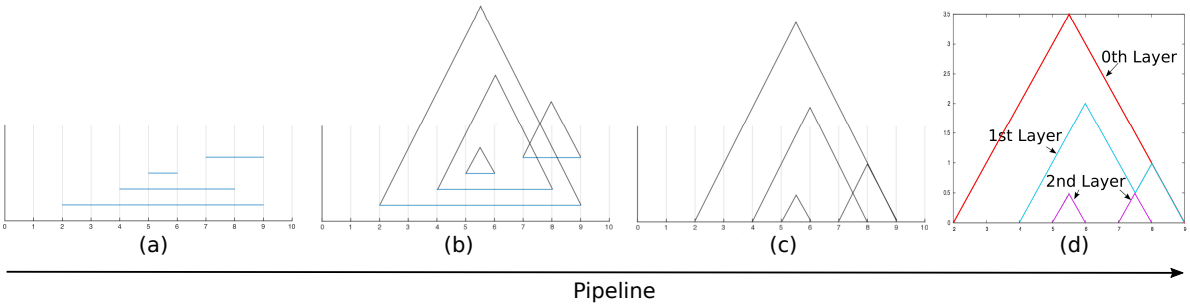


Figure 6: Schematic illustration of the steps for converting a barcode into a persistence landscape (PL). We use an example based on a weight rank clique filtration (WRCF) in dimension 1. (a) Example barcode. (b) One defines peak functions on the bars of a barcode. (c) One collapses the images of the peak functions onto the horizontal axis. (d) The PL consists of the collection of layers q (with $q = 0$, $q = 1$, and $q = 2$ in this example), which indicate the qth-largest values of the collection of peak-function values. To visualize the third dimension, we show the different layers using different colors. (This figure is a modified version of a figure used in [20].)

An advantage of PLs is that one can construct a mean PL for a set of landscapes. A mean landscape no longer corresponds to a barcode or a persistence diagram. However,

one can define pairwise distances between two or more mean landscapes and use them to quantify the difference between two sets of barcodes. We will use the L^2 distance. One can also use a variety of statistical tools on PLs [67]. Such calculations have been used for applications such as conformational changes in protein binding sites [71], phase separation in binary metal alloys [72], classification of music audio signals [73], and human motor learning [20].

2.4.4. Persistence images Another representation of topological features in PH calculations are *persistence images* (PIs), which are based on persistence diagrams and take the form of real-valued vectors[+] that one can use as an input into a variety of machine-learning approaches. The transformation from a PD to a PI is stable with respect to the 1-Wasserstein distance and maintains a clear and interpretable connection to the original PD [69].

We show a schematic of the mapping from a PD to a PI in Fig. 7, which depicts the various stages involved in the transformation. Recall that the output of a PH computation is usually a set of points (or intervals) corresponding to the birth and death times of each topological feature for a specified homological dimension. In Fig. 7(b), we show a sample PD[*] for the sample point cloud from Fig. 7(a). In Fig. 7(c), we show the PD in the birth–persistence coordinate system. In Fig. 7(d), we show an overlay of the surface generated by centering a two-dimensional (2D) Gaussian on each point in the rotated PD in Fig. 7(c). Finally, in Fig. 7(e), we show an example PI produced by computing the volume under the surface in Fig. 7(d) over a uniformly-spaced grid. (We set the resolution so that we have a 50×50 grid of elements.) One can then reshape this final PI into a vector by stacking the columns (or, equivalently, the rows), as is often done in image processing. As described in [69], the generation of a PI involves the choice of (1) a 2D probability density function to center at each point in the birth–persistence PD, (2) a resolution, and (3) a weighting function. The role of the weighting function is, when necessary, to suppress points in a PD that lie very close to the diagonal and are often construed as "noisy" features. For all PIs in this paper, we use the default settings for the code: 2D Gaussian probability density functions, a linear weighting function, and a 50×50 grid of elements. We choose additional parameters that are associated with these choices (e.g., the variance of the Gaussians) according to the defaults in [74]. When analyzing and comparing multiple PIs, there is an additional pair of values that one must choose based on the data; these are the maximum birth and persistence values. We will see in Subsection 3.3 that this choice can influence our results.

2.4.5. Software employed For our PH calculations, we implement MATLAB code constructed using JAVAPLEX [75], a software package for PH[♯]. For a given filtration

[+] We use the term 'vectorization' for the production of such vectors.

[*] This corresponds to the output after running a WRCF for dimension 1 on the functional network of the first sibling from the first time regime in our data set.

[♯] For an overview of available PH software and additional references, see [39].

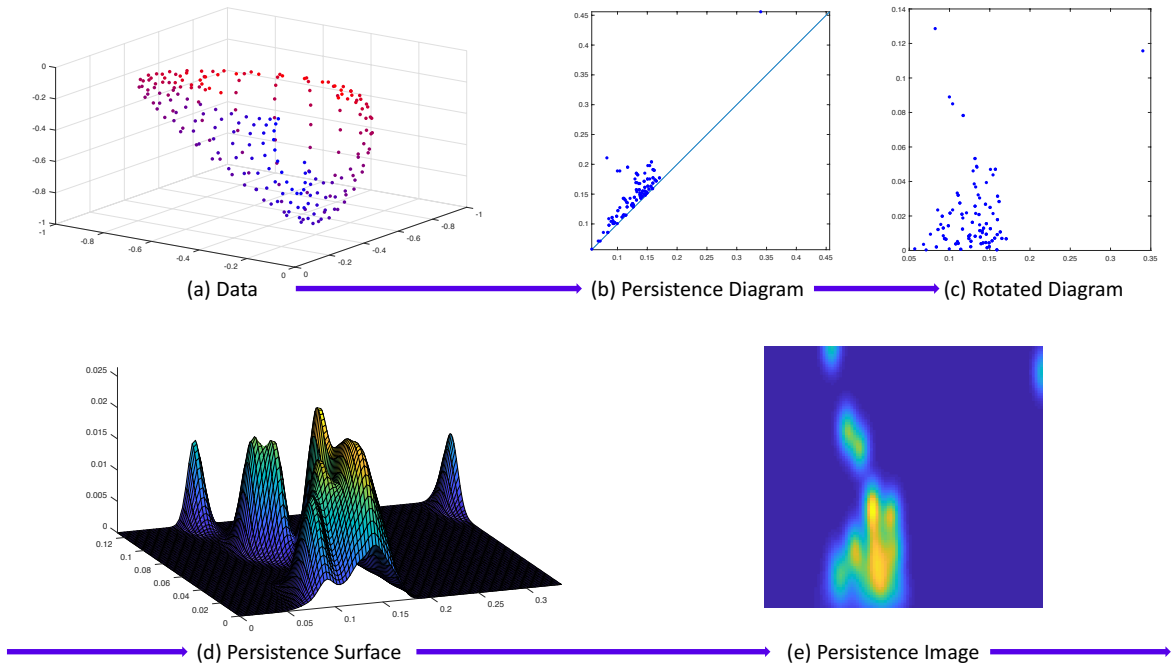


Figure 7: Schematic illustrating the primary steps for converting a persistence diagram (PD) to a persistence image (PI). (a) Sample point cloud in $\mathbb{R}^4$ plotted in $\mathbb{R}^3$ where the coloring corresponds to the fourth coordinate value. (b) PD in birth–death coordinates (i.e., the standard choice), with the diagonal identity line in blue. (c) PD in birth–persistence coordinates. (d) The process of generating a surface by centering 2D Gaussian distributions at each point in panel (c). (e) One generates a PI by summing the volume under 2D Gaussian distributions over the area of a pixel (i.e., the area of a square) in a uniformly-spaced grid overlay.

of a simplicial complex, JAVAPLEX can output [birth, death) barcode intervals, representatives for each topological feature, and PDs. JAVAPLEX outputs PDs in the standard birth–death coordinates, from which one computes birth–persistence coordinates as (birth, death − birth). For the WRCF, we also use a maximal clique-finding algorithm (that is based on the Bron–Kerbosch algorithm [76]) from the Mathworks library [77]. For the analysis and interpretation of our barcodes, we use the PERSISTENCE LANDSCAPES TOOLBOX [68]. We create PIs using the code available at [74] with the default parameters.

2.5. Clustering methods from data mining and network analysis

Given output of PH calculations, one can use clustering methods. There are myriad ways to proceed. In this paper, we use a few different approaches. First, we apply the k-means clustering algorithm and community detection to examine whether we can separate the three subject groups based on the topological features of their functional networks. Second, we apply a linear sparse support vector machine (SSVM) to identify pixels in PIs to discriminate between the subject groups and examine which brain regions are

generators of loops that help discriminate between groups. We describe these techniques in the following subsections.

2.5.1. Employing k-means clustering for subject-group separation The method of k-means clustering aims to produce a partition of a metric space into k clusters of points [78]. Suppose that there are μ data points in the metric space. One selects k of the μ points as "centers" and assigns all other points of a data set into clusters based on their closest center point. The "score" of such a clustering is the sum of the distances from each point to its nearest center. The desired output of k-means clustering is an assignment of points to clusters with the minimum clustering score. An exhaustive search for a global minimum is often prohibitively expensive. A typical approach to search for a global minimum is to choose a large selection of k initial centers uniformly at random, improve each selection of centers iteratively until the clustering score stabilizes, and then return the identified final clustering with the lowest score for each initialization. One iteratively updates the centers by setting the new center equal to the mean of the points assigned to the center from the current iteration. One can apply k-means on either a distance matrix (which one can calculate for either PDs or PLs) or on a set of input vectors (such as those obtained from a PI).

2.5.2. Community detection for persistence-landscape classification Community detection is a method from network analysis that attempts to partition a network into sets (called "communities") of nodes that are more densely connected to themselves than to other sets of networks [36, 79, 80]. One can detect communities in either weighted or unweighted networks. In a weighted network, one finds larger total edge weight within communities than between them.

One can also use community detection to partition data (e.g., for classification) by studying a given distance matrix of data objects such as (mean) PLs. One interprets the n PLs as n nodes of a network and converts the pairwise distances into edge weights, where a large edge weight signifies closeness in the distance matrix and a small edge weight signifies a long distance between two landscapes. We convert the distance $d(i, j)$ between landscapes i and j into an edge weight A_{ij} between nodes i and j with the following formula:

$$A_{ij} = 1.01 - \frac{d(i,j)}{\max_{i,j\in\{1,\ldots,n\}}\{d(i,j)\}} \,. \tag{4}$$

This yields an adjacency matrix A with elements A_{ij}. Naturally, there are many choices for converting from pairwise distances to pairwise weights, and one has to be careful about how that influences community structure and other computations.

There are numerous methods that one can use for community detection in networks [80]. One approach for decomposing a network into communities (i.e., for performing a "hard partitioning") is to seek a partition that maximizes an objective function Q. The

quality function that we use is modularity

$$Q = \sum_{i,j} [A_{ij} - \gamma P_{ij}] \delta(g_i, g_j) \,, \tag{5}$$

where P (with elements P_{ij}) is a null-model matrix (which specifies the expected edge weight between nodes i and j), the resolution parameter γ is a factor that determines how much weight one gives to the null model, and $\delta(g_i, g_j) = 1$ if nodes i and j are in the same community g (i.e., if $g_i = g_j$) and $\delta(g_i, g_j) = 0$ otherwise [79, 80].

For our computations, we use the GenLouvain package [81, 82], which maximizes Q using a variant of the Louvain algorithm [83] to algorithmically detect communities in our (mean) PLs. We vary the weighting factor γ (which is often called a "resolution parameter") to compare results for different values of γ.

2.5.3. Linear sparse support vector machines for discriminatory feature selection The 1-norm, regularized, linear support vector machine (i.e., SSVM) classifies data by generating a separating hyperplane between data points that depends on very few input-space features [84–86]. A hyperplane is a flat surface that cuts an ambient space into two parts. One can use an SSVM to identify discriminatory features between different groups of data points. One implements linear SSVM feature selection on data points in the form of vectors, so we can use it on our PIs to select "distinguishing pixels" during classification. In a PI, a *distinguishing pixel* is a bounded region in the birth–persistence coordinate system. For clarity, we use the term "distinguishing pixel' to signify a region selected by SSVM and a "feature" to refer to a topological feature from a PH computation. During the analysis of our results (see Subsection 3.3), we aim to match distinguishing pixels to their corresponding features.

We apply a "one-against-all" (OAA) SSVM to dimension-1 PIs from each subject to identify pixels in PIs that can discriminate between the subject groups. In a one-against-all SSVM, there is one binary SSVM for each class to separate members of that class from members of all other classes. In our case, this amounts to defining three hyperplanes: one that separates patients from controls and siblings, one that separates siblings from patients and controls, and one that separates controls from patients and siblings. We use a 5-fold cross-validated SSVM. One specifies an optimal separating hyperplane by a normal vector. The values of the components of the normal vector are called *SSVM weights*. We select distinguishing pixels for each classifier by retaining the vector components (which are pixels in this application) with nonzero SSVM weights, ordering the nonzero SSVM weights by decreasing magnitude, and discarding SSVM weights when the ratio of successive SSVM weights drops below a user specified tolerance; for details, see [87]. Given a set of distinguishing pixels, we can see for each subject whether there are any loops in the associated functional network that are born and persist in the corresponding bounded PI region. If there are loops in this region, we can identify a set of brain regions that are representative of that loop in the network. We are thereby able to leverage PIs to obtain (biologically) interpretable

information about the involvement of different brain regions in the task (as measured with fMRI) for different subject types.

3. Results

We now present our results of our PH computations to examine loops in functional brain networks. We focus exclusively on topological features in dimension 1 and, except in Subsection 3.4, we perform our computations on all four times regimes as part of one data set rather than separating the data for each time regime. Aside from the aforementioned exception, we run our PH computations on four functional networks per subject. From the PH output, we create either PLs or PIs. We then perform our computations either on (i) the full data set of PLs or PIs of 281 subjects and four time regimes (which gives 1124 landscapes or images, respectively, for the data set) or on (ii) the 12 subject-group means of the landscapes/images (from three subject groups with four time regimes each). We indicate which case we are examining in the relevant subsections. In Subsection 3.4, we consider one full time series for each subject; in other words, we study one functional network per subject.

For both PLs and PIs, we find that there seem to be differences in the topological features of the functional networks between subject groups, although we only observe these for PLs when examining means across groups. To illustrate limitations of the methods, we also discuss results in which we were unable to find differences between subject groups.

3.1. Results of k-means clustering on PLs

Using k-means clustering on mean PLs, we are able to separate siblings of schizophrenia patients from controls and patients. For these calculations, recall that we use all four time regimes in each of the 12 mean landscapes.

We construct mean PLs from the 1D barcodes (i.e., the barcodes that represent loops in the networks) for each time regime and each subject group. We obtain 12 mean landscapes and exclude infinitely-persisting bars, because all of our landscapes include (persistent) infinite features, and these tend to dominate the first several layers of the landscapes. Other researchers have excluded layers of landscapes (e.g., the first twenty) to filter out "topological noise" [88]. Although we threshold our weighted networks prior to analyzing them, this does not necessarily imply that we lose significant information by disregarding the infinite features. Additionally, infinitely persisting features do not necessarily correspond to the most persistent features in the barcodes, as even features that are born in the last filtration steps are infinitely persisting if they do not die over the course of the filtration. In our case, the presence of infinite features prevented us from discriminating between landscapes based on pairwise distances between them. When we considered infinite features separately, we did not observe any noticeable differences between the three subject groups.

We calculate a pairwise L^2 distance matrix of the mean landscapes, and we then perform k-means clustering on the distance matrix (which has 12×12 entries). For $k = 3$, we obtain the expected division of the mean landscapes into patients, controls, and siblings. Although the fact that one can separate the three cohorts based on fMRI data is not a new finding — see, for example, [6, 7, 22–26] for patients versus controls and [89] for patients versus siblings — the novelty of our work is that k-means clustering successfully distinguishes between the three different cohorts based on topological information (in the form of loops) in the functional networks.

We also perform k-means clustering for $k = 2$. Surprisingly, we find that the patients and controls are grouped in one cluster for all time regimes, whereas the siblings are in a separate cluster for all time regimes. We show the mean landscapes and clusters in Fig. 8.

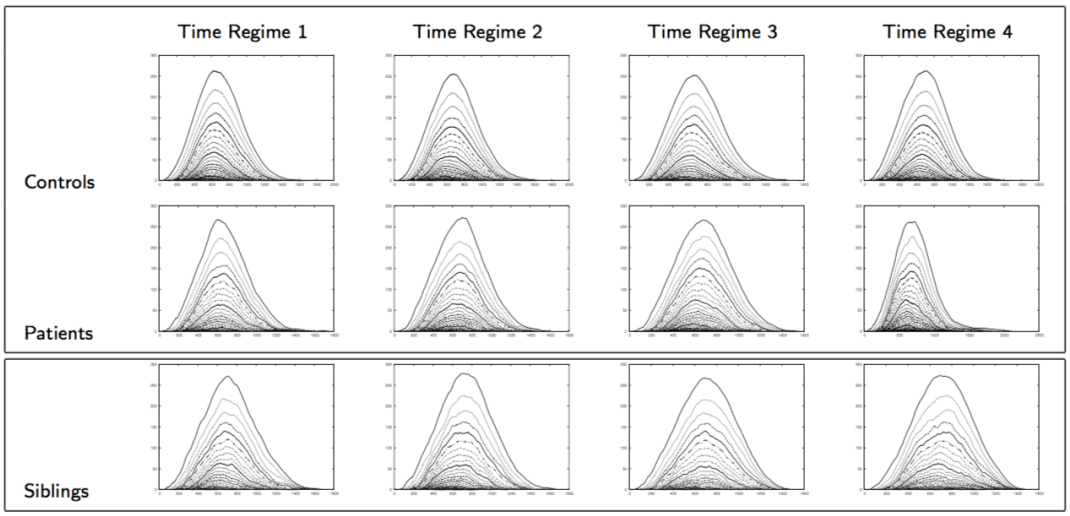


Figure 8: Mean PLs for each of the four time regimes and subject groups. Using k-means clustering with $k = 2$ on the set of 12 PLs (which consists of all subject-group means and time regimes as one data set) assigns patients and controls to one group. We show the mean PLs and their k-means-clustering grouping for the four time regimes separately.

For larger values of k, we do not observe a clear subject-group separation. To compare our results with ones from other clustering methods, we also apply average linkage clustering to the distance matrix and perform community detection on networks that we construct from the distance matrices (as described in Subsection 2.5.2). We obtain the same qualitative result for these two methods. For community detection, we observe a clear separation for resolution-parameter values $\gamma = 0.82, 0.83, \ldots, 1.14$ into two communities (the siblings versus the patients and controls). Our results appear to indicate that the sibling cohort is particularly distinct from the other two cohorts, as compared to any other pairwise comparison among the three cohorts, with respect to their loop topology in the functional networks.

We also perform a permutation test on the mean PLs for each time regime to determine the significance of the landscape distances, as suggested in [68]. In the

permutation test, we regroup the individual landscapes into three groups uniformly at random, create a new mean landscape for each newly assigned group, and calculate the pairwise L^2 distances between them. We then count how many of the L^2 distances of the new groups are larger than the ones that we observe when using the mean landscapes of the three subject groups. We use 10000 permutations to obtain our results, which we summarize in Table 1.

Table 1: Using a permutation test, we calculate p-values for the mean landscape distances between the three subject groups in each time regime.

p-values for	Controls versus Patients	Controls versus Siblings	Patients versus Siblings
time regime 1	0.302	0.200	0.051
time regime 2	0.460	0.009	0.052
time regime 3	0.477	0.102	0.270
time regime 4	0.736	0.110	0.229

Interestingly, for time regimes 1 and 2, we find significant distances between the patient and sibling mean landscapes, whereas the p-values for time regime 3 and 4 suggest that the distance is not significant (even though the p-values are comparably small). The distance between the mean landscapes of the controls and the siblings appears to be significant for time regime 2, but this does not appear to be the case for the other time regimes, although the p-values are again much smaller than for the distances between the mean landscape of the patients and controls. Thus, for the controls and the patients, there are many other divisions into two groups that lead to more extreme distances between the mean landscapes than what one obtains by simply assigning them to a control group and a patient group.

To see if we can further support our result from k-means clustering for $k = 2$, we artificially group the controls and patients into one group to create a mean landscape and again perform a permutation test to verify whether the distance between the mean landscapes for the two groups is significant. In Table 2, we show the p-values that we obtain with 10000 permutations.

Table 2: Using a permutation test, we calculate p-values for the controls-and-patients mean landscape versus the siblings mean landscape.

Time regime 1	Time regime 2	Time regime 3	Time regime 4
0.112	0.008	0.092	0.110

For time regime 2, we obtain a significant distance, but the p-values for time regimes 1, 3, and 4 are about 0.1. Given the artificial grouping of the two subject groups, we construe these values as small, although they are not statistically significant.

3.2. Results of community detection on a distance matrix from individual PLs

We construct PLs from each of the 1D barcodes, which we calculate by examining each subject in each of the four time regimes, and we calculate the L^2 distance matrix for the resulting 1124 PLs. We again use the distance matrix to construct a network between the PLs, and we detect communities in this network by maximizing modularity. For $\gamma = 0.92, 0.93, \ldots, 1$, we obtain a separation into two communities. The partition that is closest to what we observe with 2-means clustering for the mean landscape distance occurs for the resolution-parameter value $\gamma = 0.93$. We summarize our results in Table 3.

Table 3: Number of subjects from each subject group that are assigned to communities 1 and 2 by community detection using modularity maximization.

Subject group	Number of subjects in community 1	Number of subjects in community 2
Patients	122	94
Controls	418	290
Siblings	93	107

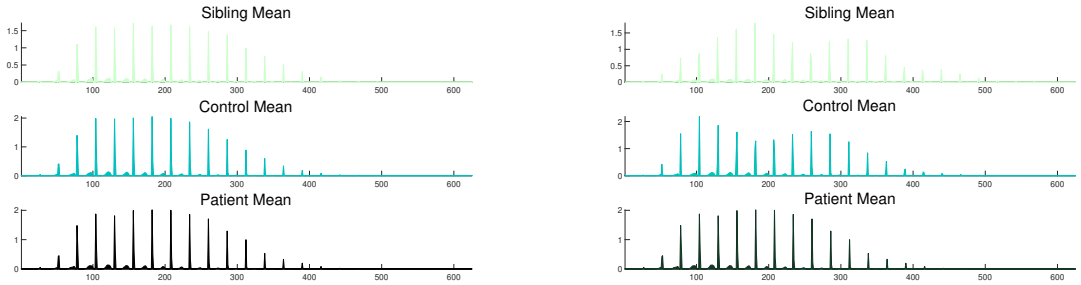
We also apply k-means clustering and average linkage clustering to the distance matrix from the individual PLs (results not shown). Of all classification methods that we perform on these distance matrices, community detection appears to perform best at "separating" the subject groups, although we do not observe a very clear separation.

3.3. Results from analysis of PIs

We find that PIs can identify discriminatory topological features across the three subject groups considered. We generate PIs for each of the subjects for each of the four time regimes for the 1D persistence diagrams. We set the resolution, probability density function, and weighting function to the defaults in the PI code available from [74].

However, there is an additional pair of values that one must choose from the data that one is analyzing using PIs. This pair of values are the maximum birth and persistence values; they determine the discretization of the pixel boundaries in the images once one sets the resolution. Possibilities include taking the maximum birth and persistence values across all PDs or normalizing each PD relative to its individual maximum. In the original paper on PIs [69], the maximum values were chosen across all PDs under consideration. although no theoretical rationale was provided for this choice. We were unable to obtain clear results using either of these conventions. For example, in the left image of Fig. 9, we see the mean vectorized PI for each subject group when we generate the PIs using the maximum birth time and persistence across all subjects. (We create the mean vectorized PI for each subject group by taking the mean of each vector entry.) Observe that, other than a slight amplitude variation, the means look very similar. These mean PIs are the mean of the vectorized PIs for all samples from

each group. The top row of Fig. 10 contains the mean PIs in image form when one selects the maximum birth and persistence across all subjects.



(a) **(b)**

Figure 9: (a) Mean vectorized PI — the horizontal axis corresponds to individual pixels in the PIs, and the vertical axis indicates their intensity values — for each subject group generated using the maximum values of birth and persistence across all subjects to create all PIs. We then take the means over the PIs of each group. (b) Mean vectorized PI for each subject group generated using maximum values of birth and persistence determined by calculating the maximum birth and persistence for each of the three groups separately and using this group-specific information to create the PIs for each subject within its group. We then take the means over the PIs of each group.

Alternatively, if we use a priori knowledge of subject-group membership and fix the maximum birth values separately for each subject group (based on the collection of PDs that we computed separately for each subject group), we can discriminate between the three subject groups. This provides a first interesting observation from PIs: the maximum birth time which corresponds (or almost corresponds, in exceptional

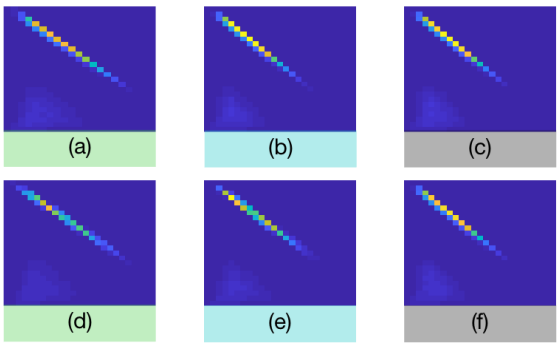
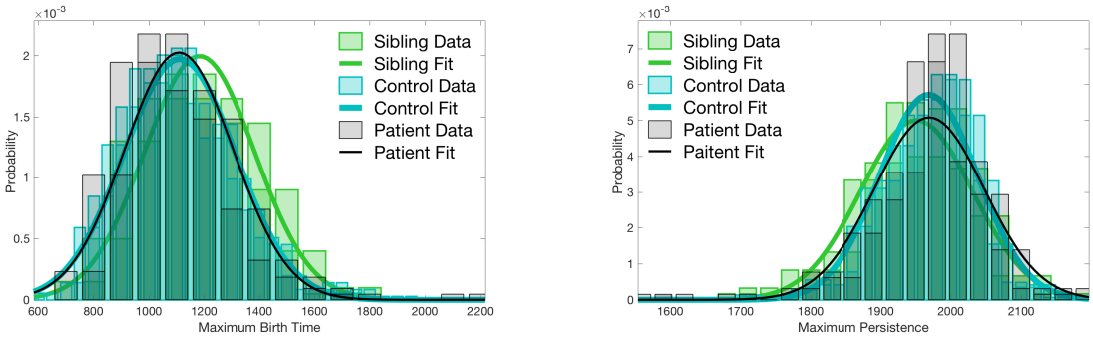


Figure 10: Mean PI for each subject group. We generate panels (a)–(c) using the maximum values of birth and persistence determined across all subjects to compute all PIs before creating the depicted means over the PIs in each subject group. We generate panels (d)–(f) using maximum values of birth and persistence determined using maximum birth and persistence based on subject-group membership to compute all PIs within each group before creating the depicted means over the PIs in each group. The color axis is the same across rows. From left to right across each row are sibling, control, and patient averages.

cases in which multiple edges have exactly the same weight) to the number of pairs of regions in the brain with positive functional connectivity, appears to contain nontrivial information. (Recall that we do not add edges that correspond to negative Pearson correlations.) On the right of Fig. 9, we show the mean vectorized PI for each subject group, where we set the maximum birth and persistence values separately for each subject group (instead of setting the maximum birth and persistence values to be the same for all subjects). Observe that the sibling and control means both have two humps, whereas the patients have one that is clearly discernible. Similarly, in Fig. 10, we observe two patches along the prominent diagonal with high intensity for the sibling and control means; however, in the bottom row, we only observe one clear (and elongated) hot spot for the patient mean. Therefore, there are multiple, smaller regions where loops often occur in the filtrations of the functional networks of siblings and controls, whereas there is seemingly a single, larger region of loops in the filtrations of the networks of the patients.

It is also worth noting the locations of the local maxima for each subject type. Relative to the maximum values across each class, groupings of loops occur at different locations. From the values of the vectors, we see that the controls and patients have more similar maximum magnitudes than do the patients and their siblings. Based on these similarities and differences, we conclude that we are able to accurately separate the populations using PIs. Surprisingly, despite the pronounced difference in PI performance when we use different maximum values for each class, the distributions of the maximum birth times and persistences for each subject type are not statistically-significantly different from each other. In Fig. 11, we show Gaussian fits to the set of maximum birth times and maximum persistences for each subject type. Observe the strong similarity across all classes and the especially close similarity between the control and patient distributions. Because the maximum values are linked closely to the preprocessing of the data, it is important to conduct further research into how to account for these observations. The results that we discuss subsequently are based on the PIs that we generate using a priori membership knowledge.



(a) **(b)**

Figure 11: The distribution of (a) the maximum birth times across all samples for each subject type and (b) the maximum persistences across all samples for each subject type.

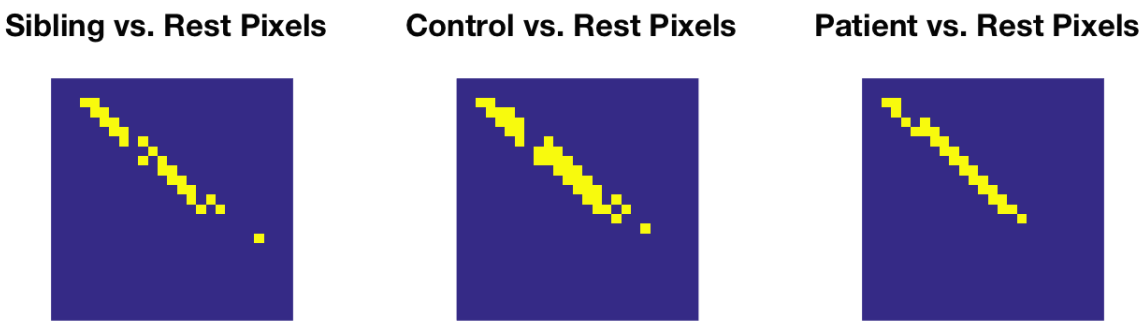


Figure 12: The set of distinguishing pixels determined via SSVM as critical for obtaining 100% classification accuracy on the testing set.

As we discussed in Subsection 2.5.3, it is possible to apply a linear SSVM to the set of PIs to identify distinguishing pixels that allow interpretation of our classification results. Using a one-against-all SSVM with 5-fold cross validation, we obtain a 100% classification accuracy. In Fig. 12, we show the distinguishing pixels from each of the three binary classifiers. We obtain the complete accuracy using the set of 41 unique pixels from the total of 625 pixels in the PIs. Again, we emphasize that each of these pixels corresponds to a bounded region in the birth–persistence plane.

Interpreting these distinguishing pixels requires discussing relations with particular regions in the brain. We make these connection as follows: For each subject, it is possible to determine whether or not a topological feature (in our case, a loop) in a filtration of a network exists in the bounded region of the birth–persistence plane that corresponds to a particular distinguishing pixel. If a loop does exist, one can identify a set of brain regions that comprise the loop (i.e., representatives of this loop; see Section 4). We are particularly interested in brain regions that are consistently involved in the generation of particular loops across subjects. We identify the set of nodes, which we call *top node(s)*††, that are involved in the generation of loop(s) for each distinguishing pixel in each of the four of the time regimes for each subject. We then create histograms of the union of the nodes that we select in this fashion to examine the relative importances of top nodes across each subject type.

In Fig. 13, we present the relative importances of different brain regions for each pixel. In the left panel, we show the top nodes for each subject type based on frequency (proportion of the subject type for which that top node is involved in the generation of a loop in the distinguishing-pixel region). In the right panel, we show the proportion of the subjects for which the top node(s) is (are) present. The vertical gaps in each plot signify that there are no nodes that are consistently involved in loops for that distinguishing pixel. We can make several observations from the left panel of Fig. 13. First, there are only five distinguishing pixels for which we find top nodes for the patients. We

†† One can construe our calculation of top nodes in a similar spirit as calculations of node centralities [36].

are thus unable to predict which brain regions are involved in loops in the functional networks during the given task for schizophrenia patients. By contrast, there are many distinguishing pixels for which we find top nodes for the siblings. The control group lies between the other two in terms of its number of distinguishing pixels with top nodes, but there are still few top nodes, relative to the number of distinguishing pixels that have top nodes. In Tables 4–7 (see also Figs. A1–A3 of Appendix Appendix A), we indicate which brain regions (as well as their locations) we identify as top nodes. We include only the distinguishing pixels at which top nodes exist within a cohort.

An equivalent way to identify a top node is to calculate the percentage of a given subject class that has a topological feature in the corresponding pixel region (see the bar graph in Fig. 13) and determine if a specific node is in the group of representatives for all of the subjects that have a topological feature in the pixel region. If a node occurs in the list of representatives for a topological feature for every subject of the class with a topological feature in the pixel region, then it is identified as a top node. Thus, when considering Tables 4–7, it is possible to see the same brain regions listed for more than one distinguishing pixel index. This is also reflected in Fig. 13 by the occurrence of multiple markers along the same horizontal line.

3.4. Results from Betti curves

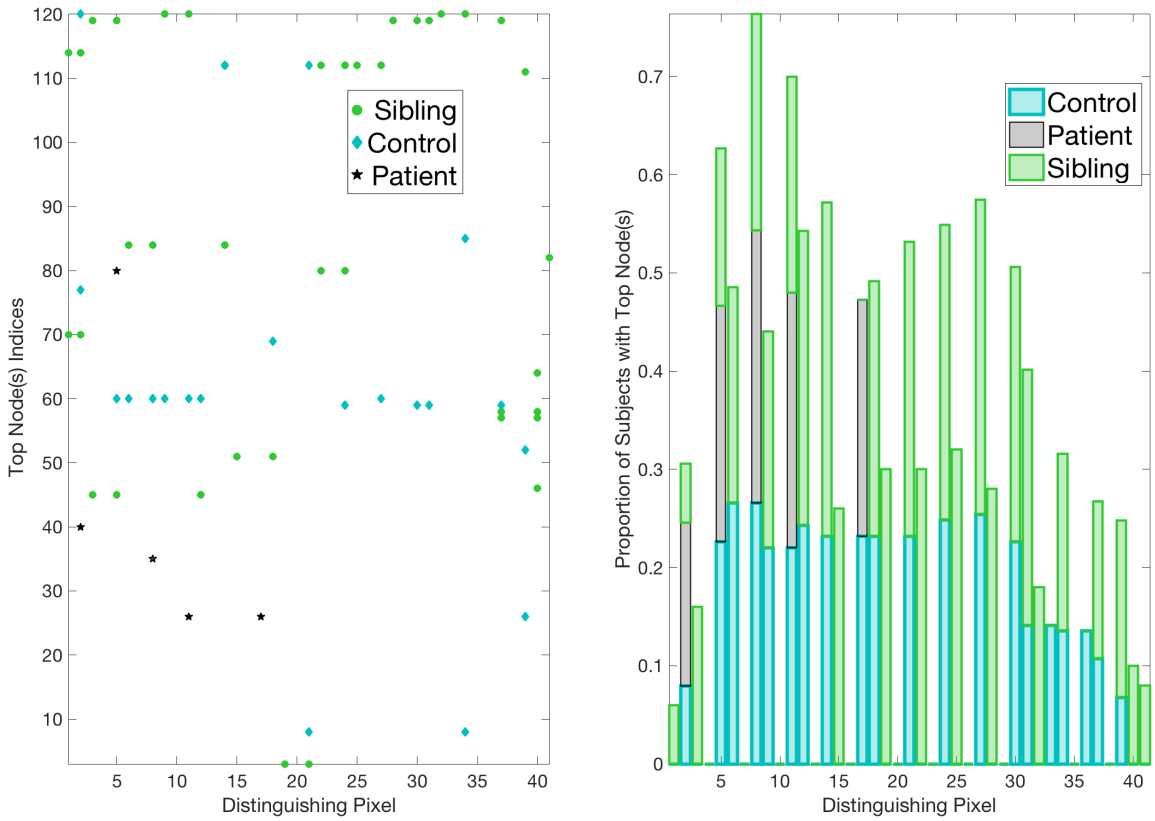
Finally, we also study *Betti curves*, introduced in [50], which describe Betti numbers and their changes across a filtration. We use the entire time period (i.e., one time regime, rather than four separate ones) of the experiment. In all other respects, we construct the functional networks as we described previously (see Subsection 2.2). We compute the mean and standard deviation across the Betti numbers for dimension 1 (i.e., the number of loops) for each cohort in each filtration step. We find that, apart from a slightly larger standard deviation in the patient cohort, the Betti curves for the three groups look essentially the same. We show our results in Fig. A4.

4. Discussion

We applied methods from persistent homology to analyze loops in functional brain networks of schizophrenia patients, siblings of schizophrenia patients, and healthy controls. We constructed both PLs and PIs, and we analyzed them using several clustering techniques.

We observed topological differences in the functional brain networks of schizophrenia patients, siblings of schizophrenia patients, and healthy controls with respect to the loops in their networks. We also found that PLs and PIs have different practical advantages and disadvantages when applied to the same data set, and these insights may useful for interpreting the results of persistent homology computations in networks in diverse applications.

Computing PLs gave interesting results when comparing mean PLs of the cohorts



(a) **(b)**

Figure 13: (a) The index (indices) of the top node(s) associated with each distinguishing pixel that we determine via SSVM. (b) A stacked bar graph of the proportion of each subject type with the corresponding node(s). Pale green indicates siblings, greenish blue indicates controls, and black indicates patients.

but not when comparing individual landscapes of the subjects. Using mean PLs, we were able to separate the sibling cohort from the other two subject groups in each of the four time regimes. This is supported by the p-values that we obtained for the distances between the mean landscapes of the sibling cohort versus the controls and patient cohorts, though not all of our p-values are statistically significant. The shape of the mean PLs seems to suggest that loops that occur in the functional brain networks of siblings are on average more persistent than those in the functional networks of controls or patients. This could imply either that loops in the networks of siblings tend to be longer or that the third edge between three nodes has a small edge weight and thus that three brain regions with a large pairwise Pearson correlation between one region and two of the other regions do not necessarily imply that there is a large correlation between the other two brain regions; this facilitates the creation of a loop structure in the filtration. (Recall that we need at least four nodes for our loops.) To examine

Table 4: Top nodes involved in loop representatives over distinguishing pixel birth–persistence bounds (part I). We include only distinguishing pixels for which there is (are) top node(s) within a cohort. 'Left' and 'Right' refer to the hemispheres of the brain. We use the following abbreviations: superior frontal gyrus medial segment (MSFG), superior temporal gyrus (STG), opercular part of the inferior frontal gyrus (OpIFG), transverse temporal gyrus (TTG), frontal operculum (FO), gyrus rectus (GRe), middle frontal gyrus (MFG), orbital part of the inferior frontal gyrus (OrIFG), precuneus (PCu), cuneus (CC), anterior insula (AIns), superior parietal lobule (SPL), lingual gyrus (LiG), cerebellum exterior (CE), parahippocampal gyrus (PHG), medial frontal cortex (MFC), medial orbital gyrus (MOrG), and posterior cingulate gyrus (PCgG).

Pixel Indices	Siblings		Controls		Patients	
	Node	Location	Node	Location	Node	Location
1	70	Left MSFG	–	–	–	–
	114	Left STG	–	–	–	–
2	70	Left MSFG	77	Left OpIFG	40	Left FO
	114	Left STG	120	Left TTG	–	–
3	45	Right GRe	–	–	–	–
	119	Right TTG	–	–	–	–
5	45	Right GRe	60	Left MFG	80	Left OrIFG
	119	Right TTG	–	–	–	–
6	84	Left PCu	60	Left MFG	–	–
8	84	Left PCu	60	Left MFG	35	Right CC
9	120	Left TTG	60	Left MFG	–	–
11	120	Left TTG	60	Left MFG	26	Left AIns
12	45	Right GRe	60	Left MFG	–	–

this issue further, it may be useful to analyze cross links in the functional networks, as in [90]. For the above computations and their interpretation, we need to take into account that we did not include infinitely-persisting loops (which persist until the end of a filtration). We also only include positive edge weights in our networks, so we only analyzed loops that arise from brain regions with positive pairwise Pearson correlations.

Although we were able to obtain interesting insights about the data using mean PLs, we did not find interpretable results from comparing individual landscapes, and only being able to use the mean landscapes reduces the amount of information the we can obtain from this approach. By contrast, using individual PIs and SSVMs allowed us to separate the entire set of subjects (with 100% accuracy) in each of the four time regimes. In previous work, Anderson and Cohen [91] obtained 65% accuracy for

Table 5: Top nodes involved in loop representatives over distinguishing pixel birth–persistence bounds (part II). We include only distinguishing pixels for which there is (are) top node(s) within a cohort. 'Left' and 'Right' refer to the hemispheres of the brain. We use the following abbreviations: superior frontal gyrus medial segment (MSFG), superior temporal gyrus (STG), opercular part of the inferior frontal gyrus (OpIFG), transverse temporal gyrus (TTG), frontal operculum (FO), gyrus rectus (GRe), middle frontal gyrus (MFG), orbital part of the inferior frontal gyrus (OrIFG), precuneus (PCu), cuneus (CC), anterior insula (AIns), superior parietal lobule (SPL), lingual gyrus (LiG), cerebellum exterior (CE), parahippocampal gyrus (PHG), medial frontal cortex (MFC), medial orbital gyrus (MOrG), and posterior cingulate gyrus (PCgG).

Pixel Indices	Siblings		Controls		Patients	
	Node	Location	Node	Location	Node	Location
14	84	Left PCu	112	Left SPL	–	–
	112	Left SPL	–	–	–	–
15	51	Right LiG	–	–	–	–
17	–	–	112	Left SPL	26	Left AIns
18	51	Right LiG	69	Right MSFG	–	–
19	3	Right Amyg.	–	–	–	–
21	3	Right Amyg.	8	Left CE	–	–
	–	–	112	Left SPL	–	–
22	80	Left OrIFG	–	–	–	–
	112	Left SPL	–	–	–	–
24	80	Left OrIFG	59	Right MFG	–	–
	112	Left SPL	–	–	–	–
25	112	Left SPL	–	–	–	–

schizophrenia classification by applying machine-learning techniques to functional brain networks. A study on Alzheimer's disease using PLs [92] and machine learning attained a 73% separation of diseased and healthy subjects. It is important to note, however, our results are based on using a priori knowledge of group membership, including specifically the maximum birth times of loops within subject groups. These birth times seem to include nontrivial information, which is important to pursue further in future studies. This a priori knowledge is tied closely to the choice of statistical thresholding when preprocessing fMRI data. Developing a statistical model that can classify a novel subject based on a PI representation thus also requires further exploration into how to choose such a threshold.

Computing PIs also allowed us to identify brain regions with consistent involvement

Table 6: Top nodes involved in loop representatives over distinguishing pixel birth–persistence bounds (part III). We include only distinguishing pixels for which there is (are) top node(s) within a cohort. 'Left' and 'Right' refer to the hemispheres of the brain. We use the following abbreviations: superior frontal gyrus medial segment (MSFG), superior temporal gyrus (STG), opercular part of the inferior frontal gyrus (OpIFG), transverse temporal gyrus (TTG), frontal operculum (FO), gyrus rectus (GRe), middle frontal gyrus (MFG), orbital part of the inferior frontal gyrus (OrIFG), precuneus (PCu), cuneus (CC), anterior insula (AIns), superior parietal lobule (SPL), lingual gyrus (LiG), cerebellum exterior (CE), parahippocampal gyrus (PHG), medial frontal cortex (MFC), medial orbital gyrus (MOrG), and posterior cingulate gyrus (PCgG).

Pixel Indices	Siblings		Controls		Patients	
	Node	Location	Node	Location	Node	Location
27	112	Left SPL	60	Left MFG	–	–
28	119	Right TTG	–	–	–	–
30	119	Right TTG	59	Right MFG	–	–
31	119	Right TTG	59	Right MFG	–	–
32	120	Left TTG	–	–	–	–
33	–	–	59	Right MFG	–	–
34	120	Left TTG	8	Left CE	–	–
	–	–	85	Right PHG	–	–
36	–	–	8	Left CE	–	–
	–	–	85	Right PHG	–	–

in loops in the functional networks within subject cohorts. Of the three cohorts, we found that siblings have the highest level of consistent brain-region involvement in the performance of the mental task in this study across the four time regimes. That is, regions that are involved in loops for siblings in one of the time regimes are more likely to also be involved in loops in other time regimes than is the case for patients or controls. It is particularly noteworthy that the number of brain regions that are consistently involved in the separation of the three cohorts is larger in the siblings of schizophrenia patients than in the healthy controls. We view variable involvement of brain regions in loops as a notion of neurological 'flexibility'. Various works have studied concepts of brain flexibility using community structure [63,93]. In those studies, flexibility was defined differently — based on how often a brain region changes its allegiance to a community of nodes over time, so it does not use loops directly — but it is noteworthy that Braun *et al.* [93] observed that relatives of schizophrenia patients have large flexibility than healthy controls. In our work, we found that a specific group of brain regions leads to the separation of the three subject groups when

Table 7: Top nodes involved in loop representatives over distinguishing pixel birth–persistence bounds (part IV). We include only distinguishing pixels for which there is (are) top node(s) within a cohort. 'Left' and 'Right' refer to the hemispheres of the brain. We use the following abbreviations: superior frontal gyrus medial segment (MSFG), superior temporal gyrus (STG), opercular part of the inferior frontal gyrus (OpIFG), transverse temporal gyrus (TTG), frontal operculum (FO), gyrus rectus (GRe), middle frontal gyrus (MFG), orbital part of the inferior frontal gyrus (OrIFG), precuneus (PCu), cuneus (CC), anterior insula (AIns), superior parietal lobule (SPL), lingual gyrus (LiG), cerebellum exterior (CE), parahippocampal gyrus (PHG), medial frontal cortex (MFC), medial orbital gyrus (MOrG), and posterior cingulate gyrus (PCgG).

Pixel Indices	Siblings		Controls		Patients	
	Node	Location	Node	Location	Node	Location
37	57	Right MFC	59	Right MFG	–	–
	58	Left MFC	–	–	–	–
	119	Right TTG	–	–	–	–
39	111	Right SPL	26	Left AIns	–	–
	–	–	52	Left LiG	–	–
40	46	Left GRe	–	–	–	–
	57	Right MFC	–	–	–	–
	58	Left MFC	–	–	–	–
	64	Left MOrG	–	–	–	–
41	82	Left PCgG	–	–	–	–

using PIs and observed for the schizophrenia patients that the regions that lead to a separation consistently in each of the four time regimes are fewer in number than for the siblings and controls. Braun *et al.* [93] reported that there is larger node flexibility in network organization of schizophrenia patients than in healthy controls. Additionally, Siebenhühner *et al.* [94] observed a greater variability in temporal networks constructed from Magnetoencephalography (MEG) data of schizophrenia patients than those from in healthy controls.

We did not observe any differences between the four time regimes, which each consist of responses during a 0-back task and a 2-back task, in any of our calculations. No significant changes seem to be occurring in the persistence or appearance of loops in the networks over the course of the data measurement. Additionally, when studying experiments as a single regime using Betti curves, we did not observe a clear difference between the cohorts.

Schizophrenia has a high genetic determinism, so siblings of schizophrenia patients have a significant genetic risk of developing the disease themselves [95], and it has been demonstrated that they have abnormalities in their structural neuronal

networks [96]. Although our results that functional brain networks constructed from fMRI measurements of siblings differ both from patients and from healthy controls do not agree completely with the current standard in the literature, other studies have also reported that the features of fMRIs of siblings of schizophrenia patients differ from both schizophrenia patients and controls. For example, Callicott *et al.* [97] observed in an fMRI study that there was no difference in task performance between healthy siblings of schizophrenia patients and healthy controls, yet they detected a physiological similarity between the sibling cohort and the schizophrenia patients in the corresponding fMRI data. Similarly, Sepede *et al.* [89] observed using fMRI data from a different data set that healthy siblings of schizophrenia patients exhibit differences in brain function to schizophrenia patients, although they did not differ significantly in task performance.

It was demonstrated recently that schizophrenia patients undergo a cortical normalization process over the course of the disease [98], and a current study on blood samples of schizophrenia patients [99] has also observed that the measurements for patients who have had the disease for a long time are more similar to the measurements of healthy controls than to those of early stage patients. We would need further phenotypic information to assess whether any of the aforementioned studies can be connected more directly to our observations.

As our results are somewhat inconsistent with prior observations, it is also possible that our data set contains experimental noise that is beyond our control. Using standard network-analysis techniques, we do not observe any differences between the three subject groups. Nevertheless, we believe that our comparison of PLs to PIs and the different types of results from these techniques provide a valuable example of a TDA approach to functional brain networks.

To give another cautionary note, one needs to take into account that there are difficulties when interpreting the information about node participation in loops from computations of PH, as the software used for such computations (including, specifically, javaPlex, which is what we used) only finds representatives of the loops. These representatives are not determined optimally, and they need not be 'geometrically nice' [100]. For example, in these calculations, one often encounters double loops or even triple loops as generators for one loop in a functional network. Selecting a basis of homology generators that behaves in a biologically representative way corresponds mathematically to solving a problem known as the "optimal homology-basis problem", which is not trivial and can be NP-hard [101].

Despite these difficulties, our list of discriminating nodes provides a useful starting point for further investigations into neuronal abnormalities in functional networks of schizophrenia patients.

Another important issue is that we preprocessed the data for our study. This is very common when working with fMRI data, but such steps are not uncontroversial, and studies on functional connectivity in schizophrenia patients have found contradictory results depending on whether or not one performed global signal correction [9, 102]. It is also relevant to keep in mind that the choice of functional connectivity measure can

influence results [61]. We chose to use a Pearson correlation due to its simplicity and the fact that it is a widely used measure of functional connectivity [103, 104]. Many other choices are also available.

Acknowledgements

We thank Alessandro Bertolino, Fabio Sambataro, and the Bari psychiatric neuroscience group for permission to study their data. We also thank Pawel Dłotko for his help with the Persistence Landscapes Toolbox and for providing us with new versions of the code during our work. We are also grateful to Danielle Bassett, Peter Bubenik, Carina Curto, and Parker Edwards for helpful comments; and we thank Florian Lipsmeier and Franziska Mech from Roche for useful discussions. We also acknowledge Advanced Research Computing (ARC) at University of Oxford for resources that we used in carrying out this work. BJS thanks the EPSRC and MRC (grant number EP/G037280/1) and F. Hoffmann-La Roche AG for funding her doctoral studies. HAH acknowledges funding from an EPSRC Fellowship (EP/K041096/1) and a Royal Society University Research Fellowship.

[1] World Health Organization 19 September 2015 Schizophrenia http://www.who.int/mental_health/management/schizophrenia/en/
[2] Bertolino A and Blasi G 2009 *Neuroscience* **164** 288–299
[3] Dawson N, Xiao X, McDonald M, Higham D J, Morris B J and Pratt J A 2014 *Cerebral Cortex* **24** 452–464
[4] Bullmore E T, Frangou S and Murray R M 1997 *Schizophrenia Research* **28** 143–156
[5] Peled A, Geva A B, Kremen W S, Blankfeld H M, Esfandiarfard R and Nordahi T E 2001 *International Journal of Neuroscience* **106** 47–61
[6] Bassett D S, Bullmore E T, Verchinski B A, Mattay V S, Weinberger D R and Meyer-Lindenberg A 2008 *The Journal of Neuroscience* **28** 9239–9248
[7] Fornito A, Zalesky A, Pantelis C and Bullmore E T 2012 *NeuroImage* **62** 2296–2314
[8] Zalesky A, Fornito A, Egan G F, Pantelis C and Bullmore E T 2012 *Human Brain Mapping* **33** 2535–2549
[9] Fornito A and Bullmore E T 2015 *Current Opinion in Neurobiology* **30** 44–50
[10] Bullmore E T and Sporns O 2009 *Nature Reviews* **10** 186–198
[11] Bullmore E T and Bassett D 2011 *Annual Review of Clinical Psychology* **7** 113–140
[12] Sporns O 2014 *Nature Reviews Neuroscience* **17** 652–660
[13] Papo D, Zanin M, Pineda-Pardo J A, Boccaletti S and Buldú J M 2014 *Philosophical Transactions of the Royal Society B* **369** 20130525
[14] Papo D, Buldú J M, Boccaletti S and Bullmore E T 2014 *Philosophical Transactions of the Royal Society B* **369** 20130520
[15] Betzel R F and Bassett D S 2017 *NeuroImage* **160** 73–83 ISSN 1053-8119 functional Architecture of the Brain
[16] Bassett D S and Sporns O 2017 *Nature Neuroscience* **20** 353–364
[17] Bassett D S, Zurn P and Gold J I 2018 *Nature Reviews Neuroscience* Available at https://doi.org/10.1038/s41583-018-0038-8
[18] Sporns O 2015 Graph-theoretical analysis of brain networks *Brain Mapping: An Encyclopedic Reference* vol 1 ed Toga A W (Cambridge, Massachusetts: Academic Press: Elsevier) pp 629–633
[19] Petersen S E and Sporns O 2015 *Neuron* **88** 207–219
[20] Stolz B J, Harrington H A and Porter M A 2017 *Chaos* **27** 047410

[21] Eklund A, Nichols T E and Knutsson H 2016 *Proceedings of the National Academy of Sciences of the United States of America* **113** 7900–7905
[22] Lynall M E, Bassett D S, Kerwin R, McKenna P J and Kitzbichler M 2010 *The Journal of Neuroscience* **30** 9477–9487
[23] Rubinov M and Bullmore E T 2013 *Dialogues in Clinical Research* **15** 339–349
[24] Alexander-Bloch A F, Lambiotte R, Roberts B, Giedd J, Gogtay N and Bullmore E T 2012 *NeuroImage* **15** 3889–3900
[25] Liu Y, Linag M, Zhou Y, He Y, Hao Y, Song M, Yu C, Liu H, Liu Z and Jiang T 2008 *Brain* **131** 945–961
[26] Singh M and Bagler G 2016 *arXiv:1602.01191*
[27] Flanagan R, Lacasa L, Towlson E K, Lee S H and Porter M A 2019 *Journal of Complex Networks* Advanced access, doi:10.1093/comnet/cnz013
[28] Alexander-Bloch A F, Gogtay N, Meunier D, Birn R, Clasen L, Lalonde F, Lenroot R, Giedd J and Bullmore E T 2010 *Frontiers in Systems Neuroscience* **4** 1–16
[29] Towlson E K, Vértes P E, Müller U and Ahnert S E 2018 Brain networks reveal the effects of antipsychotic drugs on schizophrenia patients and controls arXiv:1806.00128
[30] Edelsbrunner H, Letscher D and Zomorodian A 2002 *Discrete and Computational Geometry* **28** 511–533
[31] Edelsbrunner H and Harer J L 2008 Persistent homology — A survey *Surveys on Discrete and Computational Geometry. Twenty years later* (*Contemporary Mathematics* vol 453) ed Goodman J E, Pach J and Pollak R (American Mathematical Society) pp 257–282
[32] Ghrist R 2008 *Bulletin of the American Mathematical Society* **45** 61–75
[33] Carlsson G 2009 *Bulletin of the American Mathematical Society* **46** 255–308
[34] Edelsbrunner H and Harer J L 2010 *Computational Topology* (Providence R. I.: American Mathematical Society)
[35] Sizemore A E, Phillips-Cremins J E, Ghrist R and Bassett D S 2018 *arXiv:1806.05167*
[36] Newman M E J 2018 *Networks* 2nd ed (Oxford, UK: Oxford University Press)
[37] Bollobás B 1998 *Modern Graph Theory* (Heidelberg, Germany: Springer-Verlag)
[38] Bassett D S, Yang M, Wymbs N F and Grafton S T 2015 *Nature Neuroscience* **18** 744–751
[39] Otter N, Porter M A, Tillmann U, Grindrod P and Harrington H A 2017 *European Physical Journal — Data Science* **6** 1–38
[40] Patania A, Vaccarino F and Petri G 2017 *European Physical Journal — Data Science* **6** 7
[41] Kramár M, Goullet A, Kondic L and Mischaikow K 2013 *Physical Review E* **87** 042207
[42] Taylor D, Klimm F, Harrington H A, Kramár M, Mishchaikow K, Porter M A and Mucha P J 2015 *Nature Communications* **6** 7723
[43] Bhattacharya S, Ghrist R and Kumar V 2015 *IEEE Transactions on Robotics* **31** 578–590
[44] Topaz C M, Ziegelmeier L and Halverson T 2015 *PloS ONE* **10** e0126383
[45] Bendich P, Marron J S, Miller E, Pieloch A and Skwerer S 2016 *Annals of Applied Statistics* **10** 198–218
[46] Curto C and Itskov V 2008 *PLoS Computational Biology* **4** e000205
[47] Dabaghian Y, Mémoli F, Frank L and Carlsson G E 2012 *PLoS Computational Biology* **8** e1002581
[48] Petri G, Scolamiero M, Donato I and Vaccarino F 2013 *PloS ONE* **8** e66505
[49] Lee H, Chung M K, Kang H, Kim B N and Lee D S 2011 Discriminative persistent homology of brain networks *IEEE International Symposium on Biomedical Imaging: From Nano to Macro* pp 841–844
[50] Giusti C, Pastalkova E, Curto C and Itskov V 2015 *Proceedings of the National Academy of Sciences of the United States of America* **112** 13455–13460
[51] Spreemann G, Dunn B, Botnan M B and Baas N A 2015 Using persistent homology to reveal hidden information in neural data arXiv: 1510.06629v1
[52] Curto C 2017 *Bulletin of the American Mathematical Society* **54** 63–78
[53] Giusti C, Ghrist R and Bassett D S 2016 *Journal of Computational Neuroscience* 1–14

[54] Dłotko P, Hess K, Lavi R, Nolte M, Reimann M, Scholamiero M, Turner K, Muller E and Markram H 2016 Topological analysis of the connectome of digital reconstructions of neural microcircuits arXiv: 1601.01580v1
[55] Babichev A and Dabaghian Y 2017 Persistent memories in transient networks *Emergent Complexity from Nonlinearity, in Physics, Engineering and the Life Sciences* (Springer-Verlag) pp 179–188
[56] Croom F H 1978 *Basic Concepts of Algebraic Topology* (Heidelberg, Germany: Springer-Verlag)
[57] Talairach J and Tournoux P 1988 *Co-planar stereotaxic atlas of the human brain. 3-D proportional system: An approach to cerebral imaging* 1st ed (Thieme)
[58] Bertolino A, Taurisano P, Pisciotta N M, Blasi G, Fazio L, Romano R, Gelao B, Biancho L L, Lozupone M, Giorgio A D, Grazia, Sambataro F, Niccoli-Asabella A, Papp A, Ursini G, Sinibaldi L, Popoloizo T, Sadee W and Rubini G 2010 *PloS one* **5** e9348
[59] Sambataro F, Blasi G, Fazio L, Caforio G, Taurisano P, Romano R, Giorgio A D, Gelao B, Biancho L L, Papazacharias A, Popolizio T, Nardini M and Bertolino A 2010 *Neuropsychopharmacology* **35** 904–912
[60] Rampino A, Walker R M, Torrance H S, Anderson S M, Fazio L, Di Giorgio A, Taurisano P, Gelao B, Romano R, Masellis R, Ursini G, Caforio G, Blasi G, Millar J K, Porteous D J, Thomson P A, Bertolino A and Evans K L 2014 *PloS ONE* **9** e99892
[61] Smith S M, Miller K L, Salimi-Khorshidi G, Webster M, Beckmann C F, Nichols T E, Ramsay J D and Woolrich M W 2011 *NeuroImage* **54** 875–891
[62] Zhou D, Thompson W K and Siegle G 2009 *NeuroImage* **47** 1590–1607
[63] Bassett D S, Wymbs N F, Porter M A, Mucha P J, Carlson J M and Grafton S T 2011 *Proceedings of the National Academy of Sciences of the United States of America* **108** 7641–7646
[64] Kosniowski C 1980 *A First Course in Algebraic Topology* (Cambridge, UK: Cambridge University Press)
[65] Lee H, Kang H, Chung M K, Kim B N and Lee D S 2012 Weighted functional brain network modeling via network filtration *NIPS Workshop on Algebraic Topology and Machine Learning*
[66] Petri G, Expert P, Turkheimer F, Carhart-Harris R, Nutt D, Hellyer P J and Vaccarino F 2014 *Journal of the Royal Society Interface* **11** 20140873
[67] Bubenik P 2015 *Journal of Machine Learning Research* **16** 77–102
[68] Bubenik P and Dłotko P 2017 *Journal of Symbolic Computation* **78** 91–114
[69] Adams H, Chepushtanova S, Emerson T, Hanson E, Kirby M, Motta F, Neville R, Peterson C, Shipman P and Ziegelmeier L 2017 *Journal of Machine Learning Research* **18** 218–252
[70] Cohen-Steiner D, Edelsbrunner H and Harer J 2007 *Discrete Computational Geometry* **27** 103–120
[71] Kovacev-Nikolic V, Bubenik P, Nikolic D and Heo G 2016 *Statistical Applications in Genetics and Molecular Biology* **15** 1–27
[72] Dłotko P and Wanner T 2016 *Physica D* **334** 60–81
[73] Liu J Y, Jeng S K and Yang Y H 2016 Applying topological persistence in convolutional neural network for music audio signals arXiv:1608.07373
[74] Adams H, Chepushtanova S, Emerson T, Hanson E, Kirby M, Motta F, Neville R, Peterson C, Shipman P and Ziegelmeier L 2016 Persistence images https://github.com/CSU-TDA/PersistenceImages
[75] Adams H, Tausz A and Vejdemo-Johansson M 2014 JavaPlex: A research software package for persistent (co)homology (2011) *Mathematical Software—ICMS 2014* vol 8592 ed Hong H and Yap C pp 129–136 software available at http://javaplex.github.io/
[76] Bron C and Kerbosch J 1973 *Communications of the ACM* **16** 575–577
[77] Wildmann J 2011 Bron–Kerbosch maximal clique finding algorithm code available at: http://www.mathworks.co.uk/matlabcentral/fileexchange/30413-bron-kerbosch-maximal-clique-finding-algorithm
[78] Gan G, Ma C and Wu J 2007 *Data Clustering: Theory, Algorithms, and Applications*

(Philadelphia, PA: Society for Industrial and Applied Mathematics)
[79] Porter M A, Onnela J P and Mucha P J 2009 *Notices of the American Mathematical Society* **56** 1082–1097, 1164–1166
[80] Fortunato S and Hric D 2016 *Physics Reports* **659** 1–44
[81] Jeub L G S, Bazzi M, Jutla I S and Mucha P J 2011–2016 A generalized Louvain method for community detection implemented in MATLAB, version 2.0 https://github.com/GenLouvain/GenLouvain
[82] Mucha P J, Richardson T, Macon K, Porter M A and Onnela J P 2010 *Science* **328** 876–878
[83] Blondel V D, Guillaume J L, Lambiotte R and Lefebvre E 2008 *J. Stat. Mech: Theory Exp.* **2008** P10008
[84] Bradley P S and Mangasarian O L 1998 Feature selection via concave minimization and support vector machines *Machine Learning Proceedings of the Fifteenth International Conference* ICML 1998 pp 82–90
[85] Zhu J, Rosset S, Hastie T and Tibshirani R 2004 *Advances in Neural Information Processing Systems* **16** 49–56
[86] Zhang L and Zhou W 2010 *Neural Networks* **23** 373–385
[87] Chepushtanova S, Gittins C and Kirby M 2014 Band selection in hyperspectral imagery using sparse support vector machines *Algorithms and Technologies for Multispectral, Hyperspectral, and Ultraspectral Imagery XX* vol 9088 (International Society for Optics and Photonics) p 90881F
[88] Patrangenaru V, Bubenik P, Paige R L and Osborne D 2018 *arXiv preprint arXiv:1804.10255*
[89] Sepede G, Ferretti A, Perrucci M G, Gambi F, Di Donato F, Nuccetelli F, Del Gratta C, Tartaro A, Salerno R M, Ferro F M and Romani G L 2010 *NeuroImage* **49** 1080–1090
[90] Bassett D S, Wymbs N F, Porter M A, Mucha P J and Grafton S T 2014 *Chaos* **24** 013112
[91] Anderson A and Cohen M S 2013 *Frontiers in Human Neuroscience* **7** 1–18
[92] Bubenik P 2016
[93] Braun U, Schäfer A, Bassett D S, Rausch F, Schweiger J, Bilek E, Erk S, Romanczuk-Seiferth N, Grimm O, Haddad L, Otto K, Mohnke S, Heinz A, Zink M, Walter H, Meyer-Lindenberg A and Tost H 2016 *Proceedings of the National Academy of Sciences of the United States of America* **113** 12568–12573
[94] Siebenhühner F, Weiss S A, Coppola R, Weinberger D R and Bassett D S 2013 *PloS ONE* **8** e72351
[95] Bertolino A, Fazio L, Giorgio A D, Blasi G, Romano R, Taurisano P, Caforio G, Sinibaldi L, Ursini G, Popoloizo T, Tirotta E, Papp A, Dallapiccola B, Borrelli E and Sadee W 2009 *The Journal of Neuroscience* **29** 1224–1234
[96] Collin G, Kahn R S, de Reus M A, Cahn W and van den Heuvel M P 2014 *Schizophrenia Bulletin* **40** 438–448
[97] Callicott J H, Egan M F, Mattay V S, Bertolino A, Bone A D, Verchinksi B and Weinberger D R 2003 *American Journal of Psychiatry* **160** 709–719
[98] Guo S, Palaniyappan L, Liddle P F and Feng J 2016 *Psychological Medicine* **46** 2201–2214
[99] Scolamiero M 2016
[100] Adams H and Tausz A 2015 JAVAPLEX tutorial pDF version available online: http://javaplex.googlecode.com/svn/trunk/reports/javaplex_tutorial/javaplex_tutorial.pdf
[101] Erickson J 2012 Combinatorial optimization of cycles and bases *Advances in Applied and Computational Topology* vol 70 ed Zomorodian A (American Mathematical Society) pp 195–228
[102] Fox M D, Zhang D, Snyder A Z and Raichle M E 2009 *Journal of Neurophysiology* **101** 3270–3283
[103] Wang L, Metzak P D, Honer W G and Woodward T S 2010 *The Journal of Neuroscience* **30** 13171–13179
[104] Bassett D S, Nelson B G, Mueller B A, Camchong J and Lim K O 2012 *NeuroImage* **59** 2196–2207

Appendix A. Appendix

We now give some additional details on a few calculations and results.

Appendix A.1. Top brain regions in the distinguishing pixel birth–persistence bounds

In this section, we illustrate the top nodes (i.e., top brain regions) in the distinguishing pixel birth–persistence bounds for the three cohorts. Recall that each pixel in the birth–persistence plane corresponds to a bounded region of the original PD (i.e., in the *birth–death* plane). We show results for siblings in Fig. A1, controls in Fig. A2, and patients in Fig. A3.

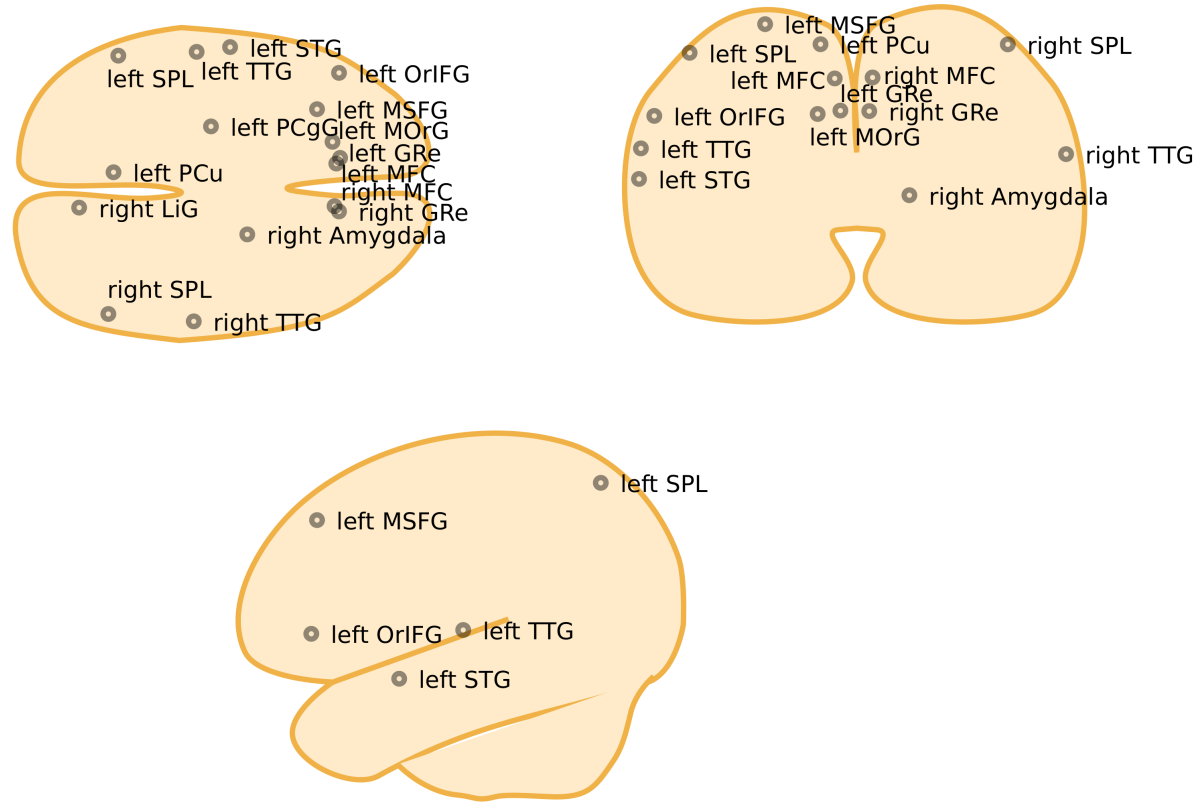


Figure A1: Top nodes in representatives of loops in the distinguishing pixel birth–persistence bounds for siblings.

Appendix A.2. Supplementary Tables

In Tables A1–A5, we give the numbering of the brain regions and their corresponding IDs.

Table A1: Node numbers (NNs) of brain regions (BRs) and their corresponding IDs (part I).

NN	1	2	3	4	5	6
ID	23	30	31	32	36	37
BR	Right accumbens area	Left accumbens area	Right amygdala	Left amygdala	Right caudate	Left caudate
NN	7	8	9	10	11	12
ID	38	39	47	48	55	56
BR	Right cerebellum exterior	Left cerebellum exterior	Right hippocampus	Left hippocampus	Right pallidum	Left pallidum
NN	13	14	15	16	17	18
ID	57	58	59	60	61	62
BR	Right putamen	Left putamen	Right thalamus proper	Left thalamus proper	Right ventral diencephalon	Left ventral diencephalon
NN	19	20	21	22	23	24
ID	71	72	75	76	100	101
BR	Cerebellar vermal lobules I–V	Cerebellar vermal lobules VI–VII	Left basal forebrain	Right basal forebrain	Right anterior cingulate gyrus	Left anterior cingulate gyrus
NN	25	26	27	28	29	30
ID	102	103	104	105	106	107
BR	Right anterior insula	Left anterior insula	Right anterior orbital gyrus	Left anterior orbital gyrus	Right angular gyrus	Left angular gyrus

Table A2: Node numbers (NNs) of brain regions (BRs) and their corresponding IDs (part II).

NN	31	32	33	34	35	36
ID	108	109	112	113	114	115
BR	Right calcarine cortex	Left calcarine cortex	Right central operculum	Left central operculum	Right cuneus	Left cuneus
NN	37	38	39	40	41	42
ID	116	117	118	119	120	121
BR	Right entorhinal area	Left entorhinal area	Right frontal operculum	Left frontal operculum	Right frontal pole	Left frontal pole
NN	43	44	45	46	47	48
ID	122	123	124	125	128	129
BR	Right fusiform gyrus	Left fusiform gyrus	Right gyrus rectus	Left gyrus rectus	Right inferior occipital gyrus	Left inferior occipital gyrus
NN	49	50	51	52	53	54
ID	132	133	134	135	136	137
BR	Right inferior temporal gyrus	Left inferior temporal gyrus	Right lingual gyrus	Left lingual gyrus	Right lateral orbital gyrus	Left lateral orbital gyrus

Table A3: Node numbers (NNs) of brain regions (BRs) and their corresponding IDs (part III).

NN	55	56	57	58	59	60
ID	138	139	140	141	142	143
BR	Right middle cingulate gyrus	Left middle cingulate gyrus	Right medial frontal cortex	Left medial frontal cortex	Right middle frontal gyrus	Left middle frontal gyrus
NN	61	62	63	64	65	66
ID	144	145	146	147	148	149
BR	Right middle occipital gyrus	Left middle occipital gyrus	Right medial orbital gyrus	Left medial orbital gyrus	Right postcentral gyrus medial segment	Left postcentral gyrus medial segment
NN	67	68	69	70	71	72
ID	150	151	152	153	154	155
BR	Right precentral gyrus medial segment	Left precentral gyrus medial segment	Right superior frontal gyrus medial segment	Left superior frontal gyrus medial segment	Right middle temporal gyrus	Left middle temporal gyrus
NN	73	74	75	76	77	78
ID	156	157	160	161	162	163
BR	Right occipital pole	Left occipital pole	Right occipital fusiform gyrus	Left occipital fusiform gyrus	Right opercular part of the inferior frontal gyrus	Left opercular part of the inferior frontal gyrus

Table A4: Node numbers (NNs) of brain regions (BRs) and their corresponding IDs (part IV).

NN	79	80	81	82	83	84
ID	164	165	166	167	168	169
BR	Right orbital part of the inferior frontal gyrus	Left orbital part of the inferior frontal gyrus	Right posterior cingulate gyrus	Left posterior cingulate gyrus	Right precuneus	Left precuneus
NN	85	86	87	88	89	90
ID	170	171	172	173	174	175
BR	Right parahippocampal gyrus	Left parahippocampal gyrus	Right posterior insula	Left posterior insula	Right parietal operculum	Left parietal operculum
NN	91	92	93	94	95	96
ID	176	177	178	179	180	181
BR	Right postcentral gyrus	Left postcentral gyrus	Right posterior orbital gyrus	Left posterior orbital gyrus	Right planum polare	Left planum polare
NN	97	98	99	100	101	102
ID	182	183	184	185	186	187
BR	Right precentral gyrus	Left precentral gyrus	Right planum temporale	Left planum temporale	Right subcallosal area	Left subcallosal area

Table A5: Node numbers (NNs) of brain regions (BRs) and their corresponding IDs (part V).

NN	103	104	105	106	107	108
ID	190	191	192	193	194	195
BR	Right superior frontal gyrus	Left superior frontal gyrus	Right supplementary motor cortex	Left supplementary motor cortex	Right supramarginal gyrus	Left supramarginal gyrus
NN	109	110	111	112	113	114
ID	196	197	198	199	200	201
BR	Right superior occipital gyrus	Left superior occipital gyrus	Right superior parietal lobule	Left superior parietal lobule	Right superior temporal gyrus	Left superior temporal gyrus
NN	115	116	117	118	119	120
ID	202	203	204	205	206	207
BR	Right temporal pole	Left temporal pole	Right triangular part of the inferior frontal gyrus	Left triangular part of the inferior frontal gyrus	Right transverse temporal gyrus	Left transverse temporal gyrus

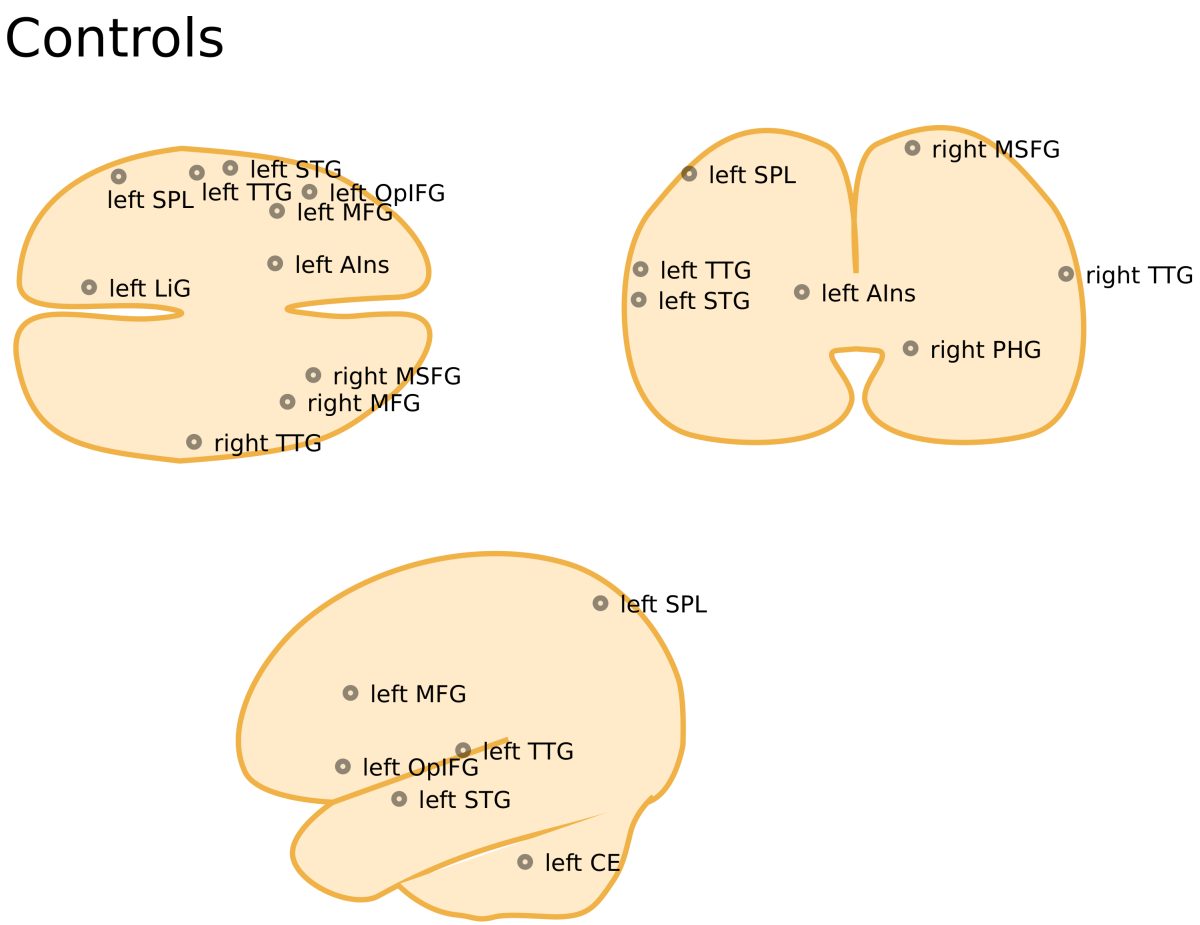
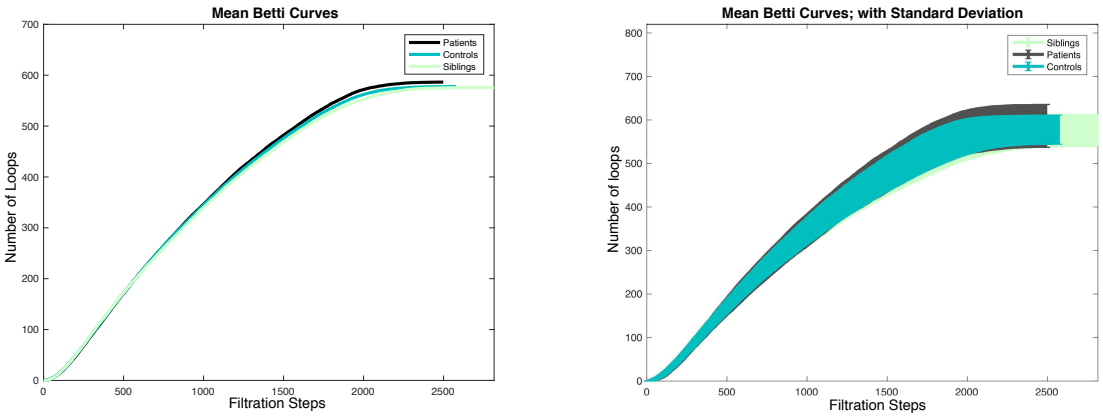


Figure A2: Top nodes in representatives of loops in the distinguishing pixel birth–persistence bounds for controls.

Appendix A.3. Betti curves

In Fig. A4, we show the results of computing Betti curves.

Figure A3: Top nodes in representatives of loops in the distinguishing pixel birth–persistence bounds for patients.



(a) **(b)**

Figure A4: (a) Mean Betti curves for the patients, controls, and siblings. (b) Mean Betti curves and their standard deviations for patients, controls, and siblings.