

Date of publication xxxx 00, 0000, date of current version xxxx 00, 0000.

Digital Object Identifier 10.1109/ACCESS.2024.0429000

Topology assisted clustering of temporal fMRI brain networks with use-case in mitigating non-neural multi-site variability

AHMEDUR RAHMAN SHOVON¹, SIDHARTH KUMAR², and GOPIKRISHNA DESHPANDE³

¹Department of Computer Science, University of Illinois Chicago, Chicago, IL 60607 USA (e-mail: ashov@uic.edu)

²Department of Computer Science, University of Illinois Chicago, Chicago, IL 60607 USA (e-mail: sidharth@uic.edu)

³Auburn University Neuroimaging Center, Department of Electrical and Computer Engineering, Auburn University, Auburn, AL, USA

Department of Psychological Sciences, Auburn University, Auburn, AL, USA

Alabama Advanced Imaging Consortium, Birmingham, AL, USA

Center for Neuroscience, Auburn University, Auburn, AL, USA

Department of Heritage Science and Technology, Indian Institute of Technology Hyderabad, India

Department of Psychiatry, National Institute of Mental Health and Neurosciences, Bangalore, India

Corresponding author: Gopikrishna Deshpande (e-mail: gzd0005@auburn.edu).

ABSTRACT Using temporal analysis of fMRI (functional Magnetic Resonance Imaging) data, we can characterize dynamic changes in brain connectivity over time. However, dynamic temporal analysis of fMRI data is challenging due to the high dimensionality of the datasets. Another fundamental challenge of dynamic temporal analysis of fMRI is the presence of non-neural artifacts that add sources of variation in the data that are not directly related to brain activity. For example, when data are acquired at different scanners at different temporal sampling rates and later analyzed as a single dataset, we have to contend with different number of image snapshots for different subjects. Also, high-frequency scans lead to more fine-grained temporal snapshotting than low-frequency scans. These factors can obscure true neural signals and lead to inconsistent characterization of dynamic brain connectivity across scans. Existing graph-based solutions often struggle with parameter sensitivity, since their outcomes depend heavily on selecting an arbitrary correlation threshold for defining network edges. In contrast, topological data analysis (TDA) sweeps across all threshold values to track the persistence of connectivity features, making it more robust for capturing fine-grained temporal dynamics. Clustering methods become imperative in this context as they offer a powerful means to uncover underlying structures within the high-dimensional temporal data. We address these challenges by developing a topological data analysis based temporal clustering pipeline targeted for dynamic functional connectivity derived from fMRI datasets that can preserve the dynamics of the temporal datasets and mask out the non-neural variability induced by varying sampling rates. The TDA-based pipeline extracts robust features that are invariant to non-neural noise and uses them to perform temporal clustering.

We evaluate our framework by performing temporal clustering of resting-state fMRI-derived dynamic functional connectivity brain networks obtained from 316 subjects, each of whom was scanned thrice using different temporal sampling periods. The efficacy of our TDA-based pipeline is compared against three alternative approaches: direct time-series clustering, PCA-based dimensionality reduction and clustering, and a traditional fully connected network analysis pipeline with MDS-based dimensionality reduction. Additionally, we demonstrate that for a majority of cases, the number of clusters remains consistent for the same subjects scanned at different temporal sampling rates—showcasing the greater robustness of our TDA-based pipeline compared to other pipelines. The TDA pipeline achieved higher overlaps (59 %) in optimal cluster numbers across sampling cohorts, as well as higher pairwise similarity (74-77 %) between subjects' cluster solutions. This indicates that incorporating network topology via TDA enables more robust clustering of temporal fMRI datasets despite changes in sampling rates. Furthermore, we validate our method on a clinical dataset (ADHD-200). The TDA-based pipeline successfully captures consistent clustering patterns across different sites and scanning protocols, with higher stability of cluster assignments (> 80% similarity) and better separation of subject-level dynamics compared to existing approaches. This reinforces the method's robustness in multi-site, multi-condition settings. Our results demonstrate that incorporating network topology via TDA significantly enhances the reliability of temporal clustering in fMRI studies, offering a robust framework for studying brain dynamics across heterogeneous acquisition settings.

I. INTRODUCTION

Temporal fMRI (functional Magnetic Resonance Imaging) analysis [1], [2] is a technique used to study brain activity over time, with a focus on capturing changes in neural activity as they unfold during various cognitive tasks or in resting state conditions. Resting-state functional MRI (rs-fMRI), in particular, examines spontaneous fluctuations in brain activity that occur when a subject is at rest and not performing a specific task. This approach aims to identify intrinsic functional connectivity patterns within the brain, which can reveal networks and interactions that are present even in the absence of external stimuli. Temporal resting-state fMRI analysis is crucial for understanding how the brain processes information dynamically and how the interactions between different brain regions change with time, referred to as dynamic functional connectivity (DFC). It has been shown that DFC is very important for characterizing the healthy brain [3], [4], as well as in various brain disorders [5]–[7].

Temporal fMRI data typically has a temporal resolution on the order of hundreds of milliseconds to a couple of seconds, with each time point representing a snapshot of brain activity at a particular moment during the scan. DFC calculated from rs-fMRI data sampled at different rates tends to capture underlying dynamics evolving at different scales. Also, rs-fMRI data with higher sampling frequency tends to have more time points compared to that sampled at lower frequency for the same experimental duration. More time points tend to increase the robustness of DFC estimates. Given these factors, DFC calculated from rs-fMRI data sampled at different frequencies are not comparable. This is problematic since different groups tend to have different sampling rates owing to other factors, such as the capabilities of the MRI scanner as well as SNR available for a given field strength and field of view. Given the ever-increasing demands on sample size, pooling data acquired from multiple sites has become a priority, but that seems impractical as data across sites is not acquired using the same sampling frequency. Here, we seek to address this problem. Our reasoning is that a video captured at two different frame rates must convey the same content, even though they might not be of the same quality. Our quest is to devise a method that would capture the underlying content correctly, irrespective of how fast or slow the temporal dynamics are sampled. Obviously, DFC is not doing this job. Instead, we propose to use topological data analysis (TDA), which has been shown to characterize the structure underlying the data, which may be invariant to how the data is acquired [8].

Within the domain of functional brain imaging research, there is a growing trend towards the adoption of topological data analysis (TDA), an algebraic topology-based mathematical approach [9]. More recently, topological data analysis tools such as persistent homology (PH) have been utilized to study complex networks [10] including fMRI network dynamics [11]–[14]. The persistence of topological features over a range of spatial scales provides insight into the robustness and stability of network architecture. Persistent homol-

ogy generates topological barcodes that serve as a quantitative fingerprint for complex networks. Specifically, persistent homology tracks the emergence and disappearance of topological features such as connected components, loops, and higher-dimensional cavities across a range of thresholds applied to the network. Importantly, statistical distances between barcodes, such as the Wasserstein distance(WD) [15], [16], can be computed to quantitatively compare network topologies robustly. Thus, persistent homology barcodes provide a distinctive topological signature and metric for interrogating complex network architecture over the traditional graph-based tools [8]. Recent studies have further extended the use of TDA to unsupervised learning and clustering applications. For example, TDA-based clustering of functional brain networks has been successfully applied to Alzheimer's disease cohorts, revealing significant associations between functional topology and brain morphometry using Wasserstein distance kernels [17]. In task-based fMRI, TDA pipelines have outperformed traditional vectorization methods in classifying condition-specific activity, highlighting their ability to capture individualized functional profiles [18]. More broadly, TDA-based classifiers have shown promising results across domains like trajectory classification and imbalanced multi-class datasets, underscoring the versatility and power of topological representations in complex data settings [19], [20]. We harness this capability to compare the similarities of temporal dynamics of brain networks extracted from different sampling periods. We develop a TDA-based statistical data processing pipeline targeted for temporal fMRI datasets that preserves the temporal dynamics of rs-fMRI datasets with the ability to mask out the non-neural variability induced by varying temporal sampling rates.

We evaluate the efficacy of our framework using three data cohorts, each of which corresponds to rs-fMRI data acquired for a subject at a different temporal sampling rate. The input to our pipeline comprises functional connectivity networks (FCNs) derived from rs-fMRI data acquired from 316 subjects, scanned at three temporal frequencies: $f_{high} = 645ms$, $f_{medium} = 1400ms$, and $f_{low} = 2500ms$. High temporal frequency ($f_{high} = 645ms$) corresponds to fine-grained snapshotting, yielding a total of 754 time-steps; medium temporal frequency ($f_{medium} = 1400ms$) scan yields 336 time-steps; and on the other end of the spectrum is low-frequency scans of $f_{low} = 2500ms$, which yields 86 time-steps. The total number of adjacency matrices (FCNs) used in our study is therefore $371,616 (= 316 \times 754 \text{ for } f_{high} + 316 \times 336 \text{ for } f_{medium} + 316 \times 86 \text{ matrices for } f_{low})$. In essence, our statistical pipeline deals with a complex high-dimensional space, the three cohorts are each 4-dimensional spaces, of resolution $113 \times 113 \times \text{timestep\#} \times \text{subject\#}$. 113×113 corresponds to the spatial resolution of each of the individual FCN (adjacency matrix), capturing the pairwise connectivity strengths among the 113 spatial regions of the brain. The timestep\# is respectively 754, 336, and 86 for the three temporal frequencies f_{high}, f_{medium} , and f_{low} . subject\# is 316. Demonstrating any notion of similarity across such a high-dimensional space

is a challenging task, which is further exacerbated by the fact that the resolution across the three spaces varies (along the *timestep#* dimension). To solve this challenge, we rely on a well-established notion that the resting-state brain typically oscillates between a handful of discrete states. We develop a statistical pipeline that uses persistent homology from TDA to gradually reduce the high-dimensional temporal data to two dimensions – which is then reduced further to a one-dimensional scalar quantity that captures the total number of clusters. The novelty of this paper is twofold: (1) developing a statistical framework based on TDA techniques to effectively reduce the complex high dimensional space into simple 1D space (corresponding to the number of clusters), and (2) demonstrating that the number of temporal clusters of same subjects across the three temporal sampling rates is *indeed* the same, unlike existing methods.

To further evaluate the performance of our TDA pipeline, we developed three alternative statistical data analysis pipelines that utilize the same datasets. These pipelines represent different approaches to data processing and dimensionality reduction that have been used traditionally, allowing for a comprehensive comparison of our TDA pipeline against other established methods.

- 1) **Direct time-series clustering pipeline:** The first pipeline employs a direct clustering across time, where the original time-series data is reshaped into a suitable format for clustering algorithms. This direct approach bypasses the need for dimensionality reduction, retaining the full temporal dynamics of the data.
- 2) **PCA-Based dimensionality reduction and clustering pipeline:** The second pipeline utilizes principal component analysis (PCA) to reduce the dimensionality of the time-series data before applying a clustering algorithm. PCA identifies the principal components, which are linear combinations of the original features that capture the maximum variance in the data. By retaining only the most significant principal components, PCA reduces the dimensionality while preserving the essential information for clustering.
- 3) **Traditional dynamic FCN analysis pipeline with MDS-based dimensionality reduction:** The third pipeline adopts a traditional dynamic FCN (dFCN) analysis approach similar to our TDA pipeline with multidimensional scaling (MDS) for dimensionality reduction. This pipeline calculates the pairwise correlation coefficients between all 113 brain regions at every time point, resulting in a temporally varying correlation matrix. MDS is then applied to the correlation matrix to reduce its dimensionality while preserving the underlying connectivity patterns. This approach retains the inter-subject connectivity information, which differs from the aforementioned pipelines.

To determine the appropriate number of clusters for our TDA and the three nonTDA pipelines, we employ the k-means clustering algorithm in conjunction with the silhouette

criterion. Since the number of clusters is insufficient to effectively capture the similarities between different temporal sampling periods, we conduct cluster distance comparisons both across the data cohorts (f_{high} , f_{medium} , and f_{low}) and between individual data cohort pairs ($f_{high} - f_{medium}$, $f_{low} - f_{medium}$, and $f_{high} - f_{low}$). We further evaluate the efficacy of the TDA pipeline with a real clinical dataset. Detailed descriptions of these pipelines and statistical methods can be found in Section III and Section IV.

Our key contributions include:

- 1) **Novel TDA pipeline for preserving temporal dynamics:** We introduce a novel topological data analysis (TDA) pipeline specifically designed for processing resting state fMRI datasets that preserves the temporal dynamics of functional connectivity networks, enabling more accurate and robust analysis of brain connectivity patterns.
- 2) **Evaluate pipeline accuracy:** We evaluate the performance of our TDA pipeline on a resting state fMRI dataset comprising 371,616 adjacency matrices from 316 subjects at three temporal frequencies. To ensure the efficacy of the TDA-based pipeline, we compare it with three alternative approaches: direct time-series clustering, PCA-based dimensionality reduction and clustering, and traditional fully connected network analysis pipeline with MDS-based dimensionality reduction technique.
- 3) **Robustness to non-neural variability:** Additionally, we investigate the robustness of the TDA pipeline to variations in data acquisition parameters, specifically examining the impact of temporal sampling rates mitigating the influence of non-neural variability on temporal fMRI datasets.
- 4) **Validation on multi-site ADHD dataset:** We further validate our method on the ADHD-200 dataset to test generalizability in a multi-site setting with heterogeneous scanning protocols. The TDA pipeline achieves higher inter-site consistency in cluster structure and better subject-level reproducibility, confirming its applicability to broader clinical and developmental neuroscience settings.
- 5) **Open-source implementation for reproducibility:** We open-source the source code, documentation, and data for all components of our work on GitHub (<https://github.com/harp-lab/TemporalBrainPH/>), ensuring reproducibility and accessibility.

The remainder of this paper proceeds as follows. First, Section II provides background on related work and the progression of topology-based functional connectivity network analysis. Next, Section III introduces our proposed end-to-end topological data analysis pipeline for temporal brain rs-fMRI data, and Section IV describes the nonTDA-based data processing pipelines for comparison. Section V then presents an evaluation of our TDA pipeline by applying it to analyze temporal dynamics in resting-state fMRI datasets. Finally,

Section VI offers a discussion of the results and implications, and Section VII concludes with a summary of our contributions and directions for future work.

II. RELATED WORK

The application of topological data analysis (TDA) in network analysis extends beyond conventional graph theory, harnessing computational topology tools to characterize network or data structure architectures with greater adaptability [21]. TDA-driven technique has also demonstrated encouraging outcomes in modeling transitions among brain states within fMRI datasets [22]. Persistent homology, as an advanced tool of TDA, is being applied to analyze the topological features of data, providing a powerful framework for studying the evolution and persistence of structural patterns across various scales [23]–[26]. The use of persistent homology to analyze functional connectivity networks (FCNs) from resting state fMRI data is on the rise [12]–[14], [27]. Persistent homology can quantify structural changes in time-varying graphs, providing both topological summaries and visualizations to identify temporal patterns and anomalies [28]. Recently, we demonstrated that FCN metrics are statistically similar across varied sampling periods [8] only for TDA and not for traditional pipelines. This suggests persistent homology provides a robust topological representation of FCNs invariant to acquisition parameters, potentially removing noise in multi-site studies and improving group comparison effect sizes. Multi-site fMRI studies, despite their advantages in increasing sample size and generalizability [29], [30], introduce variability across scanners and protocols that can undermine statistical power and validity. Differences in acquisition parameters and processing methods across sites can lead to non-biological variability in functional connectivity metrics, posing challenges in large-scale fMRI research [31]. Furthermore, multi-site designs can potentially impact the measurement of temporal dynamics in fMRI studies.

The study of temporal variability in rs-fMRI data has gained significant attention in recent years, as it provides valuable insights into the dynamic nature of functional brain networks and their relevance to human cognition [32]. Growing evidence suggests that functional brain networks exhibit temporal variability, reflecting the dynamic interplay between different cognitive states, arousal levels, and external stimuli [33]. One of the widely used techniques for characterizing temporal variability in rs-fMRI data is the sliding-window correlation analysis. This approach involves dividing the time series into overlapping temporal windows and estimating FCNs for each window [33], [34]. By tracking the changes in functional connectivity patterns across these windows, the dynamic nature of brain networks can be captured, and their evolution over time can be investigated. However, this method relies on predefined window lengths and may not fully capture the inherent temporal dynamics of the data. To address these challenges, existing techniques for analyzing FCNs derived from rs-fMRI have largely relied on graph theory to compute metrics like clustering coefficient and

node degree [35], [36]. These measures summarize individual weighted networks and enable comparisons between FCNs, either at the level of nodes/links or using whole-network summaries. However, graph-based methods have significant limitations. First, common graph metrics depend heavily on the parcellation scheme used to define network nodes [37]. Second, graph analysis requires binarizing network links based on an arbitrary threshold, discarding valuable weighting information [38]. While weighted graph analysis has been proposed, such measures still vary with network density [39]. Multi-threshold analysis can mitigate this issue but remains constrained by density effects [40].

Principal component analysis (PCA) is another a useful technique for reducing the high dimensionality of rs-fMRI datasets before applying clustering [41]. By transforming the data into a lower dimensional space of uncorrelated principal components, PCA enables more efficient and robust clustering of large-scale rs-fMRI data [42]. Studies have shown PCA preprocessing can remove noise and highlight the most informative spatial and temporal features in rs-fMRI for improved clustering performance [43], [44] compared to direct clustering approaches [45]. Another approach to characterizing temporal variability is through temporal independent component analysis (ICA) [46]. Temporal ICA aims to identify temporally distinct functional modes or networks from the rs-fMRI data, revealing the underlying temporal structure and dynamics of brain activity. The applicability of temporal principal component analysis (PCA) as a preprocessing step for sliding-window spatial independent component analysis (sICA) of rs-fMRI data is evaluated by analyzing the consistency of PCA-retained subspaces across overlapping time windows [47]. While PCA-based dimensionality reduction has proven useful for rs-fMRI clustering, it is limited to capturing linear relationships and may miss critical topological features in complex functional connectivity networks. Advanced network science techniques are needed to overcome these limitations and characterize topology more comprehensively in a manner invariant to connectivity density or regional parcellation. For example, topological data analysis (TDA) provides a powerful framework for analyzing weighted networks across scales using methods like persistent homology. TDA yields topological summaries that are stable to network density variation. Further, TDA can be applied in a resolution-invariant manner using expansive parcellations. By moving beyond graph theory, TDA-based network analysis can uncover robust and informative topological signatures while avoiding common density and parcellation dependencies. More recently, TDA techniques, such as Mapper and persistent homology, have been introduced as promising tools for studying the temporal variability of rs-fMRI data. Mapper offers a low-dimensional representation of the high-dimensional rs-fMRI data while preserving its topological features [22]. This technique can be particularly useful for visualizing and exploring the temporal variability and overall structure of functional connectivity patterns. Persistent homology, on the other hand, captures

the evolution and persistence of topological features (e.g., connected components, loops, voids) across different scales or time points, providing a quantitative measure of the stability and dynamics of functional brain networks [8]. Recently, Chung et al. demonstrated the use of a TDA technique that estimates the state spaces of dynamically changing functional brain networks during resting state by clustering based on the Wasserstein distance metric [48] and proposed a dynamic-TDA framework to distinguish topological patterns of gender-specific brain networks [49]. Furthermore, TDA is well-suited for characterizing changes in network topology over time, providing new avenues for investigating temporal dynamics in fMRI connectivity [50], [51].

III. TOPOLOGICAL DATA ANALYSIS BASED TEMPORAL CLUSTERING PIPELINE

In this section, we present our end-to-end topological data analysis-based clustering pipeline targeted for temporal resting state fMRI datasets that preserve the temporal dynamics of the high dimensional brain networks. This pipeline masks out non-neural variability in temporal fMRI datasets and demonstrates similarity across temporal cohorts acquired at different sampling frequencies. Our TDA pipeline integrally relies on persistent homology, which, unlike traditional graph analytic approaches, permits analyzing a range of thresholds to gather connectivity information from FCNs.

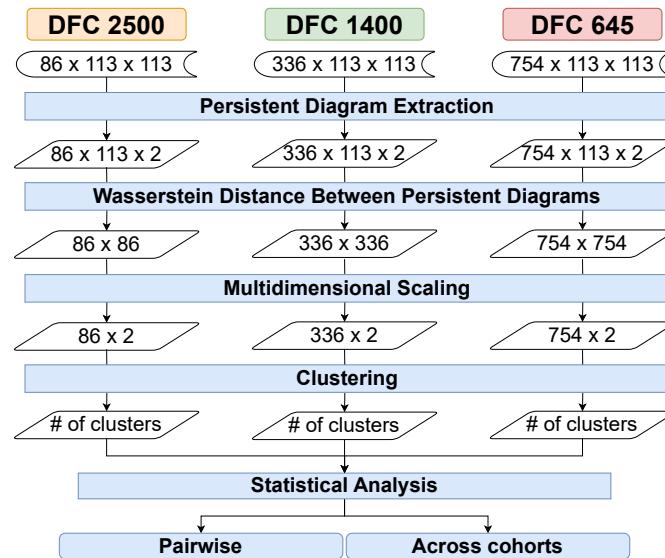


FIGURE 1: Topological data analysis based temporal clustering pipeline for evaluating the robustness of persistent homology applied to dynamic functional connectivity (DFC) matrices calculated from rs-fMRI data acquired using three different sampling periods for the same subjects ($f_{low} = 2500ms$ (left), $f_{medium} = 1400ms$ (center), and $f_{high} = 645ms$ (right)).

Our processing pipeline aims to exploit the commonly known fact that the resting-state brain typically oscillates between a handful of discrete states [52], [53]. This implies that it is potentially possible to group the temporal timesteps

into a discrete number of states. This inherently is a data clustering problem. Therefore, at the heart of our data processing pipeline, we have data clustering across time, which is applied to topological features extracted from the rs-fMRI datasets. A schematic representation of our data processing pipeline can be seen in figure 1. Here we can see the inputs are the three temporal cohorts acquired from sampling 316 subjects at three temporal frequencies ($f_{low} = 2500ms$, $f_{medium} = 1400ms$, and $f_{high} = 645ms$).

A. MATHEMATICAL FOUNDATIONS AND APPLICATION OF PERSISTENT HOMOLOGY TO RESTING-STATE FMRI

Persistent homology (PH) is an algebraic topology method from Topological Data Analysis (TDA) domain used to study qualitative features of data across multiple scales. It operates on the principle of building simplicial complexes, which are geometric constructs consisting of vertices, edges, triangles, and their higher-dimensional counterparts. These complexes evolve over a range of parameters, providing insights into the topology of the underlying dataset.

Formally, for a given set of points (point cloud) X , persistent homology tracks the birth and death of topological features as a parameter ϵ varies. At each threshold ϵ , a Vietoris-Rips complex $\mathcal{R}_\epsilon(X)$ is constructed by connecting points $x_i, x_j \in X$ if the distance between them $d(x_i, x_j)$ is less than or equal to ϵ . As ϵ grows, simplices (vertices, edges, triangles) appear and merge, reflecting changes in topology. Persistent homology captures these changes as persistent barcodes or persistence diagrams, encoding the lifespan (birth to death intervals) of topological features.

In the context of rs-fMRI, PH provides a robust methodology to analyze FCNs, which represent interactions between brain regions as correlation-based adjacency matrices. Direct analysis of these networks can be challenging due to high dimensionality and inherent noise. Persistent homology addresses these challenges by interpreting each brain region as a vertex, forming Vietoris-Rips complexes based on distances derived from correlations. By varying the threshold ϵ , persistent homology captures and quantifies the emergence and merging of connected components, offering robust summaries of complex temporal dynamics.

Our pipeline specifically employs 0-dimensional persistent homology to characterize the evolution of connected components within dynamic FCNs. This approach facilitates robust and interpretable topological signatures suitable for temporal clustering analyses across diverse data acquisition parameters and sampling frequencies.

Our end-to-end TDA pipeline has the following steps:

- 1) **Generate FCNs from rs-fMRI data:** We applied dynamic-windowed Pearson correlation on the fMRI dataset (as used by us before in Jia et al [4]) to generate the dFCNs from data acquired with different acquisition parameters ($f_{high} = 645ms$, $f_{medium} = 1400ms$, and $f_{low} = 2500ms$) as a data pre-processing step (Section III-B). This provides an FCN at each time point.

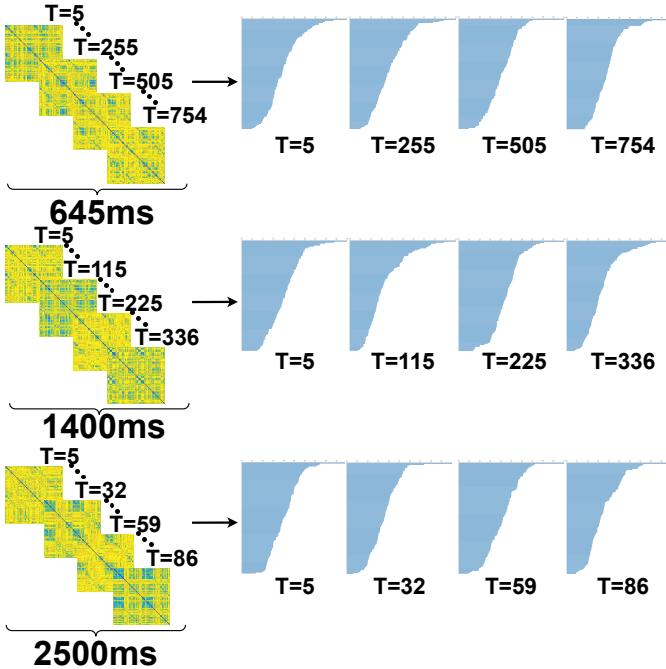


FIGURE 2: The FCN and extracted topological feature as 0-dimensional barcodes for subject 32 at different time points for temporal sampling periods, i.e. high temporal frequency ($f_{high} = 645\text{ms}$) (top), medium temporal frequency ($f_{medium} = 1400\text{ms}$) (center), and low temporal frequency $f_{low} = 2500\text{ms}$ (bottom). The matrix size is 113×113 , corresponding to 113 brain regions.

- 2) **Create distance matrix from FCNs:** The extracted FCNs at each time point are then converted into distance matrices as a weighted graph (Section III-C).
- 3) **Extract persistent barcodes from distance matrix:** In this step, we extract persistent diagrams (0-dimensional barcodes) from the distance matrices using persistent homology to identify the topological features from the matrices (Section III-D).
- 4) **Temporal clustering on PD to prove the resiliency of different data acquisition parameters:** Finally, we apply temporal clustering on the extracted barcodes to see the similarity between the topological features extracted using different data acquisition parameters (Section III-E).

B. FCN GENERATION FROM RS-FMRI DATA

We sourced the structural T1-weighted and rs-fMRI data from the freely accessible Enhanced Nathan Kline Institute Rockland Sample database (NKI-RS) [54]. The MRI data was collected using a 3T Siemens Magnetom Tim Trio scanner. The acquisition parameters for the T1-weighted structural data included: isotropic voxels of 1.0 mm with 176 slices, a repetition time (TR) of 1900 ms, an echo time (TE) of 2.52 ms, and a field of view (FOV) of 250×250 . Resting state fMRI data was gathered using multiband echo-planar imaging

(EPI) [55] from each participant using three distinct acquisition protocols with varying parameters. The first protocol used 3.0 mm isotropic voxels with 40 slices, a TR of 645 ms, a TE of 30 ms, a FOV of 222×222 mm, 900 volumes, and a multi-band factor of 4. The second protocol used 2.0 mm isotropic voxels with 64 slices, a TR of 1400 ms, a TE of 30 ms, a FOV of 224×224 mm, 404 volumes, and a multi-band factor of 4. The third protocol used 2.0 mm isotropic voxels with 38 slices, a TR of 2500 ms, a TE of 30 ms, a FOV of 216×216 mm, 120 volumes, and a multi-band factor of 1. Even though the three datasets from each participant are identified with the corresponding TR, they differ in several other scan parameters, such as the number of volumes, multi-band factor, FOV, and voxel size. The MRI data underwent a standard pre-processing pipeline, which included the removal of the first five volumes, slice time correction, and motion correction. T1-weighted anatomical images were aligned to the mean functional images, which were then spatially registered to a standard MNI152 template. Nuisance variables such as low-frequency drifts, and motion parameters were regressed out. Unwanted physiological fluctuations (signals from white matter and cerebrospinal fluid) were eliminated using aCOMPCor (anatomical component-based noise correction). After excluding subjects that did not pass quality control, we identified 316 subjects with usable data from all three acquisition protocols. We then obtained mean time series from 113 brain regions (using the Yeo parcellation template [56]) for each subject and acquisition protocol. Using Pearson's correlation, we estimated FCN matrices from these mean time series

At this stage, we generate one FCN for each timepoint using dynamic-windowed Pearson's correlation as in our previous work [4]. Briefly, this method uses sliding temporal windows to calculate Pearson's correlation to be assigned to each time point. The width of the temporal window is dynamically determined by the stationarity of the statistical properties of the time series within the window, determined by the Augmented Dickey-Fuller test. Each FCN is stored as a symmetric adjacency matrix M with size 113×113 , where M_{ij} represents the correlation coefficient between brain nodes i and j . The dataset consists of three temporal frequencies ($f_{high} = 645\text{ms}$, $f_{medium} = 1400\text{ms}$, and $f_{low} = 2500\text{ms}$). There are 316 subjects for each temporal frequency. High temporal frequency ($f_{high} = 645\text{ms}$) yields a total of 754 time-steps; medium temporal frequency ($f_{medium} = 1400\text{ms}$) scan yields 336 time-steps; and on the other end of the spectrum is low-frequency scans of $f_{low} = 2500\text{ms}$ yields 86 time-steps. The total number of adjacency matrices is $371,616$ (316×754) + (316×336) + (316×86) with a dimension of 113×113 . Figure 2 (left column) shows an example of the extracted FCNs for subject 32, for the three sampling periods at different time points.

C. CREATION OF MATRICES FROM FCNS

Usually, topological data analysis uses point cloud data in metric configuration. We confine the weighted networks from

fMRI data in distance matrices in our TDA pipeline before applying TDA techniques. Then, we extract persistent barcodes from the distance matrices.

We use popular Pearson's correlation coefficients $pcc(p_t, q_t)$ to measure the linear correlation between any two data points (nodes) p_t and q_t at time t in the fMRI data [57]. However, correlation coefficients do not directly represent distances, which are often required for clustering algorithms or other analyses. Therefore, we transform the correlation coefficients into a distance metric using the following formula:

$$d(p_t, q_t) = \sqrt{1 - pcc(p_t, q_t)^2}$$

This transformation ensures that higher positive correlations (closer to 1) are mapped to smaller distances (closer to 0), while lower correlations (closer to 0) or negative correlations are mapped to larger distances. The temporal indexing t indicates this is a temporally dynamic dataset, and the correlations and distances are calculated between node pairs at each time point.

D. EXTRACTION OF PERSISTENT BARCODES FROM DISTANCE MATRICES

Unlike traditional graph analytic approaches, persistent homology permits analyzing a range of thresholds to gather connectivity information from a given FCN. Figure 3 shows an example of using persistent homology to record the changes of topological features over the changes of distances using zero-dimensional barcodes. Vietoris-Rips filtration on the given 5×5 FCN is applied to capture the changes in the number of connected components for different parameters of d . Using persistent homology, we capture topological features from the distance matrices extracted from the fMRI FCNs. This section overviews topological feature extraction using persistent barcodes and the distance metric we have used. The existing literature contains the details on these topics [25], [58].

1) Extraction of topological features using persistent homology

Persistent homology can extract topological features from a topological space. The homology of the space can be divided into groups based on the dimensions of the features. A topological space \mathbb{X} can be divided into homology groups $H_i(\mathbb{X})$ for $i = 0, 1, 2, \dots$ where $H_i(\mathbb{X})$ represents the i^{th} homology group. Each homology group $H_n(\mathbb{X})$ denotes the number of n -dimensional holes in the topological space \mathbb{X} . For example, the $H_0(\mathbb{X})$ homology group shows the number of connected components, $H_1(\mathbb{X})$ homology group shows the number of holes, $H_2(\mathbb{X})$ homology group shows the number of voids in the topological space \mathbb{X} .

In this paper, we use $H_0(\mathbb{X})$ homology group (0-dimensional) to extract the number of connected components from the rs-fMRI FCNs (at each time point) as topological features. Figure 3 shows a simple example of using persistent homology to extract the topological features from a given

FCN. The table at the bottom right in Figure 3(b) represents the adjacency matrix of an FCN with five nodes. Persistent homology captures the changes in topological features over distance thresholds (ϵ) between the nodes. In a point cloud, F with p nodes, two nodes (x, y) are marked as connected with an edge if the distance $d(x, y)$ is less than threshold ϵ . In this scenario, they form a 1-Simplex. When three nodes are connected with each other for some value of ϵ , they form 2-Simplex and so on. For a given ϵ , the graph is called a Rips complex represented by $Rips(F, \epsilon)$. These continuous changes in the value of ϵ result in the changes in topological features. Vietoris-Rips filtration captures the increasing value of ϵ for which a new Rips-complex, in other words, a new topological feature, is being generated [59]. Figure 3(a) shows the extraction of different topological features in various thresholds of ϵ using Vietoris-Rips filtration.

In this example, for each real number t where topological features are changed, we consider them important events and store these values of t . Here, t represents the filtration value used in the persistent homology analysis, which is derived from the adjacency matrix (bottom right table of Figure 3) representing the FCN. As the filtration value t increases, topological features (such as connected components) can appear, merge, or disappear in the simplicial complex constructed from the data. For instance, at some time t_0 , a topological feature, a component is being created, and at time t_1 , it is merged with another component. We keep track of the *birth* as $t_{\text{birth}} = t_0$ and *death* as $t_{\text{death}} = t_1$ for each component. The time of the $(t_{\text{birth}}, t_{\text{death}})$ of the topological features is visualized as *barcodes*. The span of time for each feature $t_{\text{death}} - t_{\text{birth}}$ is called the persistence of that feature. Figure 3(b)(left) shows the barcodes for the given FCN. At $t = 0$, five topological features are born as five independent (connected) components. At $t = 3.6$, two components are merged; thus, the death of a component is recorded at $t = 3.6$. Therefore, the persistence of that component is 3.6. Similarly, at $t = 6.32$, another component is merged with a persistence of 6.32. For 0-dimensional persistence barcodes, this process continues until there is only one connected component. This last component never dies; thus, the persistence of this component is ∞ . The 0-dimensional persistence barcodes in Figure 3(b)(left) represent the *birth* and *death* of the topological features, which is the changes in the number of connected components of the FCN. Each horizontal bar begins at the birth of a component and ends at the death of each component in the barcodes representation. While higher-order features such as 1-dimensional homology (loops) can in principle be extracted, we restrict our analysis to H_0 for two reasons: (i) computational tractability on large, time-varying FCNs, and (ii) prior evidence that connected components provide stable and interpretable topological features for functional brain networks.

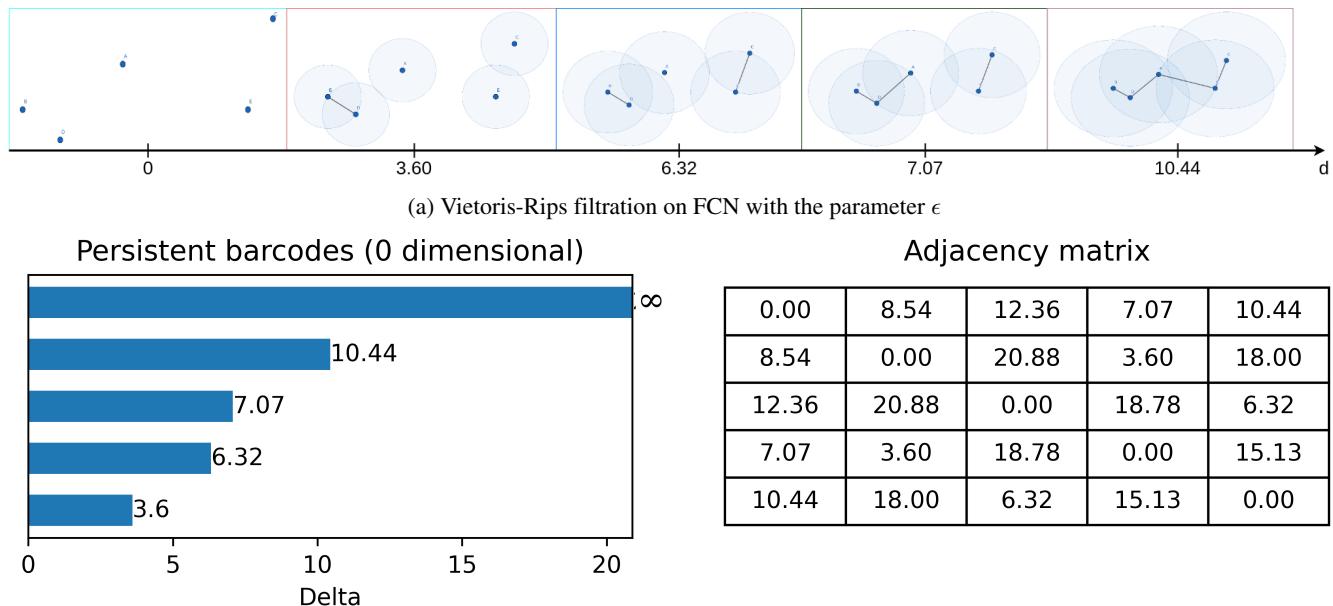


FIGURE 3: An example of persistent homology to extract topological features using 0-dimensional barcodes. The adjacency matrix is of size 5×5 .

2) Formation of persistent diagrams as signatures for the FCNs

The 0-dimensional barcodes extracted from the functional connectivity networks (FCNs) represent the evolution of connected components over distance thresholds. In our pipeline, we generate 0-dimensional persistent barcodes for each FCN. Each FCN has 113 vertices that form the finite set of points F where the value d_i represents the pairwise distance between the points. We extract 0-dimensional barcodes using persistent homology for all the 371,616 FCNs $((316 \times 86) + (316 \times 336) + (316 \times 754))$. Figure 2 displays illustrations of the extracted barcodes at various time points across three different temporal sampling periods ($f_{high} = 645ms$, $f_{medium} = 1400ms$, and $f_{low} = 2500ms$) for a single subject (subject 32). After this stage, we get timestep# \times 113 \times 2 as 0-dimensional barcodes for each of the sampling periods.

A persistent barcode can be represented with a persistent diagram without information loss where the birth and death of a component (a topological feature) are represented as a point on the X-axis and Y-axis, respectively. These points on a two-dimensional surface as a *persistent diagram* can be used for statistical inference to prove that persistent homology is resilient to different data acquisition parameters. By forming the persistent diagrams from the 0-dimensional barcodes of each FCN, we obtain topological signatures that quantify the evolution of connected components over distance thresholds in the functional brain networks. The distance between points in the persistent diagram provides a stability measure, allowing us to compare topological signatures across subjects. Thus, these signatures are used as features for statistical

analysis in our experiments.

E. TEMPORAL CLUSTERING ANALYSIS

The literature shows the usability of earth moving distance, also known as Wasserstein distance (WD), for statistical inference of persistent diagrams [15], [16]. We use WD as a metric to compute the distance between two persistence diagrams extracted from two FCNs. WD represents the minimum value that is computed in the match calculation between the points of two persistent diagrams. The WD value of two similar persistent diagrams is smaller than two dissimilar persistent diagrams. This WD metric assists in proving the hypothesis of similarity between the persistent diagrams extracted from different data acquisition parameters, such as sampling rates.

The temporal clustering analysis is to evaluate our hypothesis of the similarity of the fMRI FCNs obtained from different data acquisition parameters, such as different temporal sampling rates (TR). Our dataset includes three data cohorts: $f_{high} = 645ms$, $f_{medium} = 1400ms$, and $f_{low} = 2500ms$. High temporal frequency ($f_{high} = 645ms$) has a total of 754 time-steps; medium temporal frequency ($f_{medium} = 1400ms$) scan yields 336 time-steps; and low temporal frequency $f_{low} = 2500ms$ yields 86 time-steps for each of the subjects. Figure 1 represents the TDA pipeline we develop for the TDA framework. We calculate pairwise WD between the persistent diagrams of the timesteps for all subjects within the same data cohort.

Let,

$$\begin{aligned} S_i &= \text{Subject } i, \text{ where } i \in \{1, 2, \dots, 316\} \\ T_k &= \text{Number of time steps for data cohort } k, \\ &\quad \text{where } k \in \{754, 336, 86\} \\ P_{S_i,t} &= \text{Persistence diagram for subject } S_i \text{ at time } t, \\ &\quad \text{where } t \in \{1, 2, \dots, T_k\} \\ d_W(P_{S_i,t}, P_{S_j,t}) &= \text{Wasserstein distance between } P_{S_i,t} \text{ and } P_{S_j,t} \end{aligned}$$

The TDA distance matrix for a given subject S_i in cohort k is defined as:

$$D(S_i)_{t_1, t_2} = d_W(P_{S_i, t_1}, P_{S_i, t_2}) \quad (1)$$

for $t_1 \in \{1, 2, \dots, T_k\} \times t_2 \in \{1, 2, \dots, T_k\}$

In Eq. 1, we compute the pairwise Wasserstein distance between persistence diagrams at times t_1 and t_2 , which gives a $T_k \times T_k$ distance matrix for each subject S_i . We have three data cohorts ($f_{high} = 645ms$, $f_{medium} = 1400ms$, and $f_{low} = 2500ms$) with different timesteps (754, 336, 86). Thus, we get 316 adjacency matrices for each data cohort with size (754×754) , (336×336) , and (86×86) , respectively.

These high-dimensional adjacency matrices are complex and cannot be analyzed by statistical methods. To make it interpretable by the statistical methods and for better visualization, we apply the multidimensional scaling (MDS) technique to reduce the dimensionality of the matrices. After this stage, we get 316 adjacency matrices for each data cohort with sizes (754×2) , (336×2) , and (86×2) . The two-dimensional MDS results are plotted using scatter plots that give an intuition to use clustering to calculate the similarity between the TRs. We apply the k-means clustering technique to the MDS results to get the number of clusters for the reduced-sized matrices. As k-means clustering requires configuring the number of clusters n before the cluster computation, we choose n using a well-known approach called Silhouette analysis. Using Silhouette analysis, we choose n that gives the maximum Silhouette score for the given adjacency matrices between the range from 2 to 16 [60], [61]. After calculating the number of clusters for all three data cohorts, we get the number of clusters for all 316 subjects across these data cohorts. A similar number of clusters for the same subject across three data cohorts will indicate the similarity between the subjects for different data acquisition parameters (different TRs in our case). If we get significant similarity between the TRs, we can conclude the resiliency of persistent homology for dynamic functional connectivity derived from rs-fMRI with different data acquisition parameters.

We use Matlab during the data preprocessing steps and Python for statistical analysis. For the topological data analysis framework, we use *Gudhi* library [62], [63] to compute the 0-dimensional barcodes from FCNs and to calculate WD distances between the persistent diagrams. We also use *scikit-learn* library for multidimensional scaling and cluster calculation [60].

IV. NONTDA-BASED DATA PROCESSING PIPELINES FOR RS-FMRI DATASET

To assess the effectiveness of our TDA pipeline presented in Section III, we implement three alternative nonTDA-based data processing pipelines for comparative analysis. Section IV-A provides a detailed walkthrough of a direct time-series clustering pipeline. In Section IV-B, we show a pipeline that employs Principal Component Analysis (PCA) for dimensionality reduction and clustering. Section IV-C outlines a traditional functional connectivity network (FCN) analysis pipeline tailored for brain network datasets. As the datasets are in *mat* format, for data preprocessing, we utilize Matlab to process the raw fMRI time series, handle missing values, and construct functional connectivity networks. Python is used for all subsequent analyses, including dimensionality reduction, clustering, and statistical testing. We use *scikit-learn* library for principal component analysis, multidimensional scaling, and k-means clustering [60].

A. DIRECT TIME-SERIES CLUSTERING PIPELINE

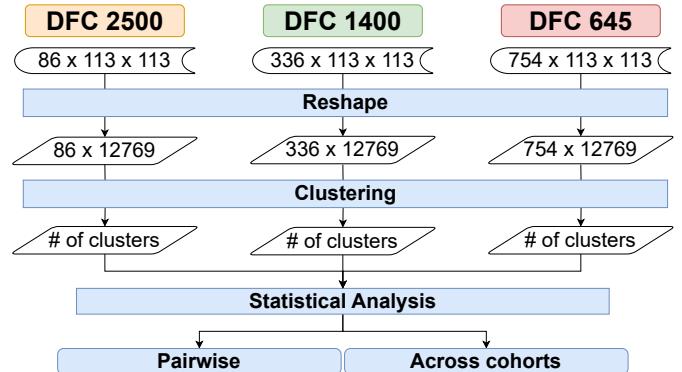


FIGURE 4: Direct time-series clustering pipeline that bypasses dimensionality reduction of the high dimensional temporal datasets.

Our first baseline pipeline is the nonTDA-based direct clustering pipeline. It implements a direct clustering approach on the temporal rs-fMRI datasets without any dimensionality reduction or graph construction steps. As shown in Figure 4 the input of this pipeline is the $timestep\# \times 113 \times 113 \times$ where 113×113 corresponds to the spatial resolution of each of the individual FCN capturing the pairwise connectivity strengths among the 113 spatial regions of the brain. The $timestep\#$ is respectively 86, 336, and 754 for the three temporal frequencies f_{low} , f_{medium} , and f_{high} respectively.

Directly clustering multivariate FCNs poses challenges due to the high dimensionality of the data. To mitigate this, we flatten the 113×113 matrices into one-dimensional arrays of length 12,769, resulting in $timestep\# \times 12,769$ sized matrices for each of the subjects for each data cohort prior to clustering. This reshaping transforms the data into a format amenable to traditional clustering algorithms. Similar to the TDA pipeline, we then apply the k-means clustering technique to reshaped matrices. We utilize the silhouette analysis

method to determine the optimal number of clusters, denoted as n , within the range of 2 to 16, ensuring the selection of the cluster configuration with the highest silhouette score.

After calculating the number of clusters for all three data cohorts, we get the number of clusters for all 316 subjects across these data cohorts. The subsequent step defines the statistical analysis phase, where we perform pairwise and cohort-wide set overlaps of the number of clusters. Similar to the TDA pipeline, the key hypothesis is that robust clustering solutions should exhibit consistency across subjects and sampling rates. To evaluate this, we statistically compare the identified cluster numbers across cohorts using pairwise and group-wise similarity scores. Higher overlaps in the optimal cluster numbers between the three temporal sampling periods and higher pairwise set overlaps will indicate higher similarity scores, thereby affirming the robustness of the direct clustering approach to variability in sampling rates.

B. PCA-BASED DIMENSIONALITY REDUCTION AND CLUSTERING PIPELINE

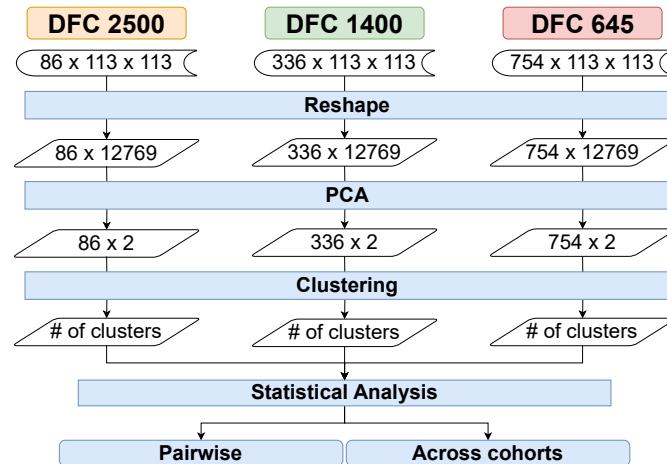


FIGURE 5: PCA-based dimensionality reduction and clustering pipeline where the high dimensional data is reduced to 2 principal components capturing the most variance before applying clustering algorithm.

After establishing the baseline with the direct clustering approach, we introduce a principal component analysis (PCA)-based dimensionality reduction and clustering pipeline to address the high dimensionality challenges in the temporal FCN datasets. Figure 5 displays PCA-based dimensionality reduction and clustering pipeline. This pipeline incorporates a three-step process involving PCA for dimensionality reduction, k-means clustering, followed by statistical analysis.

Like the prior pipelines, this pipelines takes as input matrices of size $timestep\# \times 113 \times 113$. where the $timestep\#$ values are 86, 336, and 754 for the temporal frequencies f_{low} , f_{medium} , and f_{high} , respectively. We flatten the 113×113 functional connectivity matrices into one-dimensional arrays of length 12,769 to reorganize our data into a 2D matrix suitable for PCA, where each row represents a timestep and each column

represents an element of the flattened connectivity matrix. After flattening, the input to PCA would indeed be a 2D matrix with dimensions: for f_{low} : $86 \times 12,769$, for f_{medium} : $336 \times 12,769$, and for f_{high} : $754 \times 12,769$. This reshaping transforms the higher-dimensional tensor data into a 2D matrix format required by PCA to identify the top principal components capturing the most variance in the connectivity patterns across time. Applying PCA on the flattened 12,769 dimensional data then allows us to reduce the high dimensionality down to the most informative 2 principal components before clustering. In this stage, we apply PCA to reduce the dimensionality of the flattened $timestep\# \times 12,769$ matrices. Specifically, we use PCA to project the data onto a lower-dimensional subspace while retaining as much variance as possible. For each temporal frequency ($f_{low}, f_{medium}, f_{high}$), we reduce the dimensions from 86×12769 to 86×2 for f_{low} , 336×12769 to 336×2 for f_{medium} , and 754×12769 to 754×2 for f_{high} . The resulting PCA-transformed matrices, now of size $timestep\# \times 2$, capture the most salient features of the original data. This dimensionality reduction facilitates subsequent clustering by focusing on the most informative components while significantly reducing computational complexity.

Following the PCA-based dimensionality reduction, we apply k-means clustering to identify patterns and group subjects based on the reduced feature space. Similar to the direct clustering approach, we leverage silhouette analysis to determine the optimal number of clusters (n) within the range of 2 to 16 for each temporal frequency that maximizes clustering quality. The resulting cluster assignments provide a compact representation of the original data while capturing meaningful variations across subjects and temporal sampling rates.

To assess the performance of the PCA-based pipeline, we follow a similar statistical analysis phase as in the direct clustering approach. The number of clusters obtained for all subjects across the three temporal frequencies undergoes pairwise and cohort-wide set overlap analysis. This comparison helps evaluate the consistency of clustering solutions across different temporal sampling periods. The hypothesis remains that a robust clustering solution should exhibit coherence in identified clusters across subjects and temporal frequencies. We use pairwise and group-wise similarity scores to statistically compare the optimal cluster numbers. Higher overlaps in cluster assignments between temporal sampling periods and increased pairwise set overlaps indicate greater stability and reliability in the face of variability in sampling rates.

C. TRADITIONAL DFCN CLUSTERING PIPELINE WITH MDS-BASED DIMENSIONALITY REDUCTION

We develop a traditional dynamic FCN (dFCN) analysis pipeline with similar steps to the aforementioned TDA-based pipeline for the rs-fMRI dataset. The traditional dFCN analysis pipeline is illustrated in Figure 6. Analysis steps include extracting subject-specific DFC (dynamic functional connectivity) matrices, calculating graph metrics, dimensionality reduction via MDS, clustering with k-means, and computing cluster overlaps. Instead of the persistent diagrams or apply-

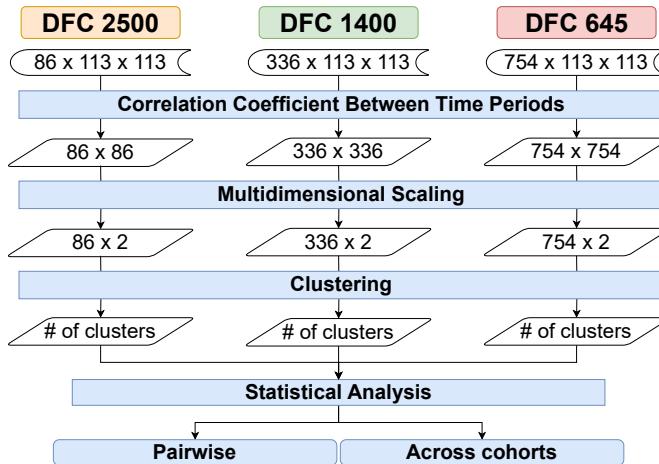


FIGURE 6: Traditional dynamic functional connectivity network (dFCN) clustering pipeline. This non-TDA pipeline has a similar statistical analysis structure to the TDA pipeline, allowing comparison of approaches for assessing the robustness of persistent homology on temporal rs-fMRI data.

ing any persistent homology methods, we use a correlation coefficient between the timesteps for all three data cohorts.

Let,

$$S_i = \text{Subject } i, \text{ where } i \in \{1, 2, \dots, 316\}$$

T_k = Number of time steps for data cohort k ,

$$\text{where } k \in \{f_{high} = 754, f_{medium} = 336, f_{low} = 86\}$$

$A_{S_i,t}$ = Adjacency matrix for subject S_i at time t ,
where $t \in \{1, 2, \dots, T_k\}$

The distance matrix for a given subject S_i in cohort k is defined as:

$$D(S_i)_{t_1, t_2} = \sqrt{\sum_{m=1}^M \sum_{n=1}^N |a_{mn}^{t_1} - a_{mn}^{t_2}|^2} \quad (2)$$

In Eq. 2, M, N are the dimensions of the adjacency matrices. We compute the Euclidean distance [64] between adjacency matrices at times t_1 and t_2 , which gives 316 distance matrices of sizes (86×86) , (336×336) , and (754×754) for the three cohorts respectively. In this stage, we acquire 316 matrices for each of the data cohorts with the size of (86×86) for $f_{low} = 2500ms$, (336×336) for $f_{medium} = 1400ms$, and (754×754) for $f_{high} = 645ms$. Then, we follow a similar pipeline of the TDA framework to keep the comparison uniform. We reduce the dimension of the matrices using two-component multidimensional scaling (MDS) and then calculate the number of clusters (n) on the reduced matrices using k-means clustering. We also use the maximum Silhouette score to choose the value of n within the range from 2 to 16. Finally, this pipeline will also produce the number of clusters of the MDS for every 316 subjects for all three data cohorts.

Our hypothesis will be proven right if the TDA pipeline gives a better similarity score than the non-TDA pipelines.

V. RESULTS

All of the TDA-based and nonTDA-based pipelines start with embedding one FCN for each rs-fMRI scan as an adjacency matrix (section III-B). In this stage, we get 371, 616 adjacency matrices $((316 \times 86) + (316 \times 336) + (316 \times 754))$ for three temporal sampling periods ($f_{high} = 645ms$, $f_{medium} = 1400ms$, and $f_{low} = 2500ms$). Each matrix has a dimension of 113×113 . In the second stage of the pipelines, we embed the FCN using Pearson's correlation coefficients and range the values between 0 and 1. In the TDA-based pipeline, the third stage extracts 0-dimensional persistent barcodes from the matrices using persistent homology (Section III-D). In the nonTDA-based pipelines, instead of using persistent homology, we use correlation coefficients between the timesteps for the three temporal sampling periods (TRs) for traditional FCN analysis(Section IV-C). Both the direct clustering and PCA-based dimensionality and clustering pipeline reshape the input matrices from $timestep\# \times 113 \times 113$ to $timestep\# \times 12,769$ where the later one applies PCA based dimensionality reduction before applying clustering. For all of the TDA-based and nonTDA-based pipelines, we continue to the statistical analysis phase, where we compute the cohort-wide and pairwise cluster intersections of the subjects for all data cohorts.

In the TDA-based pipeline, we use the Wasserstein distance metric on the persistent diagrams for all the subjects for all three temporal sampling periods. On the contrary, in the nonTDA-based traditional FCN analysis pipeline, we directly use the correlation coefficient on the extracted FCNs. Adjacency matrices generated after this stage in these pipelines are similar in size for respective temporal sampling periods. For temporal sampling period $f_{low} = 2500ms$ with 86 timesteps in the TDA-based pipeline, each subject yields adjacency matrix WD of size (86×86) where WD_{ij} represents the pairwise Wasserstein distance between timestep i and j . In the traditional FCN analysis pipeline for the same data cohort, each subject yields adjacency matrix A of size (86×86) where A_{ij} represents the pairwise norm between timestep i and j . Similarly, $f_{medium} = 1400ms$ and $f_{high} = 645ms$ yield adjacency matrix of size (336×336) and (754×754) , respectively, for each of the subjects during TDA and traditional FCN analysis. This high dimensionality of the matrix size makes it challenging to apply statistical analysis. For this reason, we applied multidimensional scaling (MDS) and reduced the size of the matrices to fit into a two-dimensional surface for all the data cohorts(f_{low} : (86×2) , f_{medium} : (336×2) , f_{high} : (754×2)). Then, we applied clustering on the MDS data using the k-means clustering algorithm with Silhouette analysis to select the number of clusters. Finally, we get the number of clusters for all 316 subjects for both of these pipelines.

For the other two nonTDA-based data processing pipelines, we first reshape the input matrices. In the direct time-series clustering pipeline, we flatten the 113×113 matrices into 12,769-dimensional vectors and calculate the optimal num-

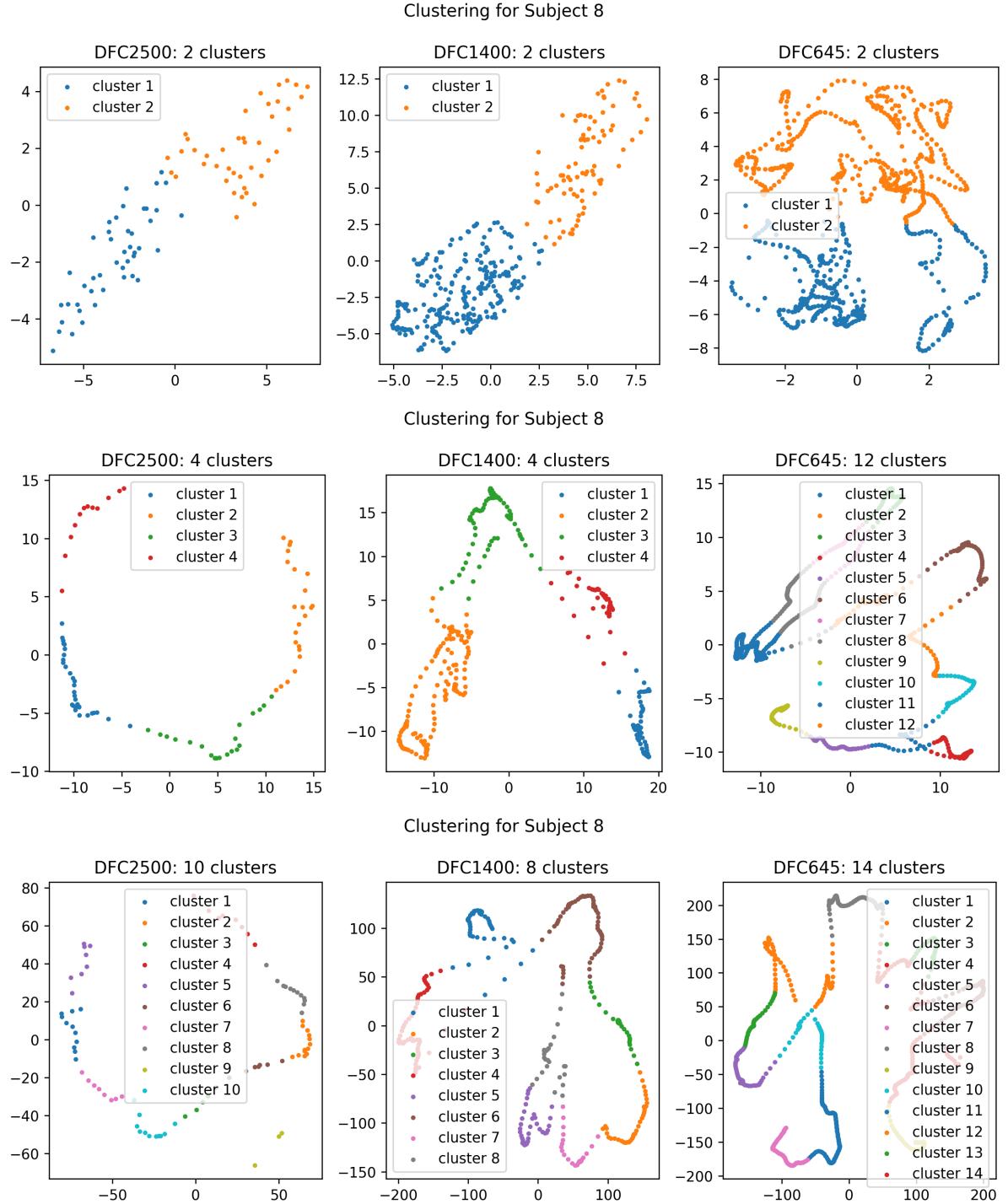


FIGURE 7: Clustering result for subject 8 across different temporal periods: $f_{low} = 2500ms$ (left), $f_{medium} = 1400ms$ (center), $f_{high} = 645ms$ (right) using TDA-based temporal clustering pipeline (top row). Comparative results for the same subject employing nonTDA-based pipelines, including PCA-based dimensionality reduction and clustering (middle row), and the traditional DFCN clustering pipeline (bottom row).

ber of clusters for each data cohort directly on this reshaped high-dimensional data. In contrast, for the PCA-based dimensionality reduction and clustering pipeline, we flatten the FCNs and then apply 2-component PCA to reduce the dimen-

sionality down to 2 principal components before clustering. This PCA step mitigates the challenge of directly clustering high-dimensional data. After PCA reduction to 2D, we determine the optimal cluster numbers for each subject on the

low-dimensional PCA-reduced data. Both pipelines reshape the data as a preprocessing step, but the PCA pipeline has an additional dimensionality reduction phase prior to clustering.

Figure 7 shows the clustering result for a single subject (subject 8) for all three data cohorts ($f_{low} = 2500ms$, $f_{medium} = 1400ms$, $f_{high} = 645ms$). The top row of the figure represents the plotted clusters using the TDA-based pipeline, and we see that each data cohort here has two clusters. The second row shows the clustering result for same subject using the PCA-based dimensionality reduction and clustering pipeline. While the number of clusters remains consistent for f_{low} and f_{medium} in this pipeline, there is a notable discrepancy in the number of clusters for f_{high} . This suggests that the pipeline can not effectively preserve robustness across different data cohorts. The bottom row of the figure shows the plotted clusters for the same subject using the nonTDA-based traditional FCN analysis pipeline, and the number of clusters varies for the data cohorts. As the number of clusters remains unchanged for different temporal sampling periods with similar shape using the TDA-based pipeline and varies largely for the nonTDA-based pipelines, it shows the invariant of the TDA pipeline. We cannot plot the clustering result for direct clustering pipeline due to the elevated size of the data in this clustering analysis ($\#timestep \times 12, 769$), and the absence of any dimensionality reduction techniques (MDS or PCA) applied on the original data. Thus, this illustration gives an intuition towards our hypothesis of the resiliency of persistent homology-based methods to different data acquisition parameters (temporal sampling periods) in brain rs-fMRI data analysis.

To statistically compare the consistency of identified cluster patterns across subjects for both the TDA-based and nonTDA-based pipelines, we compute two similarity metrics - the cohort-wide and pairwise cluster distances. For the cohort-wide analysis, we calculate the absolute difference in number of clusters between each subject's optimal solution. This provides a distribution of cluster number distances indicating the spread/variability across subjects. For pairwise analysis, we compute cluster distances between each pair of data cohorts for the same subject. The pairwise distances are aggregated to produce a distribution showing the overall pairwise consistency. Lower cohort-wide and pairwise distances indicate higher similarity and consistency in optimal cluster numbers. This quantifies the robustness of each method to individual variations and its ability to extract connectivity patterns that are generalizable across populations. The metrics provide crucial insights into the stability and reproducibility of the clustering solutions.

A. COHORT-WIDE CLUSTER DISTANCE COMPARISON

We capture the number of clusters for all 316 subjects for all three data cohorts for all of the pipelines. We calculate the distance between the number of clusters for each subject using:

$$\begin{aligned} \text{distance}(\text{subject}_i) = & \text{abs}(\text{subject}_{i_{2500ms}} - \text{subject}_{i_{1400ms}}) \\ & + \text{abs}(\text{subject}_{i_{1400ms}} - \text{subject}_{i_{645ms}}) \\ & + \text{abs}(\text{subject}_{i_{645ms}} - \text{subject}_{i_{2500ms}}) \end{aligned} \quad (3)$$

where $\text{subject}_{i_{2500ms}}$, $\text{subject}_{i_{1400ms}}$, $\text{subject}_{i_{645ms}}$ are the number of clusters for subject_i for the data cohorts $f_{low} = 2500ms$, $f_{medium} = 1400ms$, $f_{high} = 645ms$ respectively and abs denotes the absolute difference.

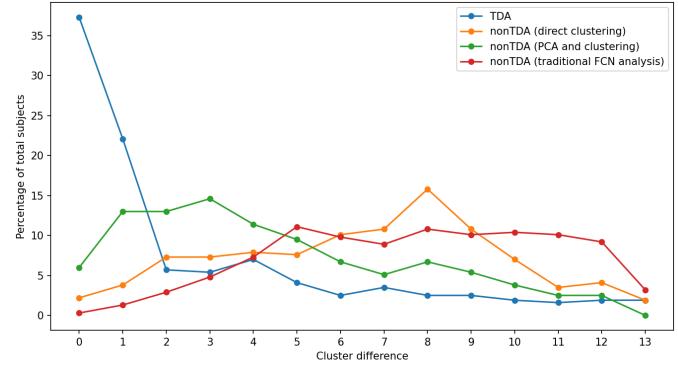


FIGURE 8: Cohort-wide cluster distance comparison between TDA-based statistical data processing pipeline and nonTDA-based pipelines for all temporal sampling periods $f_{low} = 2500ms$, $f_{medium} = 1400ms$ and $f_{high} = 645ms$. In the TDA-based pipeline, more than 59% of the total subjects exhibit a cluster difference less or equal to 1, indicating the highly robust cluster patterns consistent across cohorts.

The cohort-wide cluster distance analysis in Figure 8 reveals striking differences between the TDA-based and nonTDA-based pipelines. For the TDA pipeline, the majority of subjects (59%) exhibit a tight cluster number distance of less than or equal to one across cohorts. This indicates TDA identifies highly robust cluster patterns consistent across individuals. In contrast, for the direct clustering pipeline, only 6% of subjects have a cohort-wide distance less than or equal to one. For the PCA-based pipeline, this number rises to 19% of subjects and for the traditional FCN analysis plummets to 2% of subjects. The significantly lower consistency highlights the inability of these nonTDA-based techniques to extract stable brain states generalizable across the population. Unlike the topological approach, these methods are heavily influenced by individual variations. Overall, the cohort-wide analysis affirms the resilience of TDA for rs-fMRI analysis, which is able to mitigate differences in data acquisition parameters.

B. PAIRWISE CLUSTER DISTANCE COMPARISON

Additionally, we perform a pairwise comparison of the number of clusters for the data cohorts for all of the pipelines. The value of $\text{abs}(\text{subject}_{i_{2500ms}} - \text{subject}_{i_{1400ms}})$ represents the pairwise distance on the number of clusters for subject_i for the data cohorts $f_{low} = 2500ms$ and $f_{medium} = 1400ms$. Similarly, the value of $\text{abs}(\text{subject}_{i_{1400ms}} - \text{subject}_{i_{645ms}})$ and

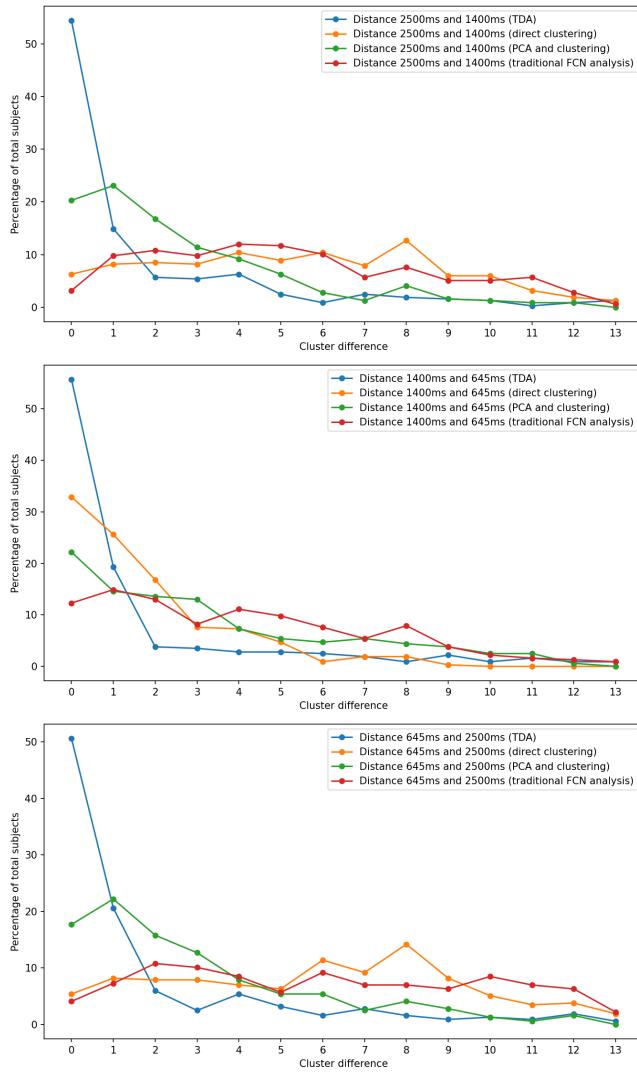


FIGURE 9: Pairwise cluster distance comparison between TDA-based data processing pipeline and nonTDA-based pipelines. Comparison between temporal sampling $f_{medium} = 1400\text{ms}$ and $f_{flow} = 2500\text{ms}$ (top row), temporal sampling $f_{high} = 645\text{ms}$ and $f_{medium} = 1400\text{ms}$ (middle row), temporal sampling $f_{high} = 645\text{ms}$ and $f_{flow} = 2500\text{ms}$ (bottom row).

$\text{abs}(\text{subject}_{i_{645\text{ms}}} - \text{subject}_{i_{2500\text{ms}}})$ represent the pairwise distance on the number of clusters between the data cohorts ($f_{medium} = 1400\text{ms}, f_{high} = 645\text{ms}$) and ($f_{high} = 645\text{ms}, f_{flow} = 2500\text{ms}$) respectively. This pairwise comparison will help to identify whether there is a closer similarity between the data cohorts in the TDA-based pipeline over the nonTDA-based pipelines. Figure 9 shows the pairwise distance between the data cohorts in the TDA-based pipeline and in the nonTDA-based pipelines. In the TDA pipeline, the pairwise distance between data cohorts f_{medium} and f_{high} show the highest similarity (78% matching within distance 2). The data cohort pair f_{high} and f_{flow} has a similarity of 77% within distance two, and data cohort pair f_{flow} and f_{medium} has a similarity of 74% within

the same distance. This high similarity between the data cohort pairs proves the efficacy of persistent homology-based techniques on the rs-fMRI data analysis with different temporal sampling periods. In the nonTDA-based direct clustering pipeline, we see the maximum similarity between the data cohorts f_{medium} and f_{high} with 75% similarity within distance 2. The other data cohort pairs ((f_{flow}, f_{medium}) and (f_{high}, f_{flow})) has 23% and 21% similarity within the same distance. In the PCA-based dimensionality reduction and clustering pipeline, we see the maximum similarity between the data cohorts f_{medium} and f_{flow} with 60% similarity within distance 2. The other data cohort pairs ((f_{medium}, f_{high}) and (f_{high}, f_{flow})) has 50% and 55% similarity within the same distance. In the last nonTDA-based traditional FCN analysis pipeline, we see the maximum similarity between the data cohorts f_{medium} and f_{high} with 40% similarity within distance 2. The other data cohort pairs ((f_{flow}, f_{medium}) and (f_{high}, f_{flow})) has 23% and 22% similarity within the same distance. This low similarity between the data cohorts using nonTDA-based pipelines indicates the inefficiency of the nonTDA-based method for analysing rs-fMRI data with different data acquisition parameters.

C. EVALUATION USING A CLINICAL ADHD DATASET

To validate the robustness of our TDA-based temporal clustering pipeline, we conducted a comparative evaluation against the traditional dynamic functional connectivity network (dFCN) pipeline using the publicly available ADHD-200 dataset [65], [66]. We constructed two cohorts based on the temporal resolution (TR) of the rs-fMRI scans: TR=2s and TR=2.5s. Each cohort comprises ADHD and control subjects. The TR=2s cohort includes 290 control and 285 ADHD subjects, while the TR=2.5s cohort consists of 189 control and 68 ADHD subjects. Figure 10 presents a systematic comparison of clustering consistency between the TDA-based pipeline (left column) and traditional dFCN pipeline (right column) across ADHD and control groups. Each subplot shows the distribution of the number of clusters across all subjects within a group and TR condition. The x-axis denotes the number of clusters (ranging from 1 to 16), and the y-axis indicates the percentage of subjects exhibiting each cluster count.

In the TDA-based pipeline (left column), we observe a strong peak at 2 clusters for both ADHD and control subjects across TR=2s and TR=2.5s. This suggests a high degree of consistency in the extracted brain state patterns, with over 80% of subjects in each subgroup consistently showing two clusters. This invariance across different TRs demonstrates the robustness of persistent homology and topological features in summarizing the intrinsic structure of time-varying brain connectivity.

In contrast, the traditional dFCN pipeline (right column) exhibits significant variability in the number of clusters across TRs. For both ADHD and control groups, the cluster distributions are dispersed, with subjects assigned to a wide range of cluster numbers, especially at higher TRs. For instance, in the ADHD group (top right), the cluster counts span from 3

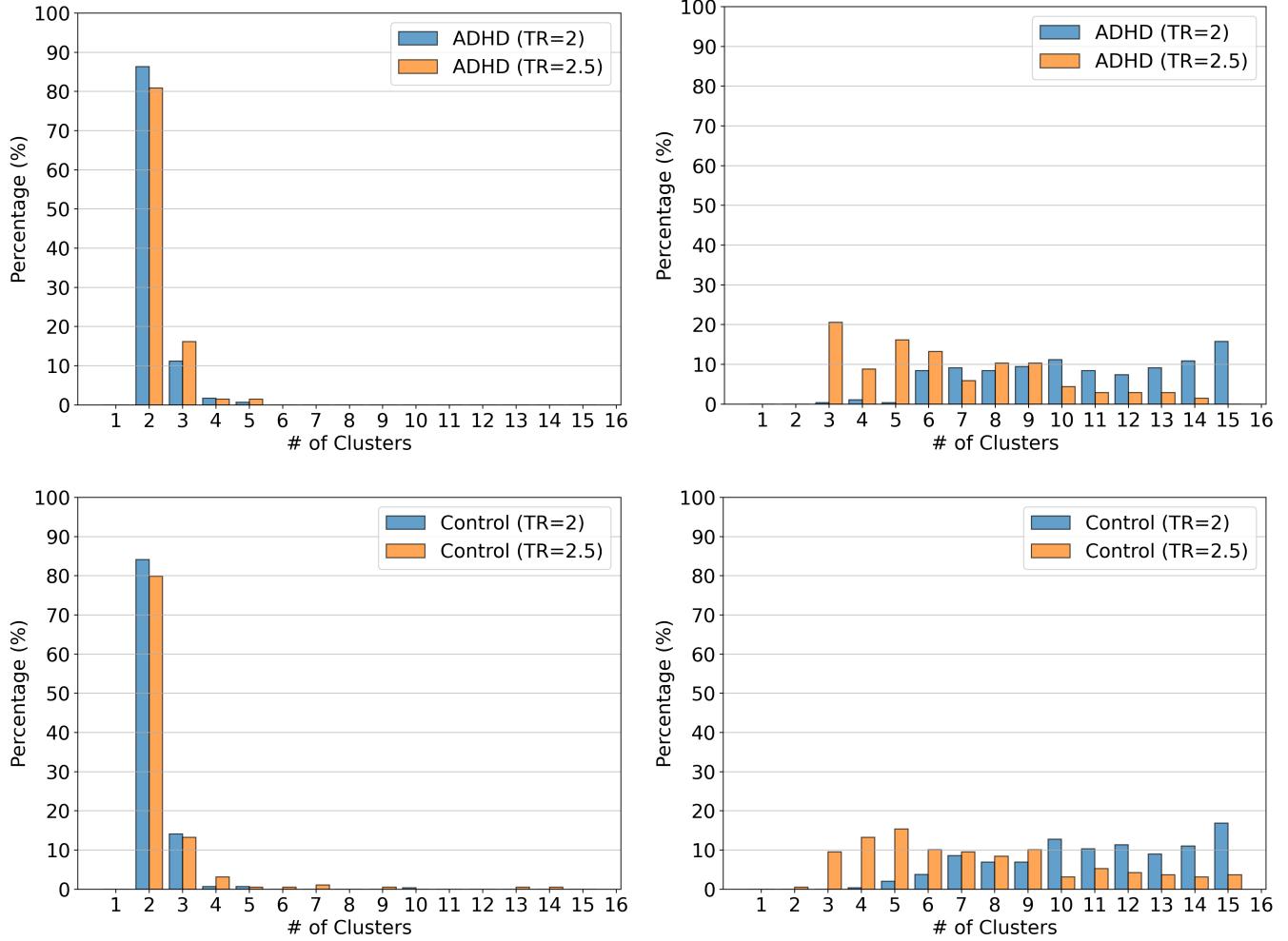


FIGURE 10: Comparison of the number of clusters obtained using TDA-based pipeline (left column) vs traditional dFCN pipeline (right column) for ADHD (top row) and Control (bottom row) groups across TR=2s and TR=2.5s. The TDA pipeline shows high consistency reflected by higher similarity ($\geq 80\%$) in the number of identified clusters. The nonTDA-based traditional dFCN pipeline exhibits variability and sensitivity to TR differences.

to 15, with no dominant mode. This inconsistency indicates the sensitivity of the dFCN pipeline to changes in temporal resolution and its reduced ability to extract reproducible brain state signatures.

This analysis highlights the advantage of using topological features derived from persistent homology over traditional correlation-based approaches. While dFCN pipelines are prone to capturing noise and suffer from over-fragmentation of temporal brain states, the TDA pipeline produces stable and interpretable cluster structures. These findings support the hypothesis that TDA provides a more reliable abstraction of temporal dynamics in brain connectivity, particularly in clinical neuroimaging datasets with acquisition variability.

D. COMPUTATIONAL COMPLEXITY ANALYSIS

We now analyze the computational complexity of the TDA-based (Sec. III) and nonTDA-based pipelines (Sec. IV) up to the statistical analysis stage. Since the statistical analysis

stage is common across all pipelines, our analysis focuses on the pipeline-specific processing steps prior to statistical testing. The computational trade-offs of the pipelines are summarized in Table 1.

Let S denote the number of subjects ($S = 316$), T the number of timepoints per subject ($T = 754$ in the largest case), D the number of regions of interest ($D = 113$), and $D^2 = 12,769$ the number of features in each flattened functional connectivity network (FCN). The number of clusters explored in KMeans is $k = 15$, and the maximum number of KMeans iterations is $T_k = 300$.

TDA-based pipeline: In the TDA-based pipeline, each subject and timepoint undergoes persistent homology computation on its FCN, resulting in a total cost of $S \cdot T \cdot C_{ph}$. The computational cost for generating a single persistence barcode is denoted by C_{ph} . For our approach, using the Gudhi library with the Vietoris–Rips complex and `max_dimension=1` on $D = 113$ ROIs, the time complex-

TABLE 1: Computational trade-offs of each pipeline up to the statistical analysis stage. The second column shows the general Big-O notation, while the third column gives the evaluated value in millions (M) for the largest values of the variables denoting $S = 316$, $T = 754$, $D = 113$ ($D^2 = 12,769$), $k = 15$, $T_k = 300$. Example: For TDA-based: for the largest case, evaluates to $316 \cdot 754 \cdot 86,800 + 316 \cdot 754^2 + 316 \cdot 754 \cdot 15 \cdot 300 = 20,871$ M

Pipeline	Computational Complexity (Big-O)	Evaluated value (in M)
TDA-based	$\mathcal{O}(S \cdot T \cdot C_{ph} + S \cdot T^2 + S \cdot T \cdot k \cdot T_k)$	20,871 M
Direct clustering	$\mathcal{O}(S \cdot T \cdot D^2 \cdot k \cdot T_k)$	13,695,004 M
PCA-based	$\mathcal{O}(S \cdot T \cdot D^4 + S \cdot T \cdot k \cdot T_k)$	38,860,145 M
Traditional dFCN	$\mathcal{O}(S \cdot T^2 \cdot D^2 + S \cdot T^2 + S \cdot T \cdot k \cdot T_k)$	2,290,525 M

ity for 0-dimensional persistent homology scales as $C_{ph} = O(D^2 \log D)$, which yields approximately 86,800 operations per timepoint. After barcode extraction, each subject requires a pairwise Wasserstein distance matrix between all timepoints ($S \cdot T^2$), two-dimensional MDS on this matrix ($S \cdot T^2$), and k-means clustering ($S \cdot T \cdot k \cdot T_k$). The total computational complexity is:

$$\mathcal{O}(S \cdot T \cdot D^2 \log D + S \cdot T^2 + S \cdot T \cdot k \cdot T_k)$$

which, for the largest case, evaluates to 20,871 M.

Direct clustering pipeline: For direct clustering, all FCNs for a subject are flattened into a $(T \times D^2)$ matrix, and k-means clustering is applied in the high-dimensional space. The complexity is:

$$\mathcal{O}(S \cdot T \cdot D^2 \cdot k \cdot T_k)$$

which evaluates to 13,695,004 M.

PCA-based pipeline: In the PCA-based pipeline, the $(T \times D^2)$ FCN matrix is first reduced to two dimensions using PCA, which has complexity $S \cdot T \cdot (D^2)^2 = S \cdot T \cdot D^4$ per subject (due to SVD on high-dimensional data), followed by k-means clustering on the reduced matrix. The total computational complexity is:

$$\mathcal{O}(S \cdot T \cdot D^4 + S \cdot T \cdot k \cdot T_k)$$

which for the largest case evaluates to 38,860,145 M.

Traditional dFCN clustering pipeline: In the traditional dFCN pipeline, for each subject we compute a pairwise Euclidean distance matrix $(T \times T)$ between all $D \times D$ adjacency matrices at all timepoints, with a cost of $S \cdot T^2 \cdot D^2$. This is followed by the same $S \cdot T^2$ MDS embedding and $S \cdot T \cdot k \cdot T_k$ k-means clustering as in the TDA-based pipeline. The total computational complexity is:

$$\mathcal{O}(S \cdot T^2 \cdot D^2 + S \cdot T^2 + S \cdot T \cdot k \cdot T_k)$$

which, for the largest case, is 2,290,525 M.

E. COMPARISON BETWEEN TDA-BASED AND NON-TDA BASED PIPELINES

We extensively evaluated the accuracy and robustness of our TDA pipeline for rs-fMRI analysis on a large-scale healthy control dataset comprising 371,616 adjacency matrices across 316 subjects as well as a clinical ADHD dataset comprising of 832 subjects. Comparisons are made to three alternative approaches - direct clustering, PCA, and traditional FCN analysis.

Table 2 summarizes the systematic comparison between TDA-based and NonTDA-based data processing pipelines in terms of methodology, dimensionality reduction, cluster interpretability, robustness (cohort-wide and pairwise similarities), as well as their advantages and limitations. Clearly, the TDA-based pipeline demonstrates significantly higher robustness, interpretability, and consistency across different temporal data acquisition parameters, albeit with higher computational overhead, when compared to the NonTDA-based methods. The NonTDA-based pipelines vary in their advantages—such as simplicity, computational efficiency, and low computational cost—but generally exhibit reduced robustness, higher noise sensitivity, and lower interpretability. The results demonstrate TDA's superior ability to extract robust and invariant topological signatures intrinsically linked to resting-state functional architectures. Remarkably, the brain states identified by TDA exhibit high consistency across the three sampling frequencies, affirming resilience to acquisition variations. This also highlights TDA's efficacy in mitigating non-neural variability and capturing fundamental dynamics as compared to conventional techniques. By applying the persistent homology technique to filter noise and reveal salient connectivity motifs, TDA provides a principled graph-free technique for preserving temporal dynamics of complex rs-fMRI data. These findings establish persistent homology as a powerful approach for analyzing temporal patterns and validating TDA as a promising pipeline for robust discovery of data-driven functional brain states.

VI. DISCUSSION

MRI scanners around the globe vary in their configurations and field strengths. This variation leads to a certain level of noise in the data collected due to non-neural differences introduced by the diverse scanner setups and data collection parameters. This noise complicates the process of combining data from different scanners into a single, large dataset for unified analysis. Consequently, most fMRI brain network research is localized, limited by the number of subjects that can be scanned at a single location. This limitation reduces the sample size and, therefore, the applicability of the results. One solution is to conduct studies across multiple sites closer to the target population. However, the noise introduced by using different scanners and parameters diminishes the neural effects of interest, thereby reducing the effectiveness of such multi-site efforts.

We addressed these issues in our previous paper [8] in the

TABLE 2: Systematic comparison between TDA-based and NonTDA-based data processing pipelines for rs-fMRI dataset

Criteria	TDA-based pipeline	NonTDA-based pipelines		
		Direct clustering	PCA pipeline	Traditional dFCN
Methodology	Persistent homology + clustering	Direct clustering	PCA + clustering	Correlation + clustering
Dimensionality reduction (Distance function)	MDS (Wasserstein distance)	None	PCA	MDS (Euclidean distance)
Cluster interpretability (Cohort-wide similarity)	High (59%)	Low (6%)	Low (19%)	Low (2%)
Robustness (Pairwise similarity)	High (78%, 77%, 74%)	Moderate (75%, 23%, 21%)	Moderate (60%, 55%, 50%)	Low (40%, 23%, 22%)
Advantages	High robustness; Effective noise filtering; Clear interpretability	Simple, straightforward; No preprocessing required	Moderate robustness; Reduced dimensions	Low computational cost; Reduced dimensions
Limitations	Additional computations	High dimensionality; Noise sensitive; Poor interpretability	Poor interpretability; Moderate robustness	Poor interpretability; Noise sensitive

context of characterizing brain networks using static functional connectivity. However, DFC is critical for understanding how the brain processes information dynamically and how the interactions between different brain regions change with time. It has been shown that DFC is very important for characterizing the healthy brain [3], [4], as well as in various brain disorders [5]–[7]. Therefore, it becomes necessary to develop a TDA-based framework for DFC so that investigations of temporal dynamics in the brain are shielded from non-neural variability in the data.

We validated the effectiveness of the proposed Topological Data Analysis (TDA)-based pipeline by contrasting it with the conventional data analysis pipelines outlined in Section 4. In the conventional pipelines, we employed direct time-series clustering, PCA-based dimensionality reduction and clustering, as well as traditional dynamic FCN pipeline with MDS-based dimensionality reduction. The outcomes of this pipeline strongly suggest that these conventional methods fail to establish similarity across dynamic FCNs of the same subjects obtained with different repetition times (TRs) and acquisition parameters.

On the contrary, for the TDA-based metric, we demonstrated both qualitatively and quantitatively that the metric remains statistically consistent across the same subjects, regardless of the sampling period used to acquire resting-state fMRI data. This underscores the usefulness of TDA-based analysis because, theoretically, data collected using different parameters from the same subject should still represent the same brain network dynamics.

VII. CONCLUSION

In this study, we have demonstrated the effectiveness of Topological Data Analysis (TDA) in uncovering temporal properties within resting-state functional magnetic resonance imaging (rs-fMRI) data. Our research highlights TDA's robustness in the presence of varying temporal sampling rates, surpassing traditional connectivity analysis methods. Key findings emphasize TDA's remarkable stability, with 59% of subjects consistently showing clustering results across dif-

ferent temporal sampling periods (2500ms, 1400ms, 645ms), compared to less than 19% using nonTDA-based methods. TDA also reveals strong pairwise similarities between sampling periods, showcasing its ability to capture temporal dynamics. The robustness of the TDA pipeline is further confirmed through evaluation on clinical ADHD datasets, where it achieves high consistency ($\geq 80\%$) in clustering outcomes across different sites and scanning conditions. In conclusion, our study establishes TDA as a valuable tool for revealing temporal nuances in rs-fMRI data, offering a level of robustness unmatched by traditional methods. Through persistent homology, TDA provides a stable, invariant representation of dynamic brain connectivity, promising valuable insights into complex temporal patterns in resting-state fMRI data across diverse acquisition parameters. To promote reproducibility, we have made all our code, scripts, data, and documentation available at <https://github.com/harp-lab/TemporalBrainPH>.

ACKNOWLEDGMENT

We thank Xi-Nian Zuo from Beijing Normal University for sharing some of the data with us. We are also grateful to the entire Enhanced Nathan Kline Institute - Rockland Sample (NKI-RS) team for providing free access to the raw MRI data. Principal support for the enhanced NKI-RS project is provided by the NIMH BRAINS R01MH094639-01 (PI Milham). Funding for key personnel also provided in part by the New York State Office of Mental Health and Research Foundation for Mental Hygiene. Funding for the decompression and augmentation of administrative and phenotypic protocols provided by a grant from the Child Mind Institute (1FDN2012-1). Additional personnel support provided by the Center for the Developing Brain at the Child Mind Institute, as well as NIMH R01MH081218, R01MH083246, and R21MH084126. Project support also provided by the NKI Center for Advanced Brain Imaging (CABI), the Brain Research Foundation, and the Stavros Niarchos Foundation.

REFERENCES

- [1] R. S. Menon and S.-G. Kim, "Spatial and temporal limits in cognitive neuroimaging with fmri," *Trends in cognitive sciences*, vol. 3, no. 6, pp. 207–216, 1999.
- [2] D. A. Hall, M. P. Haggard, M. A. Akeroyd, A. R. Palmer, A. Q. Summerfield, M. R. Elliott, E. M. Gurney, and R. W. Bowtell, "sparse" temporal sampling in auditory fmri," *Human brain mapping*, vol. 7, no. 3, pp. 213–223, 1999.
- [3] G. Deshpande and H. Jia, "Multi-level clustering of dynamic directional brain network patterns and their behavioral relevance," *Frontiers in Neuroscience*, vol. 13, p. 476054, 2020.
- [4] H. Jia, X. Hu, and G. Deshpande, "Behavioral relevance of the dynamics of the functional brain connectome," *Brain connectivity*, vol. 4, no. 9, pp. 741–759, 2014.
- [5] D. Rangaprakash, M. N. Dretsch, J. S. Katz, T. S. Denney Jr, and G. Deshpande, "Dynamics of segregation and integration in directional brain networks: illustration in soldiers with ptsd and neurotrauma," *Frontiers in neuroscience*, vol. 13, p. 442861, 2019.
- [6] D. Rangaprakash, M. N. Dretsch, A. Venkataraman, J. S. Katz, T. S. Denney Jr, and G. Deshpande, "Identifying disease foci from static and dynamic effective connectivity networks: illustration in soldiers with trauma," *Human brain mapping*, vol. 39, no. 1, pp. 264–287, 2018.
- [7] C. Jin, H. Jia, P. Lanka, D. Rangaprakash, L. Li, T. Liu, X. Hu, and G. Deshpande, "Dynamic brain connectivity is a better predictor of ptsd than static connectivity," *Human brain mapping*, vol. 38, no. 9, pp. 4479–4496, 2017.
- [8] S. Kumar, A. R. Shovon, and G. Deshpande, "The robustness of persistent homology of brain networks to data acquisition-related non-neural variability in resting state fmri," *Human Brain Mapping*, 2023.
- [9] S. Das, H. Ombao, and M. K. Chung, "Topological data analysis for functional brain networks," *arXiv preprint arXiv:2210.09092*, 2022.
- [10] M. Rubinov and O. Sporns, "Complex network measures of brain connectivity: uses and interpretations," *Neuroimage*, vol. 52, no. 3, pp. 1059–1069, 2010.
- [11] M. J. Catanzaro, S. Rizzo, J. Kopchick, A. Chowdury, D. R. Rosenberg, P. Bubenik, and V. A. Diwadkar, "Topological data analysis captures task-driven fmri profiles in individual participants: A classification pipeline based on persistence," *Neuroinformatics*, 2023. [Online]. Available: <https://api.semanticscholar.org/CorpusID:265018725>
- [12] H. Ryu, C. Habeck, Y. Stern, and S. Lee, "Persistent homology-based functional connectivity and its association with cognitive ability: Life-span study," *Human Brain Mapping*, 2023.
- [13] J. Billings, M. Saggar, J. Hlinka, S. Keilholz, and G. Petri, "Simplicial and topological descriptions of human brain dynamics," *Network Neuroscience*, vol. 5, no. 2, pp. 549–568, 06 2021. [Online]. Available: https://doi.org/10.1162/netn_a_00190
- [14] G. Petri, P. Expert, F. Turkheimer, R. Carhart-Harris, D. Nutt, P. J. Hellyer, and F. Vaccarino, "Homological scaffolds of brain functional networks," *Journal of The Royal Society Interface*, vol. 11, no. 101, p. 20140873, 2014.
- [15] S. Vallender, "Calculation of the wasserstein distance between probability distributions on the line," *Theory of Probability & Its Applications*, vol. 18, no. 4, pp. 784–786, 1974.
- [16] H. Edelsbrunner, "Persistent homology: theory and practice," 2013.
- [17] F. H. Xu, M. Gao, J. Chen, S. Garai, D. A. Duong-Tran, Y. Zhao, and L. Shen, "Topology-based clustering of functional brain networks in an alzheimer's disease cohort," *AMIA Summits on Translational Science Proceedings*, vol. 2024, p. 449, 2024.
- [18] M. J. Catanzaro, S. Rizzo, J. Kopchick, A. Chowdury, D. R. Rosenberg, P. Bubenik, and V. A. Diwadkar, "Topological data analysis captures task-driven fmri profiles in individual participants: a classification pipeline based on persistence," *Neuroinformatics*, vol. 22, no. 1, pp. 45–62, 2024.
- [19] M. Esteve and A. Falcó, "Trajectory classification through topological data analysis perspectives," *IEEE Access*, vol. 13, pp. 32458–32469, 2025.
- [20] R. Kindelan, J. Frías, M. Cerdá, and N. Hitschfeld, "A topological data analysis based classifier," *Advances in Data Analysis and Classification*, vol. 18, no. 2, pp. 493–538, 2024.
- [21] R. Ghrist, "Barcodes: the persistent topology of data," *Bulletin of the American Mathematical Society*, vol. 45, no. 1, pp. 61–75, 2008.
- [22] M. Saggar, O. Sporns, J. Gonzalez-Castillo, P. A. Bandettini, G. Carlsson, G. Glover, and A. L. Reiss, "Towards a new approach to reveal dynamical organization of the brain using topological data analysis," *Nature Communications*, vol. 9, no. 1, p. 1399, Apr 2018. [Online]. Available: <https://doi.org/10.1038/s41467-018-03664-4>
- [23] R. W. Ghrist, *Elementary applied topology*. Createspace Seattle, 2014, vol. 1.
- [24] S. Y. Oudot, *Persistence theory: from quiver representations to data analysis*. American Mathematical Society Providence, 2015, vol. 209.
- [25] H. Edelsbrunner and J. Harer, *Computational Topology: An Introduction*, 2010.
- [26] S. Weinberger, "What is... persistent homology," *Notices of the AMS*, vol. 58, no. 1, pp. 36–39, 2011.
- [27] E. Wong, S. Palande, B. Wang, B. Zielinski, J. Anderson, and P. T. Fletcher, "Kernel partial least squares regression for relating functional brain network topology to clinical measures of behavior," in *2016 IEEE 13th International Symposium on Biomedical Imaging (ISBI)*, 2016, pp. 1303–1306.
- [28] M. Hajij, B. Wang, C. Scheidegger, and P. Rosen, "Visual detection of structural changes in time-varying graphs using persistent homology," in *2018 ieee pacific visualization symposium (pacifcvis)*. IEEE, 2018, pp. 125–134.
- [29] S. Noble, D. Scheinost, E. S. Finn, X. Shen, X. Papademetris, S. C. McEwen, C. E. Bearden, J. Addington, B. Goodyear, K. S. Cadenhead *et al.*, "Multisite reliability of mr-based functional connectivity," *Neuroimage*, vol. 146, pp. 959–970, 2017.
- [30] C. Dansereau, Y. Benhajali, C. Risterucci, E. M. Pich, P. Orban, D. Arnold, and P. Bellec, "Statistical power and prediction accuracy in multisite resting-state fmri connectivity," *Neuroimage*, vol. 149, pp. 220–232, 2017.
- [31] R. T. Shinohara, J. Oh, G. Nair, P. A. Calabresi, C. Davatzikos, J. Doshi, R. G. Henry, G. Kim, K. A. Linn, N. Papinutto *et al.*, "Volumetric analysis from a harmonized multisite brain mri study of a single subject with multiple sclerosis," *American Journal of Neuroradiology*, vol. 38, no. 8, pp. 1501–1509, 2017.
- [32] R. Yuste and A. L. Fairhall, "Temporal dynamics in fmri resting-state activity," *Proceedings of the National Academy of Sciences*, vol. 112, no. 17, pp. 5263–5264, 2015.
- [33] M. G. Preti, T. A. Bolton, and D. Van De Ville, "The dynamic functional connectome: State-of-the-art and perspectives," *Neuroimage*, vol. 160, pp. 41–54, 2017.
- [34] R. M. Hutchison, T. Womelsdorf, E. A. Allen, P. A. Bandettini, V. D. Calhoun, M. Corbetta, S. Della Penna, J. H. Duyn, G. H. Glover, J. Gonzalez-Castillo *et al.*, "Dynamic functional connectivity: promise, issues, and interpretations," *Neuroimage*, vol. 80, pp. 360–378, 2013.
- [35] M. Termenon, A. JAILLARD, C. Delon-Martin, and S. Achard, "Reliability of graph analysis of resting state fmri using test-retest dataset from the human connectome project," *NeuroImage*, vol. 142, pp. 172 – 187, 2016.
- [36] M. Andellini, V. Cannata, S. Gazzellini, B. Bernardi, and A. Napolitano, "Test-retest reliability of graph metrics of resting state mri functional brain networks: A review," *Journal of Neuroscience Methods*, vol. 253, pp. 183 – 192, 2015.
- [37] C. E. Ginestet, A. P. Fournel, and A. Simmons, "Statistical network analysis for functional mri: summary networks and group comparisons," *Frontiers in Computational Neuroscience*, vol. 8, p. 51, 2014. [Online]. Available: <https://www.frontiersin.org/article/10.3389/fncom.2014.00051>
- [38] D. Papo, M. Zanin, J. H. Martínez, and J. M. Buldú, "Beware of the small-world neuroscientist!" *Frontiers in Human Neuroscience*, vol. 10, p. 96, 2016. [Online]. Available: <https://www.frontiersin.org/article/10.3389/fnhum.2016.00096>
- [39] M. Rubinov and O. Sporns, "Complex network measures of brain connectivity: Uses and interpretations," *NeuroImage*, vol. 52, no. 3, pp. 1059 – 1069, 2010, computational Models of the Brain. [Online]. Available: <http://www.sciencedirect.com/science/article/pii/S105381190901074X>
- [40] M. Drakesmith, K. Caeyenberghs, A. Dutt, G. Lewis, A. David, and D. Jones, "Overcoming the effects of false positives and threshold bias in graph theoretical analyses of neuroimaging data," *NeuroImage*, vol. 118, pp. 313 – 333, 2015. [Online]. Available: <http://www.sciencedirect.com/science/article/pii/S1053811915003912>
- [41] J. Zhang, X. Tuo, Z. Yuan, W. Liao, and H. Chen, "Analysis of fmri data using an integrated principal component analysis and supervised affinity propagation clustering approach," *IEEE Transactions on Biomedical Engineering*, vol. 58, pp. 3184–3196, 2011. [Online]. Available: <https://api.semanticscholar.org/CorpusID:11217131>
- [42] K. Vijay and K. Selvakumar, "Brain fmri clustering using interaction k-means algorithm with pca," *2015 International Conference on Communications and Signal Processing (ICCP)*, pp. 0909–0913, 2015. [Online]. Available: <https://api.semanticscholar.org/CorpusID:15453903>

- [43] B. M. S. Hasan and A. M. Abdulazeez, "A review of principal component analysis algorithm for dimensionality reduction," *Journal of Soft Computing and Data Mining*, vol. 2, no. 1, pp. 20–30, 2021.
- [44] D. Napoleon and S. Pavalkodi, "A new method for dimensionality reduction using k-means clustering algorithm for high dimensional data set," *International Journal of Computer Applications*, vol. 13, no. 7, pp. 41–46, 2011.
- [45] S. Mure, T. Grenier, C. R. G. Guttmann, and H. Benoit-Cattin, "Unsupervised time-series clustering of distorted and asynchronous temporal patterns," in *2016 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2016, pp. 1263–1267.
- [46] S. M. Smith, K. L. Miller, S. Moeller, J. Xu, E. J. Auerbach, M. W. Woolrich, C. F. Beckmann, M. Jenkinson, J. Andersson, M. F. Glasser *et al.*, "Temporally-independent functional modes of spontaneous brain activity," *Proceedings of the National Academy of Sciences*, vol. 109, no. 8, pp. 3131–3136, 2012.
- [47] J. J. Remes, A. Abou Elseoud, E. Ollila, M. Haapea, T. Starck, J. Nikkinen, O. Tervonen, and O. Silven, "On applicability of pca, voxel-wise variance normalization and dimensionality assumptions for sliding temporal window sica in resting-state fmri," *Magnetic Resonance Imaging*, vol. 31, no. 8, pp. 1338–1348, 2013.
- [48] M. K. Chung, S.-G. Huang, I. C. Carroll, V. D. Calhoun, and H. H. Goldsmith, "Topological state-space estimation of functional human brain networks," *PLOS Computational Biology*, vol. 20, no. 5, p. e1011869, 2024.
- [49] M. K. Chung, S. Das, and H. Ombao, "Dynamic topological data analysis of functional human brain networks," *Foundations of Data Science*, vol. 6, no. 1, pp. 22–40, 2024.
- [50] M. Zhang, S. Chowdhury, and M. Saggar, "Temporal Mapper: Transition networks in simulated and real neural dynamics," *Network Neuroscience*, vol. 7, no. 2, pp. 431–460, 06 2023. [Online]. Available: https://doi.org/10.1162/netn_a_00301
- [51] M. Saggar, J. M. Shine, R. Liégeois, N. U. Dosenbach, and D. Fair, "Precision dynamical mapping using topological data analysis reveals a hub-like transition state at rest," *Nature communications*, vol. 13, no. 1, p. 4791, 2022.
- [52] R. Abreu, J. Jorge, A. Leal, T. Koenig, and P. Figueiredo, "Eeg microstates predict concurrent fmri dynamic functional connectivity states," *Brain topography*, vol. 34, no. 1, pp. 41–55, 2021.
- [53] Q. Zhou, L. Zhang, J. Feng, and C.-Y. Z. Lo, "Tracking the main states of dynamic functional connectivity in resting state," *Frontiers in neuroscience*, vol. 13, p. 448068, 2019.
- [54] K. B. Noonan, S. Colcombe, R. Tobe, M. Mennes, M. Benedict, A. Moreno, L. Panek, S. Brown, S. Zavitz, Q. Li *et al.*, "The nki-rockland sample: a model for accelerating the pace of discovery science in psychiatry," *Frontiers in neuroscience*, vol. 6, p. 152, 2012.
- [55] D. A. Feinberg, S. Moeller, S. M. Smith, E. Auerbach, S. Ramanna, M. F. Glasser, K. L. Miller, K. Ugurbil, and E. Yacoub, "Multiplexed echo planar imaging for sub-second whole brain fmri and fast diffusion imaging," *PloS one*, vol. 5, no. 12, p. e15710, 2010.
- [56] B. Thomas Yeo, F. M. Krienen, J. Sepulcre, M. R. Sabuncu, D. Lashkari, M. Hollinshead, J. L. Roffman, J. W. Smoller, L. Zöllei, J. R. Polimeni *et al.*, "The organization of the human cerebral cortex estimated by intrinsic functional connectivity," *Journal of neurophysiology*, vol. 106, no. 3, pp. 1125–1165, 2011.
- [57] S. M. Smith, K. L. Miller, G. Salimi-Khorshidi, M. Webster, C. F. Beckmann, T. E. Nichols, J. D. Ramsey, and M. W. Woolrich, "Network modelling methods for fmri," *Neuroimage*, vol. 54, no. 2, pp. 875–891, 2011.
- [58] H. Edelsbrunner and J. Harer, "Surveys on discrete and computational geometry," *Contemporary Mathematics (American Mathematical Society, Providence, RI)*, vol. 453, pp. 257–282, 2008.
- [59] U. Bauer, "Ripser: efficient computation of vietoris–rips persistence barcodes," *Journal of Applied and Computational Topology*, vol. 5, no. 3, pp. 391–423, 2021.
- [60] F. Pedregosa, G. Varoquaux, A. Gramfort, V. Michel, B. Thirion, O. Grisel, M. Blondel, P. Prettenhofer, R. Weiss, V. Dubourg, J. Vanderplas, A. Pasquier, D. Cournapeau, M. Brucher, M. Perrot, and E. Duchesnay, "Scikit-learn: Machine learning in Python," *Journal of Machine Learning Research*, vol. 12, pp. 2825–2830, 2011.
- [61] P. J. Rousseeuw, "Silhouettes: a graphical aid to the interpretation and validation of cluster analysis," *Journal of computational and applied mathematics*, vol. 20, pp. 53–65, 1987.
- [62] M. Kerber, D. Morozov, and A. Nigmetov, "Geometry helps to compare persistence diagrams," *Journal of Experimental Algorithms (JEA)*, vol. 22, pp. 1–20, 2017.
- [63] T. G. Project, "Gudhi user and reference manual," <https://gudhi.inria.fr/python/latest/>, 2024.
- [64] N. Bouhmala, "How good is the euclidean distance metric for the clustering problem," in *2016 5th IIAI international congress on advanced applied informatics (IIAI-AAI)*. IEEE, 2016, pp. 312–315.
- [65] . F. C. Project, "The adhd-200 sample," 2011, https://fcon_1000.projects.nitrc.org/indi/adhd200/ [Accessed: February 20, 2025].
- [66] X. Zhao, D. Rangaprakash, B. Yuan, T. S. Denney Jr, J. S. Katz, M. N. Dretsch, and G. Deshpande, "Investigating the correspondence of clinical diagnostic grouping with underlying neurobiological and phenotypic clusters using unsupervised machine learning," *Frontiers in applied mathematics and statistics*, vol. 4, p. 25, 2018.



AHMEDUR RAHMAN SHOVON is a postdoctoral appointee in the Mathematics and Computer Science Division at Argonne National Laboratory. He received his PhD in Computer Science from the University of Illinois Chicago (UIC). He earned both his MS and BS degrees in Information Technology from Jahangirnagar University, Bangladesh. Shovon has professional experience as a software engineer, where he contributed to large-scale applications in the news media and e-governance sectors. His research interests include high-performance computing, data analytics, and cloud computing. His work has appeared in leading journals such as the *IEEE Journal of Biomedical and Health Informatics*, *Human Brain Mapping*, and *Neoplasia*, and has been presented at premier venues including ICS, ASPLOS, USENIX ATC, SC, and IEEE Cloud. In addition to his academic contributions, Shovon is an active member of the open-source community. He ranks among the top 3% of contributors on Stack Overflow, where he regularly shares technical insights with a global audience.



SIDHARTH KUMAR is an Associate Professor at the Department of Computer Science at the University of Illinois Chicago (UIC). Before moving to UIC, Dr. Kumar was first a post-doc at the Scientific Computing and Imaging (SCI) Institute at the University of Utah and then an Assistant professor at the University of Alabama at Birmingham, USA. Dr. Kumar received his Ph.D. degree in Computing from the University of Utah, USA, and his B.Tech degree in Information and Communication Technology from DAIICT, India. Dr. Kumar has published several papers in the field of high-performance computing, data visualization, big-data systems and data management. He has received best research paper awards at prestigious computer science conferences such as HiPC, ISC and LDAV. His research is funded by the National Science Foundation and National Institutes of Health.



GOPIKRISHNA DESHPANDE is a Professor at the Auburn University Neuroimaging Center, Department of Electrical and Computer Engineering, Auburn University, USA. Dr. Deshpande is also an affiliated faculty at the Department of Psychological Sciences at Auburn University. Before moving to Auburn, Dr. Deshpande was first a post-doc and then a research faculty at Biomedical Imaging Technology Center, Emory University (2007-2010), USA. Dr. Deshpande received his Ph.D. degree in Biomedical Engineering from Georgia Institute of Technology, USA (2003-2007) and M.S degree in Electrical Communication Engineering from Indian Institute of Science Bangalore, India (2001-2003). Dr. Deshpande has published more than 130 peer-reviewed journal papers in the field of signal processing, neuroimaging, brain connectivity and machine learning. He has received numerous awards including the President's Outstanding Collaborative Units Award. His research is funded by the US Department of Defense, National Science Foundation and National Institutes of Health.

• • •