Resting-State fMRI Functional Connectivity: Big Data Preprocessing Pipelines and Topological Data Analysis

Angkoon Phinyomark, *Member, IEEE,* Esther Ibáñez-Marcelo, and Giovanni Petri

Abstract—Resting state functional magnetic resonance imaging (rfMRI) can be used to measure functional connectivity and then identify brain networks and related brain disorders and diseases. To explore these complex networks, however, huge amounts of data are necessary. Recent advances in neuroimaging technologies, and the unique methodological approach of rfMRI, have enabled us to an era of Biomedical Big Data. The recent progress of big data sharing projects with their challenges are discussed. This increasing amount of neuroimaging data has greatly increased the importance of developing preprocessing pipelines and advanced analytic techniques, which are better at handling large-scale datasets. Before applying any analysis method on rfMRI data, several preprocessing steps need to be applied to reduce all unwanted effects. Three alternative ways to get access to big preprocessed rfMRI data are presented involving the minimal preprocessing pipelines. There are several commonly used methods to examine functional connectivity. However, they become limited in the analysis of big data, and a new tool to explore such data is necessary. We propose a number of novel methods rooted in algebraic topology and collectively referred to as Topological Data Analysis to rfMRI functional connectivity. Their properties for big data analysis are also discussed.

Index Terms—Big data, Brain network, Functional connectivity, Graph theory, Preprocessing pipeline, Resting-state fMRI, Topological data analysis

1 Introduction

The human brain is a complex network of functionally and structurally interconnected regions. Although each region has its own task and function, these different brain regions continuously share information with each other and then form a complex integrative network named the brain network. To understand the organization of the human brain, one can study the underlying connectivity of different functional brain regions, or functional connectivity, as well as physical or structural connectivity in the brain.

Functional connectivity is primarily explored and investigated through resting state functional magnetic resonance imaging (rfMRI or R-fMRI) and is typically analyzed in terms of correlation or spatial grouping based on temporal similarities [1]. These approaches are supported by the fact that during rest, in the absence of any explicit task, the spontaneous neuronal activity patterns of multiple brain regions observed through changes in a blood-oxygen-level dependent (BOLD) signal (or rfMRI time-series) are not random and unstructured, but, in contrast, are highly correlated. In other words, functional connectivity can be explored by measuring the level of synchronization of rfMRI time-series between anatomically separated brain regions. These approaches assume similar patterns of activation can reflect functional and neuronal communication between brain regions regardless of the apparent physical connectedness of the regions. Functional networks generated using these approaches are also termed resting-state networks [2].

Since rfMRI relies on the assumption that spontaneous low frequency BOLD fluctuations (0.01-0.1 Hz) are a measure of intrinsic activity in the brain, a group of researchers have questioned whether the fluctuations observed during the resting-state could be artifacts of other bodily functions [3]. Although the true neuronal basis of these fluctuations has not yet been fully understood, there are several supports for a possible neuronal basis of rfMRI. For instance, most of the resting-state connected activities tends to occur along structural networks in the brain [4] as well there is an association between information derived from rfMRI and from other measures of neuronal activity [5].

The first and the most fundamental resting-state network is the so-called default mode network (DMN), firstly presented in a seminal rfMRI study of Biswal and colleagues [4] and have been confirmed later by a series of studies (e.g. [6], [7]). Unlike other brain networks that can be observed and identified by their activation during tasks, DMN is a group of brain regions that is active during rest, in a baseline or default mode of the brain, and deactivated during a variety of cognitive tasks. These studies also suggest that brain networks which activate or deactivate together during tasks maintain their signature connectivity at rest. It means that neuroscientists can study the known functional brain networks of both healthy and abnormal brain without the use of specially designed tasks, which may be unable to be completed by young children or patients who cannot perform either complex cognitive tasks or long experiments.

Other advantages of employing rfMRI [8] include the simplicity of the procedure, which may offer a better signal-to-noise ratio (SNR), and its relatively short period of

- *A. Phinyomark is with ISI Foundation, Torino, 10126, Italy. E-mail: angkoon.phinyomark@isi.it, see http://www.angkoon-phinyomark.com/contact*
- *E. Ibáñez-Marcelo and G. Petri are with ISI Foundation.*

Manuscript received March 1, 2017; revised July 1, 2017.

acquisition time, which allows for increased sample size or big data. Unlike task-based imaging which typically extracts only one feature brain network, rfMRI allows us to observe many brain networks at once (or multi-purpose data sets [8]). With rfMRI, functional connectivity can also be applied to examine several hypothesized and believed functional dysconnectivity effects in brain disorders and diseases such as Alzheimers disease, amyotrophic lateral sclerosis, attention deficit hyperactivity disorder (ADHD), autism, epilepsy, Parkinson's disease, schizophrenia, multiple sclerosis, and obsessive compulsive disorder (for a review, see [8], [9]). This information will be useful for clinicians for prognosis, diagnosis and treatments. Unfortunately, clinical applications of rfMRI are still at an early stage of development.

Although functional connectivity based on rfMRI can reveal interesting new findings about the functional connections of brain regions and networks, huge amounts of data are necessary to explore these complex networks. Recent advances in neuroimaging technologies combined with the unique methodological approach of rfMRI have enabled us to an era of "Biomedical Big Data". The 1000 Functional Connectomes Project [10] and the Human Connectome Project [11], both neuroimaging databases, for instance, have publicly released over 1,000 rfMRI data sets. Here we present the recent progress of existing shared rfMRI big data sets (in Section 2). The increasing amount of shared neuroimaging datasets has greatly increased the importance of developing data preprocessing pipelines and advanced analytic techniques, which are better at handling large-scale rfMRI data.

Before applying any analytic technique on rfMRI data, several preprocessing steps are required in order to: reduce various artifacts, align the data acquired at different points in time for an individual subject, establish some correspondence between the brains of different subjects, and so on. While it is acknowledged in the literature that different methods and their order in the preprocessing pipelines can affect the results obtained from statistical group difference tests and classification models (e.g. [12], [13], [14], [15]), most studies used their own specific pipeline and no consensus across the studies has been found regarding the optimal preprocessing pipeline. Here we present state-of-the-art rfMRI preprocessing pipelines, with a focus on software packages designed for large-scale rfMRI data analysis (in Section 3).

After the rfMRI data has been preprocessed, there are several commonly used methods for examining functional connectivity such as seed-based correlation analysis (SCA), cluster analysis, principal component analysis (PCA), independent component analysis (ICA) and graph theory (for a review, see [1], [16]). However, these traditional methods encounter limits in terms of their descriptive power when faced with complex, highly-dimensional datasets describing interactions between large number of elements, as is often the case in the analysis of big data. New tools to complement the explore and analysis of such data sets are necessary. Therefore, we lastly propose a set of novel methods, which are rooted in algebraic topology and collectively referred to as "Topological Data Analysis" to rfMRI functional connectivity and their properties for big data analysis are also discussed (in Section 4).

2 Big rfMRI Data

Large shared rfMRI data sets are necessary to obtain new insights and interesting findings in the large-scale organization of complex cognitive operations in the human brain. Besides the fact that some clinical and research questions cannot be answered using a single small data set since each sub-population may exhibit different features that are not shared by others, larger samples are generally preferable in order to compensate for the large inter-subject and intra-subject variability typical in rfMRI recordings. There are many advantages of big data sharing (or *Big Value*) such as improving reliability and reproducibility of research (i.e., increasing statistical power and reducing false-positive rates), improving research practices, maximizing the contribution of research subjects, backing up valuable data and reducing the cost of research within the neuroimaging community [17].

Thanks to the unique methodological approach of rfMRI, a long-standing interest in acquiring the large-scale functional neuroimaging data sets has been increasingly fulfilled over the last decade. Recent advances in neuroimaging technologies as well data storage, management and sharing systems also enable the unrestricted sharing and open access of big neuroimaging data involving the projects with special emphasis on rfMRI data: the 1000 Functional Connectomes Project [10] and the Human Connectome Project [11]. The recent progress of these two big data sharing projects is focused and presented in this section.

2.1 The 1000 Functional Connectomes Project

The 1000 Functional Connectomes Project (FCP) was launched in 2009 by gathering rfMRI data from over 1,300 subjects collected independently at 33 international institutes and centers [10]. All datasets are fully accessible upon successful registration at http://fcon_1000.projects.nitrc.org. All datasets are anonymous and demographic information provided is limited to age, gender and handedness. No extensive data preprocessing has been performed for any of the data sets. However, scripts for further preprocessing of the data sets are provided as part of the project involving motion correction, spatial filtering with 6 mm FWHM (full width at half maximum) and 12 DOF (degrees of freedom) affine transformation to MNI152 (the Montreal Neurological Institute of McGill University Health Centre) stereotaxic space [10].

To demonstrate the feasibility of pooled rfMRI data from multiple sites, Biswal et al. [10] performed several functional connectivity analyses using two commonly used methods: SCA and ICA on 1,093 subjects from 24 sites. The results show evidence of a universal functional architecture (i.e., the consistent patterns of functional connectivity across data collection sites) as well age- and sex-related differences in rfMRI measures-based frequency-domain analysis. These findings confirm the usefulness of the high-throughput rfMRI data. Consequently, data from this project have been used as a common test bed to evaluate new methods proposed in this field of research (e.g. [18], [19]).

This project served as the parent project for many large-scale datasets under the International Neuroimaging Data-

Sharing Initiative (INDI) project[1]: for instance, the Autism Brain Imaging Data Exchange (ABIDE)[2] with 1,026 individuals with autism spectrum disorder (ASD) and 1,130 typical controls from 17 different sites [20], ADHD-200[3] with 383 children and adolescents with ADHD and 491 controls from 8 multiple sites [21], and the Consortium for Reliability and Reproducibility (CoRR)[4] with 1,652 subjects [22]. The ongoing phase of this project is to regularly release (e.g. weekly, monthly, or quarterly) prospective rfMRI data sets[5] such as the enhanced Nathan Kline Institute-Rockland Sample (NKI-RS)[6] with a current total of 973 subjects [23].

All the FCP datasets are distributed using XNAT[7], the most widely-used imaging informatics platform developed by the Neuroinformatics Research Group [24]. To support cloud computing, the FCP data is recently available for download from an Amazon Simple Storage Service (S3) bucket[8]. Further, the data from both FCP and INDI was preprocessed using different preprocessing pipelines and is openly shared under the new project, namely, the Preprocessed Connectomes Project. Unfortunately, several limitations for the FCP have been acknowledged, for instance, rfMRI data is pooled from previously collected data so there is no prior coordination of data acquisition methods [10].

2.2 The Human Connectome Project

The Human Connectome Project (HCP) was launched in 2010 led by the WU-Minn HCP consortium [11], [25]. In the first phase of this project, methods for data acquisition and analysis were developed. The standardized imaging protocols and preprocessing pipelines [26] were then applied in the second phase when data was being acquired from a target number of 1,200 subjects at three different institutes. The subjects being studied are healthy twins and their non-twin siblings ages 22-35 from varying ethnic groups. All neuroimaging data and most of the behavioral data are accessible upon successful registration at www.humanconnectome.org. This neuroimaging data includes not only rfMRI but also diffusion MRI (dMRI) with tractography analysis, task-evoked fMRI (tfMRI) and magnetoencephalography (MEG). Getting access to restricted data elements: family structure (twin or non-twin status), age and handedness requires the acceptance of the HCP Restricted Data Use Term.

The first subset of the whole target samples were released on March, 2013. To date, HCP released the entire data sets for 1,206 subjects for a total of more than 64 terabytes via ConnectomeDB [27], a data management system based on XNAT. Similar to FCP, the HCP data is also made available on Amazon S3 to allow users to process and analyze the data directly through Amazon Web Services (AWS), a cloud-based data processing. Instead of downloading all the datasets, one can also order the data on eight 8-terabyte hard drives (the so-called Connectome in a Box). In addition, a set of software packages are provided as part of the projects involving the HCP minimum preprocessing pipeline scripts [26].

This project currently served as a baseline for many new large-scale data sharing projects. The new projects are built upon the HCP by using the same data acquisition and analysis. For example, the Developing Human Connectome Project (dHCP)[9] is a study of human brain connectivity from 20 to 44 weeks post-conceptional age; the Baby Connectome Project (BCP) for children from birth through five years of age, the Lifespan Human Connectome Project (L-HCP)[10] for different age groups across the lifespan (4-6, 8-9, 14-15, 25-35, 45-55, 66-75) [28]. In addition to healthy subjects, more than ten projects are funded to study connectomes related to human disease.

2.3 Challenges of Big rfMRI Data

Data gathered for the FCP and the HCP does exhibit several big data quantities (V's definitions [29]). Although the size of rfMRI data is not as big as other forms of data (such as genome sequencing data), these shared large-scale datasets are big enough that a single computer cannot process them. In other words, this rfMRI data does exhibit *Big Volume*. Many methods for preprocessing rfMRI data and functional connectivity have been designed when the data size is not really big. These approaches thus have difficulty in handling the large-scale data (e.g. PCA [30], [31]). Considering the FCP and the HCP data has only recently been released, and only few recent methods are able to handle large-scale rfMRI data, research based on this data is considerably new. Novel methods capable of analyzing such data should be developed either by modifying traditional methods that rely on parallel computing environment or by proposing new methods that work naturally on a parallel computing or a cloud computing environment.

Big Variety refers to the diversity of information within a single big rfMRI dataset (intra-dataset variety) or the diversity of multiple rfMRI datasets (inter-dataset variety). Big Variety can also occur when rfMRI data is analyzed together with other neuroimaging data and behavioral data. This is a critical stage in Big Data research since it is widely acknowledged that no single big data set should be considered to be true, and thus cross-validation of several imaging modalities is necessary. Thanks to the HCP, it involves multiple imaging modalities (rfMRI, tfMRI, dMRI, MEG) allowing investigators to apply multimodal data integration techniques to improve the reliability and robustness of the results [32]. Big data sharing projects that are focused primarily on sharing of other MRI data types are the OpenfMRI project[11] (which are focused primarily on sharing of tfMRI) and the Open Access Series of Imaging Studies (OASIS) project[12] (which has shared more than 500 subjects worth of structural MRI data). For the OpenfMRI project, the number of currently available subjects across 63 datasets is 2,158. Furthermore, HCP also provide different types of preprocessed fMRI data ranging from unprocessed NIfTI images, minimally

1. http://fcon_1000.projects.nitrc.org/indi/IndiRetro.html
2. http://fcon_1000.projects.nitrc.org/indi/abide/
3. http://fcon_1000.projects.nitrc.org/indi/adhd200/
4. http://fcon_1000.projects.nitrc.org/indi/CoRR/html/index.html
5. http://fcon_1000.projects.nitrc.org/indi/IndiPro.html
6. http://fcon_1000.projects.nitrc.org/indi/enhanced/
7. www.xnat.org
8. https://aws.amazon.com/s3/
9. www.developingconnectome.org
10. http://lifespan.humanconnectome.org
11. https://openfmri.org
12. www.oasis-brains.org

preprocessed NIfTI images, ICA denoised rfMRI data to functional connectivity data. This increases the degree of utility and flexibility to re-analyze the data for investigators, as compared to coordinate-based data and statistical maps (which typically included in most neuroimaging papers or are available through several data sharing projects such as BrainMap[13], Neurosynth[14], SumsDB[15] and NeuroVault[16]).

Big Veracity refers to the noise, incomplete, inconsistent or erroneous in data. Although big data is very good at detecting correlation especially subtle correlations that might miss by analyzing smaller datasets, scientists are likely to find many statistically significant correlations every time looking on larger dataset and thus scientists should be very aware of which correlations are meaningful. This is due to the fact that in large-scale datasets, large deviations are more attributable to variance (or noise) than to real information (or signal). Specifically, non-neuronal fluctuation in rfMRI data can increase the apparent functional connectivity between brain regions (i.e., increasing an opportunity to find spurious and/or fluke correlations) by introducing spurious common variance across rfMRI time-series. Data preprocessing is thus necessary and is a crucial stage in Big Data research. Several preprocessing steps are progressively becoming more accepted as standard in the analysis of rfMRI data, although these advanced techniques used in data preprocessing pipelines often dramatically increase the computational burden. A new software suit that is capable of preprocessing big data using advanced analytic techniques should be developed. The reduction of data is another crucial stage especially when dealing with large-scale data sets with Big Veracity, that is, discriminating relevant and meaningful features using selection or extraction methods from the whole set of features which potentially contains irrelevant, redundant and noisy information. These tasks can also be done using Topological Data Analysis. This approach is not only reducing the effect of negative elements but also reducing the amount of storage space required.

Big Velocity could come from prospective rfMRI data in a research setting. Big Velocity also occurs when data is coming in and processing at higher speed such as a real-time monitoring of a patient's current condition in a clinical setting [33].

3 Big Data Preprocessing Pipelines

Before applying any rfMRI technique for investigating functional connectivity, several data preprocessing steps need to be performed to remove all unwanted effects in rfMRI data and also increase the possibility of observing neural effects. This large number of inter-connected preprocessing steps collectively referred to as a pipeline (or workflow). So far there is no agreement on what constitutes the optimal data preprocessing pipeline nor how to select the best pipeline given a specific intended application. Most studies use their own specific pipeline, often defined by the experimenters' personal preference, or by the defaults of the software package used. No consensus thus has been found across the studies [12]. Further, it is widely acknowledged that different versions of preprocessing pipelines can affect the results obtained from statistical group difference tests and classification models [13]. Three important characteristics that have been changed from one rfMRI study to another study are: (1) which preprocessing steps are applied; (2) in what order; and (3) their values of parameters involved in certain steps. Due to the large number of possible combinations, it is difficult to evaluate all of them on big rfMRI datasets. There have been few systematic approaches assessing the effect of different preprocessing pipelines applied in rfMRI methods to study functional connectivity of the brain, particularly in large-scale datasets [14], [15]. In this paper, we present three alternative ways to access big preprocessed rfMRI data: (1) the minimal preprocessing pipelines; (2) the Preprocessed Connectomes Project; and (3) the software packages for big rfMRI data. Each of which has its own advantages and disadvantages depending on the type of analysis.

3.1 Minimal Preprocessing Pipelines

Although the unprocessed NIfTI (Neuroimaging Informatics Technology Initiative [34]) data is available through data sharing projects, these projects anticipate that investigators will prefer to use the preprocessed data obtained from the minimal preprocessing pipelines developed by their team members. The principal goal of the minimal preprocessing pipelines is to provide rfMRI data with a minimum standard of data quality while the amount of information actually removed from the data is minimized. This minimally preprocessed data could be used as the starting point for any analysis. This is particularly advantageous for investigators who lack sufficient computational resources to preprocess large-scale datasets.

To obtain optimal results, however, it is important to apply further preprocessing steps which are dependent on the rfMRI methods used and/or characteristics of the data acquisition (in case of applying these pipelines on their own data). The notable minimal preprocessing pipelines are the ones implemented in the data sharing projects like the HCP. Since the HCP minimal preprocessing pipelines [26] are specially designed to their own specific data acquisition protocols, any study that would like to use the HCP minimal preprocessing pipelines requires their minimum data acquisition protocols. The interesting characteristic of the HCP acquisition system is the use of the fast repetition time (TR) sampling based multiband pulse sequences. Based on this approach, all slices acquired in each volume are very close together (as compared to typical fMRI acquisition system) and thus it is not necessary (but still optional) to carry out slice timing correction in the HCP pipelines.

Specifically, the HCP minimal preprocessing pipelines for functional preprocessing pipelines consist of correction of gradient-nonlinearity-induced distortion, realignment of the time-series to correct for subject head motion, registration of the fMRI data to the structural data, reduction of the bias field, normalization of the 4D image to a global mean, masking the data with the final brain mask, and the spatial smoothing using a novel geodesic Gaussian surface smoothing algorithm with 2 mm FWHM [26]. Preprocessing steps that may remove significant amounts of

13. www.brainmap.org
14. www.neurosynth.org
15. http://sumsdb.wustl.edu/
16. http://neurovault.org

information (e.g. temporal filtering, significant spatial filtering, nuisance signal regression, and movement scrubbing) are not included in these pipelines. For instance, although high frequencies have been commonly related to nuisance signals [35], some studies suggest that there is important information contained in high frequencies (0.1 to 0.5 Hz) [36]. Therefore, the preprocessing steps that still remain a topic of debate are generally excluded from the minimal preprocessing pipelines. It is interesting that the HCP minimal preprocessing pipelines include the field map distortion correction step which in practice is often neglected instead.

For the FCP data, only three simple preprocessing steps have been performed comprising of NIfTI format conversion, uniform orientation placement and the first-5-time-points removal. These few preprocessing steps may not be sufficient to yield the minimum data quality standard, and further preprocessing steps may be necessary. Further, besides the minimal preprocessing pipelines implemented in the data sharing projects, a few software packages provide the minimal preprocessing pipelines as an option, such as SPM and C-PAC. Note that the full name of software tools and packages can be found in Table 1 and 2. The contributions of these tools and packages have been presented throughout this section. More details about their principles as well as the pros and cons can be found in [37] and [38].

3.2 Preprocessed Connectomes Project

The principal goal of the Preprocessed Connectomes Project (PCP)[17] is to provide systematically preprocessed rfMRI data from the FCP and the INDI databases using different preprocessing pipelines. This is due to the fact that there is no consensus on the best preprocessing pipelines in this research field. Different preprocessing choices will allow investigators to compare the results and consequently will lead us to find the best preprocessing strategies later. Another reason behinds this project is to broaden the range of investigators who can access to the large-scale rfMRI data. Each of which was implemented using the chosen parameters and default settings of commonly used preprocessing pipeline softwares. All the preprocessed data is available on the Neuroimaging Informatics Tools and Resources Clearinghouse (NITRC) and on the Amazon S3 bucket.

It is interesting that the preprocessing steps implemented by the different common software suits are quite similar although the specific algorithms and their parameters used in each of the steps may vary, as can be observed in Table 3. This is due to the fact that most of them are developed by integrating several common brain imaging tools for functional and structural preprocessing together. A list of neuroimaging tools for general and wide-ranging purposes used by the preprocessing pipeline and functional connectivity software packages related to rfMRI analysis is presented in Table 1. For instance, CCS [49] builds upon a set of three main available tools: AFNI, FSL and FreeSurfer together with in-house developed functions while C-PAC [50] is developed by integrating many functions from three tools including AFNI, FSL and ANTS. Likewise, a general purpose software tool named SPM has been used as a basis for building many software suits with more specific purposes such as BrainVISA, CONN, cPPI, gPPI, SEM, SnPM, and TDT (for more details, see Table 2).

The first preprocessed data in this project is from the ADHD-200 data. This data was preprocessed by three different pipelines: the Athena pipeline[18] (using AFNI and FSL), the NIAK pipeline[19] (using NIAK on CBRAIN), and the Burner pipeline[20] (using SPM). The forthcoming release is the preprocessed data using the CIVET pipeline[21] [71]. It should be noted that the CBRAIN platform[22] is a web-based collaborative research platform that allows investigators to integrate large neuroimaging data resources, preprocessing and analysis software tools as well as high-performance distributed computing facilities together within a controlled, secure environment [72]. Other datasets from the FCP and the INDI databases have been added later including the Beijing Enhanced Diffusion Tensor Imaging dataset, the Neurofeedback Skull-stripped repository and ABIDE. For the preprocessed ABIDE data, four different software packages were used involving CCS, C-PAC, DPARSF and NIAK. Besides the default settings used in each software suit (Table 3), two preprocessing steps that still remain a topic of debate, i.e., temporal filtering (0.01-0.1 Hz) and global signal regression, were included and excluded, which provide four different preprocessing strategies for each pipeline. Further, statistical derivatives (e.g. amplitude of regional homogeneity (ReHo) [73], low frequency fluctuations (ALFF) [74] and fractional ALFF (fALFF) [75]) were also calculated from each of the preprocessing data sets using the C-PAC software.

3.3 Software Packages for Big rfMRI Data

A number of requirements and features are necessary to be offered by the pipeline softwares designed to handle large-scale rfMRI data such as configuration, robust, reliable, extendable and provenance tracking (e.g. [49], [66]). Currently, there are some progress toward parallelization for the three major neuroimaging software tools: SPM, FSL and AFNI. By performing them with an additional package (such as Condor) or platform (such as OpenMP), some functions can then be executed in parallel on several central processing unit (CPU) cores or on several computers. However in common neuroimaging tools (Table 1), parameters may need to be manually set step-by-step and subject-by-subject which will be time-consuming and not suitable for big data analysis. Many preprocessing pipeline software suits then have been developed to provide a user-friendly environment (Table 2). Unfortunately, only few of them have been mainly designed to preprocess and analyze big data.

Parallel computing capacity may be considered as the most important feature which developers have paid attention to. In order to preprocess a total of 418 subjects from the NKI-RS datasets, for example, the CCS pipeline

17. http://preprocessed-connectomes-project.org

18. www.nitrc.org/plugins/mwiki/index.php/neurobureau:AthenaPipeline

19. www.nitrc.org/plugins/mwiki/index.php?title=neurobureau:NIAKPipeline

20. www.nitrc.org/plugins/mwiki/index.php?title=neurobureau:BurnerPipelineg

21. www.nitrc.org/plugins/mwiki/index.php?title=neurobureau:CIVETPipeline

22. www.cbrain.mcgill.ca

TABLE 1
A List of Neuroimaging Software Tools for General Purposes

Software	Full name	Programming Languages	Availability
AFNI [39]	Analysis of Functional NeuroImages	C	https://afni.nimh.nih.gov/afni/
ANTs	Advanced Normalization Tools	C++	http://stnava.github.io/ANTs/
FreeSurfer [40]	FreeSurfer	C/C++/Shell	https://surfer.nmr.mgh.harvard.edu
FSL [41]	FMRIB Software Library	C++/Shell	https://fsl.fmrib.ox.ac.uk/fsl/fslwiki
SPM [42]	Statistical Parametric Mapping	MATLAB/C	www.fil.ion.ucl.ac.uk/spm/

TABLE 2
A List of fMRI Preprocessing Pipeline and Functional Connectivity Software Packages

Software	Full name	Programming Languages	Availability
BASCO [43]	BetA-Series COrrelation	MATLAB	www.nitrc.org/projects/basco/
BCT [44]	Brain Connectivity Toolbox	MATLAB	www.brain-connectivity-oolbox.net/
Biananes [45]	Scalable fMRI Data Analysis	Scala/C	https://github.com/rboubela/biananes
BrainNet Viewer [46]	BrainNet Viewer	MATLAB	www.nitrc.org/projects/bnv/
BrainVISA [47]	BrainVISA	Python/C++	http://brainvisa.info/web/index.html
BROCCOLI [48]	Software for Fast fMRI Analysis on Many-Core CPUs and GPUs	OpenCL/C++	https://github.com/wanderine/BROCCOLI/
CCS [49]	Connectome Computation System	MATLAB/Python /R/Shell	https://github.com/zuoxinian/CCS
C-PAC [50]	Configurable Pipeline for the Analysis of Connectomes	Python	https://fcp-indi.github.io
CONN [51]	Functional Connectivity Toolbox	MATLAB	www.nitrc.org/projects/conn/
cPPI [52]	Correlational Psychophysiological Interaction	MATLAB	www.nitrc.org/projects/cppi_toolbox/
DPABI [53]	a toolbox for Data Processing and Analysis for Brain Imaging	MATLAB	http://rfmri.org/dpabi
DPARSF [54]	Data Processing Assistant for Resting-State fMRI	MATLAB	http://rfmri.org/DPARSF
GAT [55]	Graph Analysis Toolbox	MATLAB	www.nitrc.org/projects/gat/
GIFT [56]	Group ICA Of fMRI Toolbox	MATLAB	http://mialab.mrn.org/software/gift/index.html
gPPI [57]	Generalized Psychophysiological Interactions	MATLAB	www.nitrc.org/projects/gppi
GraphVar [58]	A user-friendly toolbox for comprehensive graph analyses of functional brain connectivity	MATLAB	www.nitrc.org/projects/graphvar/
GRETNA [59]	GRaph thEoreTical Network Analysis	MATLAB	www.nitrc.org/projects/gretna/
GTG [60]	Graph Theory GLM MATLAB Toolbox	MATLAB	www.nitrc.org/projects/metalab_gtg/
NBS [61]	Network-Based Statistic	MATLAB	www.nitrc.org/projects/nbs/
NIAK [62]	Neuroimaging Analysis Kit	MATLAB/Octave	www.nitrc.org/projects/niak/
Nilearn [63]	Machine learning for Neuro-Imaging in Python	Python	https://nilearn.github.io
Nipype [64]	Neuroimaging in Python: Pipelines and Interfaces	Python	http://nipy.org/nipype
PRoNTo [65]	Pattern Recognition for Neuroimaging Toolbox	MATLAB/C++	www.mlnl.cs.ucl.ac.uk/pronto/index.html
PSOM [66]	Pipeline System for Octave and MATLAB	MATLAB/Octave	http://psom.simexp-lab.org
PyMVPA [67]	MultiVariate Pattern Analysis in Python	Python	www.pymvpa.org
REST [68]	Resting-State fMRI Data Analysis Toolkit	MATLAB	http://restfmri.net
SEM	Structural Equation Modelling	MATLAB	http://dslink333.dyndns.org/SEM.htm
SnPM [69]	Statistical NonParametric Mapping	MATLAB	http://warwick.ac.uk/snpm
TDT [70]	The Decoding Toolbox	MATLAB	https://sites.google.com/site/tdtdecodingtoolbox/

TABLE 3
Chosen Parameters and Default Settings of Four Different Functional Preprocessing Pipelines for the ABIDE Data (PC: Principal Components, WM: White Matter, CSF: CerebroSpinal Fluid)

Preprocessing step	CCS	C-PAC	DPARSF	NIAK
Drop the first few volumes	4	0	4	0
Slice timing correction	Yes	Yes	Yes	No
Motion realignment	Yes	Yes	Yes	Yes
Intensity normalization	4D Global mean to 10,000	4D Global mean to 10,000	No	Non-uniformity correction using median volume
Nuisance signal regression (Head motion)	Friston's 24-parameter motion signal	Friston's 24-parameter motion signal	Friston's 24-parameter motion signal	Scrubbing and the 1st PC of 6 parameters and their squares
Nuisance signal regression (Tissue signals)	Mean WM and CSF signals	The first 5 PCs from WM and CSF signals	Mean WM and CSF signals	Mean WM and CSF signals
Nuisance signal regression (Low-frequency drifts)	Linear and quadratic trends	Linear and quadratic trends	Linear and quadratic trends	Discrete cosine basis with a 0.01 Hz high-pass cut-off

took approximately 15,000 CPU h in the Dell Blade Cluster System [49]. Thus pipelines that can execute jobs in parallel on a multi-core machine or a supercomputer are needed, which allow us to reduce the total time necessary to complete an analysis. C-PAC and PSOM are two common big data processing software packages. These softwares link together many functions from the common neuroimaging tools into pipelines that can execute in a single run on high-performance computing architectures, after a proper configuration has been set. Bellec et al. [66] tested the performance of the PSOM framework using the ADHD-200 datasets and showed that we could reduce the processing time for 198 subjects (with a total data size of 7.7 gigabytes and 5,153 jobs included in the NIAK pipeline) from over a week down to less than 3 h with 200 computing cores.

PSOM also offers other two important features that allow us to handle with big data, i.e., fault tolerance and smart updates. Specifically, PSOM will run each job for multiple attempts before considering it as a failed job while all the failed jobs can be automatically restarted after the pipeline termination by the investigator. Furthermore, if the restart of an analysis is needed, only the parts of the pipeline that need to be reprocessed or impacted by the changes will be executed which can be detected automatically by the toolbox. These two features are very useful particularly in the development phase (e.g. selecting the optimal algorithms and parameters of the pipeline) since the pipelines may be needed to restart multiple times at several stages. However, this framework does not focus on pipeline mapping and this key feature is performed by interfacing PSOM pipelines to another software tool with powerful pipeline mapping capabilities such as CBRAIN instead.

Another group of interesting big data processing software suits is the one designed to enable the advantages of parallel computing with a special emphasis on using inexpensive and powerful graphics processing units (GPUs). BROCCOLI [48] is one of the softwares in this group which is written in OpenCL (Open Computing Language). This makes BROCCOLI able to run the analysis in parallel. To test the parallelization efficiency of BROCCOLI, Eklund et al. [48] have run several benchmark experiments on a number of open access fMRI datasets with three different hardware configurations (i.e., an Intel CPU, an Nvidia GPU, and an AMD GPU). As compared the results for non-linear spatial normalization as an example with other three major neuroimaging tools, BROCCOLI with an Nvidia GPU can run 525 times faster than FSL and AFNI and 195 times faster than AFNI with OpenMP [48]. The results clearly support that parallel processing of the rfMRI data can lead to significantly faster analysis pipelines, which is very important for big data analysis. However, several limitations of this software suit are acknowledged. For instance, BROCCOLI does not provide a graphical user interface. Since this software suit is implemented using OpenCL, it performs best for Nvidia GPUs and thus code optimization for other hardware platforms (e.g., Intel and AMD) is necessary. Biananes [45] is another software in this group which uses GPUs to compute the voxel-wise correlation/connectivity matrix in the highest HCP resolution of all in-brain voxels. This software also provides a distributed file reader for 4D NIfTI fMRI data for use in an Apache Spark environment. By using a scalable platform [45], [76], [77], we can move data analysis and computational tasks to cloud service providers, for example the AWS cloud which can run the Spark Framework with the GPU accelerated computation.

3.4 Challenges of Data Preprocessing Pipelines

The first two alternative approaches could be consecutively used as the first and the second starting points for investigators who would like to perform functional connectivity analyses on big rfMRI data but do not have enough sufficient computational resources to acquire or preprocess large-scale data, or those who prefer to focus on data analysis rather than data acquisition and preprocessing. As previously mentioned the minimally preprocessed data provides a minimum standard of data quality while greater amount of information is still contained in the data. If further preprocessing steps are necessary, preprocessed data from the PCP that was prepared using the default chosen parameters and settings of several common preprocessing software suits would be a safe bet for further data analyses as they would represent peer-reviewed accepted preprocessing implementations. Investigators can choose one of the pipelines that is appropriate to their application or even compare the results across the different pipelines.

On the other hand, if investigators have enough resources to preprocess large-scale rfMRI data, they could use one of the software packages designed for preprocessing large-scale rfMRI data as discussed in the third alternative approach. They could also consider to preprocess their own data using the minimum preprocessing pipelines and/or the default preprocessing pipelines from common software packages as the starting points. However, several modifications and additional steps may be required to make the pipeline more suitable to the challenge of unique characteristics of specific rfMRI data and functional connectivity analysis methods proposed. For example, the dHCP minimal preprocessing pipelines which are developed based on the HCP have modified several preprocessing steps in order to preprocess the data with low and variable contrast and high levels of head motion in neonate acquisition [78]. To yield the valid and optimal results for specific application, a comprehensive investigation of optimal preprocessing steps and parameter values is necessary.

There are several specific areas in which the preprocessing pipelines need to be improved, and novel methods will continue to be developed. Since there is currently no solution to find the best preprocessing pipelines, data preprocessing steps that are consensus across common software pipelines and/or high-quality peer-reviewed research studies by using a systematic review and meta-analysis could be one of the solutions. For instance, most recently, Caballero and Reynolds [79] suggested some guidelines to choose the preprocessing steps and their order. Specifically, the preprocessing pipeline could start by despiking the fMRI data and then applying a block of operations involving physiological noise correction, slice timing correction, volume registration and correction of magnetic field distortions. The choice of the order within this block is still controversial and they recommend to integrate these four operations into a unified framework. Next, the alignment of the subject's anatomical image to the functional data could be performed. The final steps consist of spatial smoothing, and the combination of nuisance regression, temporal filtering and censoring. The nuisance regressors can be defined either on anatomical masks or by data decomposition techniques such as PCA, kernel PCA and ICA. An additional advantage of data driven approaches is that it can also reduce multiple noise fluctuations simultaneously. However, it has been suggested that for example spatial ICA cannot completely separate physiological noise components and denoising physiological noise based on external recordings is necessary prior to ICA decomposition. In future studies, more comprehensive investigations are still needed to determine better evidence-based recommendations and best practices for minimal and/or optimal preprocessing pipelines.

Further, as it is acknowledged that data preprocessing pipelines can affect the final results obtained from statistical group difference tests and classification models, and there have been very few systematic studies investigating these effects, a better understanding of whether and which preprocessing steps and parameters affect the results derived from any analysis method is warranted. This is also very important to determine the best, or the optimal, preprocessing pipelines. For example, Vergara et al. [14] evaluated the effect of several preprocessing pipelines in the detection of abnormal functional network connectivity and the classification of patient and control using group ICA methods. Four different pipelines were tested with special emphasis on the effects of (1) the order of head motion correction: before or after group ICA applied, and (2) temporal filtering to remove relatively high frequency content. Both experimental and simulation data was used. For real data, two different cohorts were included in the study: one cohort is mild traumatic brain injury patients with controls and the other cohort is smokers and non-smokers. The results of this study show that data preprocessing pipeline can change the final results. That is, if motion correction is applied before group ICA, patient-control group differences are increased as well as correlation with behavioral assessments are stronger.

Andronache et al. [15] evaluated the effect of several preprocessing pipelines in the detection of the DMN using the SCA and ICA methods. Five different pipelines were tested by adding several preprocessing steps (e.g. removal of co-variance with movement parameters, band-pass filtering, etc.) to the minimum preprocessing pipelines (i.e., realignment, slice timing correction, normalization to MNI space, and spatial smoothing). Only the real data was used in this study including patients with disorders of consciousness and their control counterparts. The results support the study of Vergara et al. [14] that data preprocessing pipeline can change the final results. The results of this study also show that different functional connectivity methods (SCA and ICA) are affected by data preprocessing pipelines differently. Although the effect is reduced when extensive preprocessing steps are applied, it may be due to the fact that some meaningful variability in the data is removed and the valid results are not obtained. The effect of preprocessing pipelines on other commonly used or novel analysis methods should be investigated in future studies.

4 RFMRI TECHNIQUES

Functional Magnetic Resonance provides complex signals to study the highly variable and entangled activity of the brain. Being able to parse it and extract meaningfulx information is one of the great challenges of neuroimaging research. We can broadly identify two main types of analysis: one focuses on identifying functionally independent brain regions, or functional subnetworks, usually associated to specific functions; a second one focused instead on the relational among the activities of sets of regions. Classic examples of the first approach as decomposition techniques, like ICA and PCA which we already mentioned in previous sections. Here we put our focus on the second type, with particular attention The most relevant examples are techniques that produce simplified topological representation (e.g. Mapper [80], [81]), graph-theoretic and network tools amenable to statistical mechanical treatments [82], and finally full fledged topological data analysis tools, in particular persistent homology [83]. In the following, we briefly illustrate the merits of each and their relevance for big data analysis.

4.1 Mapper Algorithms and Data-Driven Methods

Mapper, first introduced by Singh et al. [81], is one of the most used topological tools for direct data exploration. Its

TABLE 4
An Overview of Available Softwares for Mapper and Persistent Homology. We provide here a minimal overview and a list of references to existing softwares for TDA, with a short description of the respective advantages and limits. We refer to [84] for a thorough review including computational performances and scalings with dataset size relevant to big data analysis.

Software	Programming Languages	Main Features	Ref./URL
MapperTools	Python	Clean implementation, two-dimensional filters, easy access to metadata.	[85]
PythonMapper	R	Clean implementation, ease of use, recently revamped	[86]
Javaplex	Java	Persistent and zigzag homology, various filtrations available, provides homology generators. Lims.: rather slow and memory-taxing	[87]
Perseus	C++	Based on Morse reductions, various filtrations available. Lims.: no generators	[88], [89]
jHoles	Java	Fast preprocessing, weighted network homology. Lims.: single application, not very versatile	[90], [91]
Dionysus	Python/C++	Persistent and zigzag homology, vineyards, various filtrations available, provides homology generators Lims.: sparsely documented and compilation issues	[92]
Phat/Dipha	C++	Fast, parallel implementation of simplicial and cubical homology. Lims.: no generators	[93], [94]
Gudhi	C++	Multi-field homology, various filtrations available. Lims.: no generators	[95], [96]
Ripser	C++	Fast computation of VietorisRips persistence barcodes. Lims.: no generators	[97]

fundamentally new character, shared with persistent homology, comes from its algebraic foundation: its recovers the shape of topological spaces at the mesoscopic scale by going beyond the standard measures defined on data points' pair. Given a point cloud dataset, typically in high dimensions, one begins by dividing the space into a set of overlapping slices. Within each of thses a local clustering algorithm is performed to partition the points in a set of separate clusters. Since the slices are overlapping, there will be common points between adjacent ones. One can then build a topologically simplified skeleton of the original dataset by joining together clusters that belong to adjacent slices and that have non-empty intersection (i.e., that contain some of the same points across the two slices) [80]. This type of approach is guaranteed to preserve the overall topology via the gluing of local clusterings.

Mapper lends itself to the analysis of very large datasets, because the complete problem (e.g. the overall clustering structure) is subdivided in any number of smaller local problems (i.e., the clusterings within slices), which can be run in parallel and that are merged only at the final step. Moreover, the local clusterings depend only on the distances between the points in the slices, hence also high-dimensional data are projected effectively down to a (typically small) distance matrix. These properties make Mapper a very good tool for the analysis of large-scale data as this approach can be naturally performed in a framework of big data analysis such as the Google's MapReduce paradigm [98].

Despite the useful properties of Mapper, to our knowledge only one recent study has leveraged it for the study of rfMRI data. Kyeong et al. [99] used the Mapper algorithm to investigate the relationship between brain functional connectivity and characteristics of ADHD (from the ADHD-200 datasets). Because ADHD is defined as a single disorder without subtypes [100], thus the topological network obtained from the Mapper algorithms is presented as a long gradual progression. Although this study does not show the clustering potential of the Mapper algorithm to identify meaningful subtypes, the resulting topological network of the Mapper algorithm can significantly distinguish patients with ADHD from normal control subjects (P-value < 0.0005). Moreover, the results obtained using the Mapper algorithm should be the same either the rfMRI data was preprocessed with or without scrubbing the time points that showed large head motions since the values of the chosen objective measure are almost the same ($r = 0.99$). This study supports the useful properties of the Mapper algorithms, and warrants the potential of Mapper for future studies of brain function connectivity and characteristics of many brain disorders and diseases.

To discuss this in more details, standard clustering approaches for rfMRI work by constructing a series of spatially (or ICA-) coherent coarse-grained regions [1] that are then thought as nodes for a similarity or correlation network. However, Zuo and Xing [101] strongly recommend voxel-wise analysis because the analysis of the signal averaged from multiple voxel based on anatomical structure can lead to difficulties in the reliability and interpretation of derived results. The clustering of activity time-series obtained during rfMRI is the direct and natural application of the Mapper algorithm. Thanks to their scalability, Mapper approaches would be able to address directly high-resolution voxel-level datasets without the need for any preliminary coarse-graining of the regions or resampling data to a lower isotropic resolution and would be able to yield a fully functional representation. Thus we can use clustering-based mapper algorithms instead of existing slower methods used for rfMRI studies: hierarchical clustering [102], spectral clustering, k-means clustering, or fuzzy clustering [103].

Further, clustering is considered as an exploratory data-driven approach which is used to overcome the limitation of model-based analyses (e.g. SCA, ReHo, ALFF and fALFF). Despite serving similar purposes as other common data-driven methods such as ICA and PCA, a comparison between several different clustering and ICA methods in a systematic fMRI study [104] showed that clustering outperforms ICA (i.e., the most frequency used method for rfMRI studies [105]) for classification purposes. While the efficacy of PCA is strongly dependent on assumptions of linearity,

normality, and high SNR of the rfMRI data, clustering-based mapper algorithms are free from these assumptions and have achieved to extract non-trivial qualitative information from large-scale datasets (e.g. extracting a previously unknown subtype of breast cancer with a unique mutational profile and excellent survival [106]).

Note also that the output of Mapper depends critically on the chosen slicing of the original dataset. In other words, choosing the slicing defines what will be the interpretation of the resulting network. This opens the door to combining the full set of existing data-reduction and data-analysis techniques with Mapper. For example, by using the projections of the dataset along main directions obtained by (group) PCA, ICA, or similar decomposition techniques [30], [31], [105], that is using information that is fully contained within the dataset itself; it is however also possible to augment this information by including in the slicing function meta-information about the subjects under study, making this tool extremely versatile for both data exploration and feature extraction in large complex datasets.

4.2 Graph Theory and Networks

Graph theory is the mathematics of networks which describe pairwise relationships [107], as sets of nodes and links, usually equipped with a weight. Networks, thanks to their expressive power and simplicity, have become over the last decade into one of the most popular tools to describe both the brain's physical structure and its patterns of activity [108]. Indeed, via network representations it has been possible to uncover a large set of properties of brain function that previously could hardly be described: among others, for example, we now know that specific functional subnetworks correspond to known cognitive and sensory modalities [109], that the observed robustness of the brain to lesions and perturbations is rooted in the combination of small-worldness and strong local clustering coefficient displayed by real-world networks [110], or that information in the brain is processed in tightly integrated modules and then shared across longer distances via long-range links [111]. Until recently, most of the research in functional network however focused on small-size parcellations because they provided anatomically interpretable descriptions and also facilitated the computation of graph metrics, which can often be rather cumbersome computationally. This trend is changing however due to the combined effect of increased computational power, optimised network analysis libraries [112], [113] and accurate measurements. For example, the first tools to analyze large-scale neural network data over Spark architectures [114], as well as scalable techniques able to process, analyze, correlate fMRI data at the full-voxel matrix level [115], are being developed, allowing de facto the scaling of network techniques to the scale of big data. Despite their success, networks however can only describe many-body interactions as the sum of pairwise interactions, an assumption that is not always verified and that, in some applications, can provide a biased representation of the system under study.

4.3 Persistent Homology

One progressively more popular answer to the need to describe higher-order interactions is given by another TDA technique, persistent homology. It yields deeper, quantitative information about the shape of a dataset than that obtained through Mapper, and allows richer descriptions than those provided by networks, at the cost of increased interpretative complexity. Persistent homology works by building a multi-scale summary of a whole dataset via a series of progressively finer approximations, called filtration, of the relation between neighbourhoods of points. Filtration is the key point in order to consider all possible thresholds, avoiding one of the main cons in graph theory. In addition, persistent homology is phrased in the language of simplicial complexes that, by construction, describe many-body interaction patterns and thus go beyond the network description based on two-point interactions (i.e., edges defined on two points, simplices are generic sets of points) [1]. For this reason, it has found wide application in neuroscience with direct applications to the study of rfMRI correlation networks for healthy [116], [117], [118] and altered [119] or pathological [120] brain states, models of spatial learning [121], [122], and dynamical functional connectivity [123].

Indeed, even when starting directly from network data, persistent homology is able to provide information that is not easily –or sometimes at all– available from the standard combinatoric or statistical mechanical point of view, e.g. topological distances defined via persistence diagram useful in discriminating between brain network [124] and multi-scale network descriptions, i.e., that do not require choosing a threshold, of the functional network yielding discrimination power that was absent from a pure graph-theoretic perspective [119], [125].

Interestingly, once topological features are detected, statistical mechanical methods can give an important contribution to their interpretation, e.g. via projections to simpler representation (e.g. scaffolds [116]), and the modeling of what should be considered significant structure and what noise, e.g. by constructing minimal topological random null models [126], [127], [128].

One of the main limits for the application to large datasets is however that persistent homology can be computationally cumbersome if computed naively. However, recent algorithmic advances have significant reduced its complexity and parallel algorithms have become available (such as a spectral sequence algorithm [129], a chunk algorithm [93], [94] and a number of others (e.g. [130], [131], [132], [133])). As a result, persistent homology can be now used to approach very large, high-dimensional data sets, for example fMRI data.

Furthermore, there have been recent advances in methods to compare the information obtained from persistent homology across subjects and groups: the persistence landscape, introduced by Bubenik et al. [134], allows the direct comparison of the persistence profiles of different subjects, while kernelization techniques [135], [136] will allow to apply machine-learning techniques to the persistent homology. Persistent homology, while very promising, is still in its infancy as a branch of data science. It provides a radically new perspective on how we approach data and brings with itself a new language grounded in algebraic topology. However, there are still open challenges in order to fully leverage its potential in the study of large rfMRI datasets. The first and most obvious one is the necessity to keep improving

the computational scalability of persistent homology. While topological simplification via Mapper is cheap and scalable, it also does not directly yield the quantitative output that persistent homology provides. It is then paramount to improve further on the existing implementations, in particular in the direction of effective simplicial complex reduction schemes preserving not only the topological information at the global level, but also the actual localization of homology classes [130]. A second challenge is lowering the entry cost for practitioners coming from outside the TDA community and seeking to apply these techniques to their specific case studies. Although the required mathematical background is significant, having user-friendly and well documented software packages dedicated to the fMRI analysis would already go a long way in this direction.

5 Conclusion

The era of "Biomedical Big Data" has arrived for the rfMRI research, thanks to the unrestricted sharing and open access of big neuroimaging data: the 1000 Functional Connectomes Project and the Human Connectome Project. These large-scale rfMRI data does exhibit the 5 V's of Big Data: Volume, Veracity, Variety, Velocity and Value. Thus, there is an urgent need to develop data preprocessing pipelines and analyses methods for big rfMRI data.

For data preprocessing pipelines, three alternative approaches to get access to big preprocessed rfMRI data were presented. If investigators would like to perform analyses on big rfMRI data but lack sufficient resources to acquire or preprocess them, or prefer to focus on data analysis rather than data acquisition and preprocessing, the first two approaches: the minimal preprocessing pipelines and the Preprocessed Connectomes Project are the good starting points for their own analysis. If investigators have enough resources to preprocess large-scale data, they can choose one of the software suits designed for preprocessing big data. However, a comprehensive investigation of the effects of data preprocessing steps on the results obtained from functional connectivity analyses as well as an extensive development of the new preprocessing software packages for large-scale data is highly necessary in future studies.

After rfMRI data has been preprocessed, there are several methods commonly used in rfMRI studies to examine functional connectivity such as SCA, PCA, ICA and clustering methods. To enable these approaches to identify large-scale brain networks, recently more sophisticated studies have been performed. However, we still should consider some limitations of the existing common methods, and a novel method is essential for big rfMRI data analysis. We proposed a technique called Topological Data Analysis to rs-fMRI functional connectivity. Many TDA properties clearly show the potential of different TDA methods to be used as big rfMRI data analyses methods. Clinical applications of rfMRI-based TDA should be explored in future studies.

Acknowledgments

The authors acknowledge the support of the ADnD project by Compagnia San Paolo.

References

[1] M. P. van den Heuvel and H. E. H. Pol, "Exploring the brain network: A review on resting-state fMRI functional connectivity," *Eur. Neuropsychopharmacol.*, vol. 20, no. 8, pp. 519–534, Aug. 2010.

[2] J. S. Damoiseaux *et al.*, "Consistent resting-state networks across healthy subjects," *Proc. Natl. Acad. Sci. U S A*, vol. 103, no. 37, p. 1384813853, Sep. 2006.

[3] H. H. Shen, "Core concept: Resting-state connectivity," *Proc. Natl. Acad. Sci. U S A*, vol. 112, no. 46, pp. 14 115–14 116, Nov. 2015.

[4] B. Biswal, F. Z. Yetkin, V. M. Haughton, and J. S. Hyde, "Functional connectivity in the motor cortex of resting human brain using echo-planar MRI," *Magn. Reson. Med.*, vol. 34, no. 4, pp. 537–541, Oct. 1995.

[5] Y. Nir *et al.*, "Interhemispheric correlations of slow spontaneous neuronal fluctuations revealed in human sensory cortex," *Nat. Neurosci.*, vol. 11, no. 9, pp. 1100–1108, Sep. 2008.

[6] M. E. Raichle, A. M. MacLeod, A. Z. Snyder, W. J. Powers, D. A. Gusnard, and G. L. Shulman, "A default mode of brain function," *Proc. Natl. Acad. Sci. U S A*, vol. 98, no. 2, pp. 676–682, Jan. 2001.

[7] M. D. Greicius, B. Krasnow, A. L. Reiss, and V. Menon, "Functional connectivity in the resting brain: A network analysis of the default mode hypothesis," *Proc. Natl. Acad. Sci. U S A*, vol. 100, no. 1, pp. 253–258, Jan. 2003.

[8] M. Fox and M. Greicius, "Clinical applications of resting state functional connectivity," *Front. Syst. Neurosci.*, vol. 4, p. 19, Jun. 2010.

[9] M. H. Lee, C. D. Smyser, and J. S. Shimony, "Resting-state fMRI: A review of methods and clinical applications," *AJNR Am. J. Neuroradiol.*, vol. 34, no. 10, pp. 1866–1872, Oct. 2013.

[10] B. B. Biswal *et al.*, "Toward discovery science of human brain function," *Proc. Natl. Acad. Sci. U S A*, vol. 107, no. 10, pp. 4734–4739, Mar. 2010.

[11] D. C. V. Essen *et al.*, "The Human Connectome Project: A data acquisition perspective," *NeuroImage*, vol. 62, no. 4, pp. 2222–2231, Oct. 2012.

[12] N. W. Churchill *et al.*, "Optimizing preprocessing and analysis pipelines for single-subject fMRI. I. Standard temporal motion and physiological noise correction methods," *Hum. Brain Mapp.*, vol. 33, no. 3, p. 609627, Mar. 2012.

[13] N. W. Churchill, G. Yourganov, A. Oder, F. Tam, S. J. Graham, and S. C. Strother, "Optimizing preprocessing and analysis pipelines for single-subject fMRI: 2. Interactions with ICA, PCA, task contrast and inter-subject heterogeneity," *PLoS ONE*, vol. 7, no. 2, p. e31147, Feb. 2012.

[14] V. M. Vergara, A. R. Mayer, E. Damaraju, K. Hutchison, and V. D. Calhoun, "The effect of preprocessing pipelines in subject classification and detection of abnormal resting state functional network connectivity using group ICA," *NeuroImage*, vol. 145, Part B, pp. 365–376, Jan. 2017.

[15] A. Andronache, C. Rosazza, D. Sattin, M. Leonardi, L. D'Incerti, and L. Minati, "Impact of functional mri data preprocessing pipeline on default-mode network detectability in patients with disorders of consciousness," *Front. Neuroinform.*, vol. 7, p. 16, Feb. 2013.

[16] K. Li, L. Guo, J. Nie, G. Li, and T. Liu, "Review of methods for functional brain connectivity detection using fMRI," *Comput. Med. Imaging Graph.*, vol. 33, no. 2, pp. 131–139, Mar. 2009.

[17] R. A. Poldrack and K. J. Gorgolewski, "Making big data open: Data sharing in neuroimaging," *Nat. Neurosci.*, vol. 17, no. 11, pp. 1510–1517, Nov. 2014.

[18] D. Tomasi and N. D. Volkow, "Functional connectivity density mapping," *Proc. Natl. Acad. Sci. U S A*, vol. 107, no. 21, pp. 9885–9890, May 2010.

[19] C. G. Yan, R. C. Craddock, X. N. Zuo, Y. F. Zang, and M. P. Milham, "Standardizing the intrinsic brain: Towards robust measurement of inter-individual variation in 1000 functional connectomes," *NeuroImage*, vol. 80, pp. 246–262, Oct. 2013.

[20] A. D. Martino *et al.*, "The autism brain imaging data exchange: towards a large-scale evaluation of the intrinsic brain architecture in autism," *Mol. Psychiatry*, vol. 19, no. 6, pp. 659–667, Jun. 2014.

[21] P. Bellec, C. Chu, F. Chouinard-Decorte, D. S. Margulies, and C. R. Craddock, "The Neuro Bureau ADHD-200 Preprocessed repository," *NeuroImage*, vol. 144, Part B, p. 275286, Jan. 2017.

[22] X. N. Zuo *et al.*, "An open science resource for establishing reliability and reproducibility in functional connectomics," *Sci. Data*, vol. 1, p. 140049, Dec. 2014.

[23] K. B. Nooner *et al.*, "The NKI-Rockland sample: a model for accelerating the pace of discovery science in psychiatry," *Front. Neurosci.*, vol. 6, p. 152, Oct. 2012.

[24] D. S. Marcus, T. R. Olsen, M. Ramaratnam, and R. L. Buckner, "The Extensible Neuroimaging Archive Toolkit: An informatics platform for managing, exploring, and sharing neuroimaging data," *Neuroinformatics*, vol. 5, no. 1, pp. 11–33, Mar. 2007.

[25] D. C. V. Essen *et al.*, "The WU-Minn Human Connectome Project: An overview," *NeuroImage*, vol. 80, pp. 62–79, Oct. 2013.

[26] M. F. Glasser *et al.*, "The minimal preprocessing pipelines for the Human Connectome Project," *NeuroImage*, vol. 80, pp. 105–124, Oct. 2013.

[27] D. S. Marcus *et al.*, "Human Connectome Project informatics: Quality control, database services, and data visualization," *NeuroImage*, vol. 80, pp. 202–219, Oct. 2013.

[28] M. D. Tisdall, A. T. Hess, M. Reuter, E. M. Meintjes, B. Fischl, and A. J. W. van der Kouwe, "Volumetric navigators for prospective motion correction and selective reacquisition in neuroanatomical MRI," *Magn. Reson. Med.*, vol. 68, no. 2, pp. 389–399, Aug. 2012.

[29] Y. Demchenko, Z. Zhao, P. Grosso, A. Wibisono, and C. de Laat, "Addressing big data challenges for scientific data infrastructure," in *Proc. IEEE 4th International Conference on Cloud Computing Technology and Science (CloudCom)*, Taipei, Taiwan, Dec. 2012, pp. 614–617.

[30] S. M. Smith, A. Hyvrinen, G. Varoquaux, K. L. Miller, and C. F. Beckmann, "Group-PCA for very large fMRI datasets," *NeuroImage*, vol. 101, p. 738749, Nov. 2014.

[31] S. Rachakonda, R. F. Silva, J. Liu, and V. D. Calhoun, "Group-PCA for very large fMRI datasets," *NeuroImage*, vol. 101, p. 738749, Nov. 2014.

[32] M. F. Glasser *et al.*, "A multi-modal parcellation of human cerebral cortex," *Nature*, vol. 536, no. 7615, pp. 171–178, Aug. 2016.

[33] E. Williams *et al.*, "Clinical language fMRI with real-time monitoring in temporal lobe epilepsy: Online processing methods," *Epilepsy Behav.*, vol. 25, no. 1, pp. 120–124, Sep. 2012.

[34] R. W. Cox *et al.*, "A (sort of) new image data format standard: NIfTI-1," in *Proc. 10th Annual Meeting of Organisation of Human Brain Mapping (OHBM 2004)*, Budapest, Hungary, Jun. 2004.

[35] D. Cordes *et al.*, "Frequencies contributing to functional connectivity in the cerebral cortex in."

[36] J. E. Chen and G. H. Glover, "BOLD fractional contribution to resting-state functional connectivity above 0.1 Hz," *Neuroimage*, vol. 107, pp. 207–218, Feb. 2015.

[37] J. M. Soares *et al.*, "A hitchhiker's guide to functional magnetic resonance imaging," *Front. Neurosci.*, vol. 10, p. 515, Nov. 2016.

[38] M. Y. Man *et al.*, "A review on the bioinformatics tools for neuroimaging," *Malays. J. Med. Sci.*, vol. 22, no. Spec Issue, pp. 9–19, Dec. 2015.

[39] R. W. Cox, "AFNI: Software for analysis and visualization of functional magnetic resonance neuroimages," *Comput. Biomed. Res.*, vol. 29, no. 3, pp. 162–173, Jun. 1996.

[40] B. Fischl, "FreeSurfer," *Neuroimage*, vol. 62, no. 2, p. 774781, Aug. 2012.

[41] M. Jenkinson, C. F. Beckmann, T. E. Behrens, M. W. Woolrich, and S. M. Smith, "FSL," *Neuroimage*, vol. 62, no. 2, p. 782790, Aug. 2012.

[42] K. J. Friston, *Statistical Parametric Mapping: The Analysis of Functional Brain Images*. Boston, MA: Elsevier/Academic Press, 2007.

[43] M. Göttlich, F. Beyer, and U. Krämer, "BASCO: A toolbox for task-related functional connectivity," *Front. Syst. Neurosci.*, vol. 9, p. 126, Sep. 2015.

[44] M. Rubinov and O. Sporns, "Complex network measures of brain connectivity: Uses and interpretations," *NeuroImage*, vol. 52, no. 3, p. 10591069, Sep. 2010.

[45] R. N. Boubela, K. Kalcher, W. Huf, C. Našel, and E. Moser, "Big data approaches for the analysis of large-scale fMRI data using apache spark and GPU processing: A demonstration on resting-state fMRI data from the Human Connectome Project," *Front. Neurosci.*, vol. 9, p. 492, Jan. 2016.

[46] M. Xia, J. Wang, and Y. He, "BrainNet Viewer: A network visualization tool for human brain connectomics," *PLoS ONE*, vol. 8, no. 7, p. e68910, Jul. 2013.

[47] D. Geffroy, D. Rivire, I. Denghien, N. Souedet, S. Laguitton, and Y. Cointepas, "BrainVISA: A complete software platform for neuroimaging," in *Proc. Python in Neuroscience workshop*, Paris, France, Aug. 2011.

[48] A. Eklund, P. Dufort, M. Villani, and S. LaConte, "BROCCOLI: Software for fast fMRI analysis on many-core CPUs and GPUs," *Front. Neuroinform.*, vol. 8, p. 24, Mar. 2014.

[49] S. M. Smith, A. Hyvärinen, G. Varoquaux, K. L. Miller, and C. F. Beckmann, "Group-PCA for very large fMRI datasets," *NeuroImage*, vol. 101, p. 738749, Nov. 2014.

[50] C. Craddock *et al.*, "Towards automated analysis of connectomes: The configurable pipeline for the analysis of connectomes (C-PAC)," *Front. Neuroinform.*, no. 42, Jul. 2013.

[51] S. Whitfield-Gabrieli and A. Nieto-Castanon, "Conn: A functional connectivity toolbox for correlated and anticorrelated brain networks," *Brain Connect.*, vol. 2, no. 3, pp. 125–141, Aug. 2012.

[52] A. Fornito, B. J. Harrison, A. Zalesky, and J. S. Simons, "Competitive and cooperative dynamics of large-scale brain functional networks supporting recollection," *Proc. Natl. Acad. Sci. U S A*, vol. 109, no. 31, pp. 12 788–12 793, Jul. 2012.

[53] C. G. Yan, X. D. Wang, X. N. Zuo, and Y. F. Zang, "DPABI: Data processing & analysis for (resting-state) brain imaging," *Neuroinformatics*, vol. 14, no. 3, pp. 339–351, Jul. 2016.

[54] Y. Chao-Gan and Z. Yu-Feng, "DPARSF: a matlab toolbox for."

[55] S. M. H. Hosseini, F. Hoeft, and S. R. Kesler, "GAT: A graph-theoretical analysis toolbox for analyzing between-group differences in large-scale structural and functional brain networks," *PLoS ONE*, vol. 7, no. 7, p. e40709, Jul. 2012.

[56] V. D. Calhoun, T. Adali, G. D. Pearlson, and J. J. Pekar, "A method for making group inferences from functional MRI data using independent component analysis," *Hum. Brain Mapp.*, vol. 14, no. 3, pp. 140–151, Nov. 2011.

[57] D. G. McLaren, M. L. Ries, G. Xu, and S. C. Johnson, "A generalized form of context-dependent psychophysiological interactions (gPPI): A comparison to standard approaches," *Neuroimage*, vol. 61, no. 4, pp. 1277–1286, Jul. 2012.

[58] J. D. Kruschwitz, D. List, L. Waller, M. Rubinov, and H. Walter, "GraphVar: A user-friendly toolbox for comprehensive graph analyses of functional brain connectivity," *J. Neurosci. Methods*, vol. 245, p. 107115, Apr. 2015.

[59] J. Wang, X. Wang, M. Xia, X. Liao, A. Evans, and Y. He, "GRETNA: A graph theoretical network analysis toolbox for imaging connectomics," *Front. Hum. Neurosci.*, vol. 9, p. 386, Jun. 2015.

[60] J. M. Spielberg, "Graph theoretic general linear model: A MATLAB toolbox," *Brain Connect.*, vol. 4, no. 9, p. A120, Nov. 2014.

[61] A. Zalesky, A. Fornito, and E. T. Bullmore, "Network-based statistic: Identifying differences in brain networks," *Neuroimage*, vol. 53, no. 4, pp. 1197–1207, Dec. 2010.

[62] P. Bellec *et al.*, "A neuroimaging analysis kit for Matlab and Octave," in *Proc. 17th International Conference on Functional Mapping of the Human Brain*, Quebec, QC, Canada, 2011.

[63] A. Abraham *et al.*, "Machine learning for neuroimaging with scikit-learn," *Front. Neuroinform.*, vol. 5, p. 13, Aug. 2011.

[64] K. Gorgolewski *et al.*, "Nipype: A flexible, lightweight and extensible neuroimaging data processing framework in Python," *Front. Neuroinform.*, vol. 8, p. 14, Feb. 2014.

[65] J. Schrouff *et al.*, "PRoNTo: pattern recognition for neuroimaging toolbox," *Neuroinformatics*, vol. 11, no. 3, pp. 319–337, Jul. 2013.

[66] P. Bellec, S. Lavoie-Courchesne, P. Dickinson, J. P. Lerch, A. P. Zijdenbos, and A. C. Evans, "The pipeline system for Octave and Matlab (PSOM): A lightweight scripting framework and execution engine for scientific workflows," *Front. Neuroinform.*, vol. 6, p. 7, Apr. 2012.

[67] M. Hanke, Y. O. Halchenko, P. B. Sederberg, S. J. Hanson, J. V. Haxby, and S. Pollmann, "PyMVPA: A python toolbox for multivariate pattern analysis of fMRI data," *Neuroinformatics*, vol. 7, no. 1, pp. 37–53, Mar. 2009.

[68] X. W. Song *et al.*, "REST: A toolkit for resting-state functional magnetic resonance imaging data processing," *PLoS ONE*, vol. 6, no. 9, p. e25031, Sep. 2012.

[69] T. E. Nichols and A. P. Holmes, "Nonparametric permutation tests for functional neuroimaging: A primer with examples," *Hum. Brain Mapp.*, vol. 15, no. 1, pp. 1–25, Jan. 2012.

[70] M. N. Hebart, K. Görgen, and J. D. Haynes, "The decoding toolbox (TDT): A versatile software package for multivariate analyses of functional imaging data," *Front. Neuroinform.*, vol. 8, p. 88, Jan. 2015.

[71] Y. Ad-Dab'bagh *et al.*, "The CIVET image processing environment: A fully automated comprehensive pipeline for anatomical neuroimaging research," in *Proc. 12th Annual Meeting of the*

Human Brain Mapping Organization (OHBM 2006), Florence, Italy, 2006.

[72] T. Sherif *et al.*, "CBRAIN: A web-based, distributed computing platform for collaborative neuroimaging research," *Front. Neuroinform.*, vol. 8, p. 54, May 2014.

[73] Y. Zang, T. Jiang, Y. Lu, Y. He, and L. Tian, "Regional homogeneity approach to fMRI data analysis," *NeuroImage*, vol. 22, no. 1, pp. 394–400, May 2004.

[74] Y. F. Zang *et al.*, "Altered baseline brain activity in children with ADHD revealed by resting-state functional MRI," *Brain Dev.*, vol. 29, no. 2, p. 8391, Mar. 2007.

[75] Q. H. Zou *et al.*, "An improved approach to detection of amplitude of low-frequency fluctuation (ALFF) for resting-state fMRI: fractional ALFF," *J. Neurosci. Methods*, vol. 172, no. 1, p. 137141, Jul. 2008.

[76] X. Li *et al.*, "Scalable fast rank-1 dictionary learning for fMRI big data analysis," in *Proc. 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining (KDD '16)*, San Francisco, CA, USA, Aug. 2006.

[77] M. Makkie *et al.*, "Hafni-enabled largescale platform for neuroimaging informatics (HELPNI)," *Brain Inform.*, vol. 2, no. 4, p. 225238, Dec. 2015.

[78] P. Fitzgibbon *et al.*, "The developing Human Connectome Project (dHCP): Minimal functional preprocessing pipeline for neonates," in *Proc. IEEE 5th Biennial Conference on Resting State and Brain Connectivity*, Vienna, Austria, Sep. 2016.

[79] C. Caballero-Gaudes and R. C. Reynolds, "Group-PCA for very large fMRI datasets," *NeuroImage*, Dec. 2016.

[80] P. Y. Lum *et al.*, "Extracting insights from the shape of complex data using topology," *Sci. Rep.*, vol. 3, p. 1236, Feb. 2013.

[81] G. Singh, F. Mémoli, and G. E. Carlsson, "Topological methods for the analysis of high dimensional data sets and 3D object recognition." in *SPBG*, 2007, pp. 91–100.

[82] E. Bullmore and O. Sporns, "Complex brain networks: graph theoretical analysis of structural and functional systems," *Nature Reviews Neuroscience*, vol. 10, no. 3, pp. 186–198, 2009.

[83] R. Ghrist, "Barcodes: the persistent topology of data," *Bull. Amer. Math. Soc.*, vol. 45, no. 1, pp. 61–75, 2008.

[84] N. Otter, M. A. Porter, U. Tillmann, P. Grindrod, and H. A. Harrington, "A roadmap for the computation of persistent homology," *arXiv preprint arXiv:1506.08903*, 2015.

[85] A. Patania, "Mappertools," available on line. [Online]. Available: https://alpatania.github.io/MapperTools/

[86] D. Muellner, "Tdamapper: Topological data analysis using mapper," available on line. [Online]. Available: https://github.com/paultpearson/TDAmapper/

[87] A. Tausz, M. Vejdemo-Johansson, and H. Adams, "JavaPlex: A research software package for persistent (co)homology," in *Proceedings of ICMS 2014*, ser. Lecture Notes in Computer Science 8592, H. Hong and C. Yap, Eds., 2014, pp. 129–136, software available at http://appliedtopology.github.io/javaplex/.

[88] K. Mischaikow and V. Nanda, "Morse Theory for Filtrations and Efficient Computation of Persistent Homology," *Discrete Comput. Geom.*, vol. 50, no. 2, pp. 330–353, Jul. 2013.

[89] V. Nanda, "Perseus, the persistent homology software," available on line. [Online]. Available: http://www.sas.upenn.edu/~vnanda/perseus

[90] "jholes website," available on line. [Online]. Available: http://www.jholes.eu/

[91] J. Binchi, E. Merelli, M. Rucco, G. Petri, and F. Vaccarino, "jholes: A tool for understanding biological complex networks via clique weight rank persistent homology," *Electronic Notes in Theoretical Computer Science*, vol. 306, pp. 5–18, 2014.

[92] D. Morozov, "Dionysus website," available on line. [Online]. Available: http://www.mrzv.org/software/dionysus/

[93] U. Bauer, M. Kerber, J. Reininghaus, and H. Wagner, "Phat–persistent homology algorithms toolbox," in *Proc. International Congress on Mathematical Software*. Springer, 2014, pp. 137–143.

[94] U. Bauer, M. Kerber, and J. Reininghaus, "Clear and compress: Computing persistent homology in chunks," in *Topological Methods in Data Analysis and Visualization III*. Springer, 2014, pp. 103–117.

[95] C. Maria, J.-D. Boissonnat, M. Glisse, and M. Yvinec, "The Gudhi Library: Simplicial Complexes and Persistent Homology," in *The 4th International Congress on Mathematical Software (ICMS)*, Hanyang University, Seoul, Korea, France, Aug. 2014. [Online]. Available: https://hal.inria.fr/hal-01108461

[96] "Gudhi," 2015. [Online]. Available: http://gudhi.gforge.inria.fr/

[97] U. Bauer, "Ripser github repository," 2016, available on line. [Online]. Available: https://github.com/Ripser/ripser

[98] J. Dean and S. Ghemawat, "MapReduce: Simplified data processing on large clusters," in *Proc. 6th Conference on Symposium on Operating Systems Design & Implementation (OSDI'04)*, vol. 6, San Francisco, CA, Dec. 2004, p. 10.

[99] S. Kyeong, S. Park, K. A. Cheon, J. J. Kim, D. H. Song, and E. Kim, "A new approach to investigate the association between brain functional connectivity and disease characteristics of attention-deficit/hyperactivity disorder: Topological neuroimaging data analysis," *PLoS ONE*, vol. 10, no. 9, p. e0137296, Sep. 2015.

[100] B. B. Lahey and E. G. Willcutt, "Predictive validity of a continuous alternative to nominal subtypes of attention-deficit/hyperactivity disorder for DSM-V," *J. Clin. Child. Adolesc. Psychol.*, vol. 39, no. 6, pp. 761–775, 2010.

[101] X. N. Zuo and X. X. Xing, "Test-retest reliabilities of resting-state FMRI measurements in human brain functional connectomics: A systems neuroscience perspective," *Neurosci. Biobehav. Rev.*, vol. 45, pp. 100–118, Sep. 2014.

[102] D. Cordes, V. Haughton, J. D. Carew, K. Arfanakis, and K. Maravilla, "Hierarchical clustering to measure connectivity in fMRI resting-state data," *Magn. Reson. Imaging.*, vol. 20, no. 4, pp. 305–317, May 2012.

[103] M. J. Fadili, S. Ruan, D. Bloyet, and B. Mazoyer, "A multistep unsupervised fuzzy clustering analysis of fMRI time series," *Hum. Brain Mapp.*, vol. 10, no. 4, pp. 160–178, Aug. 2000.

[104] A. Meyer-Baese, A. Wismueller, and O. Lange, "Comparison of two exploratory data analysis methods for fMRI: Unsupervised clustering versus independent component analysis," *IEEE Trans. Inf. Technol. Biomed.*, vol. 8, no. 3, pp. 387–398, Sep. 2004.

[105] V. D. Calhoun, J. Liu, and T. Adalı, "A review of group ICA for fMRI data and ICA for joint inference of imaging, genetic, and ERP data," *Neuroimage*, vol. 45, no. 1 Supp, pp. S163–S172, Mar. 2009.

[106] M. Nicolau, A. J. Levine, and G. Carlsson, "Topology based data analysis identifies a subgroup of breast cancers with a unique mutational profile and excellent survival," *Proc. Natl. Acad. Sci. U S A*, vol. 108, no. 17, pp. 7265–7270, Apr. 2011.

[107] M. Newman, "Networks: an introduction. 2010," *United Slates: Oxford University Press Inc., New York*, pp. 1–2.

[108] O. Sporns, "Structure and function of complex brain networks," *Dialogues Clin Neurosci*, vol. 15, no. 3, pp. 247–262, 2013.

[109] G. Deco, V. K. Jirsa, and A. R. McIntosh, "Emerging concepts for the dynamical organization of resting-state activity in the brain," *Nature Reviews Neuroscience*, vol. 12, no. 1, pp. 43–56, 2011.

[110] M. Kaiser, R. Martin, P. Andras, and M. P. Young, "Simulation of robustness against lesions of cortical networks," *European Journal of Neuroscience*, vol. 25, no. 10, pp. 3185–3192, 2007.

[111] D. Meunier, R. Lambiotte, A. Fornito, K. D. Ersche, and E. T. Bullmore, "Hierarchical modularity in human brain functional networks," *Frontiers in neuroinformatics*, vol. 3, 2009.

[112] J. Leskovec and A. Krevl, "{SNAP Datasets}:{Stanford} large network dataset collection," 2015.

[113] G. Csardi and T. Nepusz, "The igraph software package for complex network research," *InterJournal, Complex Systems*, vol. 1695, no. 5, pp. 1–9, 2006.

[114] J. Freeman, N. Vladimirov, T. Kawashima, Y. Mu, N. J. Sofroniew, D. V. Bennett, J. Rosen, C.-T. Yang, L. L. Looger, and M. B. Ahrens, "Mapping brain activity at scale with cluster computing," *Nature methods*, vol. 11, no. 9, pp. 941–950, 2014.

[115] Y. Wang, M. J. Anderson, J. D. Cohen, A. Heinecke, K. Li, N. Satish, N. Sundaram, N. B. Turk-Browne, and T. L. Willke, "Full correlation matrix analysis of fmri data on intel® xeon phi coprocessors," in *Proceedings of the International Conference for High Performance Computing, Networking, Storage and Analysis*. ACM, 2015, p. 23.

[116] L.-D. Lord *et al.*, "Insights into brain architectures from the homological scaffolds of functional connectivity networks," *Front. Syst. Neurosci.*, vol. 10, p. 85, 2016.

[117] B. Cassidy, C. Rae, and V. Solo, "Brain activity: Conditional dissimilarity and persistent homology," in *Proc. IEEE 12th International Symposium on Biomedical Imaging (ISBI 2015)*, 2015, pp. 1356–1359.

[118] H. Lee, H. Kang, M. K. Chung, B.-N. Kim, and D. S. Lee, "Weighted functional brain network modeling via network fil-

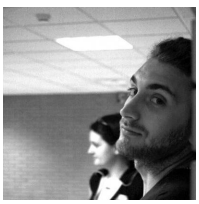
tration," in *NIPS Workshop on Algebraic Topology and Machine Learning*. Citeseer, 2012.
[119] G. Petri, P. Expert, F. Turkheimer, R. Carhart-Harris, D. Nutt, P. J. Hellyer, and F. Vaccarino, "Homological scaffolds of brain functional networks," *J. R. Soc. Interface*, vol. 11, no. 101, p. 20140873, Dec. 2014.
[120] D. Pachauri, C. Hinrichs, M. K. Chung, S. C. Johnson, and V. Singh, "Topology-based kernels with application to inference problems in alzheimer's disease," *IEEE transactions on medical imaging*, vol. 30, no. 10, pp. 1760–1770, 2011.
[121] M. Arai, V. Brandt, and Y. Dabaghian, "The effects of theta precession on spatial learning and simplicial complex dynamics in a topological model of the hippocampal spatial map," *PLoS Comput. Biol.*, vol. 10, no. 6, p. e1003651, Jun. 2014.
[122] Y. Dabaghian, F. Mémoli, L. Frank, and G. Carlsson, "A topological paradigm for hippocampal spatial map formation using persistent homology," *PLoS Comput. Biol.*, vol. 8, no. 8, p. e1002581, Aug. 2012.
[123] J. Yoo, E. Y. Kim, Y. M. Ahn, and J. C. Ye, "Topological persistence vineyard for dynamic functional brain connectivity during resting and gaming stages," *J. Neurosci. Methods*, vol. 267, pp. 1–13, Jul. 2016.
[124] H. Lee, Z. Ma, Y. Wang, and M. K. Chung, "Topological distances between networks and its application to brain imaging," *arXiv preprint arXiv:1701.04171*, 2017.
[125] H. Kim, J. Hahm, H. Lee, E. Kang, H. Kang, and D. S. Lee, "Brain networks engaged in audiovisual integration during speech perception revealed by persistent homology-based network filtration," *Brain connectivity*, vol. 5, no. 4, pp. 245–258, 2015.
[126] O. T. Courtney and G. Bianconi, "Generalized network structures: The configuration model and the canonical ensemble of simplicial complexes," *Physical Review E*, vol. 93, no. 6, p. 062311, 2016.
[127] J.-G. Young, G. Petri, F. Vaccarino, and A. Patania, "Construction of and efficient sampling from the simplicial configuration model," *arXiv preprint arXiv:1705.10298*, 2017.
[128] G. Bianconi and C. Rahmede, "Emergent hyperbolic network geometry," *Scientific Reports*, vol. 7, 2017.
[129] H. Edelsbrunner and J. Harer, *Computational topology: An introduction*. American Mathematical Soc., 2010.
[130] O. Busaryev, S. Cabello, C. Chen, T. K. Dey, and Y. Wang, "Annotating Simplices with a Homology Basis and Its Applications," in *Algorithm Theory – SWAT 2012*. Berlin, Heidelberg: Springer Berlin Heidelberg, Jul. 2012, pp. 189–200.
[131] H. Edelsbrunner and S. Parsa, "On the computational complexity of betti numbers: reductions from matrix rank," in *Proc. 25th annual ACM-SIAM symposium on Discrete algorithms*, Portland, Oregon, Jan. 2014, pp. 152–160.
[132] J. D. Boissonnat, t. Dey, and c. Maria, "The compressed annotation matrix: An efficient data structure for computing persistent cohomology," in *Algorithms ESA 2013*, Apr. 2013, pp. 695–706.
[133] D. Gunther, A. Jacobson, J. Reininghaus, H. P. Seidel, O. Sorkine-Hornung, and T. Weinkauf, "Fast and memory-efficienty topological denoising of 2D and 3D scalar fields," *IEEE Trans. Vis. Comput. Graph.*, vol. 20, no. 12, pp. 2585–2594, Oct. 2014.
[134] P. Bubenik, "Statistical topological data analysis using persistence landscapes," *J. Mach. Learn. Res.*, vol. 16, no. 1, pp. 77–102, Jan. 2015.
[135] J. Reininghaus, S. Huber, U. Bauer, and R. Kwitt, "A stable multiscale kernel for topological machine learning," in *proc. 2015 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2015, pp. 4741–4748.
[136] R. Kwitt, S. Huber, M. Niethammer, W. Lin, and U. Bauer, "Statistical topological data analysis - a kernel perspective," in *Advances in Neural Information Processing Systems*, 2015, pp. 3070–3078.

Angkoon Phinyomark (M'09) was born in Thailand, in 1986. He received the B.Eng. degree, with First Class Honors in computer engineering from Prince of Songkla University, Thailand, in 2008, where he pursued Ph.D. degree in electrical engineering, in 2012. His collective dissertation research also received the "Best PhD Thesis Award" from the National Research Council of Thailand.

From 2012 to 2013, he was a Postdoctoral Research Fellow with the GIPSA-lab and the LIG lab, University Joseph Fourier, Grenoble, France and from 2013 to 2016, he was a Postdoctoral Research Fellow with the Human Performance Laboratory, the University of Calgary, Canada. Since 2016, he has been a Researcher with the ISI Foundation, Turin, Italy. He has developed an H-index of 18 (i10-index of 29) and his 90 published refereed journals, book chapters, and conference proceedings have been cited 1,466 times. Thirty papers have been published in ISI indexed journals (21 as the first author). His publications are in the areas of biomedical signal processing notably electromyogram (EMG), gait biomechanics, neuroimaging analysis, and machine learning.

He is an Associate Editor of the IRBM journal. Based on his research expertise, he has served as an invited reviewer for 40 different peer-reviewed journals in ISI Web of Science. He has also invited to serve as a reviewer for 33 different international conferences.

Esther Ibáñez-Marcelo was born in Barcelona (Spain), in 1987. She received the Mathematics degree from Universitat Politècnica de Catalunya (UPC), Barcelona (Spain), in 2010, where she also obtained her Master's Degree in Mathematical Engineering, specialized in Biomedical sciences modelling in 2011. Her Master Thesis project was awared with Évariste Galois prize 2013 by Societat Catalana de Matemàtiques. She got her Ph.D. degree in Applied Mathematics by the UPC and Centre de Recerca Matemàtica, Barcelona (Spain) at the end of 2014. She has also experience as a Data Scientist in the private sector, participating and leading projects on time-series, prediction analysis and modelling client segmentation. Since May 2016, she is a Postdoctoral Researcher at ISI Foundation, Turin, Italy. Her research interests include network science, topological data analysis and Machine Learning applied mainly to biomedice and her publications are in the areas of biomathematics, genotype-phenotype networks, phylogenetic reconstruction and topological data analysis. Dra Esther Ibáñez is a member of the Societat Catalana de Matemàtiques since 2011.



Giovanni Petri was born in Italy in 1983. He received a B.Sc. degree in General Physics in 2005 and M.Sc. degree in Theoretical Physics in 2008 from the University of Pisa. He then obtained a PhD degree from Imperial College London in 2012, working on traffic and dynamics on networks coupled to information.

In 2005 he was a summer research fellow at Stanford Linear Accelerator Center working in the GLAST collaboration. From 2010 to 2012 he was a teaching assistant in the Business School of Queen Mary's University London. From 2012 to 2015 he was a Junior Researcher at ISI Foundation working on the extension and application of methods from algebraic topology to the study of large complex systems. From 2013 to 2014 he was also a long-term visitor at the Institute for Mathematical Sciences at University of Minnesota. From 2015 he is Principal Researcher at ISI Foundation and leads the data-driven neuroscience group. His research interests include statistical mechanics of complex networks, topological data analysis and their applications to biological, cognitive and neuroimaging problems.