Comparisons of Null Models with Topological Data Analysis

Paul Sitoleux, Lucrezia Carboni, Lucrezia Carboni, Hanâ Lbath, Sophie Achard

Comparisons of null models with topological data analysis

Paul Sitoleux
Univ. Grenoble Alpes, CNRS, Inria, Grenoble INP, LJK
38000 Grenoble, France
paul.sitoleux@ens-paris-saclay.fr

Lucrezia Carboni
Univ. Grenoble Alpes, CNRS, Inria, Grenoble INP, LJK
Univ. Grenoble Alpes, Inserm, U1216, Grenoble Institut Neurosciences, GIN,
38000 Grenoble, France
lucrezia.carboni@univ-grenoble-alpes.fr

Hanâ Lbath, and Sophie Achard
Univ. Grenoble Alpes, CNRS, Inria, Grenoble INP, LJK
38000 Grenoble, France
first.last@inria.fr

***Abstract*—Functional magnetic resonance imaging (fMRI) allows the construction of functional brain networks, offering a tool to probe the organization of neural activity. Null models have been proposed in this framework to evaluate the accuracy of new proposed approaches to discriminate network features coming from the data themselves in comparison to randomize procedure. Several models have been recently proposed and it is still complicated to choose one. We propose in this paper to compare null models and real datasets using Persistent homology (PH). PH is part of topological data analysis (TDA), and offers a framework for building multiscale summaries of networks. PH is first applied to a density-based filtration. We propose a procedure to extract label information from persistent homology summaries of labeled graphs. We then investigate its ability to discriminate between real data and surrogate data generated from null models. Interestingly, our new proposed label-informed approach is able to discriminate very accurately real datasets and classical null models opening the way to the design of new null models.**

***Index Terms*—Toppological data analysis, graph distances, persistent homology, null models, function brain connectivity.**

I. Introduction

Analyses of real data is a complex task, where no ground truth is available. Many sources of noise, disturbances, preprocessing can affect the results, with few tools able to quantify them. Lots of efforts are now built to make research reproducible, repeatable, generalisable, reliable ...(see for exemple The Turing Way[1]). In brain analyses, these challenges are present in all the research conducted nowadays. Statisticians have developed a whole framework to validate and promote responsible research [1]–[3]. This includes quantification of uncertainties, statistical tests for fMRI data, machine learning dealing with the high-dimensional and low sample size nature of the datasets. Among all this, we are interested in the generation of null datasets to be able to validate the statistical results. As very nicely described in [3], this is a difficult task because of the complex nature of brain fMRI data. A widely used approach to study the functioning of the brain is to construct brain connectivity matrices. After thresholding, purely graph-based methods are used. They are often limited to global or local statistics, only presenting information at the scale of the whole graph or of a single node neighborhood [4]. Persistent homology is a topological data analysis (TDA) approach that produces multiscale summaries from a point cloud or a distance matrix. Persistent homology tracks homological features, in short, the number of connected components and the number of topological cavities (equivalent to a circle or a hollow sphere in dimensions 1 and 2, or to higher dimensional cavities), by building simplices. This is a way to look at structures in the graph that are orders higher than the dyadic relationships the edges of the graph represent [5]. However, inherent to the use of new methods, we need to ask whether this method is robust, reliable, generalisable ....To that end, we consider null models that generate surrogate connectivity matrices as benchmarks of real data. This permits the identification of relevant features captured by a persistent homology approach in brain connectivity. Null models are ubiquitous in network neuroscience [3]. For instance, they allow one to test the statistical significance of graph features of empirical networks against a null hypothesis. Furthermore, null models which can be tuned to reproduce a specific graph layout offer a tool for comparing graph distances and understanding which properties they capture. This paper introduces a novel label-informed distance on persistence diagrams, useful in non-permutation invariant settings. Then a validation is proposed with application to null-models and fMRI brain connectivity datasets. This allows us to gain insights on the construction of null models from real brain data.

II. State of the art

The interest in higher-order interactions, with persistent homology for example, has taken off in recent years, whether in the broader context of complex systems modeling [6] or in structural [7] or functional brain networks [8], with notable findings in aging and neurodegenerative disorders. It is easily applicable to biological data, which is often highly dimensional and lacks a natural concept of distance [9]. For instance, in neuroscience network data, it has been

This work was partly funded by French National Research Agency project ANR-20-NEUC-0003-02, and by MIAI@Grenoble Alpes (ANR 19-P3IA-003).

[1]https://www.turing.ac.uk/research/research-projects/turing-way

applied to differentiate healthy subjects and subjects with neural disorders [10], [11], study the influence of psychotropic substances and sedatives [12] or to analyze neuronal network simulations [13]. Null models for network neuroscience have been reviewed recently in [3]. Different strategies are proposed, either randomising the weights of the connectivity matrices [14], generating graphs with same degree or any other characteristics [15], and other possibilities as detailed in [3]. It is now well admitted in the community that validation of results should include a comparison with null models [16]. In this paper, we propose for the first time to compare the classically used null models proposed in the literature through the lens of TDA.

III. Data

A. Real Data

We consider a set of resting-state functional networks from the Human Connectome Project (HCP). The measurements were parcellated into $N = 90$ regions using the Automated Anatomical labeling (AAL). For each region, the time series were estimated by averaging over all voxels, weighted by the proportion of gray matter in each voxel (estimated through structural MRI) and corrected for head motion. Correlations were estimated between Daubechie's wavelets coefficient between all pairs of time series. Analysis was restrained to the frequency interval 0.04-0.08 Hz. See [17] for additional details. A graph $G = (V, E)$ is a collection of nodes or vertices V and edges linking those vertices, i.e., each element in E is an element of $V \times V$. There is a range of ways to represent a graph, here it will be represented by its adjacency matrix, usually denoted A. An adjacency matrix is a symmetric matrix where each off-diagonal element a_{ij} is a weight, representing a chosen attribute of the relation between nodes i and j.

Using the correlation matrix between the regional wavelet coefficients C, we construct the graph adjacency matrix $A = 1 - |C|$.

B. Null models for functional connectivity

Null models are an ubiquitous tool in network neuroscience [3]. By offering a way to generate networks according to a simplified model, they allow benchmarking of the properties of empirical networks against the null hypothesis provided by a given model.

Comparing null models and empirical functional connectivity allows for the appreciation of null model properties and improved understanding of the information captured by various distances. Here, four null models for correlation networks are presented, each conserving some features from an initial functional connectivity matrix [3]. First, we consider the Zalesky matching algorithm, which generates correlation matrices with a given average correlation and variance between correlations [18]. This is achieved through generating an $N + 1$ normally distributed random vectors of length T $x_i, y \sim \mathcal{N}(0, \mathbf{I})$ $(i \in \{1, \dots N\})$, with $\mathbf{I}$ the T-dimensional identity matrix. Then, the values of x_i are repeatedly adjusted as $x_i \leftarrow x_i + ay$ until the desired average correlation between the vectors is obtained. The process is repeated with a different number of time points T until the correlation variances also match.

As a second model, we consider a spatiotemporal one which generates time series imitating spatial and temporal autocorrelation from the regional time series of a given examination of a subject [19]. Here, the model is applied to the regional wavelet coefficients.

The last two null models are a phase randomization model, generating new time series where each regional time series has the same power spectrum, and an Erdős-Rényi model, where each correlation value is randomly distributed.

IV. Persistent Homology

Persistent homology is a mathematical formalism from the larger field of topological data analysis, that allows the production of multiscale summaries of point cloud or graph distance matrices [20], such as Betti numbers and persistent diagrams.

Starting from a distance matrix, simplicial complexes are built, using a filtration, tracking topological features at every scale. These features correspond to the dimension of k^{th} homology group [21], with k a non-negative integer. Essentially, the number of 0-dimensional features is the number of connected components, and higher dimensional ones represent *topological cavities*: 1-dimensional features are circle-like, 2-dimensional features are (hollow) sphere-like, and so on. In short, this means that the Betti number β_k counts the number of $k + 1$-dimensional volumes that are enclosed by, at least, a k-dimensional cycle that does not correspond to the boundary of a simplex of the given simplicial complex, with β_0 counting the number of connected components of the simplicial complex. We are, in this paper, interested in plotting the *death* values of features against their *birth* values. This yields a *persistence diagram* in the plane delimited by the $b = d$ and $d = 0$ lines. Here, the filtration is done on a density-weighted graph. This preserves the topological features, as it is a (non-linear) monotone transformation, and only re-scale their birth and death values.

A. Comparing persistence diagrams

We first recall various distances between persistence diagrams.

The *p-Wasserstein distance* between measures μ, ν, with support in $\mathcal{X}$ $p \in [1, \infty[$ is

$$W_p(\mu, \nu) = \left(\int_{\mathcal{X} \times \mathcal{X}} c(x, y)^p \, \mathrm{d}\pi(x, y) \right)^{1/p} \tag{1}$$

where $\pi(x, y)$ is the optimal coupling between μ and ν under cost $c(\cdot, \cdot)^p$. In these works, the focus is mainly on the 1-Wasserstein distance.

Another common option is the *bottleneck distance*, which is the $p \to \infty$ limit of the p-Wasserstein distance:

$$\mathrm{Bo}(\mu, \nu) = \lim_{p \to +\infty} W_p(\mu, \nu). \tag{2}$$

For a, b two discrete distributions, the *optimal transport distance* can be expressed as

$$d_{\text{OT}}(a,b) = \gamma^*_{ab} = \underset{\gamma \in \mathbb{R}^{m\times n}}{\operatorname{argmin}} \sum_{i,j} \gamma_{ij} M_{ij} \tag{3}$$

such that $\gamma \mathbf{I} = a, \gamma^T \mathbf{I} = b, \gamma \geq 0$ and with M the cost matrix defining the cost of moving the histogram bin a_i to bin b_j, $\mathbf{I}$ the identity matrix of the relevant dimension.

We now introduce a label-informed distance on persistence diagrams, by designing a novel cost matrix. First, we list edges $\{\{e_k[i]\}_{i\in\{1,\dots,n_k\}}\}$ which correspond to the birth or death of the k^{th} homological feature in the persistence diagrams, where n_k is the number features. Next we choose a simple binary cost matrix defined as follows:

$$M^*_{ij} = \begin{cases} 0 & \text{if } e[i] = e[j] \\ 1 & \text{otherwise} \end{cases} \tag{4}$$

We define the *edge comparison optimal transport* distance as the optimal transport distance d_{OT} where $M = M^*$:

$$d_{e,k\dots}(a,b) = \underset{\gamma \in \mathbb{R}^{m\times n}}{\operatorname{argmin}} \sum_{i,j} \gamma_{ij} M^*_{ij}, \tag{5}$$

where $k\cdots$ denotes the included homological features. This approach is similar to the fused Gromov-Wasserstein distance [22], applied to a labeled persistence diagram instead of a labeled graph. Fused Gromov-Wasserstein takes into account both labels and graph structure by taking a weighted sum of two terms representing each aspect. The label term is equivalent to the distance introduced here for a general cost matrix M.

In the following, the persistence diagrams and the Wasserstein distances between them are computed using `giotto-tda`, with an error tolerance of 0.01 [23]. The optimal transport distance d_{OT} between the persistence diagram is computed with the `POT` package [24].

B. Comparing edges of labeled graphs

In the context of thresholded graphs, node labels can be used to compare graphs. For $\mathcal{A}, \mathcal{B}$ two sets, the *overlap* similarity is

$$\text{Overlap}(\mathcal{A},\mathcal{B}) = \frac{|\mathcal{A}\cap\mathcal{B}|}{\min(|\mathcal{A}|,|\mathcal{B}|)} \tag{6}$$

and the overlap distance

$$d_{\text{O}}(\mathcal{A},\mathcal{B}) = 1 - \text{Overlap}(\mathcal{A},\mathcal{B}). \tag{7}$$

In our setting, $\mathcal{A}, \mathcal{B}$ correspond to two sets of node label pairs, associated to edges in the thresholded graph.

V. Betti curves of different datasets

Fig. 1 presents the Betti curves of the different rs-fMRI (resting state-fMRI) correlation matrices and their null models using the density-based filtration, which does not require any thresholding step. The latter are presented in the first two rows.

Both Erdős-Rényi and phase randomization models exhibit fast percolation, reaching a single connected component at low density (Fig. 1a). The β_1 and β_2 curves also have similar behavior, with a localized peak where the maximal value is slightly lower for the phase randomization model (Fig. 1b and Fig. 1c). The Zalesky and spatiotemporal models exhibit slightly different behavior, with a slower decrease in the β_0 curve (Fig. 1d), particularly in the case of the spatiotemporal model. The β_1 and β_2 features are on average more present in the Zalesky graphs than in the spatiotemporal graphs, and tend to appear at lower densities in the former than in the latter (Fig. 1e and Fig. 1f).

Betti numbers hence seem to capture some characteristics that are specific to each null model.

The last line of Fig. 1 presents results on real data by varying the preprocessing. The first (HCP) corresponds to only taking the wavelet correlation between time series, while the other ones include a regression on white matter and cerebrospinal fluid (WM & CSF), with the last adding global signal regression (GSR), which are common preprocessing step. For β_0, additional signal regressions make the curve fall faster, meaning that fewer regions are poorly connected to the main connected component (Fig. 1g). For β_1, β_2, the curves follow each other at low and high densities but reach higher maxima (Fig. 1h and i). However, variability is large and future work is needed.

VI. Distances to discriminate between generative models and real data

For the null models, the Wasserstein distance clearly discriminate between three groups: phase randomization, Erdős-Rényi and a third group including healthy subjects, spatiotemporal, and Zalesky models. The bottleneck distance only separates the latter from a single group containing both phase randomization and Erdős-Rényi models (Fig. 2a).

A qualitative visual inspection of the results on the considered distances (Fig. 2) shows that the Zalesky and the spatiotemporal model can roughly reproduce the persistent homology of empirical functional brain networks. Interestingly, when considering the W_1 distance the real data are grouped together with the Zalesky and spatiotemporal model. This means that real data birth and death feature distribution can be approximated by considering a dominant Gaussian signal and adjusting the noise of individual regions in order to match the correlation distribution or by capturing limited spatial and temporal auto-correlations from the fMRI data. Meanwhile, the amplitude of the persistence diagrams appears to vary across real data and these two models. This leads to high bottleneck distances and prevents grouping them all together. The opposite is observed for phase randomization and Erdős-Rényi models: they are grouped together by bottleneck distance and

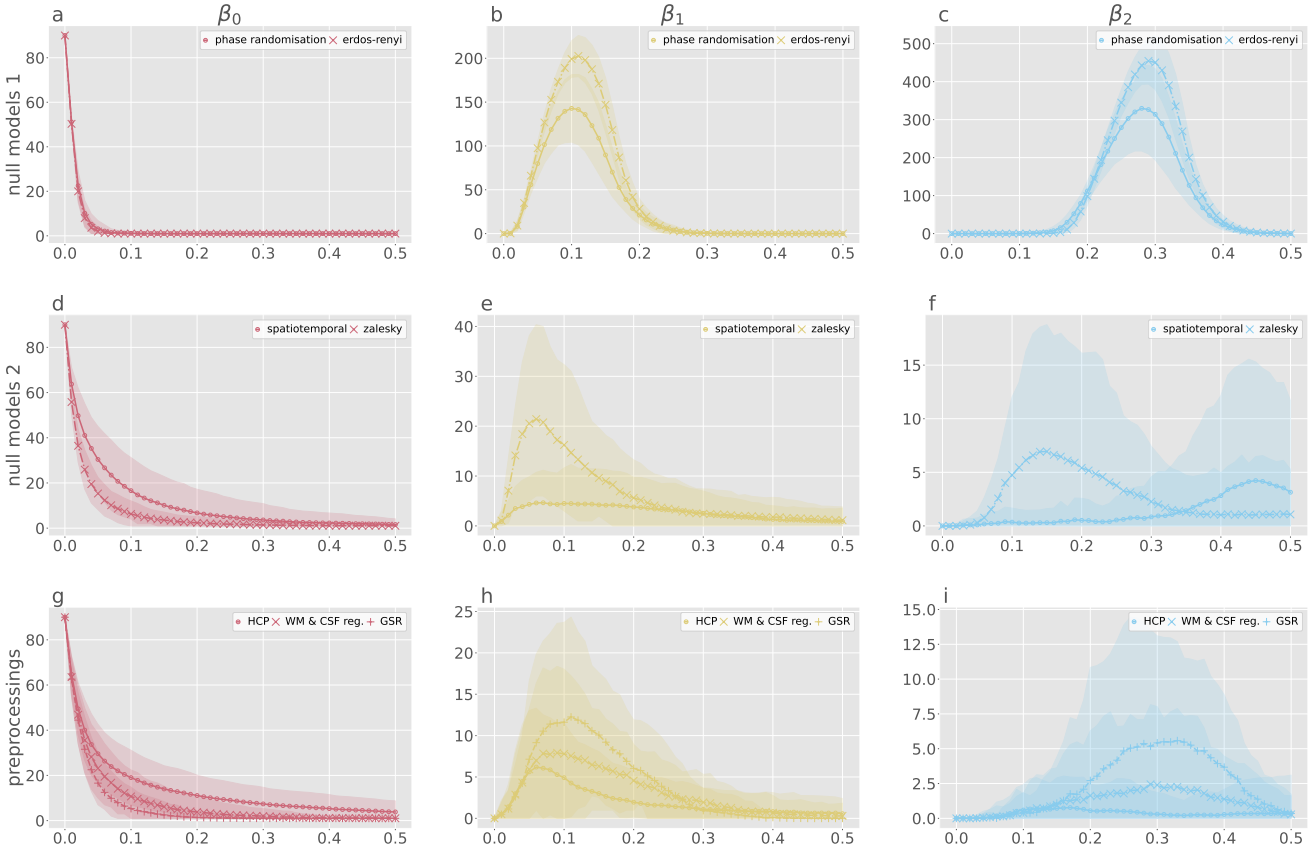

Fig. 1. From left to right: Betti curves for β_0, β_1 and β_2. (a) (b) (c) Null model Betti curves: phase randomization and Erdős-Rényi (d) (e) (f) Null model Betti curves: spatiotemporal and Zalesky (g) (h) (i) Empirical functional connectivity in HCP subjects with different preprocessings. Shaded areas give a 95% confidence interval around the mean β_i value at the given density.

not by W_1. Moreover, they are also gathered by the edge comparison optimal transport, suggesting they might produce similar label feature distribution. Both considered label-based distances differentiate the spatiotemporal and Zalesky null models from empirical data, but not from each other (Fig. 2b). This might be expected for the latter, since it does not take into account any label information to generate surrogate matrices, but is more surprising for the former. Hence, these label-dependent distances demonstrate that even null models that input some kind of spatial information do not reproduce label-dependent behavior.

None of the null models we consider manage to reproduce the real data label organization.

VII. Limitations and perspectives

For somewhat large networks, persistent homology stays limited to the lower dimensions, as the computational cost for computing higher-order features increases exponentially as it requires finding an arrangement of cliques. Although in this case, as the dependencies seem strong, finding high-order features would be surprising, it cannot yet be ruled out. Furthermore, in functional brain networks, first and second-order homology features seem to be short-lived, limiting the interest in persistent homology, where longer-lived features are the signature of particular topological invariants and are the main attribute that is targeted by persistent homology. In short, persistent homology appears to capture some of the texture of functional brain networks but does not uncover larger-scale organization.

VIII. Conclusion

To the best of our knowledge, this is the first attempt to use density levels filtration on fMRI networks. Moreover, we propose to include label information when comparing non-permutation invariant applications. Particularly, this is required in functional brain networks where nodes are associated with brain regions and are not perfectly exchangeable. We evaluate our approach on both real data and null models. These label-dependent distances show that node information included in some null models does not constrain the model enough to be close to the real data. This suggests new objectives in the design of null models for brain connectivity.

Acknowledgment

This work was partly funded by French National Research Agency project ANR-20-NEUC-0003-02, and by MIAI@Grenoble Alpes (ANR 19-P3IA-003).

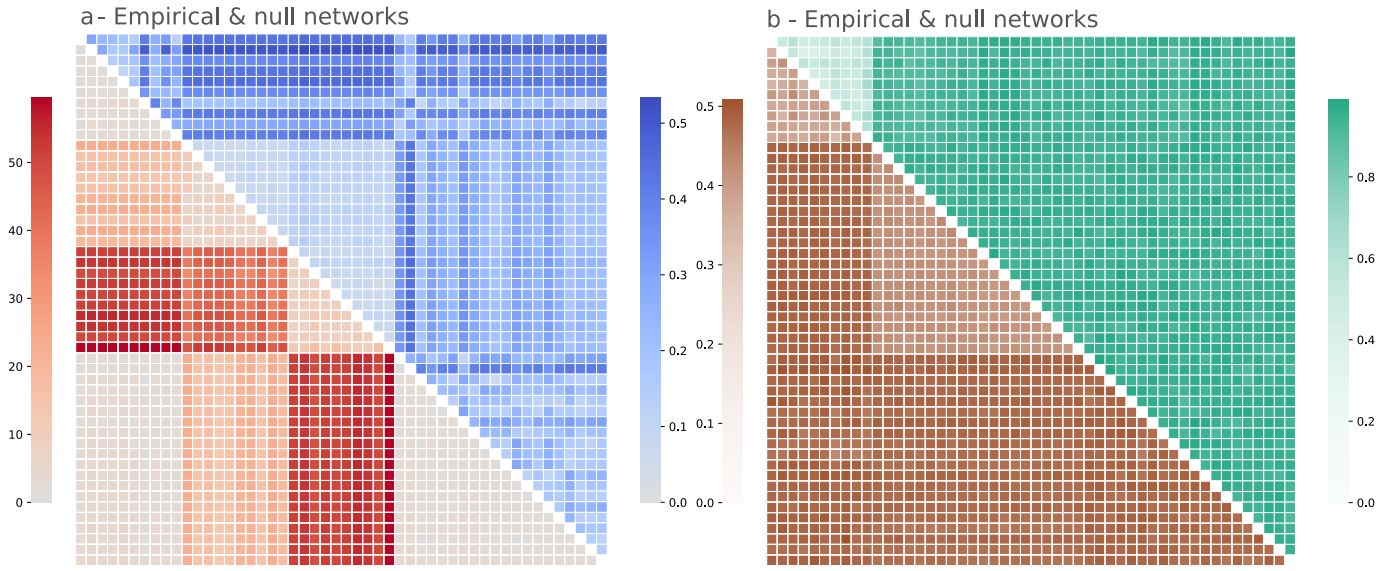

Fig. 2. 10 healthy subjects with 10 realizations of each of the four null models (phase randomization, Erdős-Rényi, spatiotemporal, Zalesky). (a) **Upper triangle:** Bottleneck distance. **Lower triangle:** 1-Wasserstein distance. (b) **Upper triangle:** overlap distances between edges at density 0.05 **Lower triangle:** edge comparison optimal transport distance $d_{e,012}$ between edges of the persistence diagrams (with $\beta_0, \beta_1, \beta_2$ features)

REFERENCES

[1] R. Gau, S. Noble, K. Heuer, K. L. Bottenhorn, I. P. Bilgin, Y.-F. Yang, J. M. Huntenburg, J. M. Bayer, R. A. Bethlehem, S. A. Rhoads, *et al.*, "Brainhack: Developing a culture of open, inclusive, community-driven neuroscience," *Neuron*, vol. 109, no. 11, pp. 1769–1775, 2021.

[2] A. Bowring, C. Maumet, and T. E. Nichols, "Exploring the impact of analysis software on task fmri results," *Human brain mapping*, vol. 40, no. 11, pp. 3362–3384, 2019.

[3] F. Váša and B. Mišić, "Null models in network neuroscience," *Nature Reviews Neuroscience*, vol. 23, pp. 493–504, May 2022.

[4] A. E. Sizemore, J. E. Phillips-Cremins, R. Ghrist, and D. S. Bassett, "The importance of the whole: Topological data analysis for the network neuroscientist," *Network Neuroscience*, vol. 3, pp. 656–673, Jan. 2019.

[5] L. Torres, A. S. Blevins, D. Bassett, and T. Eliassi-Rad, "The why, how, and when of representations for complex systems," *SIAM Review*, vol. 63, pp. 435–485, Jan. 2021.

[6] F. Battiston, E. Amico, A. Barrat, G. Bianconi, G. F. de Arruda, B. Franceschiello, I. Iacopini, S. Kéfi, V. Latora, Y. Moreno, M. M. Murray, T. P. Peixoto, F. Vaccarino, and G. Petri, "The physics of higher-order interactions in complex systems," *Nature Physics*, vol. 17, pp. 1093–1098, Oct. 2021.

[7] M. Andjelković, B. Tadić, and R. Melnik, "The topology of higher-order complexes associated with brain hubs in human connectomes," *Scientific Reports*, vol. 10, Oct. 2020.

[8] M. Gatica, F. E. Rosas, P. A. M. Mediano, I. Diez, S. P. Swinnen, P. Orio, R. Cofré, and J. M. Cortes, "High-order functional redundancy in ageing explained via alterations in the connectome in a whole-brain model," *PLOS Computational Biology*, vol. 18, p. e1010431, Sept. 2022.

[9] G. Carlsson, "Topology and data," *Bulletin of the American Mathematical Society*, vol. 46, pp. 255–308, Jan. 2009.

[10] M. K. Chung, J. L. Hanson, J. Ye, R. J. Davidson, and S. D. Pollak, "Persistent homology in sparse regression and its application to brain morphometry," *IEEE Transactions on Medical Imaging*, vol. 34, pp. 1928–1939, Sept. 2015.

[11] L. Caputi, A. Pidnebesna, and J. Hlinka, "Promises and pitfalls of topological data analysis for brain connectivity analysis," *NeuroImage*, vol. 238, p. 118245, Sept. 2021.

[12] T. F. Varley, V. Denny, O. Sporns, and A. Patania, "Topological analysis of differential effects of ketamine and propofol anaesthesia on brain dynamics," *Royal Society Open Science*, vol. 8, p. 201971, June 2021.

[13] J.-B. Bardin, G. Spreemann, and K. Hess, "Topological exploration of artificial neuronal network dynamics," *Network Neuroscience*, vol. 3, pp. 725–743, Jan. 2019.

[14] M. Rubinov and O. Sporns, "Weight-conserving characterization of complex functional brain networks," *Neuroimage*, vol. 56, no. 4, pp. 2068–2079, 2011.

[15] P. E. Vertes, A. F. Alexander-Bloch, N. Gogtay, J. N. Giedd, J. L. Rapoport, and E. T. Bullmore, "Simple models of human brain functional networks," *Proceedings of the National Academy of Sciences*, vol. 109, pp. 5868–5873, Mar. 2012.

[16] A. Rossi, S. Deslauriers-Gauthier, and E. Natale, "On null models for temporal small-worldness in brain dynamics," *Network Neuroscience*, pp. 1–30, 2024.

[17] M. Termenon, A. Jaillard, C. Delon-Martin, and S. Achard, "Reliability of graph analysis of resting state fmri using test-retest dataset from the human connectome project," *Neuroimage*, vol. 142, pp. 172–187, 2016.

[18] A. Zalesky, A. Fornito, and E. Bullmore, "On the use of correlation as a measure of network connectivity," *NeuroImage*, vol. 60, pp. 2096–2106, May 2012.

[19] M. Shinn, A. Hu, L. Turner, S. Noble, K. H. Preller, J. L. Ji, F. Moujaes, S. Achard, D. Scheinost, R. T. Constable, J. H. Krystal, F. X. Vollenweider, D. Lee, A. Anticevic, E. T. Bullmore, and J. D. Murray, "Spatial and temporal autocorrelation weave complexity in brain networks," June 2021.

[20] F. Chazal and B. Michel, "An introduction to topological data analysis: Fundamental and practical aspects for data scientists," *Frontiers in Artificial Intelligence*, vol. 4, Sept. 2021.

[21] F. H. Croom, *Basic Concepts of Algebraic Topology*. Springer New York, 1978.

[22] T. Vayer, L. Chapel, R. Flamary, R. Tavenard, and N. Courty, "Fused gromov-wasserstein distance for structured objects," *Algorithms*, vol. 13, p. 212, Aug. 2020.

[23] G. Tauzin, U. Lupo, L. Tunstall, J. B. Pérez, M. Caorsi, A. Medina-Mardones, A. Dassatti, and K. Hess, "giotto-tda: A topological data analysis toolkit for machine learning and data exploration," 2020.

[24] R. Flamary, N. Courty, A. Gramfort, M. Z. Alaya, A. Boisbunon, S. Chambon, L. Chapel, A. Corenflos, K. Fatras, N. Fournier, L. Gautheron, N. T. Gayraud, H. Janati, A. Rakotomamonjy, I. Redko, A. Rolet, A. Schutz, V. Seguy, D. J. Sutherland, R. Tavenard, A. Tong, and T. Vayer, "Pot: Python optimal transport," *Journal of Machine Learning Research*, vol. 22, no. 78, pp. 1–8, 2021.