


# frontiers

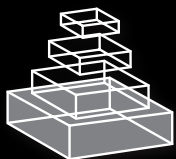
## RESEARCH TOPICS



### INFORMATION-BASED METHODS FOR NEUROIMAGING: ANALYZING STRUCTURE, FUNCTION AND DYNAMICS

Topic Editors

Jesus M. Cortés, Daniele Marinazzo and  
Miguel Angel Muñoz



**frontiers in**  
**NEUROINFORMATICS**



# frontiers

## FRONTIERS COPYRIGHT STATEMENT

© Copyright 2007-2015  
Frontiers Media SA.  
All rights reserved.

All content included on this site, such as text, graphics, logos, button icons, images, video/audio clips, downloads, data compilations and software, is the property of or is licensed to Frontiers Media SA ("Frontiers") or its licensees and/or subcontractors. The copyright in the text of individual articles is the property of their respective authors, subject to a license granted to Frontiers.

The compilation of articles constituting this e-book, wherever published, as well as the compilation of all other content on this site, is the exclusive property of Frontiers. For the conditions for downloading and copying of e-books from Frontiers' website, please see the Terms for Website Use. If purchasing Frontiers e-books from other websites or sources, the conditions of the website concerned apply.

Images and graphics not forming part of user-contributed materials may not be downloaded or copied without permission.

Individual articles may be downloaded and reproduced in accordance with the principles of the CC-BY licence subject to any copyright or other notices. They may not be re-sold as an e-book.

As author or other contributor you grant a CC-BY licence to others to reproduce your articles, including any graphics and third-party materials supplied by you, in accordance with the Conditions for Website Use and subject to any copyright notices which you include in connection with your articles and materials.

All copyright, and all rights therein, are protected by national and international copyright laws.

The above represents a summary only. For the full conditions see the Conditions for Authors and the Conditions for Website Use.

ISSN 1664-8714

ISBN 978-2-88919-208-3

DOI 10.3389/978-2-88919-208-3

## ABOUT FRONTIERS

Frontiers is more than just an open-access publisher of scholarly articles: it is a pioneering approach to the world of academia, radically improving the way scholarly research is managed. The grand vision of Frontiers is a world where all people have an equal opportunity to seek, share and generate knowledge. Frontiers provides immediate and permanent online open access to all its publications, but this alone is not enough to realize our grand goals.

## FRONTIERS JOURNAL SERIES

The Frontiers Journal Series is a multi-tier and interdisciplinary set of open-access, online journals, promising a paradigm shift from the current review, selection and dissemination processes in academic publishing.

All Frontiers journals are driven by researchers for researchers; therefore, they constitute a service to the scholarly community. At the same time, the Frontiers Journal Series operates on a revolutionary invention, the tiered publishing system, initially addressing specific communities of scholars, and gradually climbing up to broader public understanding, thus serving the interests of the lay society, too.

## DEDICATION TO QUALITY

Each Frontiers article is a landmark of the highest quality, thanks to genuinely collaborative interactions between authors and review editors, who include some of the world's best academicians. Research must be certified by peers before entering a stream of knowledge that may eventually reach the public - and shape society; therefore, Frontiers only applies the most rigorous and unbiased reviews.

Frontiers revolutionizes research publishing by freely delivering the most outstanding research, evaluated with no bias from both the academic and social point of view.

By applying the most advanced information technologies, Frontiers is catapulting scholarly publishing into a new generation.

## WHAT ARE FRONTIERS RESEARCH TOPICS?

Frontiers Research Topics are very popular trademarks of the Frontiers Journals Series: they are collections of at least ten articles, all centered on a particular subject. With their unique mix of varied contributions from Original Research to Review Articles, Frontiers Research Topics unify the most influential researchers, the latest key findings and historical advances in a hot research area!

Find out more on how to host your own Frontiers Research Topic or contribute to one as an author by contacting the Frontiers Editorial Office: [researchtopics@frontiersin.org](mailto:researchtopics@frontiersin.org)

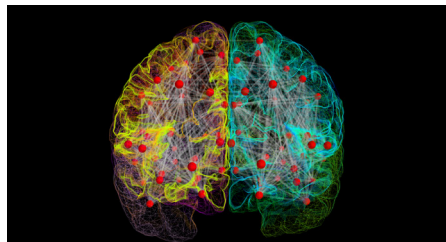
# INFORMATION-BASED METHODS FOR NEUROIMAGING: ANALYZING STRUCTURE, FUNCTION AND DYNAMICS

Topic Editors:

**Jesus M. Cortés**, Ikerbasque: The Basque Foundation for Science, and Hospital Universitario Cruces, Spain

**Daniele Marinazzo**, University of Ghent, Belgium

**Miguel Angel Muñoz**, University of Granada, Spain



The skeleton of a full brain network model in The Virtual Brain (TVB). The anatomical structure on which simulations are built in TVB represent two relevant spatial scales at which structural connectivity is defined. This separation allows for the construction of two main types of simulations: region-based and surface-based simulations. The first type uses only the connectome as spatial support and each of the nodes models the neural population activity of an entire brain region. The edges represent the long-range connections (interregional fibres in white) at the scale of centimetres. In the second type of simulations the cortex is shaped more realistically, each vertex of the surface is considered a cortical node and its dynamics are modelled by a neural population model; and, a local connectivity kernel assigns the density of local connections.

The aim of this Research Topic is to discuss the state of the art on the use of Information-based methods in the analysis of neuroimaging data. Information-based methods, typically built as extensions of the Shannon Entropy, are at the basis of model-free approaches which, being based on probability distributions rather than on specific expectations, can account for all possible non-linearities present in the data in a model-independent fashion.

Mutual Information-like methods can also be applied on interacting dynamical variables described by time-series, thus addressing the uncertainty reduction (or information) in one variable by conditioning on another set of variables.

In the last years, different Information-based methods have been shown to be flexible and powerful tools to analyze neuroimaging data, with a wide range of different methodologies, including formulations-based on bivariate vs multivariate representations, frequency vs time domains, etc. Apart from methodological issues, the information bit as a common unit

represents a convenient way to open the road for comparison and integration between different measurements of neuroimaging data in three complementary contexts: Structural Connectivity, Dynamical (Functional and Effective) Connectivity, and Modelling of brain activity. Applications are ubiquitous, starting from resting state in healthy subjects to modulations of consciousness and other aspects of pathophysiology.

Mutual Information-based methods have provided new insights about common-principles in brain organization, showing the existence of an active default network when the brain is at rest. It is not clear, however, how this default network is generated, the different modules are intra-interacting, or disappearing in the presence of stimulation. Some of these open-questions at the functional level might find their mechanisms on their structural correlates. A key question is the link between structure and function and the use of structural priors for the understanding of the functional connectivity measures.

As effective connectivity is concerned, recently a common framework has been proposed for Transfer Entropy and Granger Causality, a well-established methodology originally based on autoregressive models. This framework can open the way to new theories and applications.

This Research Topic brings together contributions from researchers from different backgrounds which are either developing new approaches, or applying existing methodologies to new data, and we hope it will set the basis for discussing the development and validation of new Information-based methodologies for the understanding of brain structure, function, and dynamics.



# Table of Contents

- 06 Editorial for the Research Topic: Information-Based Methods for Neuro Imaging: Analyzing Structure, Function and Dynamics**  
Jesus M. Cortes, Daniele Marinazzo and Miguel A. Muñoz
- 08 The Virtual Brain: A Simulator of Primate Brain Network Dynamics**  
Paula Sanz Leon, Stuart A. Knock, M. Marmaduke Woodman, Lia Domide, Jochen Mersmann, Anthony R. McIntosh and Viktor Jirsa
- 31 Disruption of Transfer Entropy and Inter-Hemispheric Brain Functional Connectivity in Patients with Disorder of Consciousness**  
Verónica Mäki-Marttunen, Ibai Diez, Jesus M. Cortes, Dante R. Chialvo and Mirta Villarreal
- 42 Constructing the Resting State Structural Connectome**  
Olusola Ajilore, Liang Zhan, Johnson GadElkarim, Aifeng Zhang, Jamie D. Feusner, Shaolin Yang, Paul M. Thompson, Anand Kumar and Alex Leow
- 50 Electroencephalogram Approximate Entropy Influenced by Both Age and Sleep**  
Gerick M. H. Lee, Sara Fattinger, Anne-Laure Mouthon, Quentin Noirhomme and Reto Huber
- 57 Information Gain in the Brain's Resting State: A New Perspective on Autism**  
José L. Pérez Velázquez and Roberto F. Galán
- 67 Mutual Information Spectrum for Selection of Event-Related Spatial Components. Application to Eloquent Motor Cortex Mapping**  
Alexei Ossadtchi, Platon Pronko, Sylvain Baillet, Mark E. Pflieger and Tatiana Stroganova
- 78 Local Active Information Storage as a Tool to Understand Distributed Neural Information Processing**  
Michael Wibral, Joseph T. Lizier, Sebastian Vögler, Viola Priesemann and Ralf Galuske
- 89 Reduced Predictable Information in Brain Signals in Autism Spectrum Disorder**  
Carlos Gómez, Joseph T. Lizier, Michael Schaum, Patricia Wollstadt, Christine Grützner, Peter Uhlhaas, Christine M. Freitag, Sabine Schlitt, Sven Bölte, Roberto Hornero and Michael Wibral
- 101 Energy Landscapes of Resting-State Brain Networks**  
Takamitsu Watanabe, Satoshi Hirose, Hiroyuki Wada, Yoshio Imai, Toru Machida, Ichiro Shirouzu, Seiki Konishi, Yasushi Miyashita and Naoki Masuda

- 112** *Isotropic Non-White Matter Partial Volume Effects in Constrained Spherical Deconvolution*  
Timo Roine, Ben Jeurissen, Daniele Perrone, Jan Aelterman, Alexander Leemans, Wilfried Philips and Jan Sijbers
- 121** *Variational Bayesian Causalconnectivity Analysis for fMRI*  
MartinLuessi, S. Derin Babacan, Rafael Molina, James R. Booth and Aggelos K. Katsaggelos
- 137** *Canonical Information Flow Decomposition Among Neural Structure Subsets*  
Daniel Y. Takahashi, Luiz A. Baccalá and Koichi Sameshima
- 148** *Algorithms of Causal Inference for the Analysis of Effective Connectivity Among Brain Regions*  
Daniel Chicharro and Stefano Panzeri
- 165** *Sample Entropy Reveals High Discriminative Power between Young and Elderly Adults in Short fMRI Data Sets*  
Moses O. Sokunbi
- 177** *Multi-Scale Integration and Predictability Inresting State Brain Activity*  
Artemy Kolchinsky, Martijn P. vanden Heuvel, Alessandra Griffo, Patric Hagmann Luis M. Rocha, Olaf Sporns and Joaquín Goñi



# Editorial for the research topic: information-based methods for neuroimaging: analyzing structure, function and dynamics

**Jesus M. Cortes<sup>1,2</sup>, Daniele Marinazzo<sup>3\*</sup> and Miguel A. Muñoz<sup>4</sup>**

<sup>1</sup> Biocruces Health Research Institute, Hospital Universitario Cruces, Barakaldo, Spain

<sup>2</sup> Ikerbasque: The Basque Foundation for Science, Bilbao, Spain

<sup>3</sup> Data Analysis, University of Gent, Gent, Belgium

<sup>4</sup> Departamento de Electromagnetismo y Física de la Materia and Instituto Carlos I de Física Teórica y Computacional, University of Granada, Granada, Spain

\*Correspondence: danielle.marinazzo@ugent.be

## Edited and reviewed by:

Sean L. Hill, International Neuroinformatics Coordinating Facility, Sweden

**Keywords: information theory, brain connectivity, network dynamics, neuroimaging, neuroinformatics, Granger causality, entropy, mutual information**

This Research Topic gathers different contributions highlighting novel types of analysis and methods to deal more efficiently with neuroimaging data, simulated and real, acquired with different modalities. These approaches allow us to shed light on the mechanisms of brain organization, with focus on the relationship between brain structure, function and dynamics.

The first article of this Topic (Sanz Leon et al., 2013), introduces The Virtual Brain, a Neuroinformatics platform for full brain network simulations using realistic connectivity, putting in evidence that the integration between brain structure and function is perfectly plausible by simulating realistic brain activity (more specifically, neural mass models) on the architecture of the structural connectome. The authors show that dynamical models aimed at reproducing the functional connectivity patterns observed in the resting brain exhibit a much better performance when they are tuned around a balanced state favoring the shifting between attractors. This balanced state is described in terms of energy landscape, as discussed also in Watanabe et al. (2014), where the state transitions (in terms of energy landscape) between two representative Resting State Networks—the Default Mode Network and the Fronto-Parietal Network—are addressed.

With an alternative approach, the relation between structural and functional networks is tackled in Ajilore et al. (2013), by using the functional-by-structural hierarchical (FSH) mapping. This was developed for multimodal integration of the resting state fMRI (rsfMRI) and the whole brain (tractography-derived) connectome and is based on the evidence that the level of resting-state functional correlation between any two regions (in general) decreases as the graph distance of the corresponding structural connectivity matrix between them increases. Results are reported in health and depression.

Effective connectivity methods are devised to infer directed connectivity patterns from time-series data. A new method based on Variational Bayesian Inference to infer causality from time series was proposed in Luessi et al. (2014). The method uses a vector autoregressive model for the latent variables describing neuronal activity in combination with a linear observation model based on a convolution with a hemodynamic response function.

The method is validated using both real and synthetic resting fMRI data.

Continuing with the problem of causality inference, classical methods like Granger causality were extended to the situation of time-varying signals in Chicharro and Panzeri (2014). This study also provides a graphical approach to predict dynamic statistical dependencies between the signals from the causal structure.

A different approach is presented in Kolchinsky et al. (2014), where brain regions and networks are characterized by information-theoretic measures using both functional and structural information in a complementary manner. In particular, Kolchinsky et al., quantify the amount of functional coupling between sets of regions of interest (ROIs) as well as integration within sets of ROIs. Several information-based measures are considered, and their scaling with subsystem size is explored.

Regarding consciousness and its relationship with Information Theory, two papers have been contributed to this Research Topic. In Lee et al. (2013) the Approximate Entropy (ApEn), a measure known to correlate with the level of brain consciousness, is used to characterize EEG signals in children and adults to show that the amount of ApEn is lower in children and that it correlates, in children as in adults, with consciousness; in particular, the authors show that ApEn decreases across the transition from awake to REM sleep to non-REM sleep. For patients with deficit of consciousness (DOC) after traumatic brain injury, (Mäki-Marttunen et al., 2013) show two possible markers from fMRI time series that can distinguish between DOC patients and healthy subjects. The inter-hemispheric correlations (but not the intra-hemispheric correlations) between left-right homolog areas decrease, as does the intra-hemispheric information flow, in DOC patients compared to control. For a small group of 4 patients who fully recovered from coma, the study also reports an increase of the intra-hemisphere information flow with respect to controls.

Information-based measures have been associated with altered information processing in Autism Spectrum Disorder (ASD). The close link between active information storage and general theories of cortical function has been addressed in Gómez et al. (2014), by analyzing magnetoencephalography (MEG) signals.

The authors report a significant reduction of information storage in the hippocampus in ASD patients. The amount of information and entropy of MEG signals in ASD (Asperger syndrome in this case) has been analyzed also in Pérez Velázquez and Galán (2013). The analysis, carried out at the source level, addresses the relationship between resting state activity and the brain inner processing with regards to information production, as quantified by the relative entropy. The results suggest that the brains of individuals with ASD produce more information than the age-matched participants.

The complementary role of the three components of information processing, transfer, storage and modification is investigated in Wibral et al. (2014). Local information storage is analyzed in detail in neural data and associated to neural properties such as stimulus preferences and surprise.

The connection between information content of brain activity signals and function has been explored also in Sokunbi (2014). The author investigates the power of a similarity measure (the Sample Entropy) to discriminate between young and elderly subjects emphasizing on the possible limitations arising from the reduced length of time series that are commonly encountered in fMRI studies.

Information-based measures are also useful for developing new technical tools for structural and functional analysis. Novel algorithms have been proposed to improve the construction of the structural connectome in Roine et al. (2014), by investigating the isotropic partial volume effects caused by non-white matter tissue on fiber orientation diffusion estimated with constrained spherical de-convolution. Diffusion weighted signals are simulated with varying diffusion weightings, signal-to-noise ratios, fiber configurations, and tissue fractions.

The equivalence between the information-based and model-based approaches to directed dynamical connectivity in the frequency domain was explored in Takahashi et al. (2014). To enhance the understanding of the possibly complex interaction between multiple time series the authors decompose the established approaches to Directed Coherence into different modes of interaction.

Concerning improvements to functional analysis, a novel information-theoretic approach for spatial components ranking has been proposed in Ossadtchi et al. (2013). The proposed method is based on the Mutual Information (MI) Spectrum which serves as a power-invariant measure of repetitive task-related signal in the temporal loadings of spatial components. Using realistic simulations, the authors in show that the task-relatedness measure, based on estimating the MI between a component and the expanded binary stimulus signal, allows for significantly higher detector characteristics when compared with conventional alternatives. The application of the MI Spectrum for the selection of task-related independent components is validated with real MEG data.

We hope that the reader will find in this Research Topic a useful reference for the state of the art in the emerging field of tools rooted in information theory and applied to neuroscience.

## REFERENCES

- Ajilore, O., Zhan, L., Gadelkarim, J., Zhang, A., Feusner, J. D., Yang, S., et al. (2013). Constructing the resting state structural connectome. *Front. Neuroinform.* 7:30. doi: 10.3389/fninf.2013.00030
- Chicharro, D., and Panzeri, S. (2014). Algorithms of causal inference for the analysis of effective connectivity among brain regions. *Front. Neuroinform.* 8:64. doi: 10.3389/fninf.2014.00064
- Gómez, C., Lizier, J. T., Schaum, M., Wollstadt, P., Grützner, C., Uhlhaas, P., et al. (2014). Reduced predictable information in brain signals in autism spectrum disorder. *Front. Neuroinform.* 8:9. doi: 10.3389/fninf.2014.00009
- Kolchinsky, A., van den Heuvel, M. P., Griffa, A., Hagmann, P., Rocha, L. M., Sporns, O., et al. (2014). Multi-scale integration and predictability in resting state brain activity. *Front. Neuroinform.* 8:66. doi: 10.3389/fninf.2014.00066
- Lee, G. M. H., Fattinger, S., Mouthon, A.-L., Noirhomme, Q., and Huber, R. (2013). Electroencephalogram approximate entropy influenced by both age and sleep. *Front. Neuroinform.* 7:33. doi: 10.3389/fninf.2013.00033
- Luessi, M., Babacan, S. D., Molina, R., Booth, J. R., and Katsaggelos, A. K. (2014). Variational Bayesian causal connectivity analysis for fMRI. *Front. Neuroinform.* 8:45. doi: 10.3389/fninf.2014.00045
- Mäki-Marttunen, V., Diez, I., Cortes, J. M., Chialvo, D. R., and Villarreal, M. (2013). Disruption of transfer entropy and inter-hemispheric brain functional connectivity in patients with disorder of consciousness. *Front. Neuroinform.* 7:24. doi: 10.3389/fninf.2013.00024
- Ossadtchi, A., Pronko, P., Baillet, S., Pflieger, M. E., and Stroganova, T. (2013). Mutual information spectrum for selection of event-related spatial components. Application to eloquent motor cortex mapping. *Front. Neuroinform.* 7:53. doi: 10.3389/fninf.2013.00053
- Pérez Velázquez, J. L., and Galán, R. F. (2013). Information gain in the brain's resting state: a new perspective on autism. *Front. Neuroinform.* 7:37. doi: 10.3389/fninf.2013.00037
- Roine, T., Jeurissen, B., Perrone, D., Aelterman, J., Leemans, A., Philips, W., et al. (2014). Isotropic non-white matter partial volume effects in constrained spherical deconvolution. *Front. Neuroinform.* 8:28. doi: 10.3389/fninf.2014.00028
- Sanz Leon, P., Knock, S. A., Woodman, M. M., Domide, L., Mersmann, J., McIntosh, A. R., et al. (2013). The Virtual Brain: a simulator of primate brain network dynamics. *Front. Neuroinform.* 7:10. doi: 10.3389/fninf.2013.00010
- Sokunbi, M. O. (2014). Sample entropy reveals high discriminative power between young and elderly adults in short fMRI data sets. *Front. Neuroinform.* 8:69. doi: 10.3389/fninf.2014.00069
- Takahashi, D. Y., Baccalá, L. A., and Sameshima, K. (2014). Canonical information flow decomposition among neural structure subsets. *Front. Neuroinform.* 8:49. doi: 10.3389/fninf.2014.00049
- Watanabe, T., Hirose, S., Wada, H., Imai, Y., Machida, T., Shirouzu, I., et al. (2014). Energy landscapes of resting-state brain networks. *Front. Neuroinform.* 8:12. doi: 10.3389/fninf.2014.00012
- Wibral, M., Lizier, J. T., Vögler, S., Priesemann, V., and Galuske, R. (2014). Local active information storage as a tool to understand distributed neural information processing. *Front. Neuroinform.* 8:1. doi: 10.3389/fninf.2014.00001

**Conflict of Interest Statement:** The authors declare that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.

Received: 22 October 2014; accepted: 03 December 2014; published online: 19 December 2014.

Citation: Cortes JM, Marinazzo D and Muñoz MA (2014) Editorial for the research topic: information-based methods for neuroimaging: analyzing structure, function and dynamics. *Front. Neuroinform.* 8:86. doi: 10.3389/fninf.2014.00086

This article was submitted to the journal *Frontiers in Neuroinformatics*.

Copyright © 2014 Cortes, Marinazzo and Muñoz. This is an open-access article distributed under the terms of the Creative Commons Attribution License (CC BY). The use, distribution or reproduction in other forums is permitted, provided the original author(s) or licensor are credited and that the original publication in this journal is cited, in accordance with accepted academic practice. No use, distribution or reproduction is permitted which does not comply with these terms.



# The Virtual Brain: a simulator of primate brain network dynamics

**Paula Sanz Leon<sup>1\*</sup>, Stuart A. Knock<sup>2</sup>, M. Marmaduke Woodman<sup>1</sup>, Lia Domide<sup>3</sup>, Jochen Mersmann<sup>4</sup>, Anthony R. McIntosh<sup>5</sup> and Viktor Jirsa<sup>1\*</sup>**

<sup>1</sup> Institut de Neurosciences des Systèmes, Aix Marseille Université, Marseille, France

<sup>2</sup> Department of Neurology, BrainModes Group, Charité University of Medicine, Berlin, Germany

<sup>3</sup> Codemart, Cluj-Napoca, Romania

<sup>4</sup> CodeBox GmbH, Stuttgart, Germany

<sup>5</sup> Rotman Research Institute at Baycrest, Toronto, ON, Canada

## Edited by:

Daniele Marinazzo, University of Gent, Belgium

## Reviewed by:

Ingo Bojak, University of Reading, UK

Hans Ekkehard Plesser, Norwegian University of Life Sciences, Norway  
Laurent U. Perrinet, Centre National de la Recherche Scientifique, France

## \*Correspondence:

Paula Sanz Leon and Viktor Jirsa,  
Institut de Neurosciences des Systèmes, Aix Marseille Université, 27, Bd. Jean Moulin, 13005 Marseille, France  
e-mail: paula.sanz-leon@univ-amu.fr; viktor.jirsa@univ-amu.fr

We present The Virtual Brain (TVB), a neuroinformatics platform for full brain network simulations using biologically realistic connectivity. This simulation environment enables the model-based inference of neurophysiological mechanisms across different brain scales that underlie the generation of macroscopic neuroimaging signals including functional MRI (fMRI), EEG and MEG. Researchers from different backgrounds can benefit from an integrative software platform including a supporting framework for data management (generation, organization, storage, integration and sharing) and a simulation core written in Python. TVB allows the reproduction and evaluation of personalized configurations of the brain by using individual subject data. This personalization facilitates an exploration of the consequences of pathological changes in the system, permitting to investigate potential ways to counteract such unfavorable processes. The architecture of TVB supports interaction with MATLAB packages, for example, the well known Brain Connectivity Toolbox. TVB can be used in a client-server configuration, such that it can be remotely accessed through the Internet thanks to its web-based HTML5, JS, and WebGL graphical user interface. TVB is also accessible as a standalone cross-platform Python library and application, and users can interact with the scientific core through the scripting interface IDLE, enabling easy modeling, development and debugging of the scientific kernel. This second interface makes TVB extensible by combining it with other libraries and modules developed by the Python scientific community. In this article, we describe the theoretical background and foundations that led to the development of TVB, the architecture and features of its major software components as well as potential neuroscience applications.

**Keywords:** connectome, neural masses, time delays, full-brain network model, virtual brain, large-scale simulation, web platform, python

## 1. INTRODUCTION

Brain function is thought to emerge from the interaction of large numbers of neurons, under the spatial and temporal constraints of brain structure and cognitive demands. Contemporary network simulations mainly focus on the microscopic and mesoscopic level (neural networks and neural masses representing a particular cortical region), adding detailed biophysical information at these levels of description while too often losing perspective on the global dynamics of the brain. On the other hand, the degree of assessment of global cortical dynamics, across imaging modalities, in human patients and research subjects has increased significantly in the last few decades. In particular, cognitive and clinical neuroscience employs imaging methods of macroscopic brain activity such as intracerebral measurements, stereotactic Encephalography (sEEG) (von Ellenrieder et al., 2012), Electroencephalography (EEG) (Nunez and Srinivasan, 1981; Nunez, 1995; Niedermeyer and Lopes Da Silva, 2005), Magnetoencephalography (MEG) (Hämäläinen,

1992; Hämäläinen et al., 1993; Mosher et al., 1999), and functional Magnetic Resonance Imaging (fMRI) (Ogawa et al., 1993, 1998; Logothetis et al., 2001) to assess brain dynamics and evaluate diagnostic and therapeutic strategies. Hence, there is a strong motivation to develop an efficient, flexible, neuroinformatics platform on this macroscopic level of brain organization for reproducing and probing the broad repertoire of brain dynamics, enabling quick data analysis and visualization of the results.

The Virtual Brain (TVB) is our response to this need. On the one hand, it provides a general infrastructure to support multiple users handling various kinds of empirical and simulated data, as well as tools for visualizing and analyzing that data, either via internal mechanisms or by interacting with other computational systems such as MATLAB. At the same time it provides a simulation toolkit to support a top-down modeling approach to whole brain dynamics, where the underlying network is defined by its structural large-scale connectivity and mesoscopic models that govern the nodes' intrinsic dynamics. The interaction with the



dynamics of all other network nodes happens through the connectivity matrix via specific connection weights and time delays, where the latter make a significant contribution to the biological realism of the temporal structure of dynamics.

Historically, Jirsa et al. (2002) first demonstrated neural field modeling on a spherical brain hemisphere employing EEG and MEG forward solutions to obtain simulation imaging signals. In this work, homogeneous (translationally invariant) connectivity was implemented along the lines of Jirsa and Haken (1996, 1997); Bojak and Liley (2010) yielding a neural field equation, which has its roots in classic works (Wilson and Cowan, 1972, 1973; Nunez, 1974; Amari, 1975, 1977). At that time more detailed large-scale connectivity of the full primate brain was unavailable, hence the homogeneous connectivity scaled up to the full brain was chosen as a first approximation (Nunez, 1974). The approach proved successful for the study of certain phenomena as observed in large-scale brain systems including spontaneous activity (Wright and Liley, 1995; Robinson et al., 2001, 2003; Breakspear et al., 2003; Rowe et al., 2004; Freyer et al., 2011), evoked potentials (Rennie et al., 1999, 2002), anesthesia (Liley and Bojak, 2005), epilepsy (Breakspear et al., 2006), sensorimotor coordination (Jirsa and Haken, 1996, 1997), and more recently, plasticity (Robinson, 2011) [see Deco et al. (2008) and Jirsa (2004) for a review].

Careful review of this literature though shows that these models mostly emphasize the temporal domain of brain organization, but leave the spatiotemporal organization untouched. This may be understood by the fact that the symmetry of the connectivity imposes constraints upon the range of the observable dynamics. This was pointed out early by Jirsa et al. (2002) and a suggestion was made to integrate biologically realistic DTI based connectivity into full brain modeling efforts. Large scale brain dynamics are basically expected to reflect the underlying anatomical connectivity between brain areas (Bullmore and Sporns, 2009; Deco et al., 2011), even though structural connectivity is not the only constraint, but the transmission delays play an essential role in shaping the brain network dynamics also (Jirsa and Kelso, 2000; Ghosh et al., 2008; Knock et al., 2009; Jirsa et al., 2010). Recent studies (Pinotsis et al., 2012) have systematically investigated the degree to which homogeneous approximations may serve to understand realistic connection topologies and have concluded that homogeneous approximations are more appropriate for mesoscopic descriptions of brain activity, but less well suited to address full brain network dynamics. All this underscores the need to incorporate realistic connectivity into large scale brain network models. Thus the simulation side of TVB has evolved out of a research program seeking to identify and reproduce realistic whole brain network dynamics, on the basis of empirical connectivity and neural field models (Jirsa and Stefanescu, 2010; Deco et al., 2011).

## 1.1. MODELING

In line with these previous studies, TVB incorporates a biologically realistic, large-scale connectivity of brain regions in the primate brain. Connectivity is mediated by long-range neural fiber tracts as identified by tractography based methods (Hagmann et al., 2008; Honey et al., 2009; Bastiani et al., 2012), or obtained

from CoCoMac neuroinformatics database (Kötter, 2004; Kötter and Wanke, 2005; Bakker et al., 2012). In TVB, the tract-lengths matrix of the demonstration connectivity dataset is symmetric due to the fiber detection techniques used to extract the information being insensitive to directionality. On the other hand, the weights matrix is asymmetric as it makes use of directional information contained in the tracer studies of the CoCoMac database. Such details are specific to the connectivity demonstration dataset included in the distribution packages of TVB. The symmetry (or lack thereof) is neither a modeling constraint nor an imposed restriction on the weights and tract-length matrices. The general implementation for weights and tract lengths are full  $n_{nodes} \times n_{nodes}$  matrices without any symmetry restrictions.

Two types of structural connectivity are distinguished in TVB, that is long- and short-range connectivity, given by the connectivity matrix and the folded cortical surface, respectively. The connectivity matrix defines the connection strengths and time delays via finite signal transmission speed between two regions of the brain. The cortical surface allows a more detailed spatial sampling resulting in a spatially continuous approximation of the neural activity as in neural field modeling (Deco et al., 2008; Coombes, 2010; Bressloff, 2012). When using neural mass models, building the network upon the surface allows for the application of arbitrary local connectivity kernels which represent short-range intra-cortical connections. Additionally, networks themselves can be defined at two distinct spatial scales yielding two types of simulations (or brain network models): surface-based and region-based. In the former case, cortical and sub-cortical areas are shaped more realistically, each vertex of the surface is considered a node and is modeled by a neural population model; several nodes belong to a specific brain region, and the edges of the network have a distance of the order of a few millimeters. The influence of delayed activity coming from other brain regions is added to the model via the long-range connectivity. In the latter case of nodes only per region, the connectome itself is used as a coarser representation of the brain network model. The networks comprise discrete nodes, each of which models the neural population activity of a brain region and the edges represent the long-range connectivity (interregional fibers) on the order of a few centimeters. Consequently, in surface-based simulations both types of connectivity, short- and long-range, coexist whereas in region-based simulations one level of geometry is lost: the short-range connectivity.

Neural field models have been developed over many years for their ability to capture the collective dynamics of relatively large areas of the brain in both analytically and computationally tractable forms (Beurle, 1956; Wilson and Cowan, 1972, 1973; Nunez, 1974; Amari, 1975, 1977; Wright and Liley, 1995; Jirsa and Haken, 1996, 1997; Robinson et al., 1997; Jirsa et al., 2002; Atay and Hutt, 2006; Bojak and Liley, 2010). Effectively neural field equations are tissue level models that describe the spatiotemporal evolution of coarse grained variables such as synaptic voltage or firing rate activity in populations of neurons. Some of these models include explicit spatial terms while others are formulated without an explicit spatial component leaving open the possibility to apply effectively arbitrary local connectivity kernels. The lumped representation of the dynamics of a set of similar neurons

via a common variable (e.g., mean firing rate and mean postsynaptic potential) is known as neural mass modeling (Freeman, 1975, 1992; Lopes da Silva et al., 1974). Neural mass models accounting for parameter dispersion in the neuronal parameters include Assisi et al., 2005; Stefanescu and Jirsa, 2008, 2011; Jirsa and Stefanescu, 2010. Networks of neural masses, without an explicit spatial component within the mass but with the possibility to apply local connectivity kernels (e.g., Gaussian or Laplacian functions) between masses, can be used to approximate neural field models. Both neural field and neural mass modeling approaches embody the concept from statistical physics that macroscopic physical systems obey laws that are independent of the details of the microscopic constituents of which they are built (Haken, 1983). These and related ideas have been exploited in neurosciences (Kelso, 1995; Buzsaki, 2006).

In TVB, our main interest lies in using the mesoscopic laws governing the behavior of neural populations and uncovering the laws driving the processes on the macroscopic brain network scale. The biophysical mechanisms available to microscopic single neuron approaches are absorbed in the mean field parameters on the mesoscopic scale and are not available for exploration other than through variation of the mean field parameters themselves. As a consequence, TVB represents a neuroinformatics tool that is designed to aid in the exploration of large-scale network mechanisms of brain functioning [see Ritter et al. (2013) for an example of modeling with TVB].

Furthermore, TVB's approach to multi-modal neuroimaging integration in conjunction with neural field modeling shares common features with the work of Bojak et al. (2010, 2011) and Babajani-Feremi and Soltanian-Zadeh (2010). The crucial difference is that the structure upon which TVB has been designed represents a generalized large-scale "computational neural model" of the whole brain. The components of this large-scale model have been separated as cleanly as possible, and a specific structure has been defined for the individual components. This generic structure is intended to serve the purpose of restricting the form of models enough to make direct comparison straight forward while still permitting a sufficiently large class of models to be expressed. Likewise, the paradigms presented during the last few years in this line of research (Sotero et al., 2007; Sotero and Trujillo-Barreto, 2008) could potentially be reproduced, tested and compared in TVB. The mathematics underlying our model-based approach have been partially described in various original articles (Deco et al., 2011; Deco and Jirsa, 2012) and will be reviewed in more detail in future articles.

## 1.2. INFORMATICS

From an informatics perspective, a large-scale simulation project requires a well defined yet flexible workflow, i.e., adaptable according to the users profiles. A typical workflow in TVB involves managing project information, uploading data, setting up simulation parameters (model, integration scheme, output modality), launching simulations (in parallel if needed), analyzing and visualizing, and finally storing results and sharing output data.

The web interface allows users without programming knowledge to access TVB to perform customized simulations (e.g.,

clinicians could use their patient's data obtained from DTI studies). In addition, it enables them to gain a deeper understanding of the theoretical approaches behind the scenes. On the other hand, theoreticians can design their own models and get an idea of their biophysical realism, their potential physiological applications and implications. As both kinds of users may work within the same framework, the interplay of theory and experiment or application is accelerated. Additionally, users with stronger programming skills benefit from all the advantages provided by the Python programming language: easy-to-learn, easy-to-use, scriptable and with a large choice of scientific modules (Oliphant, 2006).

TVB has been principally built in the Python programming language due to its unique combination of flexibility, existing libraries and the ease with which code can be written, documented, and maintained by non-programmers. The simulation core, originally developed in MATLAB, was ported to Python given its current significance in the numerical computing and neuroscience community and its already proven efficiency for implementing modeling tools (Spacek et al., 2008).

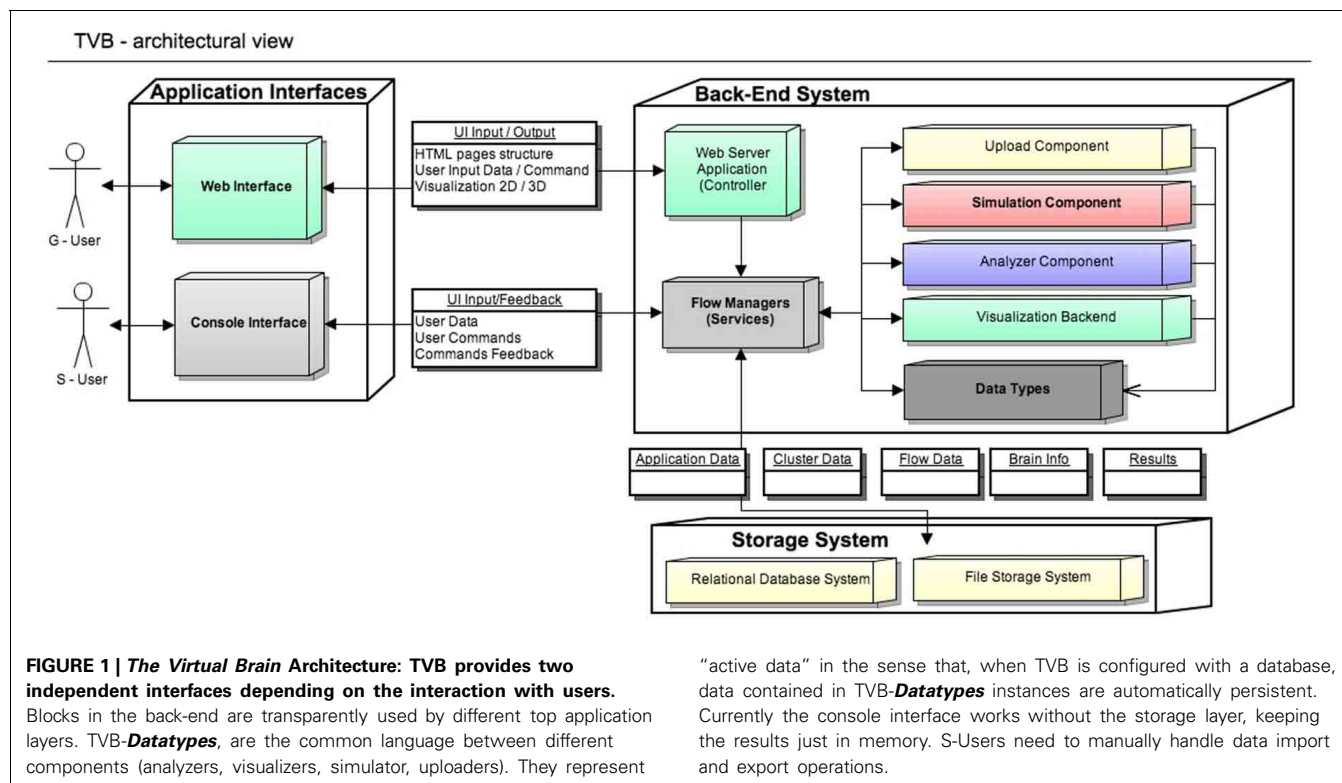
Simulations benefit from vectorized numerical computations with NumPy arrays and are enhanced by the use of the *num-expr* package. Although this allows rather efficient single simulations, the desire to systematically explore the parameter spaces of the neural dynamic models, and to compare many connectivity matrices, has lead to the implementation of code generation mechanisms for the majority of the simulator core—producing C code for both native CPU and also graphics processing units (GPU), leading to a significant speed up of parameter sweeps and parallel simulations (5x from Python to C, 40x from C to GPU). Such graphics units have become popular in scientific computing for their relatively low price and high computing power. Going forward, the GPU implementation of TVB will require testing and optimization before placing it in the hands of users.

This article intends to give a comprehensive but non-exhaustive description of TVB, from both technical and scientific points of view. It will describe the framework's architecture, the simulation core, and the user interfaces. It will also provide two examples, using specific features of the simulator, extracted from the demo scripts which are currently available in TVB's distribution packages.

## 2. TVB ARCHITECTURE

The architectural model of the system has two main components: the scientific computing core and the supporting framework with its graphical user interface. Both software components communicate through an interface represented by TVB-*Datatypes*, which are described in section 2.2. In **Figure 1** TVB's architectural details are illustrated and explained in more depth.

**General aspects:** TVB is designed for three main deployment configurations, according to the available hardware resources: (1) Stand Alone; (2) Client-Server or, (3) Cluster. In the first, a local workstation is assumed to have certain display, computing power and storage capacity resources. In the second, an instance of TVB is running on a server connected through a network link to client units, and thus accessible to a certain number of users.



In this deployment model, simulations use the back-end server's computing power while visualization tasks use resources from the client machine. The third is similar to the client-server configuration, but with the additional advantage of parallelization support in the back-end. The cluster itself needs to be configured separately of TVB.

Based on the usage scenarios and user's level of programming knowledge, two user profiles are represented: a graphical user (G-user) and scripting user (S-user). We therefore provide the corresponding main interfaces based on this classification: a graphical user interface (web) and a scripting interface (IDLE). S-users and G-users have different levels of control over different parts of the system. The profile of S-users is thought to be that of scientific developers, that is, researchers who can elaborate complex modeling scenarios, add their own models or directly modify the source code to extend the scientific core of TVB, mostly working with the scientific modules. They do, nevertheless, have the possibility to enable the database system. In contrast, G-users are relatively more constrained to the features available in the stable releases of TVB, since their profile corresponds more to that of researchers without a strong background in computational modeling. The distinction between these two profiles is mainly a categorization due to the design architecture of TVB. For instance, we could also think of other type of users who want to work with TVB's GUI and are comfortable with programming, and therefore they could potentially make modifications in the code and then see the effect of those when launching the application in a web browser.

The development of TVB is managed under Agile techniques. In accord therewith, each task is considered as *done*,

after completing a validation procedure that includes: adding a corresponding automated unit-test, labeling the task as *finished* from the team member assigned to implement the task and further tagging as *closed* from a team member responsible for the module, which means a second level of testing. Before releasing stable packages, there is a period for manual testing, that is, a small group of selected users from different institutions check the main features and functionalities through both interfaces. The navigation and workflows scenarios through the web-based interface are evaluated by means of automated integration tests for web-applications running with Selenium (<http://docs.seleniumhq.org/>) and Apache-JMeter (<http://jmeter.apache.org/>) on top of a browser engine. Special effort is being made to provide good code-coverage, including regression tests. Accordingly, the simulation engine of TVB has automated unit-tests, to guarantee the proper and coordinated functioning of all the modules, and simple programs (demonstration scripts), that permit qualitative evaluation of the scientific correctness of results.

The development version of TVB is currently hosted on a private cluster, where we use the version control system *svn* (subversion). Additionally, as any large collaborative open-source project, it is also available in a public repository, using the distributed version control system *git* (Chacon, 2009) to make accessible the scientific core and to gather, manage and integrate contributions from the community. The distribution packages for TVB come with an extensive documentation, including: a *User Guide*, explaining how to install TVB, set up models and run them; *Tutorials*, *Use Cases* and *Script Demos*, guiding users to achieve predefined simulation scenarios; and a *Developer Guide* and *API reference*. **Table 1** provides the links

**Table 1 | TVB links.**

TVB official website	<a href="http://www.thevirtualbrain.org">http://www.thevirtualbrain.org</a>
Distribution packages	<a href="http://www.thevirtualbrain.org/register">http://www.thevirtualbrain.org/register</a>
Public repository	<a href="https://github.com/the-virtual-brain">https://github.com/the-virtual-brain</a>
User group	<a href="https://groups.google.com/group/tvb-users/">https://groups.google.com/group/tvb-users/</a>

to: the official TVB website, where distribution packages for Linux and Mac OS (32 and 64 bits) and Windows (32 bits) are available for download; the active users group of TVB hosted in Google Groups, where users can ask questions, report issues and suggest improvements or new features; and the public repository, where the source code of both the framework and scientific library (which contains the simulation engine) are available.

**Installation and System Requirements:** When using the web interface, users are recommended to have a high definition monitor (at least 1600 × 1000 pixels), a WebGL and WebSockets compatible browser (latest versions of Mozilla Firefox, Apple Safari or Google Chrome), and a WebGL-compatible graphics card, that supports OpenGL version 2.0 or higher (Shreiner et al., 2005).

Regarding memory and storage capacity, for a stand alone installation a minimum of 8 GB of RAM is recommended. For multi-users environments 5 GB of space per user is suggested. This is a storage quota specified by an administrator to manage the maximum hard disk space per user. As for computing power one CPU core is needed for a simulation with a small number of nodes, while simulations with a large number of nodes, such as surface simulations, can make use of a few cores if they are available. When the number of launched simulations is larger than the number of available cores, a serialization is recommended (a serialization mechanism is provided by the supporting framework through the web user interface by specifying the maximum of simultaneous jobs allowed). In order to use the Brain Connectivity Toolbox (Rubinov and Sporns, 2010), MATLAB or Octave should be installed, activated and accessible for the current user.

## 2.1. TVB FRAMEWORK

The supporting framework provides a database back-end, workflow management and a number of features to support collaborative work. The latter feature permits TVB to be setup as a multi-user application. In this configuration, a login system enables users to access their personal sessions; by default their projects and data are private, but they can be shared with other users. The graphical user interface (GUI) is web based, making use of HTML 5, WebGL, CSS3 and Java Script (Bostock et al., 2011) tools to provide an intuitive and responsive interface that can be locally and remotely accessed.

### 2.1.1. Web-based GUI

TVB provides a web-based interactive framework to generate, manipulate and visualize connectivity and network dynamics. Additionally, TVB comprises a set of classic time-series analysis tools, structural and functional connectivity analysis tools, as well

as parameter exploration facilities which can launch simulations in parallel on a cluster or on multiple compute cores of a server. The GUI of TVB has six main working areas: **USER**, **PROJECT**, **SIMULATOR**, **ANALYZE**, **STIMULUS**, and **CONNECTIVITY**. In **USER**, the users manage their accounts and TVB settings. In **PROJECT**, individual projects are managed and navigation tools are provided to explore their structure as well as the data associated with them. A sub-menu within this area provides a dashboard with a list of all the operations along with their current status (running, error, finished), owner, wall-time and associated data, among other information. In **SIMULATOR** the large-scale network model is set up and simulations launched, additional viewers for structural and functional data are offered in 2D and 3D, as well as other displays to visualize the results of a simulation. A history of simulations is also available in this area. In **ANALYZE** time-series and network analysis methods are provided. In **STIMULUS**, users can interactively create stimulation patterns. Finally, in **CONNECTIVITY**, users are given a responsive interface to edit the connectivity matrices assisted by interactive visualization tools. **Figure 2** depicts the different working areas, as well as a number of their sub-menus, available through the web UI. A selection of screenshots illustrating the interface in a web browser is given in **Figure 3**.

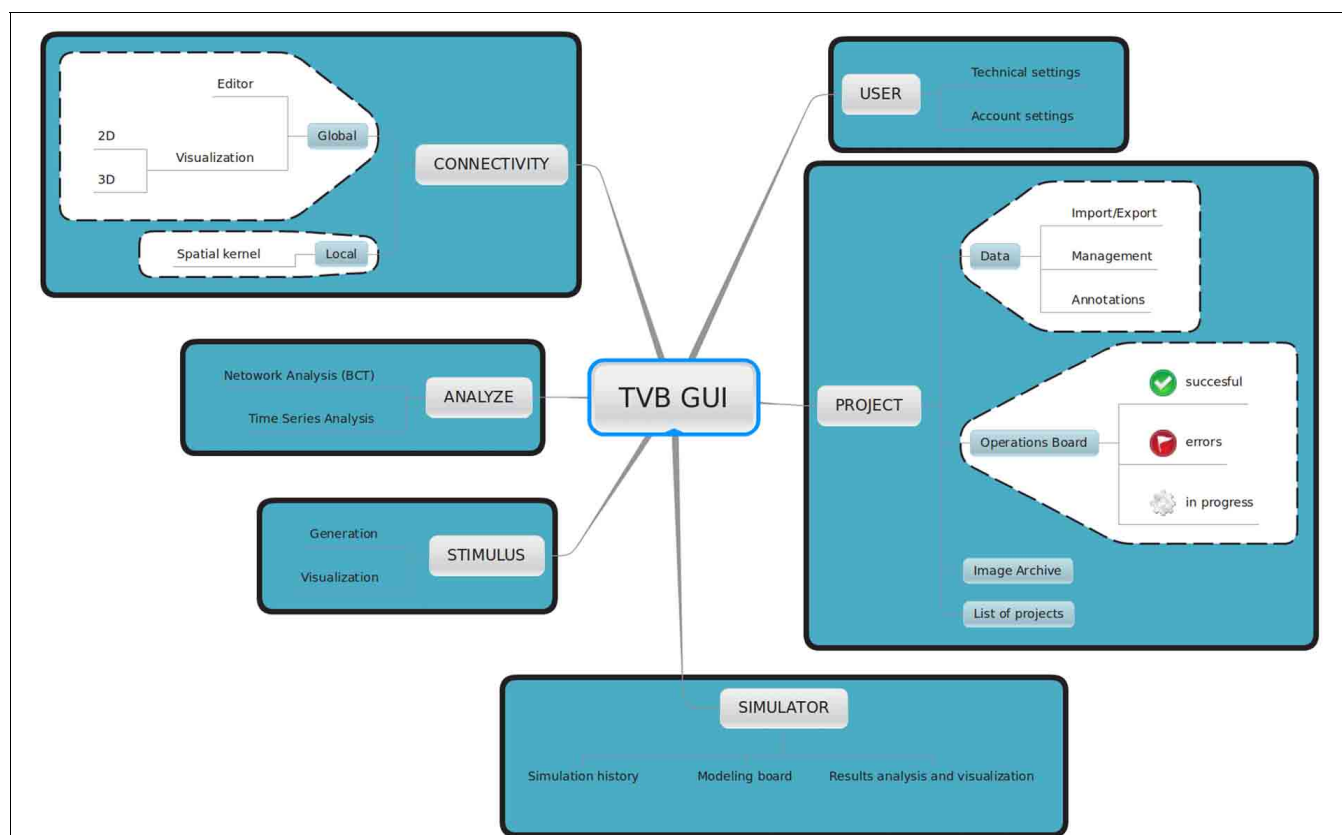
### 2.1.2. Data management and exchange

One of the goals of TVB is to allow researchers from different backgrounds and with different programming skills to have quick access to their simulated data. Data from TVB can be exchanged with other instances of TVB (copies installed on different computers) or with other applications in the neuroscientific community, e.g., MATLAB, Octave, The Connectome ToolKit (Gerhard et al., 2011).

**Export:** A project created within TVB can be entirely exported to a .zip file. Besides storing all the data generated within a particular project in binary files, additional XML files are created to provide a structured storage of metadata, especially with regard to the steps taken to set up a simulation, configuration parameters for specific operations, time-stamps and user account information. This mechanism produces a summary of the procedures carried on by researchers within a project which is used for sharing data across instances of TVB. The second export mechanism allows the export of individual data objects. The data format used in TVB is based on the HDF5 format (The HDF Group, 2010) because it presents a number of advantages over other formats: (1) huge pieces of data can be stored in a condensed form; (2) it allows grouping of data in a tree structure; (3) it allows meta-data assignment at every level; and (4) it is a widely used format, accessible in several programming languages and applications. Additionally, each object in TVB has a global unique identifier (GUID) which makes it easy to identify an object across systems, avoiding naming conflicts among files containing objects of the same type.

**Import:** A set of mechanisms (“uploaders”) is provided in TVB to import data into the framework, including neuroimaging data generated independently by other applications. The following formats are supported: NIFTI-1 (volumetric time-series), GIFTI





**FIGURE 2 | Main working areas of *The Virtual Brain*'s web interface: in **USER** personal information (account settings) as well as hardware and software preferences (technical settings) are configured. Through the **PROJECT** area users access and organize their projects, data, figures and the operations dashboard. Input and output simulated data can be exported in HDF5 format and may be used outside of the framework. Brain network models and execution of simulations are configured and launched,**

respectively in **SIMULATOR**. In this area results can be immediately analyzed and visualized to have a quick overview of the current model. A history of launched simulations is kept to have the traceability of any modifications that took place in the simulation chain. **STIMULUS** provides a collection of tools to build stimulation patterns that will be available to use in the simulations. Finally, **CONNECTIVITY** provides an interactive environment to the edit and visualize connectivity matrices.

(surfaces) and CFF (connectome file). General compression formats, such as ZIP and BZIP2 are also supported for certain data import routines that expect a set of ASCII text files compressed in an archive. Hence the use of general compression formats means that preparing datasets for TVB is as simple as generating an archive with the correct ASCII files, in contrast to some of the other neuroscientific data formats found elsewhere. For instance, a **Connectivity** dataset (*connectome*) may be uploaded as a zip folder containing the following collection of files: (1) areas.txt, (2) average\_orientations.txt, (3) info.txt, (4) positions.txt, (5) tract\_lengths.txt, and (6) weights.txt. More conventions and guidelines to use each uploader routine can be found in the *User Guide* of TVB's documentation.

### 2.1.3. File storage

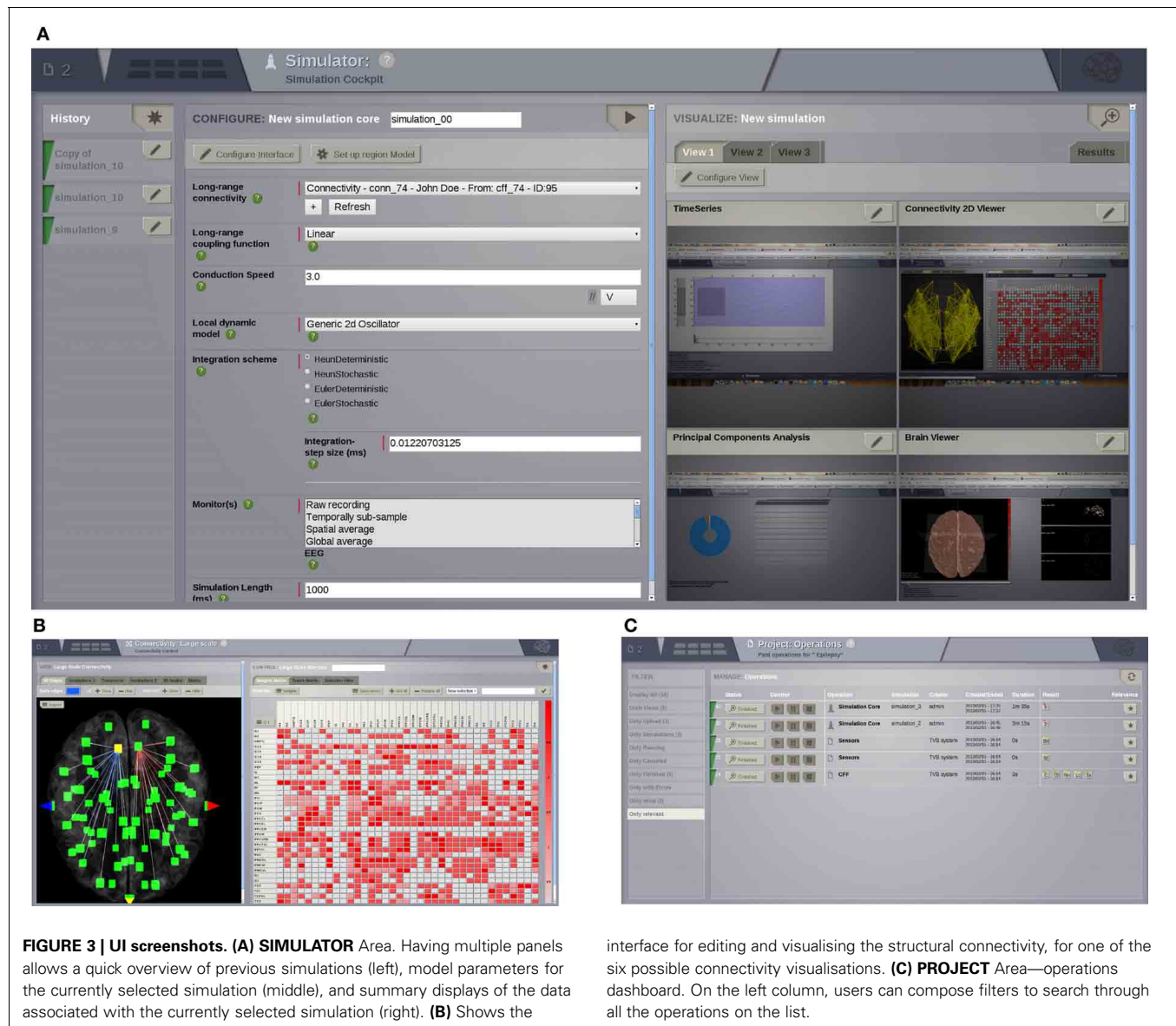
The storage system is a tree of folders and files. The actual location on disk is configurable by the user, but the default is a folder called "TVB" in the user's home folder. There is a sub-folder for each **Project** in which an XML file containing details about the project itself is stored. Then for each operation, one folder per operation is created containing a set of .h5

files generated during that particular operation, and one XML file describing the operation itself. The XML contains tags like *creation date*, *operation status* (e.g., Finished, Error), *algorithm reference*, *operation GUID*, and most importantly *input parameters dictionary*. Sufficiently detailed information is stored in the file system to be able to export data from one instance of TVB and to then import it into another instance, correctly recreating projects, including all operations and their results. Even though the amount of data generated per operation varies greatly, since it depends strongly on the Monitors used and parameters of the simulation, some rough estimates are given below:

- A 1000 ms long, region-based simulation with all the default parameters requires approximately 1 MB of disk space.
- A 10 ms long, surface-based simulation, using a precalculated sparse matrix to describe the local connectivity kernel and all the default parameters, requires about 280 MB.

Users can manually remove unused data using the corresponding controls in TVB's GUI. In this case, all files related to these





data are also deleted, freeing disk space. The amount of physical storage space available to TVB can be configured in the **USER** → **Settings** working area of the GUI—this is, of course, limited by the amount of free space available on the users hard drives.

#### 2.1.4. Database management system

Internally, TVB framework uses a relational database (DB), for ordering and linking entities and as an indexing facility to quickly look up data. At install time, users can choose between SQLite (a file based database and one of the most used embedded DB systems) and PostgreSQL (a powerful, widely spread, open-source object-relational DB system which requires a separate installation by users) as the DB engine. In the database, only references to the entities are stored, with the actual operation results always being stored in files,

due to size. A relational database was chosen as it provides speed when filtering entities and navigating entity relationship trees.

#### 2.2. TVB DATATYPES

In the architecture of TVB, a middleware layer represented by TVB-*Datatypes* allows the handling and flow of data between the scientific kernel and the supporting framework. TVB-*Datatypes* are annotated data structures which contain one or more data attributes and associated descriptive information, as well as methods for operating on the data they contain. The definition of a *Datatype* is achieved using TVB's traiting system, which was inspired by the traiting system developed by Enthought (Enthought, 2001). The traiting system of TVB, among other things, provides a mechanism for annotating data, that is, associating additional information with the data which is itself usually

a single number or an array of numbers. A complete description of TVB's traiting system is beyond the scope of this article. However, in describing TVB's **Datatypes** we will give an example of its use, which should help to provide a basic understanding of the mechanism.

A number of basic TVB-**Datatypes** are defined based on Types that are part of the traiting system, with these traited Types, in turn, wrapping Numpy data types. For instance, TVB-**FloatArray** is a datatype derived from the traiting system's Array type, which in turn wraps Numpy's **ndarray**. The traiting system's Array type has attributes or annotations, such as: `dtype`, the numerical type of the data contained in the array; `label`, a short (typically one or two word) description of what the Array refers to, this information is used by the supporting framework to create a proper label for the GUI; `doc`, a longer description of what the Array refers to, allowing the direct integration of useful documentation into array objects; and `default`, the default value for an instance of an Array type. In the case of a **FloatArray**, the `dtype` attribute is fixed as being `numpy.float64`.

More complex, higher-level, TVB-**Datatypes** are then built up with attributes that are themselves basic TVB-**Datatypes**. For example, TVB-**Connectivity** is datatype which includes multiple **FloatArrays**, as well as a number of other traited types, such

as **Integer** and **Boolean**, in its definition. An example of a **FloatArray** being used to define an attribute of a **Connectivity** can be seen in **Code 1**. The high-level **Datatypes** currently defined in TVB are summarized in **Table 2**.

An example indicating the usage and features of TVB-**Datatypes** is provided below. When a user uploads a connectivity dataset through the UI, an instance of a **Connectivity** datatype is generated. This **Connectivity** datatype is one of the required input arguments when creating an instance of **SIMULATOR**. As a result of the execution of a simulation, other TVB-**Datatypes** are generated, for instance one or more **TimeSeries** datatypes. Specifically, if the simulation is run using the MEG and EEG recording modalities then **TimeSeriesMEG**, **TimeSeriesEEG**, which are subclasses of **TimeSeries**, are returned. Both the **Connectivity**

**Code 1 | An instance of TVB's *FloatArray* Datatype being used to define the conduction speed between brain regions as an attribute of a *Connectivity* Datatype.**

```
speed = FloatArray(
    label = "Conduction_speed",
    default = numpy.array([3.0]),
    doc = """A single number or matrix of conduction speeds for the
    myelinated fibre tracts between regions.""")
```

**Table 2 | TVB Datatypes.**

Base class datatype	Description	Derived classes
Connectivity	Maps connectivity matrix data	Connectivity
Surfaces	Covers surface representations	CorticalSurface, SkinAir, BrainSkull, SkullSkin, EEGCap, FaceSurface, Cortex, RegionMapping, LocalConnectivity
Volumes	Wraps volumetric data	ParcellationMask, StructuralMRI
Sensors	Wraps sensors data used in different acquisition techniques to generate physiological recordings	SensorsEEG, SensorsMEG, SensorsInternal
ProjectionMatrix	Wraps matrices defining a linear operator to map the spatial sources into the leadfield domain	ProjectionRegionEEG, ProjectionSurfaceEEG, ProjectionRegionMEG
	It relates two datatypes: a source of type Connectivity or Surface and a set of Sensors	ProjectionSurfaceMEG
Equations	The matrix is computed using OpenMEEG. (Gramfort et al., 2010) Commonly used functions for defining local connectivity kernels and stimulation patterns	
SpatialPattern	Contains patterns mainly used as stimuli. It makes use of Equation datatypes	SpatioTemporalPattern, StimuliRegion, StimuliSurface, SpatialPatternVolume
TimeSeries	One of the most important TVB-Datatypes. Derived classes wrap measurements recorded under different acquisition modalities	TimeSeriesRegion, TimeSeriesSurface, TimeSeriesVolume, TimeSeriesEEG, TimeSeriesMEG
Graph	Wraps results from a covariance analysis or results from BCT analyzers	Covariance, ConnectivityMeasure
MappedValues	Wraps a single value computed from a TimeSeries object	
ModeDecomposition	Wraps results from matrix factorization analysis (i.e., PCA and ICA)	PrincipalComponents, IndependentComponents
Spectral	Wraps results from frequency analysis	FourierSpectrum, WaveletCoefficients, ComplexCoherenceSpectrum

Specifications about the requirements to build a TVB-Datatype can be found in the documentation of the distribution packages.

and *TimeSeries* datatypes are accepted by a range of appropriate analysis and visualization methods.

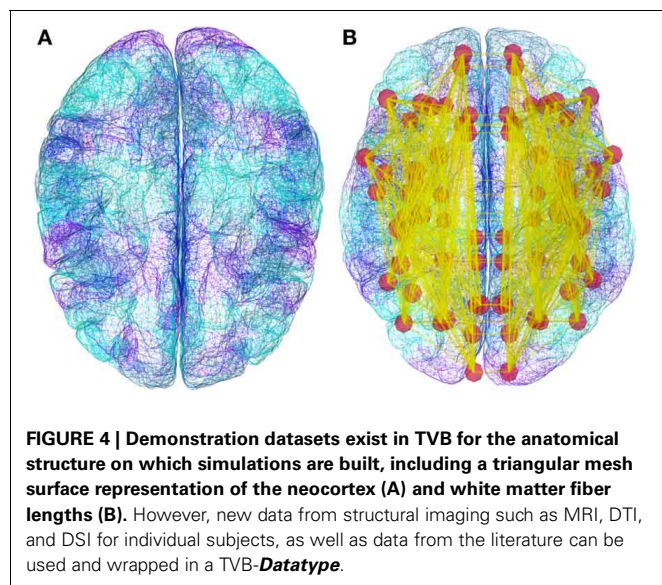
Further, TVB-*Datatypes* have attributes and metadata which remains accessible after exporting in TVB format. The meta-data includes a technical description of the data (storage size for instance) as well as scientifically relevant properties and useful documentation to properly interpret the dataset. In the shell interface, the attributes of TVB-*Datatype* can be accessed by their key-names in the same way as Python dictionaries.

### 2.3. TVB SIMULATOR

The simulation core of TVB brings together a mesoscopic model of neural dynamics with structural data. The latter defines both the spatial support (see **Figure 4**), upon which the brain network model is built, and the hierarchy of anatomical connectivity, that determines the spatial scale represented by the structural linkages between nodes (Freeman, 1975). Simulations then recreate the emergent brain dynamics by numerically integrating this coupled system of differential equations. All these entities have their equivalent representation as *classes* either in the scientific *MODULES* or *datatypes*, and are bound together in an instance of the *Simulator* class. In the following paragraphs we describe all the individual components required to build a minimal representation of a brain network model and run a simulation, as well as the outline of the operations required to initialize a *Simulator* object and the operations of the update scheme.

#### 2.3.1. Coupling

The brain activity (state variables) that has been propagated over the long-range *Connectivity* pass through these functions before entering the equations of a *Model* describing the local dynamics. A *Coupling* function's primary purpose is to rescale the incoming activity to a level appropriate to the population model. The base *Coupling* class as well as a number of different coupling functions are implemented in the *COUPLING* module, for instance *Linear* and *Sigmoidal*.



#### 2.3.2. Population models

A set of default mesoscopic neural models are defined in TVB's *MODELS*. All these models of local dynamics are classes derived from a base *Model* class.

We briefly discuss the implemented population models in order of increasing complexity. They include a generic two dimensional oscillator, a collection of classical population models and two recently developed multi-modal neural mass models. Below,  $N$  refers to the number of state variables or equations governing the evolution of the model's temporal dynamics;  $M$  is the number of modes and by default  $M = 1$  except for the multi-modal models.

The *Generic2dOscillator* model ( $N = 2$ ) is a generic phase-plane oscillator model capable of generating a wide range of phenomena observed in neuronal population dynamics, such as multistability, the coexistence of oscillatory and non-oscillatory dynamics, as well as displaying dynamics at multiple time scales.

The *WilsonCowan* model (Wilson and Cowan, 1972) ( $N = 2$ ) describes the firing rate of a neural population consisting of two subpopulations (one excitatory and the other inhibitory). It was originally derived using phenomenological arguments. This neural mass model provides an intermediate between a microscopic and macroscopic level of description of neural assemblies and populations of neurons since it can be derived from pulse-coupled neurons (Haken, 2001) and its continuum limit resembles neural field equations (Jirsa and Haken, 1996).

The *WongWang* model (Wong and Wang, 2006) represents a reduced system of  $N = 2$  coupled non-linear equations, originally derived for decision making in two-choice tasks. The *BrunelWang* model (Brunel and Wang, 2001, 2003) is a mean field model derived from integrate-and-fire spiking neurons and makes the approximation of randomly distributed interspike intervals. It is notable that this population model shows only attractor states of firing rates. It has been extensively used to study working memory. Its complexity resides in the number of parameters that it uses to characterize each population ( $N = 2$ ). These parameters correspond to physical quantities that can be measured in neurophysiology experiments. The current implementation of this model is based on the approach used in (Deco and Jirsa, 2012).

The *JansenRit* model (Jansen and Rit, 1995) is a derivative of the Wilson-Cowan model and features three coupled subpopulations of cortical neurons: an excitatory population of pyramidal cells interacting with two populations of interneurons, one inhibitory and the excitatory. This model can produce alpha activity consistent with that measured in EEG, and is capable of simulating evoked potentials (Jansen et al., 1993). It displays a surprisingly rich and complex oscillatory dynamics under periodic stimulation (Spiegler et al., 2010). Each population is described by a second order differential equation. As a consequence the system is described by a set of  $N = 6$  first order differential equations.

The *StefanescuJirsa2D* and *StefanescuJirsa3D* models (Stefanescu and Jirsa, 2008; Jirsa and Stefanescu, 2010; Stefanescu and Jirsa, 2011) are neural mass models derived from a

globally coupled population of neurons of a particular kind. The first one has been derived from coupled FitzHugh-Nagumo neurons (FitzHugh, 1961; Nagumo, 1962), which, with  $N = 2$ , are capable of displaying excitable dynamics, as well as oscillations. The second is derived from coupled Hindmarsh-Rose neurons (Hindmarsh and Rose, 1984), which are also capable of producing excitable and oscillatory dynamics, but with  $N = 3$  have the additional capability of displaying transient oscillations and bursts. The two Stefanescu-Jirsa models show the most complex repertoire of dynamics (including bursting and multi-frequency oscillations). They have been derived using mean field techniques for parameter dispersion (Assisi et al., 2005) and have an additional dimension, the mode  $M$ , which partitions the dynamics into various subtypes of population behavior. These models are therefore composed of 12 ( $N = 4$ ,  $M = 3$ ) and 18 ( $N = 6$ ,  $M = 3$ ) state variables, respectively.

### 2.3.3. Integrators

The base class for integration schemes is called *Integrator*, an INTEGRATORS module contains this base class along with a set of specific integration scheme classes for solving both deterministic and stochastic differential equations. The specific schemes implemented for brain network simulations include the *Euler* and *Heun* methods. The 4th-order *Runge-Kutta* (rk4) method is only available for solving ordinary differential equations (ODEs), i.e., deterministic integration, given that there are various variants for the stochastic version of the method, differing rates of convergence being one of the points that several attempts of creating a stochastic adaptation fail at [see Burrage et al. (2004) for an overview]. Therefore, this method is available for drawing example trajectories in the interactive phase-plane plot tool.

### 2.3.4. Noise

Noise plays a crucial role for brain dynamics, and hence for brain function (McIntosh et al., 2010). The NOISE module consists of two base classes: *RandomStream* that wraps Numpy's *RandomState* class and *Noise*. The former provides the ability to create multiple random streams which can be independently seeded or set to an explicit initial state. The latter is the base class from which specific noises, such as white and colored (Fox et al., 1988), are derived. In TVB's implementation *Noise* enters as an additional term within the stochastic integration schemes, and can be either an *Additive* or *Multiplicative* process (Klöden and Platen, 1995). As well as providing a means to generate reproducible stochastic processes for the integration schemes, the related classes in NOISE are used to set the initial conditions of the system when no explicit initial conditions are specified.

### 2.3.5. Monitors

The data from a simulation is processed and recorded while the simulation is running, that is, while the differential equations governing the system are being integrated. The base class for these processing and recording methods is the *Monitor* class in the MONITORS module. We consider two main types of online-processing: (1) raw or low-level; and (2) biophysical or high-level. The output of a *Monitor* is a 4-dimensional array (which can be wrapped in the corresponding *TimeSeries*

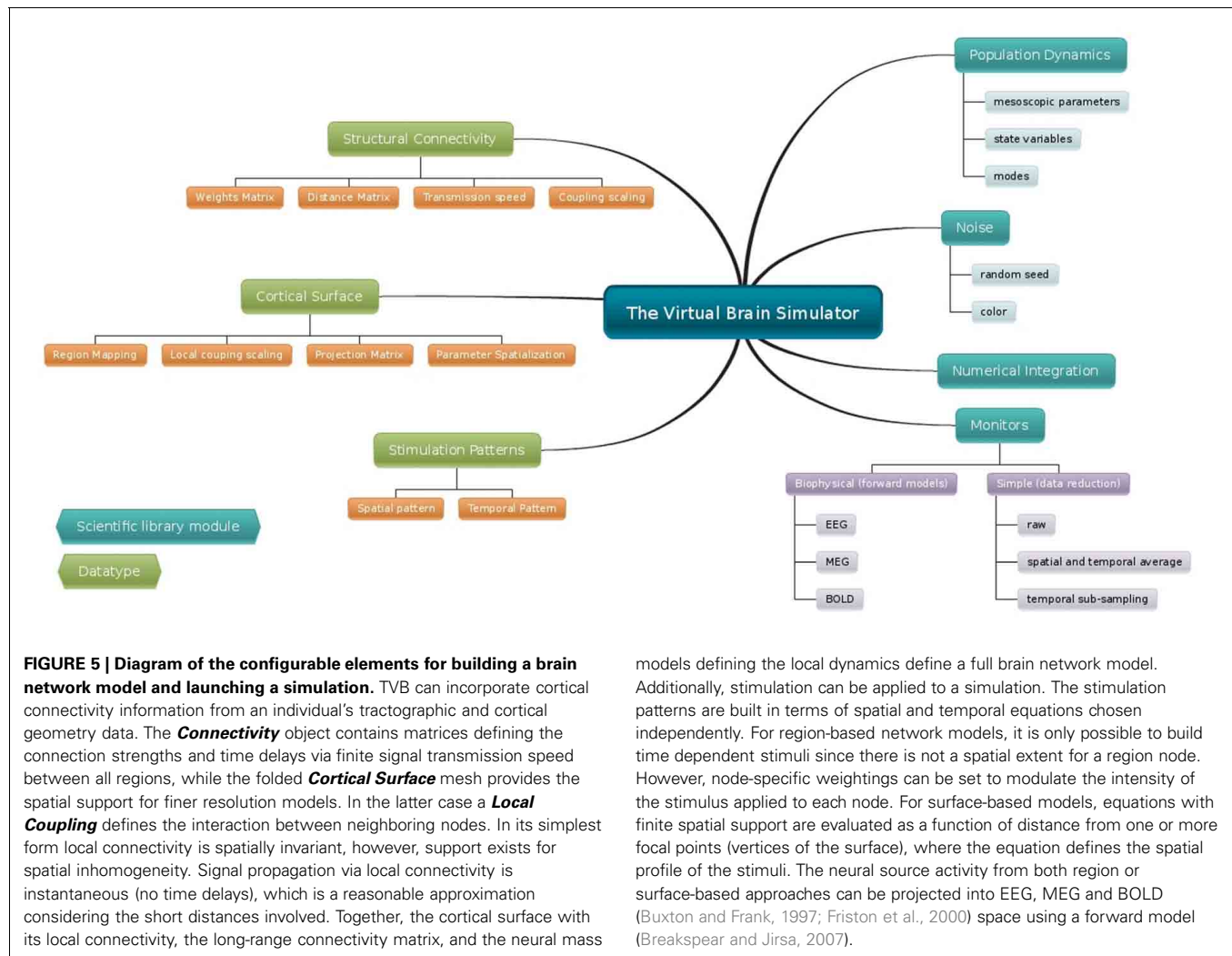
datatype), i.e., a 3D state vector as a function of time. For the first kind of *Monitors* these dimensions correspond to [*time*, *state variables*, *space*, *modes*] where “space” can be either brain regions or vertices of a cortical surface plus non-cortical brain regions. The number of state variables as well as the number of modes strictly depend on the *Model*. For the second kind of *Monitors*, the dimensions are [*time*, 1, *sensors*, 1]. The simplest form of low-level *Monitor* returns all the simulated data, i.e., time points are returned at the sampling rate corresponding to the integration scheme's step size and all state variables are returned for all nodes. All other low-level *Monitors* perform some degree of down-sampling, such as returning only a reduced set state variables (by default the *variables of interest* of a *Model*), or down-sampling in “space” or time. Some variations include temporally sub-sampled, spatially averaged and temporally sub-sampled, or temporally averaged. The biophysical *Monitors* instantiate a physically realistic measurement process on the simulation, such as *EEG*, *MEG*, *SEEG* or *BOLD*. For the first two, a *ProjectionMatrix* is also required. This matrix maps source activity (“space”) to sensor activity (“sensors”). OpenMEEG (Gramfort et al., 2010) was used to generate the demonstration projection matrix, also known as lead-field or gain matrix, that corresponds to the EEG/MEEG forward solution. The forward solution modeling the signals from depth electrodes is based on the point dipole model in homogeneous space (Sarvas, 1987). The *BOLD* monitor is based on Buxton and Frank (1997) and Friston et al. (2000). **Figure 5** summarizes the fundamental blocks required to configure a full model, launch a simulation and retrieve the simulated data.

In most neural mass models there is a state variable representing some type of neural activity (firing rate, average membrane potential, etc.), which serves as a basis for the biophysical monitors. The state variables used as source of neural activity depend both on the *Model* and the biophysical space that it will be projected onto (MEG, EEG, BOLD). Given a neural mass model with a set of state variables, G-Users can choose which subset of state variables will be fed into a *Monitor* (independently for each monitor). However, how a given *Monitor* operates on this subset of state variables is an intrinsic property of the monitor. Users with programming experience can, of course, define new monitors according to their needs. Currently, there is not a mechanism providing automatic support for general operations over state variables before they are passed to a monitor. As such, when the neural activity entering into the monitors is anything other than a summation or average over state variables then it is advised to redefine the *Model* in a way that one of the state variables actually describes the neural activity of interest.

### 2.3.6. Outline of the simulation algorithm

The *Simulator* class has several *methods* to set up the spatiotemporal dimensions of the input and output arrays, based on configurable attributes of the individual components such as integration time step (e.g., *INTEGRATORS.HeunDeterministic.dt*), structural spatial support (e.g., *connectivity.Connectivity* or *surfaces.CorticalSurface*) and transmission speed (e.g., *connectivity.Connectivity.speed*) as well as a cascade of specific





configuration methods to interface them. The *Simulator* class coordinates the collection of objects from all the modules in the scientific library needed to build the network model and yield the simulated data. To perform a simulation a *Simulator* object needs to be: (1) configured, initializing all the individual components and calculating attributes based on the combination of objects passed to the *Simulator* instance; and (2) called in a loop to obtain simulated data, i.e., to run the simulation (see **Code 2**). The next paragraphs list the main operations of the simulation algorithm.

### Initializing a Simulator

1. Check if the transmission speed was provided.
2. Configure the **Connectivity** matrix (connectome). The delays matrix is computed using the distance matrix and the transmission speed. Get the number of regions.
3. Check if a **Surface** is provided.
4. Check if a stimulus **pattern** is provided.
5. Configure individual components: *Model*, *Integrator*, *Monitors*. From here we obtain integration time

step size, number of state variables, number of modes.

6. Set the number of nodes (region-based or surface-based simulation). If a **Surface** was given the number of nodes will correspond to the number of vertices plus the number of non-cortical regions, otherwise it will be equal to the number of regions in the **Connectivity** matrix.
7. Spatialise model parameters if required. Internally, TVB uses arrays for model parameters, if the size of the array for a particular parameter is 1, then the same numerical value is applied to all nodes. If the size of the parameter array is  $N$ , where  $N$  is the number of nodes, the parameter value for each node is taken from the corresponding element of the array of parameter values.
8. If applicable, configure spatial component of stimulation **Patterns** (requires number of nodes).
9. Compute delays matrix in integration time steps.
10. Compute the horizon of the delayed state, that is the maximum delay in integration time steps.
11. Set the history shape. The history state contains the activity that propagates from the delayed state to the next.



12. Determine if the *Integrator* is deterministic or stochastic. If the latter, then configure the *Noise* and the integration method accordingly.
13. Set initial conditions. This is the state from which the simulation will begin. If none is provided, then random initial conditions are set based on the ranges of the model's state variables. Random initial conditions are fed to the initial history array providing the minimal state of the network with time-delays before  $t = 0$ . If initial conditions are user-defined but the length along the time dimension is shorter than the required horizon, then the history array will be padded using the same method of described for random initial conditions.
14. Configure the monitors for the simulation. Get variables of interest.

### Calling a Simulator

1. Get simulation length.
2. Compute estimates of run-time, memory usage and storage.
3. Check if a particular random state was provided (random seed). This feature is useful for reproducibility of results, for instance, getting the same stream of random numbers for the *Noise*.
4. Compute the number of integration steps.
5. If the simulation is surface-based, then get attributes required to compute **Local Connectivity** kernel.
6. Update state loop:
  - a. Get the corresponding coupled delayed activity. That is, compute the dot product between the weights matrix (connectome) and the delayed state of the coupling variables, transformed by a (long-range) *Coupling* function.
  - b. Update the state array. This is the numerical integration, i.e., advancing an integration time step, of the differential equations defining the neuron model. Distal delayed activity, local instantaneous activity and stimulation are fed to the integration scheme.
  - c. Update the history.
  - d. Push state data onto the *Monitors*. Yield any processed time-series data point if available.

As a working example, in **Code 2**, we show a code snippet which uses TVB's scripting interface and some of the classes and modules we have just described to generate one second of brain activity. The for loop in the example code allows scripting users to receive time-series data as available and separately for each of the monitors processing simulated raw data. In this implementation, at each time step or certain number of steps, data can be directly stored to disk, reducing the memory footprint of the simulation. Such a feature is particularly useful when dealing with larger simulations. Likewise, data can be accessed while the simulation is still running, which proves to be advantageous for modeling paradigms where one of the output signals is fed back to the network model as stimulation for instance (see the paragraph about *Dynamic modeling* in section 3).

**Code 2 | Script example to simulate 1 second of brain activity. Output is recorded with two different monitors.**

```
from tvb.simulator.lab import *

#Initialise a Model, Connectivity and Global Coupling
oscillator = models.Generic2dOscillator()
white_matter = connectivity.Connectivity()
white_matter.speed = numpy.array([4.0]) # [mm/ms]
white_matter_coupling = coupling.Linear(a=0.0042)

#Initialise an Integrator
heunint = integrators.HeunDeterministic(dt=2**-4)

#Initialise some Monitors with period in physical time
mon_raw = monitors.Raw()
mon_tav = monitors.TemporalAverage(period=2**-2)
what_to_watch = (mon_raw, mon_tav)

#Initialise a Simulator object
sim = simulator.Simulator(model = oscillator,
                          connectivity = white_matter,
                          coupling = white_matter_coupling,
                          integrator = heunint,
                          monitors = what_to_watch)

# Configure the Simulator object
sim.configure()
LOG.info("Starting_simulation...")

raw_data, raw_time = [], []
tavg_data, tavg_time = [], []

# Call the Simulator object -- Run simulation
for raw, tavg in sim(simulation_length=2**10):
    if not raw is None:
        raw_time.append(raw[0])
        raw_data.append(raw[1])
    if not tavg is None:
        tavg_time.append(tavg[0])
        tavg_data.append(tavg[1])
LOG.info("Finished_simulation.")
```

## 2.4. ANALYZERS AND VISUALIZERS

For the analysis and visualisation of simulated neuronal dynamics as well as imported data, such as anatomical structure and experimentally recorded time-series, several algorithms and techniques are currently available in TVB. Here we list some of the algorithms and methods that are provided to perform analysis and visualization of data through the GUI.

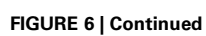
**Analyzers** are mostly standard algorithms for time-series and network analysis. The analyzers comprise techniques wrapping functions from Numpy (Fast Fourier Transform (FFT), auto-correlation, variance metrics), Scipy (cross-correlation), scikit-learn (ICA) (Pedregosa et al., 2011) and matplotlib-mlab (PCA) (Hunter, 2007). In addition, there are specific implementations of the wavelet transform, complex coherence (Nolte et al., 2004; Freyer et al., 2012) and multiscale entropy (MSE) (Costa et al., 2002, 2005; Lake and Moorman, 2011).

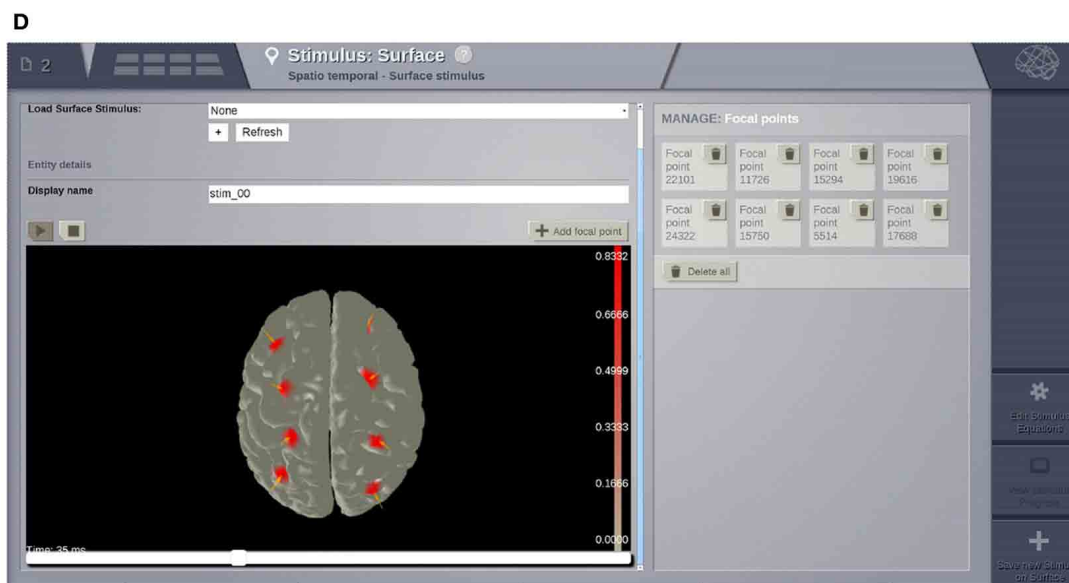
**Visualizers** are tools designed to correctly handle specific **datatypes** and display their content. Representations currently available in the GUI include: histogram plots (**Figure 6A**); interactive time-series plots, EEG (**Figure 6C**); 2D head topographic maps (**Figure 6B**); 3D displays of surfaces and animations (**Figure 6D**) and network plots. Additionally, for shell users there is a collection of plotting tools available based on matplotlib and mayavi (Ramachandran and Varoquaux, 2011).

## 3. PERFORMANCE, REPRODUCIBILITY, AND FLEXIBILITY

### 3.1. TESTING FOR SPEED

In the context of full brain models there is no other platform against which we could compare the performance results for TVB





**FIGURE 6 | Visualizers.** (A) Histogram of a graph metric as a function of nodes in the connectivity matrix. (B) A 2D projection of the head. The color map represents a graph metric computed on the connectivity matrix. (C) EEG visualizer combines a rendered head surface, an overlay

with the sensors positions and an interactive time-series display. (D) An animated display of the spatiotemporal pattern applied to the cortical surface. Red spots represent the focal points of the spatial component of the stimulus.

and define a good ratio run-time/real-time. As a first approximation a simple network of 74 nodes, whose node dynamics were governed by the equations of the *Generic2dOscillator* model (see **Code 3**) was implemented in the Brian spiking neural network simulator. The integration step size was 0.125 ms ( $dt = 2^{-3}$  ms) and the simulation length was 2048 ms. This network was evaluated without time delays and using a random sparse connectivity matrix. Execution times were about 4.5 s in Brian and 15 s in TVB. In contrast, when heterogeneous time delays were included, running times of the simulations implemented in Brian increased considerably (approximately 6.5x) whereas in TVB they hardly changed (approximately 1.2x). Simulations were run on a CPU Intel® Xeon® W3520 @ 2.67 GHz. These results, although informative, expose the fact that the architectures of TVB and the Brian simulator are different and therefore they have been optimized accordingly to serve distinct purposes from a modeling point of view.

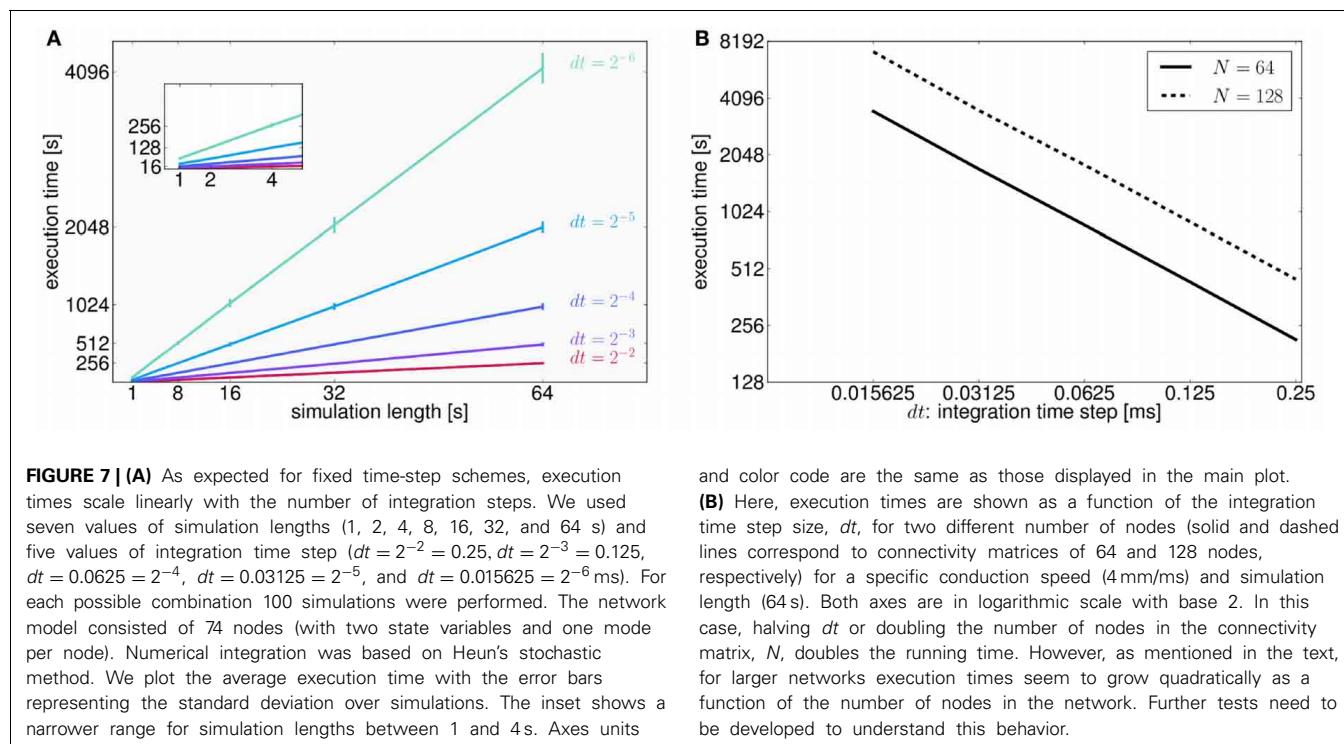
To assess the performance of TVB in terms of simulation timings, we also ran simulations for all possible combinations of two parameters: simulation length and integration time step (**Figure 7A**). We made the following estimates: it takes on average 16 s to compute 1 s of brain network dynamics [at the region level, with an integration time step of 0.0625 ms ( $dt = 2^{-4}$  ms) and including time delays of the order of 20 ms which amounts to store about 320 past states per time step] on CPUs Intel® Xeon® X5672 @ 3.20 GHz, CPU cache of 12 MB and Linux kernel 3.1.0-1-amd64 as operating system. In **Figure 7B** we quantify how running times increase as a function of the integration time step in 64 s long (region-based) simulations for two different sizes of the connectivity matrix.

**Code 3 | State equations of the generic plane oscillator as scripted to run the simulation in the Brian simulator. The description of the parameters are explained in the API documentation and will be discussed in the context of dynamical systems elsewhere.**

```
# model equations
eqs = '''
dV/dt = d * tau * (alpha * W - f * V**3 + e * V**2 + I)
dW/dt = d * (a + b * V + c * V**2 - beta * W) / tau
'''
```

In general, human cortical connectomes are derived from anatomical parcellations with a variable number of nodes, from less than 100 to over a few thousands nodes (Zalesky et al., 2010). Preliminary results of simulations (data not shown) using connectivity matrices of different sizes (16, 32, 64, 128, 256, 512, 1024, 2048, and 4096 nodes) and a supplementary parameter (transmission speed that has an effect on the size of the history array keeping the delayed states of the network) indicate that there is a quadratic growth of the running times for networks with more than 512 nodes. Since performance depends on a large number of parameters which have an effect on both memory (CPU cache and RAM) and CPU usage, and therefore resulting running times arise from the interaction between them, we see the need to develop more tests to stress in particular memory capacity and bandwidth in order to fully understand the aforementioned behavior.

In Future Work we talk about the approaches to benchmark and improve the execution times of simulations. For the present work we have restricted ourselves to present performance results looking at the parameters that have the strongest effect on simulations timings.



### 3.2. REPRODUCIBILITY OF RESULTS FROM THE LITERATURE

Ghosh et al. (2008) and Deco et al. (2009) demonstrated the important role of three large-scale parameters in the emergence of different cluster synchronization regimes: the global coupling strength factor, time-delays (introduced via the long-range connectivity fiber tract lengths and a unique transmission speed) and noise variance. They built parameter space maps using the Kuramoto synchronization index. Here, using TVB's scripting interface, we show it is easily possible to build a similar scheme and perform a parameter space exploration in the coupling strength ( $g_{cs}$ ) and transmission speed ( $s$ ) space. The *Connectivity* upon which the large-scale network is built was the demonstration dataset. It is bi-hemispheric and consists of 74 nodes, i.e., 37 regions per hemisphere. It includes all the cortical regions but without any sub-cortical structure such as the thalamic nuclei. Its weights are quantified by integer values in the range 0–3. The evolution of the local dynamics were represented by the model *Generic2dOscillator*, configured in such a way that a single isolated node exhibited 40 Hz oscillations (Figure 8). The variance of the output time-series was chosen as a simple, yet informative measure to represent the collective dynamics (Figure 9A) as a function of the parameters under study. Results are shown in Figure 9B. Parameter sweeps can also be launched from TVB web-interface (see Figure 10 for an illustration).

Currently TVB provides two scalar metrics based on the variance of the output time-series to perform data reduction when exploring a certain parameter space. These are *Variance of the nodes Variances* and *Global Variance*. The former zero-centers the output time-series and computes the variance over time of the concatenated time-series of each state variable and mode for each node and subsequently the variance of the nodes variances

is computed. This metric describes the variability of the temporal variance of each node. In the latter all the time-series are zero-centered and the variance is computed over all data points contained in the output array.

With this example we intended to expose the possibility to reproduce workflows, i.e., modeling schemes, found in the literature. TVB is a modeling platform providing a means of cross-validating scientific work by encouraging reproducibility of the results.

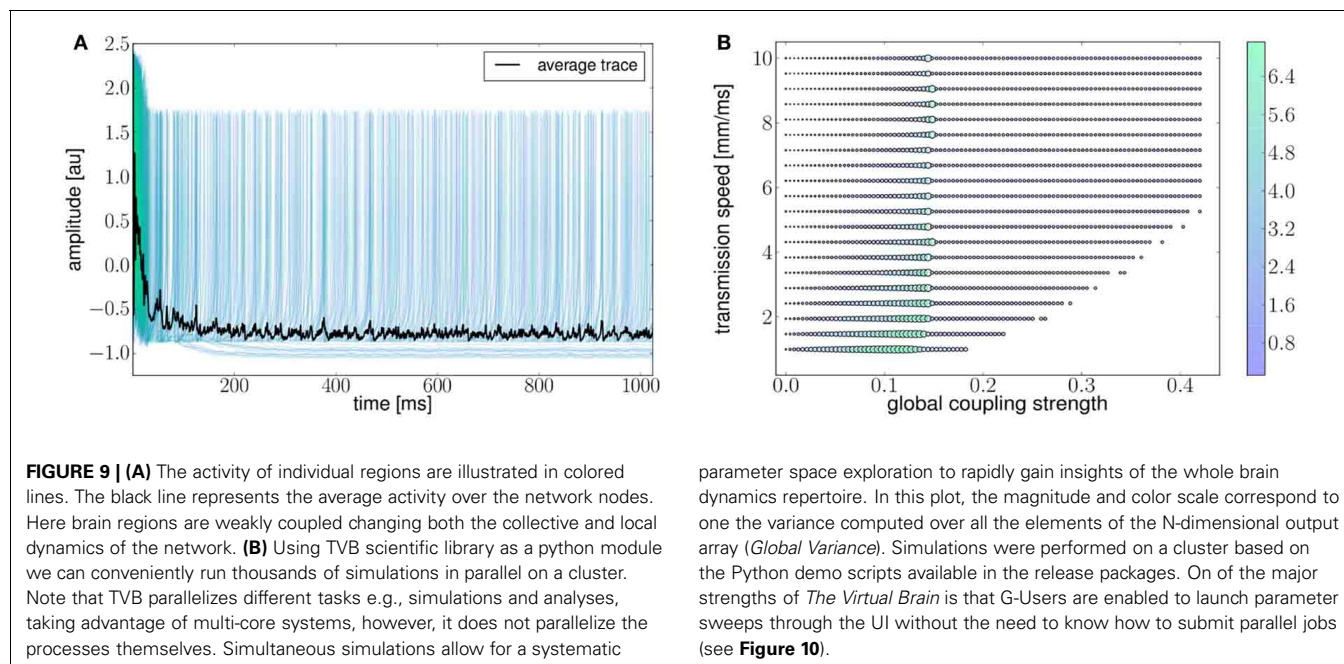
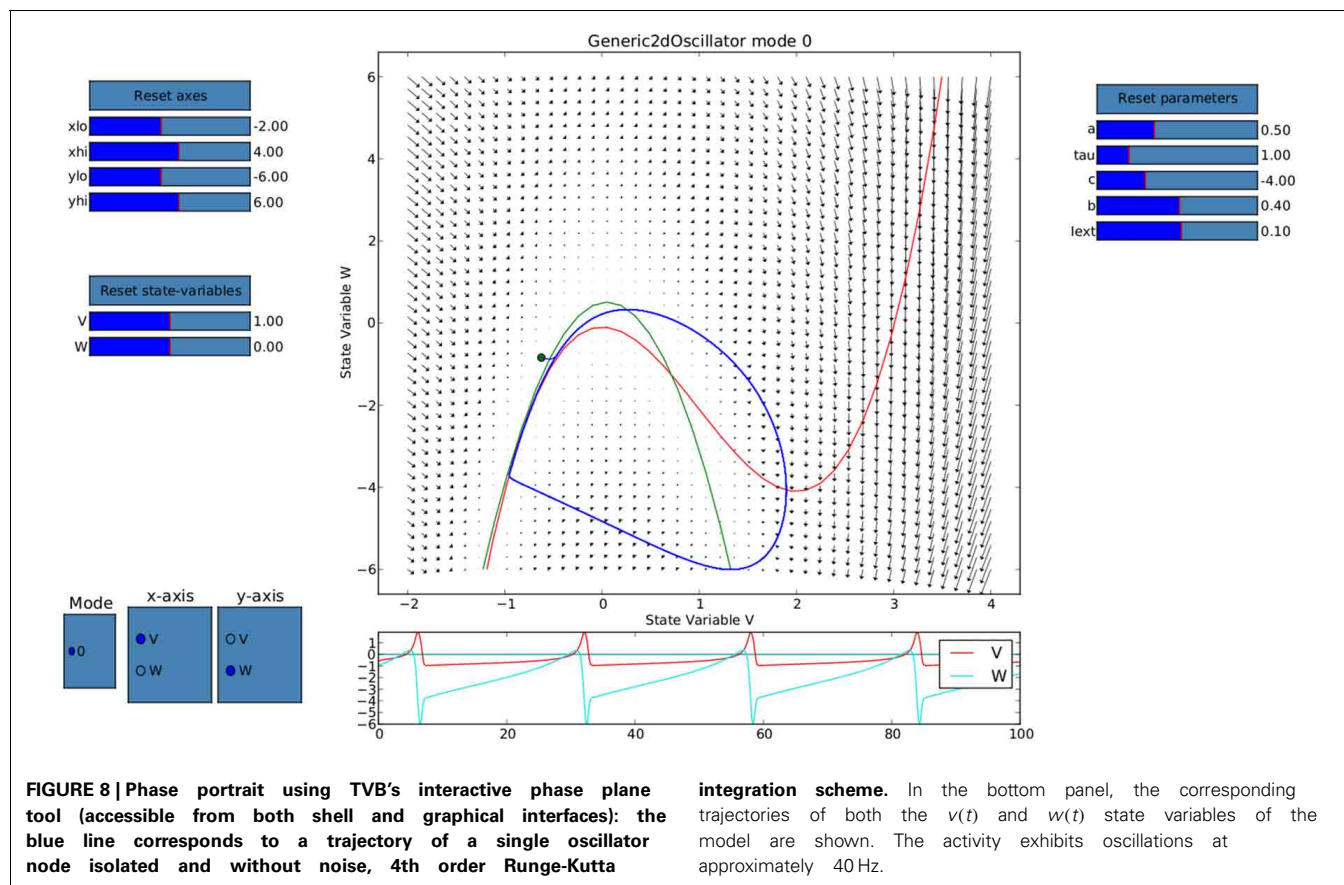
### 3.3. HIGHER-LEVEL SIMULATION SCENARIOS USING STIMULATION PROTOCOLS

As one possible use case, we have set up an example based on the scheme used in McIntosh et al. (2010). The goal is to demonstrate how to build stimulation patterns in TVB, use them in a simulation, obtain EEG recordings of both the activity similar to the resting state (RS) and to evoked responses (ER), and finally make a differential analysis of the complexity of the resulting time-series by computing MSE.

In vision neuroscience, the two-stream hypothesis (Schneider, 1969) suggests the existence of two streams of information processing, the ventral and the dorsal stream. In one of these pathways, the ventral stream, the activity from subcortical regions project to V1 and the activity propagates to the temporal cortices through V2 and V4 (Goodale and Milner, 1992). We systematically stimulated the area corresponding to the primary visual cortex (V1) to demonstrate the functioning of TVB stimulation *Patterns* and observed how the activity elicited by a periodic rectangular pulse propagates to neighboring regions, especially V2.

Benefiting from TVB's flexibility we show in Figure 11 that it is possible to systematically stimulate a specific brain region



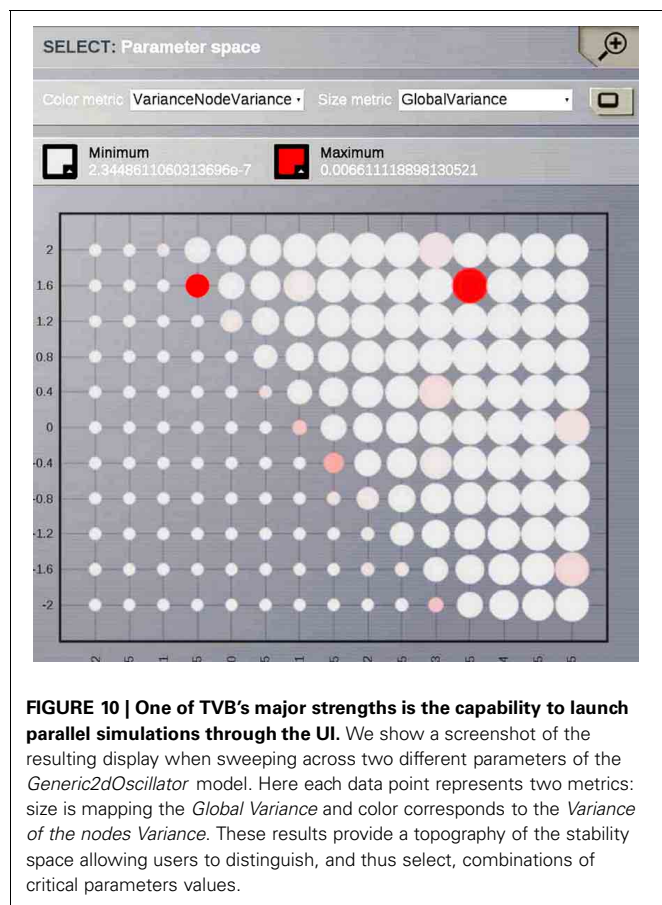


(e.g., V1) and to highlight the anatomical connection to its target region (e.g., V2) by observing the arrival of the delayed activity; analyze the responses of the model; handle multi-modal simulated data; and extract metrics from computationally expensive

algorithms to characterize both the “resting” and “evoked” states.

Currently, TVB permits the stimulation and read-out of activity from any brain area defined in the anatomical parcellation





used to derive the connectome. This modeling example was built imposing a strong restriction on the number of regions to stimulate, since global dynamics can quickly become complex. Additionally, to demonstrate the many scenarios that can be set up in TVB, we simulated the same brain network model under the influence of a stimulus, first without noise (**Figure 11A**: using Heun deterministic method) and then with white noise (**Figure 11B**: using Heun stochastic method). The first approach makes it easier to see the perturbations induced by the stimulus and the propagation of activity from one region to the other. The second approach is a more realistic representation of the neural activity.

Results of the proposed modeling protocol are presented in **Figure 12** where the EEG traces from channel Oz for the resting and evoked states are shown together with the MSE estimates.

Scripts to reproduce results from **Figures 11, 12** are available in the distribution packages of TVB.

With the availability of surface-based simulations the challenge of replicating topographic maps of different sensory systems, such as those found in the primary visual cortex (Hinds et al., 2009), could be addressed.

### 3.4. DYNAMIC MODELING

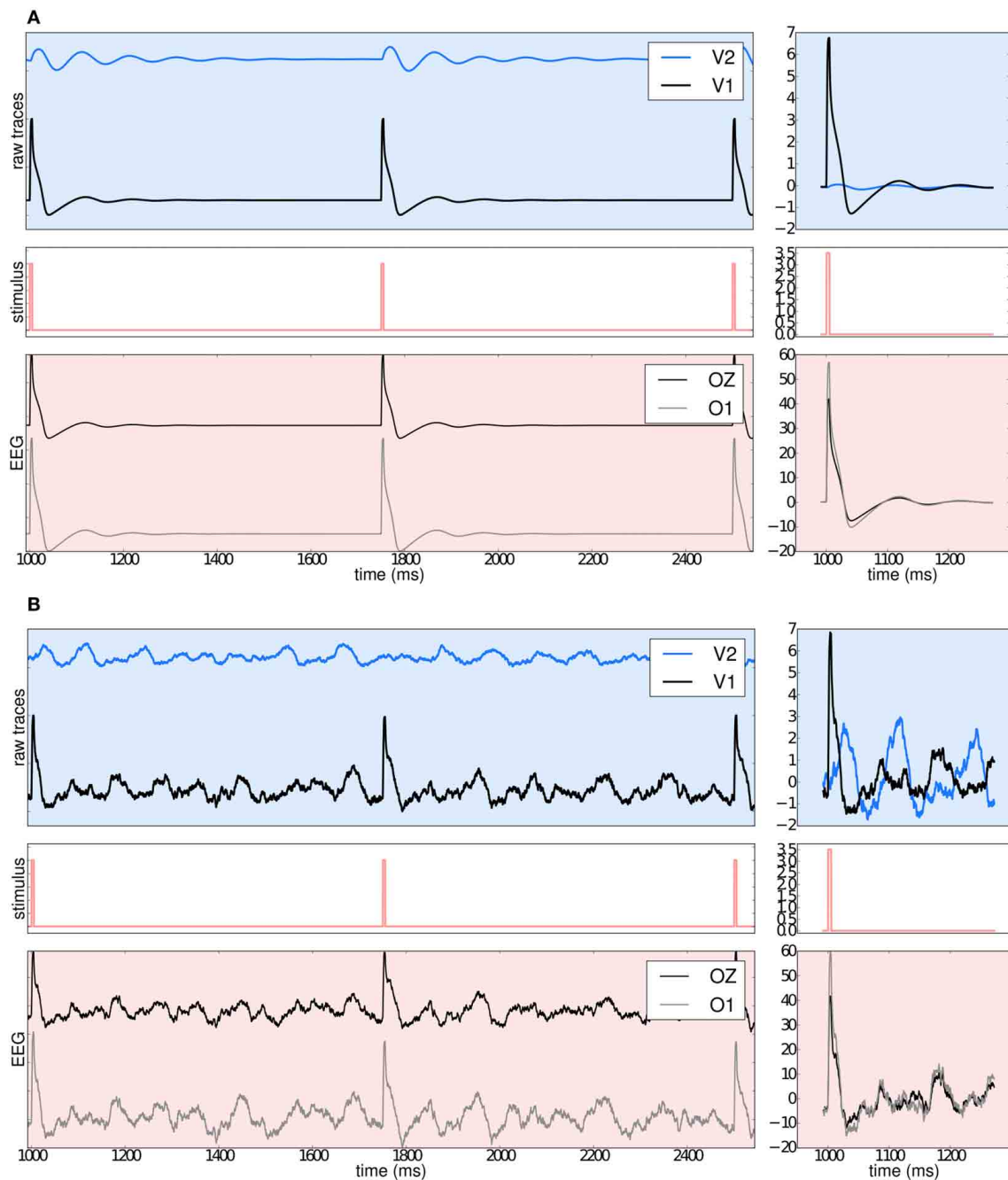
From both the shell and web interface it is possible to exploit another feature of TVB: namely, simulation continuation, i.e., a simulation can be stopped allowing users to modify model

parameters, scaling factors, apply or remove stimulation or spatial constraints (e.g., local connectivity), or make any other change that does not alter the spatiotemporal domain of the system or its output (integration step, transmission speed and spatial support) and then resumed without the need of creating a new *Simulator* instance. Furthermore, this capability opens the possibility to dynamically update the simulation at runtime. Such a dynamic approach leads toward an adaptive modeling scheme where stimuli and other factors may be regulated by the ongoing activity (this last feature can be handled only from the scripting interface for the moment).

## 4. DISCUSSION

We have presented the architecture and usage of TVB, a neuroinformatics platform developed for simulations of network models of the full brain. Its scientific core has been developed by integrating concepts from theoretical, computational, cognitive and clinical neuroscience, with the aim to integrate neuroimage modalities along with the interacting mesoscopic and macroscopic scales of a biophysical model of the brain. From a computational modeling perspective TVB constitutes an alternative to approaches such as the work of Riera et al. (2005) and more recently that of Valdes-Sosa et al. (2009), as well as other relevant studies mentioned in the main text of this article. From a neuroinformatics perspective, TVB lays the groundwork for the integration of existing paradigms in the theory of large-scale models of the brain, by providing a general and flexible framework where the advantages and limitations of each approach may be determined. It also provides the community with a technology, that until now had not been publicly available, accessible by researchers with different levels and backgrounds, enabling systematic implementation and comparison of neural mass and neural field models, incorporating biologically realistic connectivity and cortical geometry and with the potential to become a novel tool for clinical interventions. While many other environments simulate neural activity at the level of neurons (Brian simulator, MOOSE, PCSIM, NEURON, NEST, GENESIS) (Hines and Carnevale, 2001; Gewaltig and Diesmann, 2007; Goodman and Brette, 2008; Ray and Bhalla, 2008; Pecevski et al., 2009; Brette and Goodman, 2011), even mimicking a number of specific brain functions (Eliasmith et al., 2012), they, most importantly, do not consider the space-time structure of full brain connectivity constraining whole brain neurodynamics, as a crucial component in their modeling paradigm. Other approaches to multi-modal integration such as Statistical Parametric Mapping (SPM) perform statistical fitting to experimental data at the level of a small set of nodes (Friston et al., 1995, 2003; David et al., 2006; Pinotsis and Friston, 2011) [i.e., they are data-driven as in Freestone et al. (2011)], thus diverging from our approach that could be categorized as a purely “computational neural modeling” paradigm as described in Bojak et al. (2011). From this perspective, the goal is to capture and reproduce whole brain dynamics by building a network constrained by its structural large-scale connectivity and mesoscopic models governing the nodes intrinsic dynamics.

Also, the extension of neuronal level modeling to large brain structures requires vast supercomputers to emulate the large



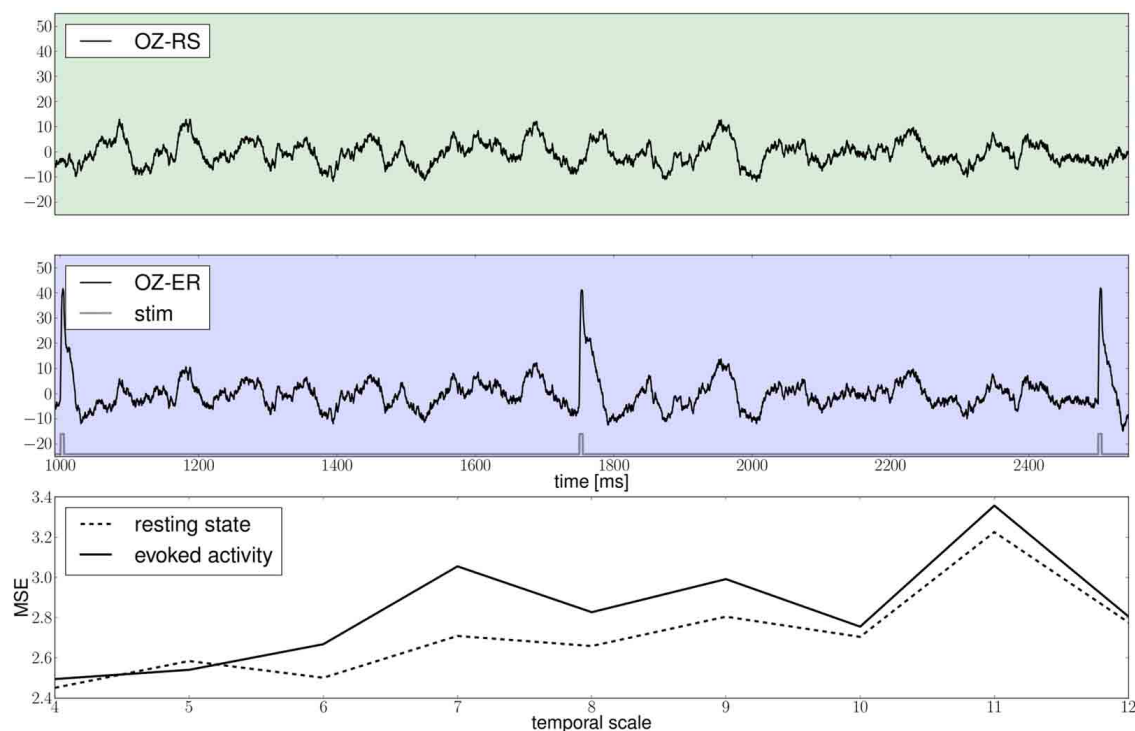
**FIGURE 11 | (A)** The upper left blue panel shows the raw traces of nodes V2 and V1; the latter stimulated with a rectangular pulse of width equal to 5 ms and repetition frequency of 1 Hz. Signals are normalized by their corresponding maximum value. The right blue panel show the signals for a shorter period of time. Amplitudes are not normalized to emphasize the relative difference between the two regions. Middle panels illustrate the stimulus pattern. Lower red panels display the activity as projected onto EEG

space and recorded from channels Oz and O1. The default EEG cap in TVB consists of 62 scalp electrodes distributed according to the 10–20 international system (Klem et al., 1999). In this simulation a deterministic integration scheme was employed to obtain the time-series of neural activity, since noise was not applied to the model's equations. **(B)** The same description as in **(A)** applies. The main difference with the previous simulation is that here white noise was added to the system.

number of complex functional units. Focusing on the brain's large-scale architecture, in addition to the dimension reduction accomplished through the mean field methods applied on the mesoscopic scale, TVB allows for computer simulations on the

full brain scale on workstations and small computing clusters, with no need to use supercomputing resources.

The simulator component of TVB has the goal of simulating mesoscopic neural dynamics on large-scale brain networks.



**FIGURE 12 |** The green and blue panels show EEG recordings from electrode Oz during the resting state, i.e., in the absence of stimulation and in the stimulated condition, respectively, notice the slow damped oscillations after stimulus onset at a approximately 10 Hz; the light gray

trace depicts the stimulation pattern. The bottom panel displays multiscale entropy estimates computed on the Oz time-series at different temporal scales using the dataset obtained by means of a stochastic integration scheme.

It does not intend to build brain models at the level of neurons (Goodman and Brette, 2009; Cornelis et al., 2012), however, it does leverage information from microscopic models to add detail and enhance the performance of the neural population models, which act as building blocks and functional units of the network. TVB thus represents a unique tool to systematically investigate the dynamics of the brain, emphasizing its large-scale network nature and moving away from the study of isolated regional responses, thereby considering the function of each region in terms of the interplay among brain regions. The primary spatial support (neuroanatomical data) on top of which the large-scale network model is built has a number of implications:

1. It constraints the type of network dynamics; dynamics that could be further related to physiology and behavior (Senden et al., 2012).
2. It permits a systematic investigation of the consequences of the particular restrictions imposed by that large-scale structure and the effect of changes to it.
3. It provides a reliable and geometrically accurate model of sources of neural activity, enabling realistic forward solutions to EEG/MEG based on implementations of boundary element methods (BEM) or other approaches such as finite difference time domain methods (FDTD).

On the basis of the literature, theoretical and clinical studies seeking to better understand and describe certain brain functions and structure use stimulation as an essential part of their protocols. Stimulation is a way to probe how the system respond under external perturbations adapting itself to the new environmental conditions or to categorize responses when stimulation represents real-life (visual, auditory, motor) sensory inputs. Among the current features of TVB, the easy generation of a variety of stimulation patterns is to be recognized as one of its great advantages and contributions to experimental protocol design. TVB permits the development of simple stimulation routines, allowing evaluation of the viability and usefulness of certain stimulation procedures.

TVB represents a powerful research platform, combining experimental design and numerical simulations into a collaborative framework that allows sharing of results and the integration of data from other applications. Naturally, this leads to the potential for an increased level of interaction among researchers of the broad neuroscience community. In the same direction, TVB is also an extensible validation platform since it supports the creation of basic modeling refinement loops, making model exploration and validation a relatively automated procedure. For instance, after generating a brain network model, exploring the system's parameter space by adjusting parameters of both the local dynamics and the large scale structure can be achieved with ease. Further, effects of local dynamics and network structure

can be disentangled by evaluating distinct local dynamic models on the same structure or the same local dynamic model coupled through distinct structures. This constrained flexibility makes it easy for modelers to test new approaches, directly compare them with existing approaches and reproduce their own and other researchers' results. Reproducibility is indeed a required feature to validate and consequently increase the reliability of scientific work (Donoho, 2010) and the extensibility of TVB's scientific components, granted by its modular design, provides a mechanism to help researchers achieve this.

The brain network models of TVB, being built on explicit anatomical structure, enable modeling investigations of practical clinical interest. Specifically, whenever a dysfunction or disease expresses itself as a change to the large scale network structure, for instance, in the case of lesions in white-matter pathways, the direct replication of this structural change in TVB's brain network models is straight forward.

## FUTURE WORK

Regarding performance, of special importance will be to evaluate all the parameters that have an effect on both memory usage and execution time for surface-based simulations. The reason is that realistic brain network models are built on top of surface meshes constructed by thousands of vertices per hemisphere ( $2^{13}$  for the TVB demonstration cortical surface) but can easily have more than 40,000.

Equally important is to develop more tests to generally evaluate the simulation engine, paying close attention to keep the consistency and stability of the algorithms currently implemented.

Another aspect that deserves careful attention is the description of our modeling approach that was largely beyond the scope of this text. Therefore, the theory underlying the different methods involved in the development of a generalized framework

for brain network models is to be presented in future scientific publications.

To allow a most optimal dissemination of knowledge in TVB we are currently developing a web-based educational platform that will allow training on the usage of TVB, as well as serve as a key reference.

As simulations in TVB are built on the large-scale anatomical structure of the human brain, continued work to integrate new, reliable, sources of structural data is essential to the progress of the platform. An obvious future resource in this regard will be the newly developed database of the Human Connectome Project (Essen and Ugurbil, 2012; Essen et al., 2012).

## INFORMATION SHARING STATEMENT (LICENSE)

The data and software in this study belong to an ongoing project; it is free software and licensed under the GNU General Public License version 2 as published by the Free Software Foundation. The latest releases of *The Virtual Brain* including the source code and demo data are free to download from <http://www.thevirtualbrain.org>. The source code available in the public repository includes the latest experimental features regarding GPU implementation.

## ACKNOWLEDGMENTS

Funding: The research reported herein was supported by the Brain Network Recovery Group through the James S. McDonnell Foundation and the FP7-ICT BrainScales. Paula Sanz Leon is supported by a doctoral fellowship from Ministere de la Recherche.

## SUPPLEMENTARY MATERIAL

The Supplementary Material for this article can be found online at: <http://www.frontiersin.org/Neuroinformatics/10.3389/fninf.2013.00010/abstract>

## REFERENCES

- Amari, S. (1975). Homogeneous nets of neuron-like elements. *Biol. Cybern.* 17, 211–220. doi: 10.1007/bf00339367
- Amari, S. (1977). Dynamics of pattern formation in lateral-inhibition type neural fields. *Biol. Cybern.* 22, 77–87. doi: 10.1007/BF00337259
- Assisi, C., Jirsa, V., and Kelso, J. (2005). Synchrony and clustering in heterogeneous networks with global coupling and parameter dispersion. *Phys. Rev. Lett.* 94:018106. doi: 10.1103/PhysRevLett.94.018106
- Atay, F., and Hutt, A. (2006). Neural fields with distributed transmission speeds and long range feedback delays. *SIAD* 5, 670–698. doi: 10.1137/050629367
- Babajani-Feremi, A., and Soltanian-Zadeh, H. (2010). Multi-area neural mass modeling of eeg and meg signals. *Neuroimage* 52, 793–811. doi: 10.1016/j.neuroimage.2010.01.034
- Bakker, R., Wachtler, T., and Diesmann, M. (2012). Cocomac 2.0 and the future of tract-tracing databases. *Front. Neuroinform.* 6:30. doi: 10.3389/fninf.2012.00030
- Bastiani, M., Shah, N. J., Goebel, R., and Roebroeck, A. (2012). Human cortical connectome reconstruction from diffusion weighted mri: the effect of tractography algorithm. *Neuroimage* 62, 1732–1749. doi: 10.1016/j.neuroimage.2012.06.002
- Beurle, R. L. (1956). Properties of a mass of cells capable of regenerating pulses. *Philos. Trans. R. Soc. Lond. B Biol. Sci.* 240, 55–94. doi: 10.1098/rstb.1956.0012
- Bojak, I., and Liley, D. T. J. (2010). Axonal velocity distributions in neural field equations. *PLoS Comput. Biol.* 6:e1000653. doi: 10.1371/journal.pcbi.1000653
- Bojak, I., Oostendorp, T., Reid, A., and Kötter, R. (2011). Towards a model-based integration of co-registered electroencephalography/functional magnetic resonance imaging data with realistic neural population meshes. *Philos. Trans. R. Soc. Lond. A* 369, 3785–3801. doi: 10.1098/rsta.2011.0080
- Bojak, I., Oostendorp, T., Reid, A., and R. K. (2010). Connecting mean field models of neural activity to eeg and fmri data. *Brain Topogr.* 23, 139–149. doi: 10.1007/s10548-010-0140-3
- Bostock, M., Ogievetsky, V., and Heer, J. (2011). D3 data-driven documents. *IEEE Trans. Visual. Comput. Graphics* 17, 2301–2309. doi: 10.1109/TVCG.2011.185
- Breakspear, M., and Jirsa, V. (2007). *Handbook of Brain Connectivity (Understanding Complex Systems) – Neuronal Dynamics and Brain Connectivity*. Berlin; Heidelberg: Springer.
- Breakspear, M., Roberts, J. A., Terry, J. R., Rodrigues, S., Mahant, N., and Robinson, P. A. (2006). A unifying explanation of primary generalized seizures through nonlinear brain modeling and bifurcation analysis. *Cereb. Cortex* 16, 1296–1313. doi: 10.1093/cercor/bhj072
- Breakspear, M., Terry, J. R., and Friston, K. J. (2003). Modulation of excitatory synaptic coupling facilitates synchronization and complex dynamics in a biophysical model of neuronal dynamics. *Network* 14, 703–732. doi: 10.1088/0954-898X/14/4/305
- Bressloff, P. C. (2012). From invasion to extinction in heterogeneous neural fields. *JMN* 2:6. doi: 10.1186/2190-8567-2-6
- Brette, R., and Goodman, D. F. M. (2011). Vectorized algorithms for spiking neural network simulation. *Neural Comput.* 23, 1503–1535. doi: 10.1162/NECO\_a\_00123
- Brunel, N., and Wang, X.-J. (2001). Effects of neuromodulation in a cortical network model of object working memory dominated by



- recurrent inhibition. *J. Comput. Neurosci.* 11, 63–85.
- Brunel, N., and Wang, X.-J. (2003). What determines the frequency of fast network oscillations with irregular neural discharges? i. synaptic dynamics and excitation-inhibition balance. *J. Neurophysiol.* 90, 415–430. doi: 10.1152/jn.01095.2002
- Bullmore, E., and Sporns, O. (2009). Complex brain networks: graph theoretical analysis of structural and functional systems. *Nat. Rev. Neurosci.* 10, 186–198. doi: 10.1038/nrn2575
- Burrage, K., Burrage, P. M., and Tian, T. (2004). Numerical methods for strong solutions of stochastic differential equations: an overview. *Proc. R. Soc. Lond. A* 460, 373–402. doi: 10.1098/rspa.2003.1247
- Buxton, R., and Frank, L. (1997). A model for the coupling between cerebral blood flow and oxygen metabolism during neural stimulation. *J. Cereb. Blood Flow Metab.* 17, 64–72. doi: 10.1097/00004647-199701000-00009
- Buzsaki, G. (2006). *Rhythms of the Brain*. Oxford: Oxford University Press.
- Chacon, S. (2009). *Pro Git*. Berkeley, CA: Apress.
- Coombes, S. (2010). Large-scale neural dynamics: simple and complex. *Neuroimage* 52, 731–739. doi: 10.1016/j.neuroimage.2010.01.045
- Cornelis, H., Rodriguez, A. L., Coop, A. D., and Bower, J. M. (2012). Python as a federation tool for genesis 3.0. *PLoS ONE* 7:e29018. doi: 10.1371/journal.pone.0029018
- Costa, M., Goldberger, A. L., and Peng, C.-K. (2002). Multiscale entropy analysis of complex physiologic time series. *Phys. Rev. Lett.* 89:068102. doi: 10.1103/PhysRevLett.89.068102
- Costa, M., Goldberger, A. L., and Peng, C.-K. (2005). Multiscale entropy analysis of biological signals. *Phys. Rev. E Stat. Nonlin. Soft Matter Phys.* 71(2 Pt 1):021906. doi: 10.1103/PhysRevE.71.021906
- David, O., Kilner, J. M., and Friston, K. J. (2006). Mechanisms of evoked and induced responses in MEG/EEG. *Neuroimage* 31, 1580–1591. doi: 10.1016/j.neuroimage.2006.02.034
- Deco, G., and Jirsa, V. (2012). Ongoing cortical activity at rest: criticality, multistability, and ghost attractors. *J. Neurosci.* 32, 3366–3375. doi: 10.1523/JNEUROSCI.2523-11.2012
- Deco, G., Jirsa, V., and McIntosh, A. (2011). Emerging concepts for the dynamical organization of resting-state activity in the brain. *Nat. Rev. Neurosci.* 12, 43–56. doi: 10.1038/nrn2961
- Deco, G., Jirsa, V., McIntosh, A., Sporns, O., and Kötter, R. (2009). Key role of coupling, delay, and noise in resting brain fluctuations. *Proc. Natl. Acad. Sci. U.S.A.* 106, 10302–10307. doi: 10.1073/pnas.0901831106
- Deco, G., Jirsa, V., Robinson, P. A., Breakspear, M., and Friston, K. (2008). The dynamic brain: from spiking neurons to neural masses and cortical fields. *PLoS Comput. Biol.* 4:e1000092. doi: 10.1371/journal.pcbi.1000092
- Donoho, D. L. (2010). An invitation to reproducible computational research. *Biostatistics* 11, 385–388. doi: 10.1093/biostatistics/kxq028
- Eliasmith, C., Stewart, T. C., Choo, X., Bekolay, T., DeWolf, T., Tang, Y., et al. (2012). A large-scale model of the functioning brain. *Science* 338, 1202–1205. doi: 10.1126/science.1225266
- Enthougt, I. (2001). The traits framework for validation and event-driven programming in python. Available online at: <http://code.enthougt.com/projects/traits/>
- Essen, D. C. V., and Ugurbil, K. (2012). The future of the human connectome. *Neuroimage* 62, 1299–1310. doi: 10.1016/j.neuroimage.2012.01.032
- Essen, D. C. V., Ugurbil, K., Auerbach, E., Barch, D., Behrens, T. E. J., Bucholz, R., et al. (2012). The human connectome project: a data acquisition perspective. *Neuroimage* 62, 2222–2231. doi: 10.1016/j.neuroimage.2012.02.018
- FitzHugh, R. (1961). Impulses and physiological states in theoretical models of nerve membrane. *Biophys. J.* 1, 445–466. doi: 10.1016/S0006-3495(61)86902-6
- Fox, R., Gatland, I., Rot, R., and Vemuri, G. (1988). Fast, accurate algorithm for numerical simulation of exponentially correlated colored noise. *Phys. Rev. A* 38, 5938–5940. doi: 10.1103/PhysRevA.38.5938
- Freeman, W. J. (1975). *Mass Action in the Nervous System*. New York; San Francisco; London: Academic press.
- Freeman, W. J. (1992). Tutorial on neurobiology: from single neurons to brain chaos. *Int. J. Bif. Chaos* 2, 451–482. doi: 10.1142/S0218127492000653
- Freestone, D. R., Aram, P., Dewar, M., Scerri, K., Grayden, D. B., and Kadirkamanathan, V. (2011). A data-driven framework for neural field modeling. *Neuroimage* 56, 1043–1058. doi: 10.1016/j.neuroimage.2011.02.027
- Freyer, F., Reinacher, M., Nolte, G., Dinse, H. R., and Ritter, P. (2012). Repetitive tactile stimulation changes resting-state functional connectivity-implications for treatment of sensorimotor decline. *Front. Hum. Neurosci.* 6:144. doi: 10.3389/fnhum.2012.00144
- Freyer, F., Roberts, J. A., Becker, R., Robinson, P. A., Ritter, P., and Breakspear, M. (2011). Biophysical mechanisms of multistability in resting-state cortical rhythms. *J. Neurosci.* 31, 6353–6361. doi: 10.1523/JNEUROSCI.6693-10.2011
- Friston, K., Harrison, L., and Penny, W. (2003). Dynamic causal modelling. *Neuroimage* 19, 1273–1302. doi: 10.1016/S1053-8119(03)00202-7
- Friston, K., Holmes, A., Worsley, K., Poline, J., Frith, C., and Frackowiak, R. (1995). Statistical parametric maps in functional imaging: a general linear approach. *Hum. Brain Mapp.* 2, 189–210. doi: 10.1002/hbm.460020402
- Friston, K. J., Mechelli, A., Turner, R., and Price, C. J. (2000). Nonlinear responses in fMRI: the balloon model, volterra kernels, and other hemodynamics. *Neuroimage* 12, 466–477. doi: 10.1006/nimg.2000.0630
- Gerhard, S., Daducci, A., Lemkaddem, A., Meuli, R., Thiran, J.-P., and Hagmann, P. (2011). The connectome viewer toolkit: an open source framework to manage, analyze, and visualize connectomes. *Front. Neuroinform.* 5:3. doi: 10.3389/fninf.2011.00003
- Gewaltig, M., and Diesmann, M. (2007). NEST (neural simulation tool). *Scholarpedia* 2:1430. doi: 10.4249/scholarpedia.1430
- Ghosh, A., Rho, Y., McIntosh, A., Kötter, R., and Jirsa, V. (2008). Noise during rest enables the exploration of the brain's dynamic repertoire. *PLoS Comput. Biol.* 4:e1000196. doi: 10.1371/journal.pcbi.1000196
- Goodale, M. A., and Milner, A. D. (1992). Separate visual pathways for perception and action. *Trends Neurosci.* 15, 20–25. doi: 10.1016/0166-2236(92)90344-8
- Goodman, D. F. M., and Brette, R. (2008). The Brian: a simulator for spiking neural networks in python. *Front. Neuroinform.* 2:5. doi: 10.3389/neuro.11.005.2008
- Goodman, D. F. M., and Brette, R. (2009). The brian simulator. *Front. Neurosci.* 3, 192–197. doi: 10.3389/neuro.01.026.2009
- Gramfort, A., Papadopoulos, T., Olivi, E., and Clerc, M. (2010). Openmeeg: opensource software for quasistatic bioelectromagnetics. *Biomed. Eng. Online* 9:45. doi: 10.1186/1475-925X-9-45
- Hagmann, P., Cammoun, L., Gigandet, X., Meuli, R., Honey, C. J., Wedeen, V. J., et al. (2008). Mapping the structural core of human cerebral cortex. *PLoS Biol.* 6:e159. doi: 10.1371/journal.pbio.0060159
- Haken, H. (1983). *Synergetics, an Introduction: Nonequilibrium Phase Transitions and Self-Organization in Physics, Chemistry, and Biology*. 3rd Edn. New York, NY: Springer Verlag.
- Haken, H. (2001). Delay, noise and phase locking in pulse coupled neural networks. *Biosystems* 63, 15–20. doi: 10.1016/S0303-2647(01)00143-5
- Hindmarsh, J., and Rose, R. (1984). A model of neuronal bursting using three coupled first order differential equations. *Proc. R. Soc. Lond. Ser. B* 221, 87–122. doi: 10.1098/rspb.1984.0024
- Hinds, O., Polimeni, J. R., Rajendran, N., Balasubramanian, M., Amunts, K., Zilles, K., et al. (2009). Locating the functional and anatomical boundaries of human primary visual cortex. *Neuroimage* 46, 915–922. doi: 10.1016/j.neuroimage.2009.03.036
- Hines, M. L., and Carnevale, N. T. (2001). Neuron: a tool for neuroscientists. *Neuroscientist* 7, 123–135.
- Hämäläinen, M. S. (1992). Magnetoencephalography: a tool for functional brain imaging. *Brain Topogr.* 5, 95–102.
- Hämäläinen, M. S., Hari, R., Ilmoniemi, R. J., Knuutila, J., and Lounasmaa, O. V. (1993). Magnetoencephalography-theory, instrumentation, and applications to noninvasive studies of the working human brain. *Rev. Modern Phys.* 65, 413–497. doi: 10.1103/revmodphys.65.413
- Honey, C. J., Sporns, O., Cammoun, L., Gigandet, X., Thiran, J. P., Meuli, R., et al. (2009). Predicting human resting-state functional connectivity from structural connectivity. *Proc. Natl. Acad. Sci. U.S.A.* 106, 2035–2040. doi: 10.1073/pnas.0811168106
- Hunter, J. D. (2007). Matplotlib: a 2d graphics environment. *Comput. Sci. Eng.* 9, 90–95. doi: 10.1109/mcse.2007.55
- Jansen, B., and Rit, V. (1995). Electroencephalogram and visual evoked potential generation in a

- mathematical model of coupled cortical columns. *Biol. Cybern.* 73, 357–366. doi: 10.1007/BF00199471
- Jansen, B., Zouridakis, G., and Brandt, M. (1993). A neurophysiologically-based mathematical model of flash visual evoked potentials. *Biol. Cybern.* 68, 275–283. doi: 10.1007/BF00224863
- Jirsa, V. (2004). Connectivity and dynamics of neural information processing. *Neuroinformatics* 2, 183–204. doi: 10.1385/NI:2:2:183
- Jirsa, V., and Haken, H. (1996). Field theory of electromagnetic brain activity. *Phys. Rev. Lett.* 77, 960–963. doi: 10.1103/PhysRevLett.77.960
- Jirsa, V., and Haken, H. (1997). A derivation of a macroscopic field theory of the brain from the quasi-microscopic neural dynamics. *Phys. D* 99, 503–526. doi: 10.1016/S0167-2789(96)00166-2
- Jirsa, V., and Kelso, J. A. (2000). Spatiotemporal pattern formation in neural systems with heterogeneous connection topologies. *Phys. Rev. E Stat. Phys. Plasmas Fluids Relat. Interdisciplin. Topics* 62(6 Pt B), 8462–8465. doi: 10.1103/PhysRevE.62.8462
- Jirsa, V., Jantzen, K., Fuchs, A., and Kelso, J. (2002). Spatiotemporal forward solution of the eeg and meg using network modeling. *IEEE Trans. Med. Imag.* 21, 493–504. doi: 10.1109/TMI.2002.1009385
- Jirsa, V., and Stefanescu, R. (2010). Neural population modes capture biologically realistic large scale network dynamics. *Bull. Math. Biol.* 73, 325–343. doi: 10.1007/s11538-010-9573-9
- Jirsa, V., Sporns, O., Breakspear, M., Deco, G., and McIntosh, A. R. (2010). Towards the virtual brain: network modeling of the intact and the damaged brain. *Arch. Ital. Biol.* 148, 189–205.
- Kelso, S. (1995). *Dynamic Patterns: The Self-Organization of Brain and Behavior (Complex Adaptive Systems)*. Cambridge, MA: MIT Press.
- Klöden and Platen (1995). *Numerical Solution of Stochastic Differential Equations*. Berlin: Springer.
- Klem, G. H., Lüders, H. O., Jasper, H. H., and Elger, C. (1999). The ten-twenty electrode system of the international federation. the international federation of clinical neurophysiology. *Electroencephalogr. Clin. Neurophysiol. Suppl.* 52, 3–6.
- Knock, S., McIntosh, A., Sporns, O., Kötter, R., Hagmann, P., and Jirsa, V. (2009). The effects of physiologically plausible connectivity structure on local and global dynamics in large scale brain models. *J. Neurosci. Methods* 183, 86–94. doi: 10.1016/j.jneumeth.2009.07.007
- Kötter, R. (2004). Online retrieval, processing, and visualization of primate connectivity data from the cocmac database. *Neuroinformatics* 2, 127–144. doi: 10.1385/NI:2:2:127
- Kötter, R., and Wanke, E. (2005). Mapping brains without coordinates. *Philos. Trans. R. Soc. Lond. B Biol. Sci.* 360, 751–766. doi: 10.1098/rstb.2005.1625
- Lake, D. E., and Moorman, J. R. (2011). Accurate estimation of entropy in very short physiological time series: the problem of atrial fibrillation detection in implanted ventricular devices. *Am. J. Physiol. Heart Circ. Physiol.* 300, H319–H325. doi: 10.1152/ajpheart.00561.2010
- Liley, D. T. J., and Bojak, I. (2005). Understanding the transition to seizure by modeling the epileptiform activity of general anesthetic agents. *J. Clin. Neurophysiol.* 22, 300–313.
- Logothetis, N. K., Pauls, J., Augath, M., Trinath, T., and Oeltermann, A. (2001). Neurophysiological investigation of the basis of the fmri signal. *Nature* 412, 150–157. doi: 10.1038/35084005
- Lopes da Silva, F. H., Hoeks, A., Smits, H., and Zetterberg, L. H. (1974). Model of brain rhythmic activity. *Biol. Cybern.* 15, 27–37. doi: 10.1007/BF00270757
- McIntosh, A., Kovacevic, N., Lippe, S., Garrett, D., Grady, C., and Jirsa, V. (2010). The development of a noisy brain. *Arch. Ital. Biol.* 148, 323–337.
- Mosher, J., Leahy, R., and Lewis, P. (1999). EEG and MEG: forward solutions for inverse methods. *IEEE Trans. Biomed. Eng.* 46, 245–259.
- Nagumo, J. (1962). An active pulse transmission line simulating nerve axon. *Proc. IRE* 50, 2061–2070. doi: 10.1109/jrproc.1962.288235
- Niedermeyer, E., and Lopes Da Silva, F. H., (eds.). (2005). *Electroencephalography: Basic Principles, Clinical Applications, and Related Fields*. Philadelphia, PA: Lippincott Williams & Wilkins.
- Nolte, G., Bai, O., Wheaton, L., Mari, Z., Vorbach, S., and Hallett, M. (2004). Identifying true brain interaction from eeg data using the imaginary part of coherency. *Clin. Neurophysiol.* 115, 2292–2307. doi: 10.1016/j.clinph.2004.04.029
- Nunez, P. (1974). The brain wave equation: a model for the EEG. *Math. Biosci.* 21, 279–297. doi: 10.1016/0025-5564(74)90020-0
- Nunez, P. L., (ed.). (1995). *Neocortical Dynamics and Human EEG Rhythms*. New York, NY: Oxford University Press.
- Nunez, P. L., and Srinivasan, R., (eds.). (1981). *Electric Fields of the Brain: The Neurophysics of EEG*. New York, NY: Oxford University Press.
- Ogawa, S., Menon, R. S., Kim, S. G., and Ugurbil, K. (1998). On the characteristics of functional magnetic resonance imaging of the brain. *Annu. Rev. Biophys. Biomol. Struct.* 27, 447–474. doi: 10.1146/annurev.biophys.27.1.447
- Ogawa, S., Menon, R. S., Tank, D. W., Kim, S. G., Merkle, H., Ellermann, J. M., et al. (1993). Functional brain mapping by blood oxygenation level-dependent contrast magnetic resonance imaging: a comparison of signal characteristics with a biophysical model. *Biophys. J.* 64, 803–812. doi: 10.1016/S0006-3495(93)81441-3
- Oliphant, T. E. (2006). *Guide to NumPy*. Trelgol Publishing.
- Pecevski, D., Natschläger, T., and Schuch, K. (2009). Pcsim: a parallel simulation environment for neural circuits fully integrated with python. *Front. Neuroinform.* 3:11. doi: 10.3389/neuro.11.011.2009
- Pedregosa, F., Varoquaux, G., Gramfort, A., Michel, V., Thirion, B., Grisel, O., et al. (2011). Scikitlearn: machine learning in Python. *JMLR* 12, 2825–2830.
- Pinotsis, D. A., and Friston, K. J. (2011). Neural fields, spectral responses and lateral connections. *Neuroimage* 55, 39–48. doi: 10.1016/j.neuroimage.2010.11.081
- Pinotsis, D. A., Moran, R. J., and Friston, K. J. (2012). Dynamic causal modeling with neural fields. *Neuroimage* 59, 1261–1274. doi: 10.1016/j.neuroimage.2011.08.020
- Ramachandran, P., and Varoquaux, G. (2011). Mayavi: 3D visualization of scientific data. *Comput. Sci. Eng.* 13, 40–51. doi: 10.1109/mcse.2011.35
- Ray, S., and Bhalla, U. S. (2008). Pymoose: interoperable scripting in python for moose. *Front. Neuroinform.* 2:6. doi: 10.3389/neuro.11.006.2008
- Rennie, C. J., Robinson, P. A., and Wright, J. J. (1999). Effects of local feedback on dispersion of electrical waves in the cerebral cortex. *Phys. Rev. E* 59, 3320–3329. doi: 10.1103/PhysRevE.59.3320
- Rennie, C. J., Robinson, P. A., and Wright, J. J. (2002). Unified neurophysical model of eeg spectra and evoked potentials. *Biol. Cybern.* 86, 457–471. doi: 10.1007/s00422-002-0310-9
- Riera, J., Aubert, E., Iwata, K., Kawashima, R., Wan, X., and Ozaki, T. (2005). Fusing eeg and fmri based on a bottom-up model: inferring activation and effective connectivity in neural masses. *Philos. Trans. R. Soc. Lond. B Biol. Sci.* 360, 1025–1041. doi: 10.1098/rstb.2005.1646
- Ritter, P., Schirner, M., McIntosh, A., and VK, J. (2013). The virtual brain integrates computational modelling and multimodal neuroimaging. *Brain Connect.* 3, 121–145. doi: 10.1089/brain.2012.0120
- Robinson, P., Rennie, C., and Wright, J. (1997). Propagation and stability of waves of electrical activity in the cerebral cortex. *Phys. Rev. E* 56, 826–840. doi: 10.1103/PhysRevE.56.826
- Robinson, P. A. (2011). Neural field theory of synaptic plasticity. *J. Theor. Biol.* 285, 156–163. doi: 10.1016/j.jtbi.2011.06.023
- Robinson, P. A., Rennie, C. J., Rowe, D. L., O'Connor, S. C., Wright, J. J., Gordon, E., et al. (2003). Neurophysical modeling of brain dynamics. *Neuropsychopharmacology* 28 (Suppl. 1), S74–S79. doi: 10.1038/sj.npp.1300143
- Robinson, P. A., Rennie, C. J., Wright, J. J., Bahramali, H., Gordon, E., and Rowe, D. L. (2001). Prediction of electroencephalographic spectra from neurophysiology. *Phys. Rev. E* 63(2 Pt 1):021903. doi: 10.1103/PhysRevE.63.021903
- Rowe, D. L., Robinson, P. A., and Rennie, C. J. (2004). Estimation of neurophysiological parameters from the waking EEG using a biophysical model of brain dynamics. *J. Theor. Biol.* 231, 413–433. doi: 10.1016/j.jtbi.2004.07.004
- Rubinov, M., and Sporns, O. (2010). Complex network measures of brain connectivity: uses and interpretations. *Neuroimage* 52, 1059–1069. doi: 10.1016/j.neuroimage.2009.10.003
- Sarvas, J. (1987). Basic mathematical and electromagnetic concepts of the biomagnetic inverse problems. *Phys. Med. Biol.* 32, 11–22. doi: 10.1088/0031-9155/32/1/004
- Schneider, G. E. (1969). Two visual systems. *Science* 163, 895–902. doi: 10.1126/science.163.3870.895

- Senden, M., Goebel, R., and Deco, G. (2012). Structural connectivity allows for multi-threading during rest: the structure of the cortex leads to efficient alternation between resting state exploratory behavior and default mode processing. *Neuroimage* 60, 2274–2284. doi: 10.1016/j.neuroimage.2012.02.061
- Shreiner, D., Woo, M., Neider, J., and Davis, T. (2005). *OpenGL(R) Programming Guide: The Official Guide to Learning OpenGL(R)*. Version 2, 5th Edn. Addison-Wesley Professional. Available online at: <http://www.openglprogramming.com/red/about.html>
- Sotero, R. C., and Trujillo-Barreto, N. J. (2008). Biophysical model for integrating neuronal activity, EEG, fMRI and metabolism. *Neuroimage* 39, 290–309. doi: 10.1016/j.neuroimage.2007.08.001
- Sotero, R. C., Trujillo-Barreto, N. J., Iturria-Medina, Y., Carbonell, F., and Jimenez, J. C. (2007). Realistically coupled neural mass models can generate eeg rhythms. *Neural Comput.* 19, 478–512. doi: 10.1162/neco.2007.19.2.478
- Spacek, M., Blanche, T., and Swindale, N. (2008). Python for large-scale electrophysiology. *Front. Neuroinform.* 2:9. doi: 10.3389/neuro.11.009.2008
- Spiegler, A., Kiebel, S. J., Atay, F. M., and Knösche, T. R. (2010). Bifurcation analysis of neural mass models: impact of extrinsic inputs and dendritic time constants. *Neuroimage* 52, 1041–1058. doi: 10.1016/j.neuroimage.2009.12.081
- Stefanescu, R., and Jirsa, V. (2008). A low dimensional description of globally coupled heterogeneous neural networks of excitatory and inhibitory. *PLoS Comput. Biol.* 4, 26–36. doi: 10.1371/journal.pcbi.1000219
- Stefanescu, R., and Jirsa, V. (2011). Reduced representations of heterogeneous mixed neural networks with synaptic coupling. *Phys. Rev. E* 83:026204. doi: 10.1103/PhysRevE.83.026204
- The HDF Group. (2000–2010). Hierarchical data format version 5. Available online at: <http://www.hdfgroup.org/>
- Valdes-Sosa, P. A., Sanchez-Bornot, J. M., Sotero, R. C., Iturria-Medina, Y., Aleman-Gomez, Y., Bosch-Bayard, J., et al. (2009). Model driven EEG/fMRI fusion of brain oscillations. *Hum. Brain Mapp.* 30, 2701–2721. doi: 10.1002/hbm.20704
- von Ellenrieder, N., Beltrachini, L., and Muravchik, C. H. (2012). Electrode and brain modeling in stereo-EEG. *Clin. Neurophysiol.* 123, 1745–1754. doi: 10.1016/j.clinph.2012.01.019
- Wilson, H., and Cowan, J. (1972). Excitatory and inhibitory interactions in localized populations of model neurons. *Biophys. J.* 12, 1–24. doi: 10.1016/S0006-3495(72)86068-5
- Wilson, H., and Cowan, J. (1973). A mathematical theory of the functional dynamics of cortical and thalamic nervous tissue. *Kybernetik* 13, 55–80. doi: 10.1007/bf00288786
- Wong, K.-F., and Wang, X.-J. (2006). A recurrent network mechanism of time integration in perceptual decisions. *J. Neurosci.* 26, 1314–1328. doi: 10.1523/JNEUROSCI.3733-05.2006
- Wright, J. J., and Liley, D. T. J. (1995). Simulation of electrocortical waves. *Biol. Cybern.* 72, 347–356. doi: 10.1007/BF00202790
- Zalesky, A., Fornito, A., Harding, I. H., Cocchi, L., Yücel, M., Pantelis, C., et al. (2010). Whole-brain anatomical networks: does the choice of nodes matter? *Neuroimage* 50, 970–983. doi: 10.1016/j.neuroimage.2009.12.027

**Conflict of Interest Statement:** The authors declare that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.

Received: 01 March 2013; accepted: 22 May 2013; published online: 11 June 2013.

Citation: Sanz Leon P, Knock SA, Woodman MM, Domide L, Mersmann J, McIntosh AR and Jirsa V (2013) The Virtual Brain: a simulator of primate brain network dynamics. *Front. Neuroinform.* 7:10. doi: 10.3389/fninf.2013.00010

Copyright © 2013 Sanz Leon, Knock, Woodman, Domide, Mersmann, McIntosh and Jirsa. This is an open-access article distributed under the terms of the Creative Commons Attribution License, which permits use, distribution and reproduction in other forums, provided the original authors and source are credited and subject to any copyright notices concerning any third-party graphics etc.



# Disruption of transfer entropy and inter-hemispheric brain functional connectivity in patients with disorder of consciousness

Verónica Mäki-Marttunen<sup>1,2†</sup>, Ibai Diez<sup>3†</sup>, Jesus M. Cortes<sup>3,4</sup>, Dante R. Chialvo<sup>2</sup> and Mirta Villarreal<sup>1,2\*</sup>

<sup>1</sup> Department of Cognitive Neuroscience, Institute for Neurological Research, FLENI, Buenos Aires, Argentina

<sup>2</sup> CONICET, Buenos Aires, Argentina

<sup>3</sup> Biocruces Health Research Institute, Hospital Universitario de Cruces, Barakaldo, Spain

<sup>4</sup> Ikerbasque, The Basque Foundation for Science, Bilbao, Spain

## Edited by:

Miguel A. Muñoz, Universidad de Granada, Spain

## Reviewed by:

Sebastiano Stramaglia, Università degli Studi di Bari, Italy

Enzo Tagliazucchi, Goethe University Frankfurt, Germany

## \*Correspondence:

Mirta Villarreal, Department of Cognitive Neuroscience, Institute for Neurological Research, FLENI, Montañeses 2325, C1428AQK, Buenos Aires, Argentina  
e-mail: mvillarreal@fleni.org.ar

<sup>†</sup> These authors have contributed equally to this work.

Severe traumatic brain injury can lead to disorders of consciousness (DOC) characterized by deficit in conscious awareness and cognitive impairment including coma, vegetative state, minimally consciousness, and lock-in syndrome. Of crucial importance is to find objective markers that can account for the large-scale disturbances of brain function to help the diagnosis and prognosis of DOC patients and eventually the prediction of the coma outcome. Following recent studies suggesting that the functional organization of brain networks can be altered in comatose patients, this work analyzes brain functional connectivity (FC) networks obtained from resting-state functional magnetic resonance imaging (rs-fMRI). Two approaches are used to estimate the FC: the Partial Correlation (PC) and the Transfer Entropy (TE). Both the PC and the TE show significant statistical differences between the group of patients and control subjects; in brief, the inter-hemispheric PC and the intra-hemispheric TE account for such differences. Overall, these results suggest two possible rs-fMRI markers useful to design new strategies for the management and neuropsychological rehabilitation of DOC patients.

**Keywords:** disorder of consciousness, resting state, functional magnetic resonance imaging, BOLD signal, transfer entropy, partial correlation, functional connectivity, brain networks

## 1. INTRODUCTION

Recent studies have shown that brain networks obtained from functional Magnetic Resonance Imaging (fMRI) recordings are altered in patients with severe disorder of consciousness (DOC) (Boveroux et al., 2010; Noirhomme et al., 2010; Heine et al., 2012; Perri et al., 2013). DOC can result from severe brain injury and is characterized by an absence of awareness of the self and the environment, either with preserved or disrupted sleep-awake cycle. DOC encompasses a wide spectrum of clinical conditions with different levels in the content of conscious awareness, ranging from the coma state (CS, patients who have a disrupted sleep-awake cycle and don't wake up), vegetative state (VS, who preserve sleep-awake cycle but are unaware of themselves and the environment), minimally consciousness state (MCS, patients who are unable to reliably communicate but show reproducible albeit fluctuating behavioral evidence of awareness), to lock-in syndrome (LI, patients who are fully conscious but are completely paralyzed except for small movements of the eyes or eyelids). For the prognosis of these patients, the clinical practice scores this graduation in DOC response by the Glasgow Coma Scale (GCS) (Teasdale and Jennett, 1974), or as we will use in this paper,

by an alternative scale such as the JFK Coma Recovery Scale-Revised (CSR-R) (Giacino et al., 2004). This scale encodes the neurological and behavioral state of the DOC patient providing a number ranging from 0 to 23, 0 for the deepest coma state, 23 for the fully recovered one. Despite the existence of such scales, there is a need for more reliable methods that based on brain neuroimaging can provide better characterization of the large-scale disturbances of brain function in DOC. Ultimately these approaches should help in understanding and eventually predicting coma outcome.

The resting state functional Magnetic Resonance Imaging (rs-fMRI) accounts for the spontaneous brain activity occurring in the high-amplitude ultra-slow (0.1 Hz) fluctuations in the Blood-Oxygen-Level-Dependent (BOLD) signal, defining networks of correlated spontaneous activity of brain Functional Connectivity (FC) (Raichle et al., 2001; Beckmann et al., 2005). The interaction between these distributed networks as well as subcortical modules is considered critical for conscious processing, and has been shown to be disrupted in DOC state (Tononi, 2004; Cauda et al., 2009; Rosanova et al., 2012). Furthermore, the rs-fMRI paradigm is a very suitable strategy for DOC patients, since they are not able to efficiently perform specific tasks. The present study addresses the question of whether the FC obtained from the rs-fMRI is altered at different brain regions as a consequence of consciousness disturbances. To this end, we investigate the FC obtained by two different measures: the Partial Correlation (PC) and the

**Abbreviations:** fMRI, functional Magnetic Resonance Imaging; rs, resting state; DOC, disorder of consciousness; BOLD, Blood-Oxygen-Level-Dependent; FC, Functional Connectivity; TE, Transfer Entropy; PC, Partial Correlation.



Transfer Entropy (TE), in two different groups: healthy adults and DOC patients.

Information theory offers an arsenal of different measures, complementing the linear correlation estimations of FC. These information tools are typically built as extensions of the Shannon Entropy, quantify the interactions between variables by measuring the information which is shared or transferred between them (Jaynes, 1957; Cover and Thomas, 2006). In the last decade, the TE method is growing in popularity as it can account for directed interactions between time-series variables (Schreiber, 2000). When applied to neuroimaging time-series, TE is a data-driven measure that assesses the functional connectivity between brain areas even for non-linear interactions. Unlike the correlations, TE reveals directionality in the interactions, allowing for determining a *directed* FC between areas.

We hypothesize that FC would be reduced in DOC patients since consciousness implies functional integration (Tononi, 2004). We anticipate that PC and TE would show different behaviors in patients with increasing level of consciousness, provided that they can be related to different mechanisms of information processing in the brain.

The paper is organized as follow: in Material and Methods, we give details on the the data acquisition and preprocessing and define the two measures PC and TE to compute FC patterns. The next section is dedicated to present the results of the analysis. The paper closes with a discussion on some consequences of the alteration of the FC patterns in DOC patients.

## 2. MATERIALS AND METHODS

### 2.1. SUBJECTS

Seventeen healthy subjects (**Group 1**) aged  $25 \pm 5$  year old (8 men, 9 women), with no history of neurological or psychiatric problems, participated in this study as a control group. The Edinburgh Handedness Inventory was used to assess handedness (Oldfield, 1971), resulting in thirteen subjects right-handed and four left-handed. Eleven DOC patients (**Group 2**) were scanned (age range, 17–44 years; 6 men, 5 women). Data from two patients were subsequently excluded because of unacceptable degrees of head and body movements. The coma severity for each patient was clinically assessed using the Revised Coma Recovery Scale

[CRS-R, (Giacino et al., 2004)]: scores range from 0 (meaning deep coma state) to 23 (full recovery). The patients were scanned the first time between 2 to 6 months after major acute brain injury, and a second time between 3 to 6 months after the first scan (**Table 1**). For better comparison, group 2 was subdivided into 2 subgroups: **Group 2a** ( $n = 12$ ) is composed by all scans of DOC patients who had a corresponding CRS-R scale. **Group 2b** ( $n = 4$ ) includes the second scans of the four patients who recovered consciousness before the second session (marked with asterisks in **Table 1**). The study protocol was approved by the Institutional Review Board of the Institute of Neurological Research FLENI. Informed consent was directly obtained from healthy participants and from the next kin of each of the patients.

### 2.2. MRI DATA ACQUISITION AND PREPROCESSING

The fMRI measurements were carried out on a 3T Signa HDxt GE scanner using an 8 channel head coil. Change in blood-oxygenation-level-dependent (BOLD)  $T2^*$  signal was measured using an interleaved gradient-echo EPI sequence. Thirty contiguous slices were obtained in the AC-PC plane with the following parameters: 2 s repetition time (TR), flip angle:  $90^\circ$ , 24 cm field of view,  $64 \times 64$  pixel matrix, and  $3.75 \times 3.75 \times 4.0$  mm voxel dimensions. During the experimental session subjects lied quietly for a period of 7 min. 220 whole brain volumes were obtained per scan session, including 5 dummy scans to allow for T1 saturation effects that were discarded from the analysis. High resolution T1-weighted 3D fast SPGR-IR were also acquired ( $TR = 6.604$  ms,  $TE = 2.796$  ms,  $TI = 450$ ; parallel imaging (ASSET) acceleration factor = 2; acquisition matrix size =  $256 \times 256$ ; FOV = 24 cm; slice thickness = 1.2 mm; 120 contiguous sections). The image data was analyzed using SPM8 (Wellcome Department of Cognitive Neurology, London, UK) implemented in MATLAB (MathWorks Inc., Natick, MA). The functional images were subjected to temporal alignment and volumes were corrected for movement using a six-parameter automated algorithm. The realigned volumes were spatially normalized to fit to the template created using the Montreal Neurological Institute reference brain based on Talairach and Tournoux's stereotaxic coordinate system (Ashburner and Friston,

**Table 1 | Clinical characteristics of DOC patients.**

Patient code	Age	Time between accident and first scan (months)	Clinical assessment at first scan	Time between first and second scan (months)	Clinical assessment at second scan
P1	34	2	VS	5	VS
P2*	18	4	MCS	4	C
P3*	44	2	MCS	3	C
P4	17	6	VS	6	MCS
P5	26	4	VS	3	MCS
P6*	26	4	EMCS	4	C
P7	29	4	MCS	3	MCS
P8	41	2	VS	6	VS
P9*	34	5	VS	5	C

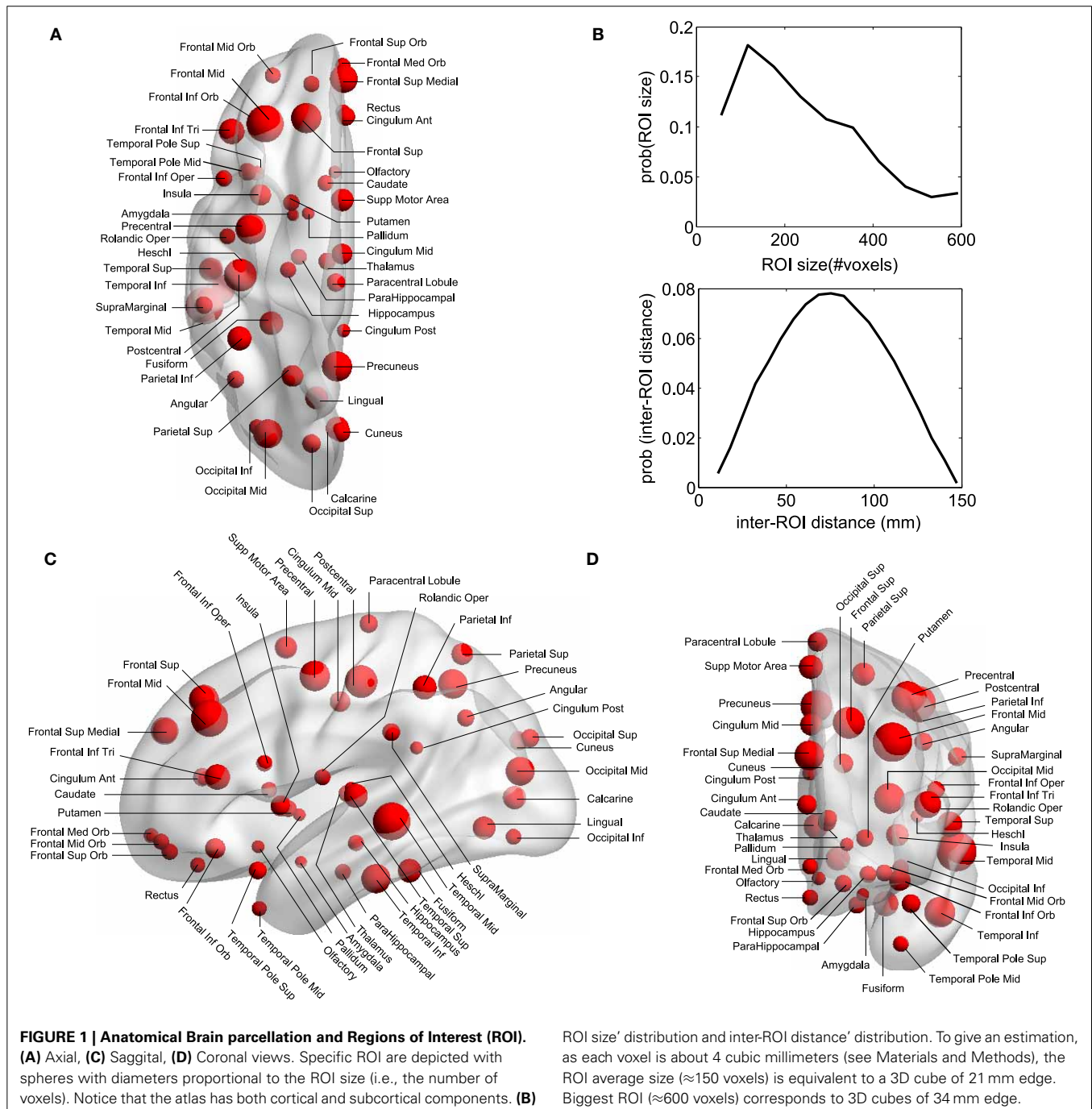
\*Patients that recovered from DOC at the second scan. VS, Vegetative State; MCS, Minimally Consciousness State; C, Conscious; EMCS, Emergence from MCS (an intermediate state between MCS and C).

1999). The spatially normalized volumes consisting of  $4 \times 4 \times 4 \text{ mm}^3$  voxels were smoothed with a 8-mm FWHM isotropic Gaussian kernel. Additionally, a linear trend removal and band pass filtering between 0.01 and 0.08 Hz was applied on the data.

### 2.3. BRAIN PARCELLATION AND REGIONS OF INTEREST

Regions of Interest (ROI) were defined following the Automatic Anatomical Labeling (AAL) atlas (Tzourio-Mazoyer et al., 2002) (see Figures 1A–D) which comprises 90 different areas, 45 on

each hemisphere (e.g., hippocampus Left, hippocampus Right, amygdala Left, amygdala Right, etc.). Importantly for the study of DOC patients, the AAL atlas includes both cortical and subcortical components (eg., hippocampus, thalamus and amygdala). Per each ROI we have extracted a mesoscopic (multi-voxel) fMRI time-series resulting from averaging over all fMRI time-series of all voxels within a given ROI (Figure 1B is showing the ROI size distribution among all areas). The MNI coordinates of the centroids in each ROI are used to calculate the Euclidean distance between each pair of regions (Figure 1B).



## 2.4. FUNCTIONAL CONNECTIVITY MATRICES

Correlated areas in the rs-fMRI time series define the Functional Connectivity (FC) matrices. Two methods have been used to the FC: The PC and TE.

### 2.4.1. The partial correlation

Matrix has dimensions  $90 \times 90$  (with 90 the ROIs number) and each element is given by the pairwise PC between any two ROIs. PC is a correlation matrix that removes for a given ROIs pair the effect of the rest of the variables, i.e., removing the correlations contribution which are coming from common neighbors interactions. Let  $C$  be a non-singular correlation matrix, then each element of the PC matrix is given by

$$PC_{ij} = -\frac{P_{ij}}{\sqrt{P_{ii}P_{jj}}} \quad (1)$$

where  $P \equiv C^{-1}$  is the inverse of the correlation matrix (i.e., the precision matrix).

Notice that PC is a symmetrical measure, i.e.,  $PC_{ij} = PC_{ji}$ . We also have computed the standard correlations  $C$ , and although  $C$  is more noisy than PC, the results we are showing here for the PC are also valid for the standard correlation.

The PC was computed by using the *partialcorr* method incorporated in MATLAB (MathWorks Inc., Natick, MA). The second argument that the function *partialcorr* outputs is a matrix of p-values for testing the hypothesis of no PC against the alternative that there is a non zero PC.

PC matrices were calculated for each subject and grouped into the following categories: inter-hemispheric (between one area on the left and all the other areas at right hemisphere, or vice versa), homologous inter-hemispheric (one area on the left hemisphere and its homologous area on the right hemisphere, or vice versa), left intra-hemispheric, right intra-hemispheric, and total.

### 2.4.2. Transfer entropy

quantifies the *directed* interaction between any two ROIs. To compute it, let define  $i^F$  as the future of the time series in ROI  $i$ . Similarly,  $i^P$  and  $j^P$  the pasts of ROIs  $i$  and  $j$ . Then, the TE from  $j$  to  $i$  is defined as

$$TE_{ji} = H(i^F|i^P) - H(i^F|i^P, j^P) \quad (2)$$

with  $H(i^F|i^P) = H(i^F, i^P) - H(i^P)$ , the conditional Shannon entropy of  $i^F$  conditioning on  $i^P$  [for details, see (Cover and Thomas, 2006)]. Similarly,  $H(i^F|i^P, j^P) = H(i^F, i^P, j^P) - H(i^P, j^P)$  is the conditional Shannon entropy of  $i^F$  conditioning on  $i^P$  and  $j^P$ .

The TE is a non-symmetrical measure, i.e.,  $TE_{ij} \neq TE_{ji}$ .

The Shannon Entropy (average uncertainty) of the random variable  $X$  is defined as  $H(X) = -\sum_x \text{prob}(x) \log \text{prob}(x)$ , where  $x$  represents a possible state in variable  $X$  (Cover and Thomas, 2006). For base 2 logarithm (as we have done here), the information is expressed as information bits.

To compute probabilities from continuous variables, we did not perform binning; alternatively, we just rounded each value in the time series to its nearest integer and computed probabilities

(number of time points in a given state divided by the total time-series length). The conditional entropies have been calculated with the function *condentropy* developed by Hanchuan Peng in C++ and plug-into MATLAB via mex. The code is available for download from Peng (Peng).

For the past of the time series it was considered the original time series. Their future were built by shifting the time series in MATLAB with the function *circshift* with a lag value of 10 time points. This lag number was previously chosen (and fixed for all simulations) in order to maximize TE values.

The statistical significance of the TE values was estimated by shuffling the time series of the target ROI (for the calculation of the TE from  $j$  to  $i$ , hereafter  $j$  will be referred as the source and  $i$  as the target). The time series was shuffled to remove the temporal information in the target variable. Next, the TE value is calculated for many repetitions of this shuffling procedure to obtain the distribution of values under the null hypothesis of zero values of TE (i.e., zero uncertainty reduction from source  $j$  to target  $i$ ).

TE matrices were calculated for each subject and grouped into the following categories: homologous inter-hemispheric (one area on the left hemisphere to its homologous area on the right hemisphere and vice versa), left intra-hemispheric, right intra-hemispheric, inter-hemispheric (from one area on the left to all the other areas at right hemisphere, and from one area on the right to all other areas at the left hemisphere) and total.

### 2.4.3. Summary of brain categories

For easy reading we have adopted the following notation:

- PC calculations: LR: inter- hemispheric (between one area on one hemisphere and all the other areas at the other hemisphere). As the PC calculation is symmetric (LR is the same than RL) we condensed the inter-hemispheric PC in only LR. HIH: homologous inter-hemispheric. LL: left intra-hemispheric. RR: right intra-hemispheric.
- TE calculations:
  - HLR: homologous inter-hemispheric from left to right (one area on the left hemisphere to its homologous area on the right hemisphere).
  - HRL: homologous inter-hemispheric from right to left (one area on the right hemisphere to its homologous area on the left hemisphere).
  - LL: intra-hemispheric from left to left.
  - RR: intra-hemispheric from right to right.
  - LR: inter-hemispheric left-right (from one area on the left to all the other areas at right hemisphere).
  - RL: inter-hemispheric right-left (from one area on the right to all the other areas at left hemisphere).

## 2.5. STATISTICAL ANALYSIS

PC and TE individual matrices were thresholded at a probability value of 0.1 (i.e., 10% confidence); these data were used for **Tables 2, 3** and all the figures shown in the paper. We also computed PC and TE matrices at different confidence values, 5% and 100% (zero threshold), and the results did not considerably change (cf. Tables S1–S4).

For comparison of PC and TE values between the different brain categories and groups, a two-ways ANOVA test was performed, using the function *anovan* from MATLAB (MathWorks Inc., Natick, MA). For *post-hoc* analysis, multi-sample *t*-tests were performed between groups for each brain category using the function *multcompare* from MATLAB which include the Bonferroni correction for multiple comparisons. To assess possible deviations from the Gaussian distribution in the data, the Kruskal–Wallis non-parametric tests were also performed using the function *kruskal-wallis* from MATLAB. The groups comparison results showed very little differences across these tests, cf. **Tables 2, 3**, in which the statistically significant differences from control were denoted using asterisks at different colors (black for ANOVA and green for Kruskal–Wallis).

## 2.6. A FURTHER TEST FOR fMRI HEAD MOTION ARTIFACTS

To reject the possibility of head motion artifacts, PC was re-computed in a matrix which included the original 90 ROIs from the AAL atlas plus two motion regressors: the translational modulus and rotational modulus. It is expected, if important correlations were introduced by head motion, that the PC results obtained from this expanded matrix must show significant

differences in comparison with the results gathered from the original 90 ROI's. However, this was not the case; no changes were observed which indicates that the data is free of head motion artifacts.

## 3. RESULTS

### 3.1. PARTIAL LINEAR CORRELATIONS (PC)

First we looked into the PC patterns (**Table 2**). ANOVA between G1 and G2 shows a significant effect of categories ( $p < 0.001$ ), and a significant interaction between categories and groups ( $p < 0.001$ ). Controls have a significantly smaller PC mean value than patients ( $p < 0.001$ ). When looking into categories, HIH PCs are significantly higher than LL, RR, and LR ( $p < 0.001$ ). In addition, LL and RR values are significantly higher than LR ( $p < 0.001$ ). To further inspect the interaction, we performed *post-hoc* multiple comparison tests between groups for the different categories. HIH PCs are significantly higher in G1 ( $p < 0.001$ ). The Kruskal–Wallis test gave the same results, with the addition of being LL and RR PCs significantly higher in G2 compared with G1 ( $p < 0.005$ ).

The comparison between G1 and G2a gives the same results. However, when comparing G1 and G2b, the effect of group still holds but is smaller than that between G1 and G2a ( $p = 0.002$ ). The effect of categories is the same as in G1 vs. G2 comparison, and there is a significant interaction effect ( $p < 0.001$ ). *Post-hoc* tests show that HIH PCs are significantly smaller in G2b with respect to G1 ( $p < 0.001$ ). Finally, the comparison including all brain categories (total) was significant between G1 vs. G2 and G1 vs. G2a ( $P = 0.018$  and  $0.044$  respectively). The same significant differences were conserved with the Kruskal–Wallis test. Results can be seen in **Table 2** and **Figure 2**.

In summary, the partial linear correlations approach allows to expose a differential functional connectivity in a healthy conscious brain in comparison with a DOC state and a recent recovery from it. A reduced inter-hemispheric connectivity is evident in DOC patients.

### 3.2. TRANSFER ENTROPY

We then examined the uncertainty reduction (information) transferred between ROIs pairs by computing the TE. ANOVA on TE values for G1 and G2 shows a significant effect of group ( $p = 0.0025$ ) and categories ( $p < 0.001$ ). Particularly there were significant differences between HLR and HRL TEs and the TE values for the other categories. In the case of HLR, TEs are significantly lower than LL, RR and inter-hemispheric (LR and RL) TEs ( $p < 0.005$ ), whereas HRL TEs are significantly lower than RR and RL TEs ( $p < 0.025$ ). There is no interaction effect between group and brain category. The *post-hoc* analysis showed that LL, LR and the total TE values differ between controls and patients.

However, when performing the ANOVA for G1 and G2a there is a significant effect of group ( $p < 0.001$ ) and categories ( $p < 0.001$ ). *Post-hoc* tests show that LL and RR TEs are significantly higher in G1 than in G2a ( $p < 0.05$ ). In addition, TE for LR is also significantly higher in G1 ( $p = 0.001$ ).

When comparing G1 and G2b, there was significant effect of group ( $p = 0.042$ ) and categories ( $p < 0.001$ ). Additionally,

**Table 2 | PC average values  $\pm$  standard deviation thresholded at 10% confidence (see Materials and Methods).**

PC	G1	G2	G2a	G2b
LR	0.11 $\pm$ 0.01	0.12 $\pm$ 0.01	0.11 $\pm$ 0.01	0.12 $\pm$ 0.01
HIH	0.40 $\pm$ 0.03	0.24 $\pm$ 0.03**	0.24 $\pm$ 0.04**	0.26 $\pm$ 0.04**
LL	0.13 $\pm$ 0.01	0.15 $\pm$ 0.01*	0.14 $\pm$ 0.09*	0.15 $\pm$ 0.01
RR	0.13 $\pm$ 0.01	0.15 $\pm$ 0.01*	0.14 $\pm$ 0.09*	0.15 $\pm$ 0.01*
Total	0.12 $\pm$ 0.01	0.13 $\pm$ 0.01**	0.13 $\pm$ 0.01**	0.13 $\pm$ 0.01**

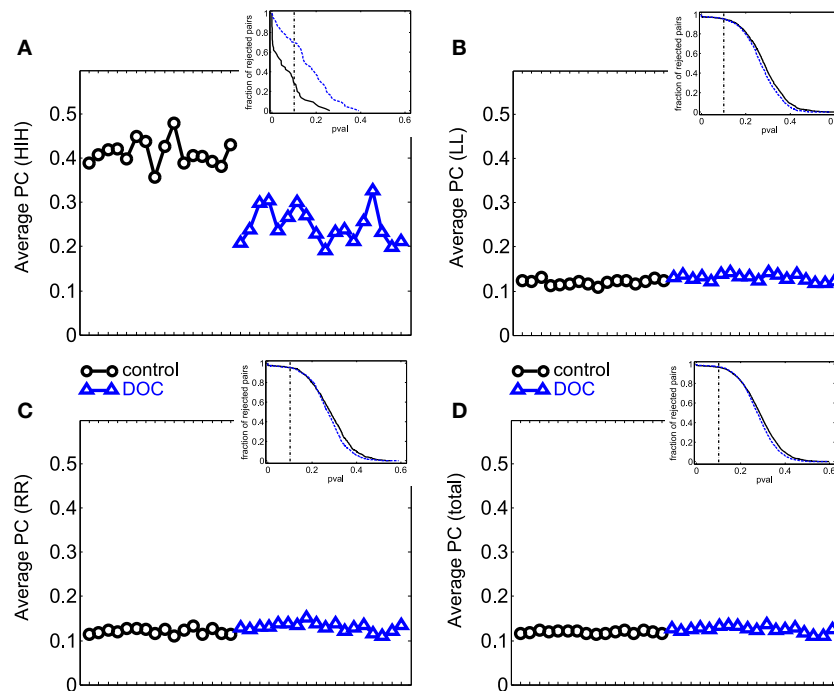
\*Significantly different from G1;  $p < 0.05$ . Significant differences are indicated with black asterisks for ANOVA and green for Kruskal–Wallis tests. LR, inter-hemispheric; HIH, homologous inter-hemispheric; LL, left intra-hemispheric; RR, right intra-hemispheric.

**Table 3 | TE average values  $\pm$  standard deviation.**

TE	G1	G2	G2a	G2b
HLR	0.009 $\pm$ 0.027	0.006 $\pm$ 0.015	0.004 $\pm$ 0.011	0.017 $\pm$ 0.030
HRL	0.011 $\pm$ 0.020	0.020 $\pm$ 0.049	0.020 $\pm$ 0.053	0.019 $\pm$ 0.032
LL	0.040 $\pm$ 0.021	0.017 $\pm$ 0.016**	0.013 $\pm$ 0.013**	0.040 $\pm$ 0.003
RR	0.039 $\pm$ 0.020	0.027 $\pm$ 0.039*	0.019 $\pm$ 0.033**	0.065 $\pm$ 0.050
LR	0.043 $\pm$ 0.024	0.021 $\pm$ 0.021**	0.016 $\pm$ 0.017**	0.047 $\pm$ 0.017
RL	0.043 $\pm$ 0.021	0.031 $\pm$ 0.045*	0.024 $\pm$ 0.042**	0.066 $\pm$ 0.049
Total	0.041 $\pm$ 0.018	0.024 $\pm$ 0.026**	0.018 $\pm$ 0.022**	0.054 $\pm$ 0.028

\*Significantly different from G1;  $p < 0.05$ . Significant differences are indicated with black asterisks for ANOVA and green for Kruskal–Wallis tests. HLR: homologous inter-hemispheric from left to right; HRL, homologous inter-hemispheric from right to left; LL, left intra-hemispheric; RR, right intra-hemispheric; LR, inter-hemispheric left to right; RL, inter-hemispheric right to left.





**FIGURE 2 | Average PC values per subject. (A)** HIH (homologue inter-hemispheric areas); **(B)** LL (left intra-hemispheric); **(C)** RR (right intra-hemispheric); **(D)** total. Insets depict the fraction of rejected pairs of areas for a given probability level. PC values were thresholded at a

probability value of 0.1 (dashed lines in the insets). Black circle: G1 (control); blue triangles: G2 (DOC). Observe the huge differences between G1 and G2 for HIH compared to LL and RR. For detailed values, see **Table 2**.

when looking into the main effect of brain sections, HLR and HRL TE values were significantly smaller than LR and RL ( $p < 0.05$ ). However the multiple-compare test did not revealed any significant difference. The Kruskal–Wallis test gave the same general results but in this case adding significant differences between G2 and G1 in the same regions were we previously found only for G2a.

The results can be seen in **Table 3** and **Figure 3**. If two time series are highly correlated, their TE is close to zero in both directions; if they are not correlated but one influences the other's behavior, TE is high in that direction and very low in the opposite direction. In our results, the significant smaller TE between homologue areas with respect to the other TE values is consistent to the fact that they are highly correlated (cf. results in 3.1).

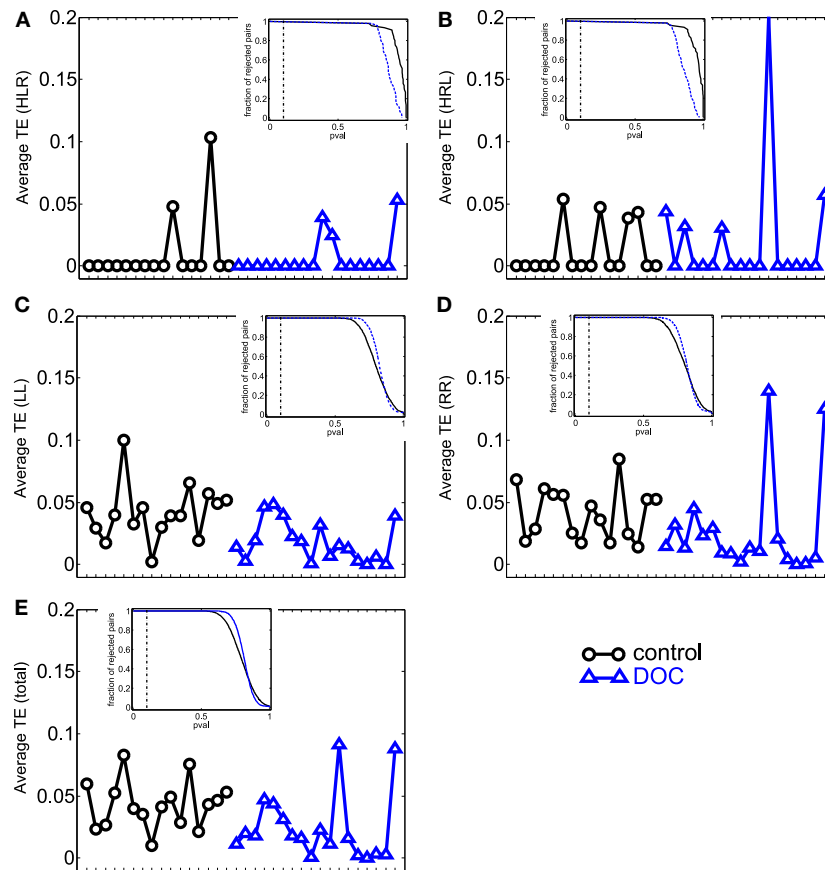
The differences found within hemispheres between the groups parallelize the increased intra-hemispheric correlations in G2 and G2a. When looking at G2b group, their averages are also biased by one patient that presented extremely high TE values (corresponding to the last case in the x-axis).

In summary, TE analysis exposes alterations in the FC exhibited by DOC patients. In particular, TE within hemispheres and between hemispheres is smaller, although no difference was found when looking at homologue areas. In contrast to the results obtained in the PC analysis, the differences found uphold irrespective of the Euclidean distance separating ROIs pairs, although when considering LL TE, a slight decrease in the statistical p value can be observed.

### 3.3. BETWEEN-HOMOLOGUE INTER-HEMISPHERIC PC AND LEFT INTRA-HEMISPHERIC TE

The results show that for all analyzed areas the best two discriminators are the between-homologue inter-hemispheric (HIH) PC (**Figures 4A–D**) and the left intra-hemispheric (LL) TE (**Figures 4E–H**). Here, colors denote group differences: black (G1), blue (G2), green (G2a) and magenta (G2b). For both PC and TE the thickness of links and arrows is proportional to the PC and TE values.

For PC there is a manifest anatomical disparity in the correlations pattern: it can be observed that homologue areas that are closer to each other show stronger correlations than farther ones (i.e., thicker connections at shorter distances in comparison with thinner connections at longer distances). To disentangle the behavior of the neural correlations regarding to a spatial factor, we look at the Euclidean distances between the centroids of homologue areas. For G1 the areas close to each other presented a high correlation, and beyond a threshold distance of 20 mm, correlations decreased, although the values remained high. Interestingly, the same behavior was found in G2. However, the correlation values there were shifted down, with lower mean value for areas closer than 20 mm, and decreasing for increasing distances. Thus, for ROIs areas distance-separated smaller than 20 mm, differences between G1 and G2 were smaller compared to areas separated at long distances, distance separation  $<20$  mm  $p\text{-val} = 10^{-6}$ , distance  $>40$  mm  $p\text{-val} = 10^{-14}$ . When inspecting G2a and G2b subgroups, there were no observable differences for anatomically



**FIGURE 3 | Average TE values per subject. (A)** HLR (homologue left-right inter-hemispheric areas); **(B)** HRL (homologue right-left inter-hemispheric areas); **(C)** LL (left intra-hemispheric); **(D)** RR (right intra-hemispheric); **(E)**

total. Insets depict the fraction of rejected pairs of areas for a given probability level. TE values were thresholded at a probability value of 0.1 (dashed lines in the insets). Black circles: G1 (control); blue triangles: G2 (DOC).

closer areas, whilst it could be detected a higher correlation of some of the anatomically further areas for G2b.

Regarding to the TE, not only the mean values of TE in LL areas were different between groups (Table 3), but the number of significant values of TE, i.e., the number of arrows plotted in Figures 4E–H varies across different groups. This number was more than 9 times bigger in G1 compared with G2 (G1 # links = 47, Figure 4E; G2 # links = 5, Figure 4F). When comparing with group G2b, this number doubled the one in group G1 (# links = 99, Figure 4H), possibly indicating a “transient” brain state in the pattern of information flows in group G2b in comparison with control.

### 3.4. CORRELATION BETWEEN fMRI MEASURES AND CRS-R SCORES

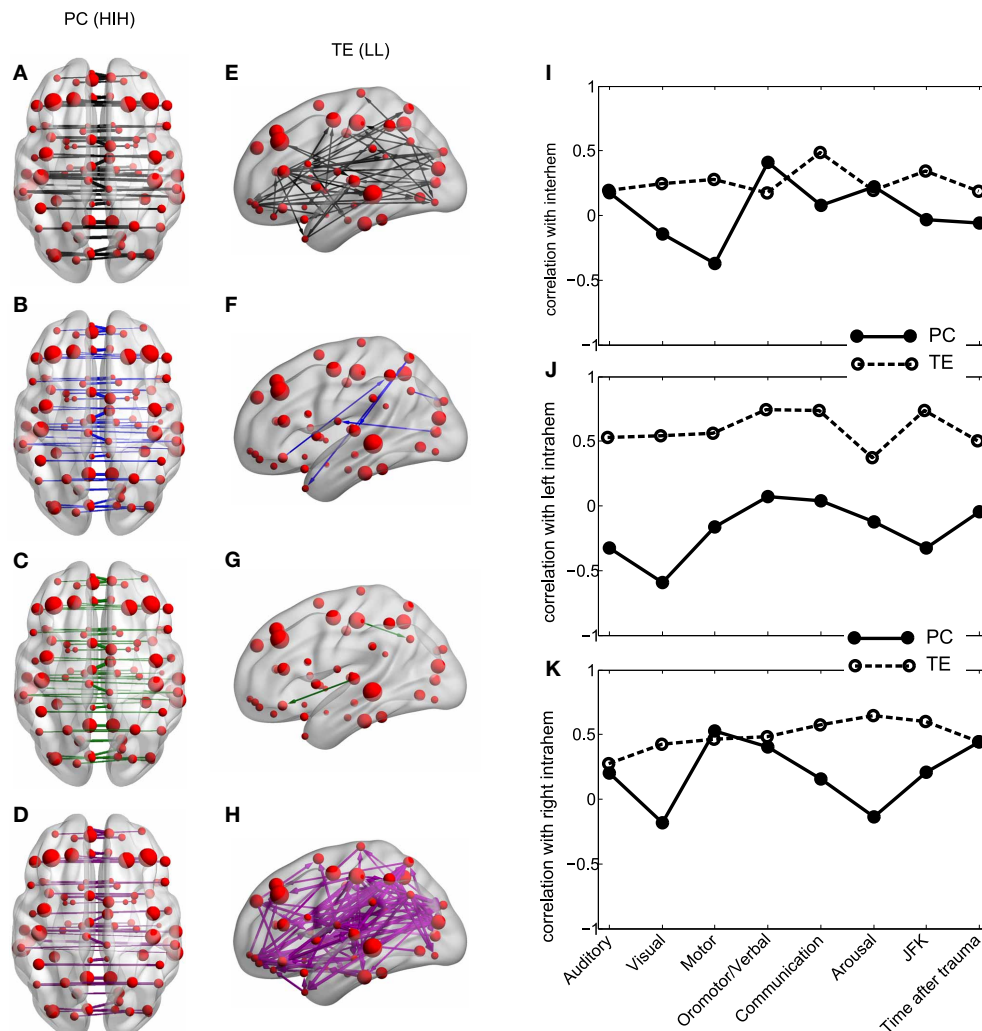
We then asked if the two fMRI measures, between-homologue inter-hemispheric PC and left intra-hemispheric TE were correlated with the neurological and behavioral scale given by the CRS-S. This is represented in Figures 4I–K. For homologue inter-hemispheric pairs we found that TE gave the biggest correlation with the corresponding value in the communication function scale. For left intra-hemispheric pairs, TE had 0.73 correlations with oromotor/verbal function scale, 0.73 with the

communication function scale and 0.73 with the total CRS-R (marked as “JFK” in Figures 4I–K).

## 4. DISCUSSION

In this study we have investigated whether the functional connectivity is altered as a consequence of consciousness disturbances. We have applied the PC and the TE approaches to analyze the FC from resting-state fMRI data. We have compared two groups, healthy subjects and Disorder of Consciousness patients. The analysis was done over the 90 anatomical brain areas, defining regions of interest from the AAL atlas. We have grouped the different pairs of ROIs in inter-hemispheric homologue regions, inter-hemispheric, left intra-hemispheric, right intra-hemispheric and total (all regions). We have found two particular markers that account for the large-scale disturbance of patients brain function: the PC calculated over homologue inter-hemispheric (HIH) regions and the TE calculated over the left intra-hemispheric (LL) ROIs.

The PC in HIH regions was found to be notably larger for control compared to DOC patients. This results holds also when comparing G1 with the recovered G2b group. The same comparison but done over the total average of the 90 regions did



**FIGURE 4 | Inter-hemispheric PC and left intra-hemispheric TE. (A–H)** PC and TE values for all the 4 different groups. The thickness of links and arrows are proportional to the PC and TE values; the thickness normalization factor is common among all the 4 groups. **(A,E)** group G1, black, **(B,F)** group G2, blue, **(C,G)** group G2a, green, **(D,H)** group G2b, magenta. **(A–D)** Visualization of the PC values HIH (homologue inter-hemispheric pairs). **(E–H)** TE in LL (left intra-hemispheric pairs). For clarity in the visualization, links have been thresholded and only TE

values bigger than  $TE = 0.2$  are depicted. **(I–K)** Correlation between PC (solid line) and TE (dashed) with the CRS-R scores at the different functional scales: Auditory, Visual, Motor, Oromotor/Verbal, Communication, Arousal and the total sum over all the function scales (JFK) as well as with the acquisition time after trauma. The correlation has been calculated over pairs which are **(I)** inter-hemispheric (HIH for PC and  $(HLR+HRL)/2$  for TE, **(J)** left intra-hemispheric (LL) and **(K)** right intra-hemispheric (RR).

not shown significant differences. Thus, one relevant result of our analysis is the finding that only by the calculation of the PC in the proposed grouping of brain regions, it was possible to detect a significant marker for the patients disturbance, results that is hidden when we looked at the PC of the total AAL brain regions.

In the case of TE, the total score did not show any significant difference either, but the brain subdivision revealed that the intra-hemispheric influences were different in control respect DOC. This happened for both LL and RR, although the TE in LL discriminated better than in RR. This is a very novel finding whose origin is still unclear and deserves further investigation.

#### 4.1. METHODOLOGICAL ISSUES

The PC is a straightforward measure able to eliminate for each specific ROIs pair, the contribution to the correlations coming from common neighbors, preserving *effective* correlations between two time series. Unlike the PC which is a symmetrical measure, the TE quantifies interaction between ROIs in a directed form, i.e., region A influences to region B but the opposite is not necessary true. In concrete, TE quantifies information bits (uncertainty reduction) flowing from one ROI to the future of the other. For the case of Gaussian data, the information bits measured by the TE coincide with the Granger causality measured from time series (Barnett et al., 2009); however for non Gaussian data, TE and causality might result in different measures.

TE emerges as a very suitable measure for the study of temporal causality in brain fMRI activity in parallel to the advantage of an accurate spatial resolution. TE assessment in a population of patients with disorder of consciousness provides the opportunity of gaining insight into brain mechanisms of information processing and the finding of possible predictors of coma outcome.

Regarding to the calculation of TE, it is well-known that the computation of the entropies with small data sets introduces some a bias (Panzeri and Treves, 1996; Paninski, 2003; Bonachela et al., 2008). Because we are performing groups comparison with the same data size in each group (i.e., the time series in each subject have the same data points), such a bias will be the same in the two groups, thus not affecting the validity of the groups comparison. Nevertheless, as far as we understand there is not any reported study analyzing either information reduction (i.e., TE) or causality in fMRI data from DOC patients.

## 4.2. INTER-RELATION BETWEEN PC AND TE IN DOC PATIENTS

To exhibit high correlations is different from having high TE between two time series. This can be clearly understood by a counter-example; two fully correlated time series have zero TE as to compute the uncertainty reduction in the future of  $i$ , conditioning on the two pasts  $i$  and  $j$  is not adding any further information to the situation of solely adding the past of  $i$ , i.e., the two terms in the right-hand side in Equation (2) are equal. As a consequence of this, the observation of having high PC for HIH pairs in healthy subjects implies to have high isolation of the information within hemispheres; thus, the TE values in both LL and RR are significantly higher than the corresponding values in HLR and HRL.

Interestingly, we found that while PC is reduced in DOC patients between inter-hemispheric homologue areas, TE shows an altered pattern at the level of general inter-hemispheric interactions. In the control group we observe that despite the coherence is high between homologue areas, their TE is low. Conversely, while PC between hemispheres is low, LR and RL TE are high. The DOC patients show the same trend, although the LR and RL TE is significantly lower than in controls. This supports the notion that consciousness arises from long-range modulation of neural activity. A disruption in long-range communication could affect mechanisms such as increase of stimulus' salience, facilitation of propagation across sparsely connected networks, and selective routing (Ganzetti and Mantini, 2013), mechanisms that are related to conscious processing (Gaillard et al., 2009).

## 4.3. rs-fMRI INTER-HEMISPHERIC CORRELATIONS AND GAMMA RHYTHMS

Recently it has been shown that the inter-hemispheric correlations in the rs-fMRI dynamics correlate with the inter hemispheric coherence exhibited by electrophysiological recordings in human sensory cortex (Nir et al., 2008), mainly with the slow modulation of the gamma rhythms in Local Field Potentials. Other studies have also found such modulation in high-level cognition tasks (Vidal et al., 2012). Thus, one could conjecture that at the functional level, a breakdown in the inter-hemispheric rs-fMRI correlations in DOC patients could be an indication of a

similar deficit in the gamma power coherence. One possibility is that low-frequency oscillatory activity is related to an underlying neuronal mechanism allowing for maintenance and consolidation of neural events across wide sections of the brain, and for the handling of incoming stimuli (de Pasquale, 2010; de Pasquale et al., 2012). Although increasing evidence points toward a property of the brain relevant for conscious processing, Vidal et al. (2012) point out that gamma-amplitude correlation would also be reflecting the parallel organization of the brain, where neural networks interact for purposeful processing of information.

## 4.4. COMPARISON WITH PREVIOUS RESULTS

As far as we know, a single study have reported that DOC patients in comparison with healthy subjects manifest a strong reduction in the inter-hemispheric correlations in the rs-fMRI time series (Ovadia-Caro et al., 2012). The authors in (Ovadia-Caro et al., 2012) did not use any atlas to compute inter-hemispheric correlations; instead they investigated specific areas such as pre- and post-central gyrus and the intra-parietal sulcus. Among other reasons, the authors selected those areas for being well separated each from the other (arguing the existence of less noise in the signal). This is consistent with our finding that DOC patients kept more similar correlations to control for ROIs separation below 20 mm. In addition to this, our study adds the novelty of having analyzed the FC obtained by the TE.

## 4.5. TE DENSITY TO MEASURE CONSCIOUSNESS ALTERATION

We have shown in Figures 4E–H how the number of TE connections can account not only for the differences between control (G1) and DOC (G2) but for the transitory brain state in the group G2b: the patients that awaked and became fully conscious at the second fMRI acquisition. Thus, we have found that the number of TE connections were 47 (G1), 5 (G2) and 99 (G2b). In a similar spirit, Seth et al. (2006) defined the causal density for measuring consciousness in brain states as the number of Granger-causality connections flowing in and out per each specific area. Interestingly, a similar behavior has been reported during recovery from anesthesia, where an increment in functional connectivity above the normal wakeful baseline is found (Hudetz, 2012).

## 4.6. DOC IMPAIRMENT AT SPECIFIC BRAIN AREAS

The aim of this analysis is not to work at the level of an individual DOC patient but to search for rs-fMRI markers that can account for groups differences in DOC patients. We have not studied yet any measure that can account for DOC impairment at specific brain areas. To this end, one could study in principle the FC graphs obtained by either PC or TE using complex networks analysis, or any other kind of graph exploration methods. In a much simpler spirit (just to illustrate that this approach is plausible), we have chosen to plot the PC values per area comparing group G2 versus G1. This is illustrated in the Figure S1. The decorrelation index per area is plotted,  $(\text{corrG1} - \text{corrG2}) / \text{corrG1}$ . Colored in blue, the five biggest decorrelation indices correspond to the following areas: Fusiform, Insula, Parietal Superior, Precentral and Temporal Superior, revealing that those areas had the major DOC impairment. Conversely the areas with less DOC impairment



(colored in red) were the Cingulum Anterior, Cingulum Middle, Frontal Superior Orbital, Superior Motor Area and Temporal Inferior.

#### 4.7. LIMITATIONS OF THE STUDY

One of the important limitation of studying DOC patients is the great amount of involuntary movements they exhibit, leading to potential artifacts in the fMRI acquisition. Techniques to overcome this issue include affine transformations to the time series creating a head-motion parameter matrix which can be used to regress out and remove the spurious variances they introduce (Fox et al., 2005). Although these methods can correct signals from movements spanning the dimensions of up to 3–4 voxels, recent work (Power et al., 2012) suggest that no technique could remove completely the effects of these artifacts over the FC. Thus especial care is necessary to tackle these problems and, eventually, discard the entire scan.

#### 4.8. FUTURE DIRECTIONS

In this study PC and TE measures were used to assess for the assessment of functional connectivity in unconscious patients. In particular we characterized their disruptions at an anatomical level, in the basis of distances between homotopic areas. Other questions that can be explored, include the integrity of FC between the areas that constitute hubs in the brain network, between areas with high *rich-clubness* (van den Heuvel and Sporns, 2011), or between associative vs. sensory areas.

#### FUNDING

Jesus M. Cortes is supported by Ikerbasque: The Basque Foundation for Science. Jesus M. Cortes acknowledges financial support from Junta de Andalucía, grant P09-FQM-4682. Verónica Mäki-Marttunen, Mirta Villarreal, and Dante R. Chialvo are partially supported by CONICET (National Council of Scientific and Technological Research) of Argentina. Additional support was provided by the Department of Neurology, and Department of Teaching and Research of FLENI, Buenos Aires, Argentina.

#### SUPPLEMENTARY MATERIAL

The Supplementary Material for this article can be found online at: <http://www.frontiersin.org/journal/10.3389/fninf.2013.00024/abstract>

#### Figure S1 | DOC impairment evaluated at specific brain areas. (A)

Decorrelation indices defined as  $(\text{corrG1}-\text{corrG2})/\text{corrG1}$  computed for each of the different 45 brain areas. In blue, top-five values of decorrelation index; in red, bottom five (positive) values. (B) Scatter of between-homologue inter-hemispheric correlations G1 vs. G2, each point represents one of the 45 brain areas.

#### REFERENCES

- Ashburner, J., and Friston, K. (1999). Nonlinear spatial normalization using basis functions. *Hum. Brain Mapp.* 7, 254–266. doi: 10.1002/(SICI)1097-0193(1999)7:4<254::AID-HBM4>3.0.CO;2-G
- Barnett, L., Barrett, A., and Seth, A. (2009). Granger causality and transfer entropy are equivalent for Gaussian variables. *Phys. Rev. Lett.* 103, 238701. doi: 10.1103/PhysRevLett.103.238701
- Beckmann, C., DeLuca, M., Devlin, J., and Smith, S. (2005). Investigations into resting-state connectivity using independent component analysis. *Philos. Trans. R. Soc. Lond. B Biol. Sci.* 360, 1001–1013. doi: 10.1098/rstb.2005.1634
- Bonachela, J., Hinrichsen, H., and noz, M. M. (2008). Entropy estimates of small data sets. *J. Phys. A Math. Theor.* 41, 202001. doi: 10.1088/1751-8113/41/20/202001
- Boveroux, P., Vanhaudenhuyse, A., Bruno, M. A., Noirhomme, Q., Lauwick, S., Luxen, A., et al. (2010). Breakdown of within- and between-network resting state functional magnetic resonance imaging connectivity. *Anesthesiology* 113, 1038–1053. doi: 10.1097/ALN.0b013e3181f697f5
- Cauda, F., Micon, B., Sacco, K., Duca, S., D'Agata, F., Geminiani, G., et al. (2009). Disrupted intrinsic functional connectivity in the vegetative state. *J. Neurol. Neurosurg. Psychiatry* 80, 429–431. doi: 10.1136/jnnp.2007.142349
- Cover, T., and Thomas, J. (2006). *Elements of Information Theory*. New York, NY: John Wiley & Sons, Inc.
- de Pasquale, F., Penna, S. D., Snyder, A., Marzetti, L., Pizzella, V., Romani, G., et al. (2012). A cortical core for dynamic integration of functional networks in the resting human brain. *Neuron* 74, 753–764.
- de Pasquale, F., Della Penna, S., Snyder, A. Z., Lewis, C., Mantini, D., Marzetti, L., et al. (2010). Temporal dynamics of spontaneous meg activity in brain networks. *Proc. Natl. Acad. Sci. U.S.A.* 107, 6040–6045. doi: 10.1073/pnas.0913863107
- Fox, M., Snyder, A., Vincent, J., Corbetta, M., Essen, D. V., and Raichle, M. (2005). The human brain is intrinsically organized into dynamic, anticorrelated functional networks. *Proc. Natl. Acad. Sci. U.S.A.* 102, 9673–9678. doi: 10.1073/pnas.0504136102
- Gaillard, R., Dehaene, S., Adam, C., Clemenceau, S., Hasboun, D., Baulac, M., et al. (2009). Converging intracranial markers of conscious access. *PLoS Biol.* 7:e1000061. doi: 10.1371/journal.pbio.1000061
- Ganzetti, M., and Mantini, D. (2013). Functional connectivity and oscillatory activity in the resting human brain. *Neuroscience* 240, 297–309. doi: 10.1016/j.neuroscience.2013.02.032
- Giacino, J., Kalmar, K., and Whyte, J. (2004). The JFK coma recovery scale-revised: measurement characteristics and diagnostic utility. *Arch. Phys. Med. Rehabil.* 85, 2020–2029. doi: 10.1016/j.apmr.2004.02.033
- Heine, L., Soddu, A., Gomez, F., Vanhaudenhuyse, A., Tshibanda, L., Thonnard, M., et al. (2012). Resting state networks and consciousness: alterations of multiple resting state network connectivity in physiological, pharmacological, and pathological consciousness states. *Front. Psychol.* 3:295. doi: 10.3389/fpsyg.2012.00295
- Hudetz, A. (2012). General anesthesia and human brain connectivity. *Brain Connect.* 2, 291–302. doi: 10.1089/brain.2012.0107
- Jaynes, E. (1957). Information theory and statistical mechanics. *Phys. Rev.* 106, 620–630. doi: 10.1103/PhysRev.106.620
- Nir, Y., Mukamel, R., Dinstein, I., Privman, E., Harel, M., Fisch, L., et al. (2008). Interhemispheric correlations of slow spontaneous neuronal fluctuations revealed in human sensory cortex. *Nat. Neurosci.* 11, 1100–1108. doi: 10.1038/nn.2177
- Noirhomme, Q., Soddu, A., Lehenbre, R., Vanhaudenhuyse, A., Boveroux, P., Boly, M., et al. (2010). Brain connectivity in pathological and pharmacological coma. *Front. Syst. Neurosci.* 4:160. doi: 10.3389/fnsys.2010.00160
- Oldfield, R. (1971). The assessment and analysis of handedness: the Edinburgh inventory. *Neuropsychologia* 9, 97–113. doi: 10.1016/0028-3932(71)90067-4
- Ovadia-Caro, S., Nir, Y., Soddu, A., Ramot, M., Hesselmann, G., Vanhaudenhuyse, A., et al. (2012). Reduction in inter-hemispheric connectivity in disorders of consciousness. *PLoS ONE* 7:e37238. doi: 10.1371/journal.pone.0037238
- Paninski, L. (2003). Estimation of entropy and mutual information. *Neural Comp.* 15, 1191–1254. doi: 10.1162/089976603321780272
- Panzeri, S., and Treves, A. (1996). Analytical estimates of limited sampling biases in different information measures. *Netw. Comput. Neural Syst.* 7, 87–107. doi: 10.1088/0954-898X/7/1/006
- Peng, H., *Mutual Information Toolbox*. Available online at: <http://home.penglab.com/software/HanchuanPengSoftware/software.html>
- Perri, C. D., Stender, J., Laureys, S., and Gosseries, O. (2013). Functional neuroanatomy of disorders of consciousness. *Epilepsy Behav.* pii: S1525-5050(13)00480-0. doi: 10.1016/j.yebeh.2013.09.014
- Power, J., Barnes, K., Snyder, A., Schlaggar, B., and Petersen, S. (2012). Spurious but systematic correlations in functional connectivity MRI networks arise from subject motion. *Neuroimage* 59, 2142–2154. doi: 10.1016/j.neuroimage.2011.10.018

- Raichle, M., MacLeod, A., Snyder, A., Powers, W., Gusnard, D., and Shulman, G. (2001). A default mode of brain function. *Proc. Natl. Acad. Sci. U.S.A.* 98, 676–682. doi: 10.1073/pnas.98.2.676
- Rosanova, M., Gosseries, O., Casarotto, S., Boly, M., Casali, A. G., Bruno, M. A., et al. (2012). Recovery of cortical effective connectivity and recovery of consciousness in vegetative patients. *Brain* 135, 1308–1320. doi: 10.1093/brain/awr340
- Schreiber, T. (2000). Measuring information transfer. *Phys. Rev. Lett.*, 85, 461–464. doi: 10.1103/PhysRevLett.85.461
- Seth, A., Izhikevich, E., Reeke, G., and Edelman, G. (2006). Theories and measures of consciousness: an extended framework. *Proc. Natl. Acad. Sci. U.S.A.* 103, 10799–10804. doi: 10.1073/pnas.0604347103
- Teasdale, G., and Jennett, B. (1974). Assessment of coma and impaired consciousness: a practical scale. *Lancet* 2, 81–84. doi: 10.1016/S0140-6736(74)91639-0
- Tononi, G. (2004). An information integration theory of consciousness. *BMC Neurosci.* 5:42. doi: 10.1186/1471-2202-5-42
- Tzourio-Mazoyer, N., Landeau, B., Papathanassiou, D., Crivello, F., Etard, O., Delcroix, N., et al. (2002). Automated anatomical labeling of activations in SPM using a macroscopic anatomical parcellation of the MNI MRI single-subject brain. *Neuroimage* 15, 273–289. doi: 10.1006/nimg.2001.0978
- van den Heuvel, M., and Sporns, O. (2011). Rich-club organization of the human connectome. *J. Neurosci.* 31, 15775–15786. doi: 10.1523/JNEUROSCI.3539-11.2011
- Vidal, J., Freyermuth, S., Jerbi, K., Hamame, C., Ossandon, T., Bertrand, O., et al. (2012). Long-distance amplitude correlations in the high gamma band reveal segregation and integration within the reading network. *J. Neuroscience.* 32, 6421–6434. doi: 10.1523/JNEUROSCI.4363-11.2012

**Conflict of Interest Statement:** The authors declare that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.

Received: 23 August 2013; paper pending published: 23 September 2013; accepted: 10 October 2013; published online: 13 November 2013.

Citation: Mäki-Marttunen V, Diez I, Cortes JM, Chialvo DR and Villarreal M (2013) Disruption of transfer entropy and inter-hemispheric brain functional connectivity in patients with disorder of consciousness. *Front. Neuroinform.* 7:24. doi: 10.3389/fninf.2013.00024

This article was submitted to the journal *Frontiers in Neuroinformatics*.

Copyright © 2013 Mäki-Marttunen, Diez, Cortes, Chialvo and Villarreal. This is an open-access article distributed under the terms of the Creative Commons Attribution License (CC BY). The use, distribution or reproduction in other forums is permitted, provided the original author(s) or licensor are credited and that the original publication in this journal is cited, in accordance with accepted academic practice. No use, distribution or reproduction is permitted which does not comply with these terms.



# Constructing the resting state structural connectome

**Olusola Ajilore<sup>1\*</sup>, Liang Zhan<sup>2</sup>, Johnson GadElkarim<sup>1,3</sup>, Aifeng Zhang<sup>1</sup>, Jamie D. Feusner<sup>4</sup>, Shaolin Yang<sup>1</sup>, Paul M. Thompson<sup>2</sup>, Anand Kumar<sup>1</sup> and Alex Leow<sup>1,5,6</sup>**

<sup>1</sup> Department of Psychiatry, University of Illinois, Chicago, IL, USA

<sup>2</sup> Laboratory of Neuro Imaging, Department of Neurology, University of California, Los Angeles, CA, USA

<sup>3</sup> Department of Electrical and Computer Engineering, University of Illinois, Chicago, IL, USA

<sup>4</sup> UCLA Semel Institute for Neuroscience and Human Behavior, University of California, Los Angeles, CA, USA

<sup>5</sup> Department of Bioengineering, University of Illinois, Chicago, IL, USA

<sup>6</sup> Community Psychiatry Associates, Sacramento, CA, USA

## Edited by:

Daniele Marinazzo, University of Gent, Belgium

## Reviewed by:

Marcel Van Gerven, Donders Institute for Brain, Cognition and Behaviour, Netherlands  
Tianming Liu, University of Georgia, USA

## \*Correspondence:

Olusola Ajilore, Department of Psychiatry, University of Illinois, 1601 W. Taylor Street, Chicago, IL 60612, USA  
e-mail: oajilore@psych.uic.edu

**Background:** Many recent studies have separately investigated functional and white matter (WM) based structural connectivity, yet their relationship remains less understood. In this paper, we proposed the functional-by-structural hierarchical (FSH) mapping to integrate multimodal connectome data from resting state fMRI (rsfMRI) and the whole brain tractography-derived connectome.

**Methods:** FSH first observes that the level of resting-state functional correlation between any two regions in general decreases as the graph distance of the corresponding structural connectivity matrix between them increases. As not all white matter tracts are actively in use (i.e., “utilized”) during resting state, FSH thus models the rsfMRI correlation as an exponential decay function of the graph distance of the rsfMRI-informed structural connectivity or rsSC. rsSC is mathematically computed by multiplying entry-by-entry the tractography-derived structural connectivity matrix with a binary white matter “utilization matrix”  $U$ .  $U$  thus encodes whether any specific WM tract is being utilized during rsfMRI, and is estimated using simulated annealing. We applied this technique and investigated the hierarchical modular structure of rsSC from 7 depressed subjects and 7 age/gender matched controls.

**Results:** No significant group differences were detected in the modular structures of either the resting state functional connectome or the whole brain tractography-derived connectome. By contrast, FSH results revealed significantly different patterns of association in the bilateral posterior cingulate cortex and right precuneus, with the depressed group exhibiting stronger associations among regions instrumental in self-referential operations.

**Discussion:** The results of this study support that enhanced sensitivity can be obtained by integrating multimodal imaging data using FSH, a novel computational technique that may increase power to detect group differences in brain connectomes.

**Keywords:** connectivity, fMRI, major depression, multimodal, neuroimaging

## INTRODUCTION

Modern imaging techniques have allowed us to study the human brain as a complex system by modeling it as a network. A brain connectivity network, also called a connectome (Sporns et al., 2005), consists of nodes (gray matter regions) and edges. Edges can represent white matter tracts in structural networks or correlations between two BOLD time series in functional networks.

In recent years, substantial research efforts have been directed toward understanding the brain at rest using resting state functional MRI (rs-fMRI). Several studies have utilized sophisticated mathematical and statistical tools to investigate the functional connectome from rs-fMRI data (Biswal et al., 1997). The “default mode network” (DMN) is a resting-state network theorized to reflect an individual’s focus on internal tasks such as daydreaming, envisioning the future, retrieving memories, and gauging

others’ perspectives. The DMN tends to negatively correlate with brain systems responsive to external signals. Anatomical regions involved include the medial temporal lobe, the medial prefrontal cortex, and the posterior cingulate cortex (Buckner et al., 2008), along with the adjacent precuneus (Zhang and Li, 2012) and the parietal cortex.

The DMN is an example of one relatively well-characterized network, among many overlapping networks that subserve different functions. Delineating these functional connections may therefore be challenging, based on the complexity of the brain. Structural white matter connectivity patterns, however, may provide a framework for understanding relevant functional relationships between regions in a network, based on direct and indirect anatomical connections. This may aid in determining information available as outputs from

certain regions and its inputs and potential influence on other regions (Saygin et al., 2011). An approach using structural to functional mapping could utilize a combination of DTI-tractography to estimate brain white matter connectivity and fMRI to estimate the neuronal activity coupled to blood flow changes in anatomical regions that comprise nodes of the network.

There have been several structural to functional mapping approaches described in the literature. While some have focused on specific but limited regional activation patterns (Johansen-Berg et al., 2004; Saygin et al., 2011), other models describe functional connections within regions comprising larger networks or systems (Passingham et al., 2002; Honey et al., 2009; Deligianni et al., 2010; Skudlarski et al., 2010; Chulwoo et al., 2011; Varkuti et al., 2011; Ng et al., 2012). Of note, these studies reviewed here all considered structural connectivity to be static, unlike their functional counterparts. However, it is highly unlikely that white matter tracts are static in relation to the brain's different functional states. Indeed, white matter tracts can be in use or engaged when the brain is performing certain tasks but disengaged during other tasks (e.g., the white matter structure subserving the DMN will be relatively disengaged when the brain is responding to external signals). In addition, some of these previously published techniques rely on statistical methods based on linear modeling, however the relationship between structural and functional connectivity may be non-linear (Deligianni et al., 2010). In other studies, sparse Gaussian graphical modeling (SGGM) is used for multimodal integration (Ng et al., 2012). There, the authors proposed to merge functional and tractography-derived structural data by casting functional connectivity estimation as a sparse inverse covariance learning problem. As functional connections with less anatomical support (i.e., fewer streamlines or fiber tracts) were more penalized via an L1 type penalty term, the resulting functional connection patterns could thus be considered structural connectivity-informed.

Here, in contrast to such SGGM models, we reversely consider functional connectivity-informed structural connectivity, thus arguing that information from fMRI can be used to infer the underlying pattern of white matter engagement specific to the brain's state at the moment of the fMRI. To address that not all white matter tracts are in use or engaged during fMRI, we will extend and adapt the functional by structural hierarchical mapping (FSH) technique, a novel framework recently proposed by our group (Leow et al., 2012) in order to estimate white matter engagement or utilization patterns that generate the functional connectome from rs-fMRI data, using structural networks derived from DTI-tractography. The resulting connectome, which we term the resting-state informed structural connectome (rsSC), encodes the structural network that underlies and facilitates the observed rs-fMRI correlation connectome. Moreover, we may detect group differences in rsSC by investigating and comparing their community or modular structures. To this end, we utilized PLACE (path length associated community estimation) (GadElkarim et al., 2013) and detected altered rsSC community structure in depressed subjects relative to controls.

## MATERIALS AND METHODS

### SUBJECT SELECTION

7 healthy comparison (HC, age:  $65.6 \pm 8.12$ , 4 males) and 7 late-life depressed (LLD, age:  $60.7 \pm 2.92$ , 4 males) subjects, were recruited via community outreach (e.g., newspaper, radio, and television advertisements) and relevant outpatient clinics. The inclusion criteria for all subjects were 55 years of age and older, medication-naïve or anti-depressant free for at least 2 weeks (in the case of our depressed subjects) and no history of unstable cardiac or neurological diseases. The exclusion criteria included: schizophrenia, bipolar or any psychotic disorders; history of anxiety disorder outside of major depressive episodes; history of head trauma; history of substance abuse; contraindications to MRI such as metal implants. This study was approved by the University of Illinois-Chicago Institutional Review Board, and written informed consent was obtained from each participant. There were no significant differences in age ( $t = 1.49$ ,  $p = 0.18$ ) and gender distribution ( $\chi^2 = 0$ ,  $p = 1$ ) between subject groups. LLD subjects had a mean HAM-D score of  $20 \pm 3.7$ .

All eligible subjects were assessed by a trained research assistant with the Structured Clinical Interview for Diagnostic and Statistical Manual of Mental Disorders, Fourth Edition (DSM-IV). The severity of depression was quantified by a board-certified/board-eligible psychiatrist (AK or OA) using the 17-item Hamilton Depression Rating Scale (Hamilton, 1960). At the time of enrollment, depressed subjects met DSM-IV criteria for MDD and required a score of 15 or greater on the HAM-D.

### MRI ACQUISITION

Brain MRI were acquired on a Philips 3.0T Achieva scanner (Philips Medical Systems, Best, The Netherlands) using an 8-channel SENSE (Sensitivity Encoding) head coil. Participants were positioned comfortably on the scanner bed and fitted with soft ear plugs; foam pads were used to minimize head movement. Participants were instructed to remain still throughout the scan. High resolution three-dimensional T1-weighted images were acquired with a MPRAGE (Magnetization Prepared Rapid Acquisition Gradient Echo) sequence (field of view: FOV = 240 mm; 134 contiguous axial slices; TR/TE = 8.4/3.9 ms; flip angle =  $8^\circ$ ; voxel size =  $1.1 \times 1.1 \times 1.1$  mm). Resting-state data were acquired with the following parameters: Single-shot gradient-echo EPI sequence, TR/TE = 2000/30 ms, Flip angle =  $80^\circ$ , EPI factor = 47, FOV =  $23 \times 23 \times 15$  cm<sup>3</sup>, in-plane resolution =  $3 \times 3$  mm<sup>2</sup>, slice thickness/gap = 5/0 mm, slice number = 30, SENSE reduction factor = 1.8, NEX = 200, total scan time = 6:52. Subjects were instructed to keep their eyes close and "not think of anything in particular". DTI images were acquired using single-shot spin-echo echo-planar imaging (EPI) sequence (FOV = 240 mm; voxel size =  $0.83 \times 0.83 \times 2.2$  mm; TR/TE = 6,994/71 ms; Flip angle =  $90^\circ$ ). Sixty seven contiguous axial slices aligned to the anterior commissure–posterior commissure (AC-PC) line were collected in 32 gradient directions with  $b = 700$  s/mm<sup>2</sup> and one acquisition without diffusion sensitization (B0 image). Parallel imaging technique was utilized with factor at 2.5 to reduce scanning time to approximately 4 min.



## DATA PREPROCESSING

Structural connectomes were generated using a pipeline which integrates multiple image analysis techniques and has been reported elsewhere (GadElkarim et al., 2012; Leow et al., 2012). In brief, DW images were eddy current corrected using the automatic image registration (AIR) tool embedded in DTIStudio software (<http://www.mristudio.org>) by registering all DW images to their corresponding b0 images with 12-parameter affine transformations. This was followed by the computation of diffusion tensors and deterministic tractography using the DTIStudio program. T1-weighted images were used to generate label maps using the Freesurfer software (<http://surfer.nmr.mgh.harvard.edu>). Brain networks formed by the 82 cortical/subcortical gray matter regions were generated using an in-house program in Matlab by counting the number of fibers connecting each pair of nodes.

Functional connectomes were generated using the resting-state fMRI toolbox, CONN (<http://www.nitrc.org/projects/conn>; Whitfield-Gabrieli and Nieto-Castanon, 2012). In brief, raw EPI images were realigned, co-registered, normalized, and smoothed before analyses. Confound effects from motion artifact, white matter, and CSF were regressed out of the signal. Using the same 82 labels as the structural brain networks, functional brain networks were derived using pairwise BOLD signal correlations, which were then converted to z scores using Fisher's r-to-z transformation.

## FUNCTIONAL BY STRUCTURAL HIERARCHICAL (FSH) MAPPING FOR CONSTRUCTING rsSC

Several assumptions and simplifications are needed in order to perform FSH mapping (Leow et al., 2013). However, in order to generalize FSH to construct rsSC, several modifications are necessary, which we outlined step-by-step as follows:

- (1) Higher level of rs-fMRI correlations will be considered evidence of strong structural interactions between two regions (either through direct or indirect structural connections in the corresponding DTI-derived structural network)
- (2) We observe that in general the level of rs-fMRI correlation between two regions decreases as the graph distance of the DTI-derived structural connectivity matrix increases between them. FSH further assumes that such a relationship is mathematically an exponential decay:

$$\text{level of rsfMRI correlation between } i \text{ and } j \approx e^{-kf_{i,j}(D)} \quad (1)$$

Here,  $k$  a rate constant to be estimated,  $D$  the DTI-derived structural connectivity matrix for the same subject, and functional  $f$  denotes the mapping of a brain connectivity matrix to its graph distance matrix (i.e., each entry denotes the shortest graph distance between node pairs). Here,  $f$  is numerically obtained by applying the Dijkstra algorithm to the entry-wise inverse of  $D$  (since stronger structural connectivity translates to shorter distance, edge lengths are then usually assumed to be the inverse of connectivity strengths).

- (3) As in the original formulation of FSH, the presence of an edge connecting any node pair in the structural connectivity matrix predicts the existence of neuroanatomical white matter connections between regions, which may or may not be actively utilized when the brain is in the resting state. In order to reduce the mathematical complexity in modeling and parameter fitting, FSH assumes an all-or-nothing edge utilization (i.e., an edge is either utilized or not at all). A connection between node  $m$  and  $n$  is considered "utilized" if including the anatomical connection between them better predicts the overall resting state fMRI correlation. This is thus mathematically represented by a binary utilization matrix  $U$  (i.e., if  $U_{(i,j)} = 1$ , then the WM structural connection between nodes  $i$  and  $j$  are utilized in the resting state; zero otherwise)
- (4) FSH now hypothesizes that a direct mathematical relationship can be established, for each node pair, between the level of rs-fMRI correlation and the modulated graph distance between the two nodes for the DTI-derived structural network according to the utilization matrix, via the following modified exponential decay equation:

$$\text{level of rsfMRI correlation between } i \text{ and } j = e^{-kf_{i,j}(U \circ D)} + \varepsilon \quad (2)$$

Here  $U$  is the utilization matrix (same dimension is  $D$ ),  $\circ$  the Hadamard entry-by-entry multiplication operator between two matrices of the same dimensions,  $\varepsilon$  the fitting error (assumed to be normally distributed).

Note the above exponential functional dictates that rsfMRI correlations exponentially decay with increasing modulated graph distance, and that when the modulated graph distance between two nodes approaches infinity (i.e., the nodes are far away from each other), the corresponding rsfMRI correlation as expected approaches zero (by contrast, if two nodes are infinitesimally close, the rsfMRI correlation is 1).

- (5) For subjects in the same diagnostic group, we fit on the group level, by minimizing the sum of squared differences between the observed and the predicted rsfMRI correlations for all node pairs (both are z-transformed), such that the group utilization matrix  $U$  is assumed to capture certain characteristics unique to this group. To this end, the resting state informed structural connectome is mathematically  $U \circ D$ . Mathematically, the minimization problem for solving group-wise  $U$  is as follows (the superscript  $n$  denotes subjects in the same diagnostic group; in this study  $n$  ranges from 1–7).

$$U = \operatorname{argmin}_n \sum_{i,j} \left( \left| \text{rsfMRI correlation}_{i,j}^n \right| - e^{-kf_{i,j}(U \circ D^n)} \right)^2 \quad (3)$$

- (6) To solve  $k$  (unique to each subject) and the utilization  $U$  (shared for each diagnostic group), we closely follow the original FSH formulation by alternating between the estimation of  $k$  and  $U$ . When fitting  $U$ , we used simulated annealing by randomly picking one element in  $U$  and changing its value

(between 0 and 1; the initial value of  $U$  was set to one for all its entries, thus indicating all edges were utilized and  $U \circ D$  simply returned the original structural connectivity matrix  $D$ ). The acceptance criterion determined whether the new state was accepted from the current state by applying the following decision rule with respect to an artificial cooling temperature ( $c$ ).

$$\text{probability of accepting a proposed new state} = \begin{cases} 1 & \text{if the new state yields a lower fitting residual} \\ \exp\left(-\frac{\text{fitting residual increase}}{c}\right) & \text{if the new state yields a higher fitting residual} \end{cases} \quad (4)$$

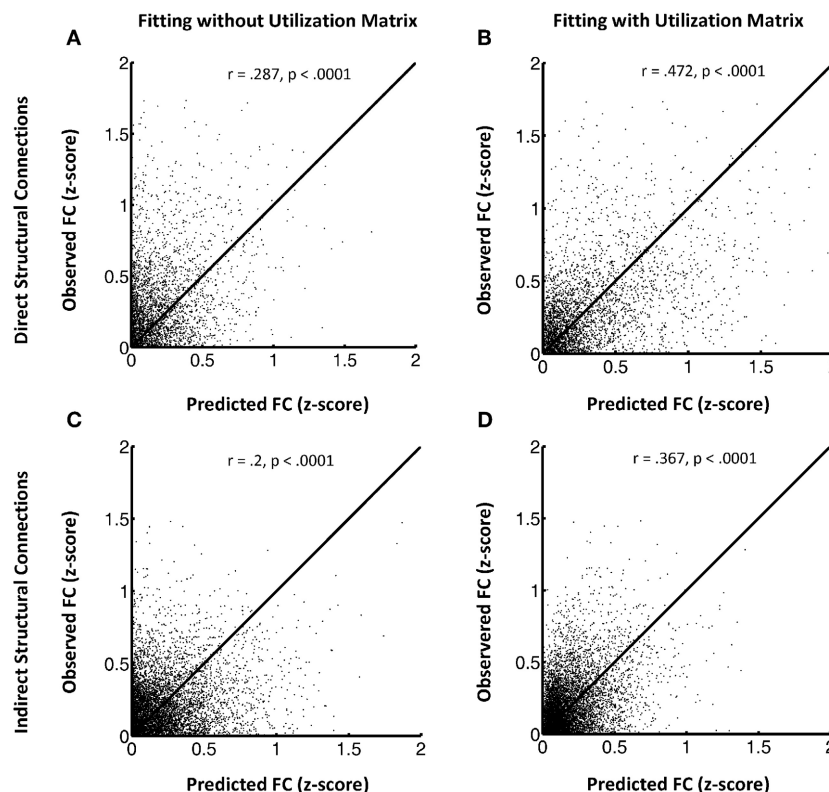
This perturbation was repeated and the temperature gradually decreased until the solution space was adequately sampled and the global minimum reached.

To assess the goodness of fit of FSH mapping, we calculated the correlation between the observed rs-fMRI z-scores and the predicted rs-fMRI z-scores according to the exponential decay function, both without (**Equation 1**) and with the utilization matrix (**Equation 2**). The effect of fitting utilization matrix was then tested by comparing the two groups of correlation coefficients using the Fisher's r-to-z transform.

## MODULAR STRUCTURE USING PLACE (PATH LENGTH ASSOCIATED COMMUNITY ESTIMATION)

After FSH mapping, we constructed the rsSC separately for the depressed and the control group, by forming the product  $U \circ D$  using group-specific utilization matrix and group-average structural connectivity matrix. We then used the PLACE (path

length associated community estimation) framework presented in (GadElkarim et al., 2012, 2013) to assess potential group differences for structural connectome (DTI-derived) alone, the functional connectome (rs-fMRI-derived) alone, and the resting state informed structural connectome. PLACE is a novel technique designed to detect and compare hierarchical modular or community structure alterations between two groups of brain networks based on shortest path lengths, and has been shown to be advantageous when compared to the



**FIGURE 1 | (A–D)** This shows the FSH mapping results for all node pairs, collected from all subjects in the HC group for region pairs with direct structural connections (**A and B**) versus those without direct structural connections (**C and D**). Left panels display the model fitting without the

utilization matrix  $U$  and the right panels show fitting with the utilization matrix in the proposed exponential decay model. The y axis indicates observed resting state fMRI correlation values and the x axis the predicted resting state fMRI correlation values.

modularity metric  $Q$  (Newman and Girvan, 2004; Blondel et al., 2008).

To summarize, in PLACE community structures are first extracted in the form of top-down hierarchical binary trees via the maximization of a path-length dependent metric  $\Psi^{PL}$ , defined as the difference between the average inter-community path-length ( $\text{inter}_{PL}$ ) and the average intra-community path-length ( $\text{intra}_{PL}$ ), for two communities  $C_1$  and  $C_2$ ,  $\Psi^{PL}$  is mathematically defined as:

$$\Psi^{PL} = \text{inter}_{PL}^{C_1, C_2} - \frac{1}{2} \left( \text{intra}_{PL}^{C_1} + \text{intra}_{PL}^{C_2} \right) \quad (5)$$

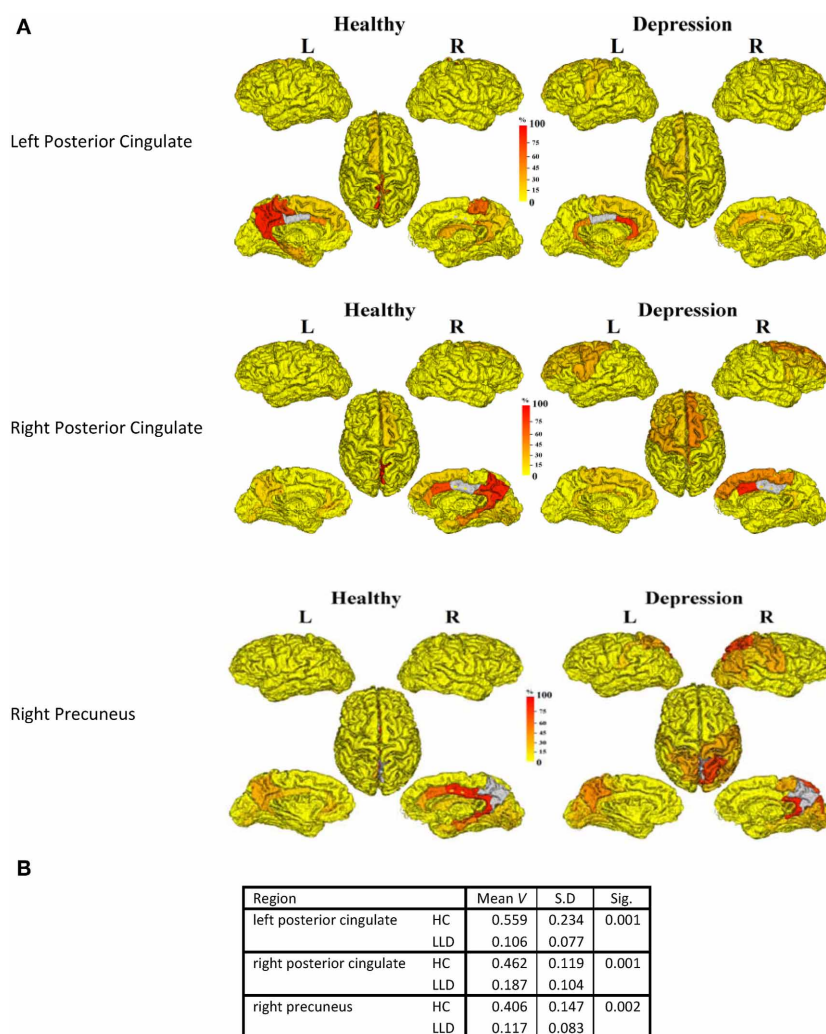
Where

$$\text{inter}_{PL}^{C_i, C_j} = \frac{\sum_{n, m \in C_i; m \in C_j} d_{nm}}{N_i N_j} \quad \text{intra}_{PL}^{C_i} = \frac{\sum_{n, m \in C_i; n > m} d_{nm}}{(N_i^2 - N_i) / 2} \quad (6)$$

where  $N_i$  is the number of nodes in community  $C_i$ ,  $d_{nm}$  is the shortest path length (i.e., graph distance) connecting nodes  $n$  and  $m$ .

To quantify node-level community differences, PLACE uses the *scaled inclusivity* metric  $V$  (Steen et al., 2011) in which a nodal consistency vector of length equal to the number of nodes in the network (82 in our case) is generated to compare nodes in a *test* tree (i.e., an individual subject's tree) to nodes in a *reference* tree. Mathematically, for each node  $k$  belonging to communities  $C_p$  and  $C_q$  in the *test* and *reference* trees respectively,  $V$  is defined as;  $V(k) = (N_c)^2 / N_p N_q$ , where  $N_c$  is the number of common nodes between  $C_p$  and  $C_q$ .

In order to examine group differences in community structures at the nodal level, one group of networks is chosen as the reference and PLACE generates the *reference* tree by extracting the community structure corresponding to the reference group's mean connectivity matrices (using node-wise averaging). Next,



**FIGURE 2 | (A)** The following three regions (in gray) exhibit significant group differences after FDR correction: 1. Left posterior cingulate. 2. Right posterior cingulate. 3. Right precuneus. The frequency of shared community membership

for these regions in HC and LLD. 100% indicates all seven subjects from the same diagnostic group have this region assigned to the same community as the gray region. **2B:** The mean consistency values ( $V$ ) for each region.

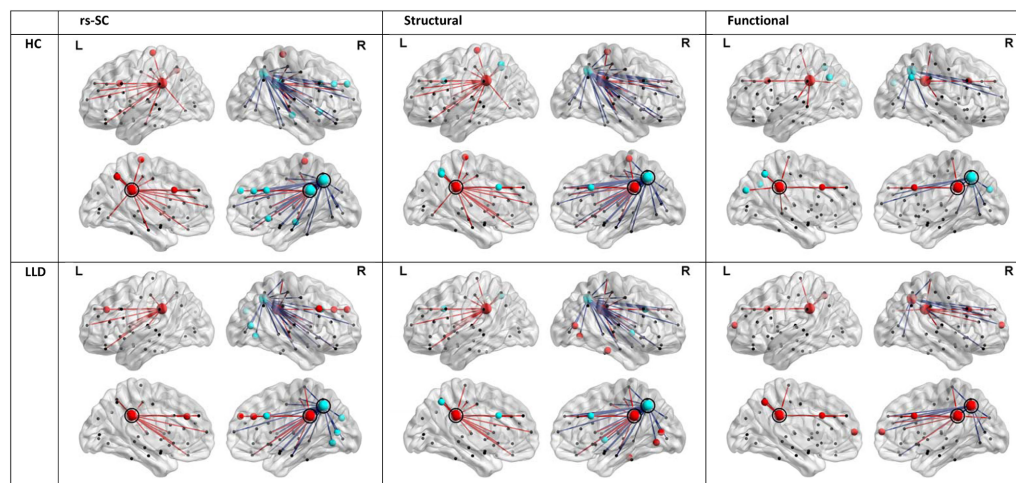
all individual subjects' trees are compared to the *reference* tree, yielding the node-level scaled inclusivity metric,  $V$ . For each node, 2-sample  $t$ -tests for  $V$  are then used to detect differences in the community structure on the nodal level (relative to the reference group), followed by multiple comparisons correction conducted using the false discovery rate (FDR) (Benjamini and Hochberg, 1995).

## RESULTS

**Figure 1** shows the FSH mapping results, which confirmed improved fitting with the additional inclusion of the utilization matrix  $U$  in the exponential decay function, in the HC group (**Figures 1A–D**) by plotting the observed rsfMRI correlations values against the predicted values. Here, we break down the results for HC subjects into two groups: region pairs with direct structural connections between them (**1A and 1B**) versus pairs without direct structural connections (i.e., only indirect structural connections; **1C and 1D**). Overall, the proposed FSH-exponential decay model significantly improved the correlation between rsfMRI and structural connectivity as all data points moved toward the line  $x = y$  after fitting (direct:  $z = -8.7$ ,  $p < 0.0001$ ; indirect:  $z = -10.3$ ,  $p < 0.0001$ ). The pattern also occurred for LLD subjects (direct:  $r = 0.246$  (without fitting  $U$ ),  $r = 0.509$  (with  $U$ ),  $z = -12.3$ ,  $p < 0.0001$ ; indirect:  $r = 0.188$  (without fitting  $U$ ),  $r = 0.267$  (with  $U$ ),  $z = -4.8$ ,  $p < 0.0001$ ). Unsurprisingly, node pairs with direct structural connections exhibited stronger associations with rsfMRI correlations compared to those with only indirect connections.

In order to understand the implications of utilization differences between groups, we examined the global modular structure of the rsSC. To this end, we compared the community structure of the rsSC between groups by applying PLACE (with the mean rsSC of the HC group as the reference tree). We also applied PLACE to connectomes derived using data from a single imaging modality of either DTI alone or rs-fMRI alone. For functional connectome PLACE results, we followed the technique of (Schwarz and McGonigle, 2011) and analyzed the functional networks formed by positive (right-tail), negative (left-tail), and absolute correlation strengths across a range of thresholds (in increments of 0.05 until one or more of the functional networks become disconnected). Results revealed that there were no significant differences in modular structure when examining connectomes from a single modality. By contrast, after applying FDR correction (with a total of 82 comparisons), rsSC community structure was significantly altered for three regions with reduced consistency in LLD subjects: bilateral posterior cingulate, and the right precuneus (**Figure 2**). Visually, the bilateral posterior cingulate was more affiliated with posterior regions such as the precuneus in HC subjects, whereas in LLD bilateral posterior cingulate was more commonly associated with the anterior cingulate. With the right precuneus, results demonstrate a strong association with a limbic lobe module in HC subjects and a parietal lobe module in LLD subjects. **Figure 3** visualizes the differential patterns of community affiliation and connectivity in the bilateral posterior cingulate and the right precuneus.

To determine whether standard community detection methods could yield similar results, we applied the modularity metric,



**FIGURE 3 | Communities and connectivities for the resting-state structural connectome (rsSC), structural connectome, and functional connectome in healthy control (HC) and late-life depressed subjects (LLD), for nodes exhibiting significant group differences in the modular structure of rsSC shown in Figure 2.** The left posterior cingulate (circled in panels indicated “L”), right posterior cingulate (circled in panels indicated “R”) and the right precuneus (caudal and posterior to the right posterior cingulate in panels indicated “R”, also circled). For each diagnostic group, nodes that are coded the same color (either red or blue) form a community or module in the average tree for that group (computed by applying PLACE to the edge-wise average of all subjects’ connectivity

matrices in the same group, see methods section). Edges linked to the bilateral posterior cingulate are indicated in red, while edges linked to the right precuneus are in blue. For the functional connectome, edges were thresholded for the level of correlation  $> 0.25$ . Of note, only the rsSC demonstrated significant differences in community structure. Visually, the pattern of associations in the rsSC are similar to those in **Figure 2A** for the left and right posterior cingulate (in that for HC there is a stronger association with ipsilateral precuneus), and for the right precuneus (in LLD there is a stronger association with occipital and posterior parietal cortices, consistent with a pattern of dorsal and anterior precuneus functional connectivity; also see discussion section).



Q to our sample. Again, there was no significant difference between groups using only structural or functional connectomes. However, there was a difference in community membership of the right fusiform gyrus using the rsSC with a significantly reduced  $V$  in LLD subjects compared to HC subjects (HC:  $0.801 \pm 0.129$ , LLD:  $0.135 \pm 0.209$ ,  $p < 0.0001$ ).

## DISCUSSION

In this study, we adapted the recently-developed FSH mapping to construct the rsSC by projecting rsfMRI time series correlation information onto the whole brain tractography-derived structural connectome. To this end, we assumed that the rsfMRI correlation exhibits an exponential decay, subject to a rate constant, with respect to the “modulated” graph distance of the structural connectivity matrix. This allowed us to compute, as in the original FSH framework, a utilization matrix in order to determine whether the inclusion of a specific structural connection better explains the relationship between rsfMRI and the structural connectivity.

As expected, including the utilization matrix significantly increased the goodness of fit of the exponential decay model in both HC and LLD subjects. Network community structure of the rsSC using PLACE was altered in LLD subjects, particularly for regions associated with the posterior DMN comprising part of the limbic lobe and sub-regions of the parietal lobe. It is important to note that in contrast to our results with the rsSC, applying PLACE to the structural connectome or the functional connectome alone failed to yield any significant group differences (the same conclusion holds even when we used more conventional community detection methods, e.g., maximizing the  $Q$  modularity). This is suggestive of enhanced sensitivity to network modular structure differences in the integrated rsSC compared to connectomes derived from a single imaging modality.

The rsSC demonstrated altered community structure in a sub-network of nodes that belong to the posterior DMN and the limbic lobe. These nodes are notable for being associated with altered structural and functional connectivity in depression. The posterior cingulate is a part of the DMN which has been shown to be altered in depression (Greicius et al., 2007; Sheline et al., 2009), while as part of the posterior medial parietal cortex, the precuneus in recent years has been shown to play a central role in wide-ranging tasks including visuospatial imagery, episodic memory retrieval, and self-referential operations. Current evidence from functional studies supports a functional partition of the precuneus into an anterior division responsible for self-referential imagery, and a posterior division related to episodic memory retrieval (Cavanna and Trimble, 2006). Recent structural brain network studies using whole-brain tractography have also consistently established the precuneus as one of the many “hub” regions in the brain (i.e., regions with the most wide-spread connections to the rest of the brain; hub regions usually exhibit high degree centrality, i.e., serving as relay centers for information transfer across the brain) (van den Heuvel and Sporns, 2011; GadElkarim et al., 2012). Tracer studies in recent years have also established cortical connections between the precuneus and the frontal, the medial parietal, and the lateral parietal cortices (Cavanna and Trimble, 2006).

The findings of higher-degree associations between the precuneus and the lateral parietal cortex, and to some extent the sensorimotor regions have two main parallels with the known literature. First, the medial parietal cortex (including the precuneus) and lateral parietal cortex (especially the inferior parietal lobule) along with the medial prefrontal cortex have been shown to be primary regions activated during first person perspective tasks (Cavanna and Trimble, 2006). Secondly, a recent resting-state functional connectivity study of the precuneus has demonstrated a transitioning pattern of functional connectivity from the posterior and most ventral part of precuneus (greater connectivity with the medial superior frontal gyrus, orbitofrontal gyrus, anterior cingulate cortex, and parahippocampus) to the more dorsal and anterior part (greater connectivity with occipital and posterior parietal cortices and somatomotor cortex, among other regions) (Zhang and Li, 2012). Thus our observed rsSC community structure group differences suggest a pattern of posterior/ventral precuneus connectivity in the control group vs. a pattern of anterior/dorsal precuneus connectivity in the depressed group.

To conclude, using our novel multimodal integration technique FSH combined with PLACE, we detected a differential pattern of functional-structural connectome integration in late-life depressed subjects relative to controls.

## ACKNOWLEDGMENTS

This work was supported by the National Institute of Mental Health (R01 MH-073989 to Anand Kumar; K23 MH-081175 to Olusola Ajilore).

## REFERENCES

- Benjamini, Y., and Hochberg, Y. (1995). Controlling the false discovery rate - a practical and powerful approach to multiple testing. *J. R. Stat. Soc. B Met.* 57, 289–300.
- Biswal, B. B., Van Kylen, J., and Hyde, J. S. (1997). Simultaneous assessment of flow and BOLD signals in resting-state functional connectivity maps. *NMR Biomed.* 10, 165–170. doi: 10.1002/(SICI)1099-1492(199704)10:4<165::AID-NBM454>3.0.CO;2-7
- Blondel, V. D., Guillaume, J. L., Lambiotte, R., and Lefebvre, E. (2008). Fast unfolding of communities in large networks. *J. Stat. Mech. Theor. Exp.* 2008, P10008. doi: 10.1088/1742-5468/2008/10/P10008
- Buckner, R. L., Andrews-Hanna, J. R., and Schacter, D. L. (2008). The brain's default network: anatomy, function, and relevance to disease. *Ann. N.Y. Acad. Sci.* 1124, 1–38. doi: 10.1196/annals.1440.011
- Cavanna, A. E., and Trimble, M. R. (2006). The precuneus: a review of its functional anatomy and behavioural correlates. *Brain* 129(Pt 3), 564–583. doi: 10.1093/brain/awl004
- Chulwoo, L., Xiang, L., Kaiming, L., Lei, G., and Tianming, L. (eds.). (2011). “Brain state change detection via fiber-centered functional connectivity analysis,” in *Biomedical Imaging: From Nano to Macro, 2011 IEEE International Symposium* (Chicago, IL).
- Deligianni, F., Robinson, E. C., Beckmann, C. F., Sharp, D., Edwards, A. D. and Rueckert, D. (eds.). (2010). “Inference of functional connectivity from structural brain connectivity,” in *Biomedical Imaging: From Nano to Macro, 2010 IEEE International Symposium* (Rotterdam).
- GadElkarim, J., Ajilore, O., Schonfeld, D., Zhan, L., Thomsson, P., Feusner, J., et al. (2013). Investigating brain community structure abnormalities in bipolar disorder using PLACE (Path Length Associated Community Estimation). *Hum. Brain Mapp.* doi: 10.1002/hbm.22324. [Epub ahead of print].
- GadElkarim, J., Schonfeld, D., Ajilore, O., Zhan, L., Zhang, A., Feusner, J., et al. (2012). A Framework for quantifying node-level community structure group differences in brain connectivity networks. *Med. Image Comp. Comp. Assist. Interv.* 2012, 196–203. doi: 10.1007/978-3-642-33418-4\_25

- Greicius, M. D., Flores, B. H., Menon, V., Glover, G. H., Solvason, H. B., Kenna, H., et al. (2007). Resting-state functional connectivity in major depression: abnormally increased contributions from subgenual cingulate cortex and thalamus. *Biol. Psychiatry* 62, 429–437.
- Hamilton, M. (1960). A rating scale for depression. *J. Neurol. Neurosurg. Psychiatry* 23, 56–62. doi: 10.1136/jnnp.23.1.56
- Honey, C. J., Sporns, O., Cammoun, L., Gigandet, X., Thiran, J. P., Meuli, R., et al. (2009). Predicting human resting-state functional connectivity from structural connectivity. *Proc. Natl. Acad. Sci. U.S.A.* 106, 2035–2040. doi: 10.1073/pnas.0811168106
- Johansen-Berg, H., Behrens, T. E., Robson, M. D., Drobjak, I., Rushworth, M. F., Brady, J. M., et al. (2004). Changes in connectivity profiles define functionally distinct regions in human medial frontal cortex. *Proc. Natl. Acad. Sci. U.S.A.* 101, 13335–13340. doi: 10.1073/pnas.0403743101
- Leow, A., Ajilore, O., Zhan, L., Arienzo, D., GadElkarim, J., Zhang, A., et al. (2013). Impaired inter-hemispheric integration in bipolar disorder revealed with brain network analyses. *Biol. Psychiatry* 73, 183–193. doi: 10.1016/j.biopsych.2012.09.014
- Leow, A., Zhan, L., Arienzo, D., GadElkarim, J., Zhang, A., Ajilore, O., et al. (2012). Hierarchical structural mapping for globally optimized estimation of functional networks. *Med. Image Comp. Comp. Assist. Interv.* 2012, 228–236. doi: 10.1007/978-3-642-33418-4\_29
- Newman, M. E., and Girvan, M. (2004). Finding and evaluating community structure in networks. *Phys. Rev. E Stat. Nonlin. Soft Matter Phys.* 69(2 Pt 2), 026113. doi: 10.1103/PhysRevE.69.026113
- Ng, B., Varoquaux, G., Poline, J.-B., Thirion, B. (2012). “A novel sparse graphical approach for multimodal brain connectivity inference,” in *Medical Image Computing and Computer-Assisted Intervention – MICCAI (2012). Lecture Notes in Computer Science 7510*, eds N. Ayache, H. Delingette, P. Golland, and K. Mori (Berlin; Heidelberg: Springer), 707–714.
- Passingham, R. E., Stephan, K. E., and Kotter, R. (2002). The anatomical basis of functional localization in the cortex. *Nat. Rev. Neurosci.* 3, 606–616. doi: 10.1038/nrn893
- Saygin, Z. M., Osher, D. E., Koldewyn, K., Reynolds, G., Gabrieli, J. D., and Saxe, R. R. (2011). Anatomical connectivity patterns predict face selectivity in the fusiform gyrus. *Nat. Neurosci.* 15, 321–327. doi: 10.1038/nn.3001
- Schwarz, A. J., and McGonigle, J. (2011). Negative edges and soft thresholding in complex network analysis of resting state functional connectivity data. *Neuroimage* 55, 1132–1146. doi: 10.1016/j.neuroimage.2010.12.047S1053-811901633-2
- Sheline, Y. I., Barch, D. M., Price, J. L., Rundle, M. M., Vaishnavi, S. N., Snyder, A. Z., et al. (2009). The default mode network and self-referential processes in depression. *Proc. Natl. Acad. Sci. U.S.A.* 106, 1942–1947. doi: 10.1073/pnas.0812686106
- Skudlarski, P., Jagannathan, K., Anderson, K., Stevens, M. C., Calhoun, V. D., Skudlarska, B. A., et al. (2010). Brain connectivity is not only lower but different in schizophrenia: a combined anatomical and functional approach. *Biol. Psychiatry* 68, 61–69. doi: 10.1016/j.biopsych.2010.03.035
- Sporns, O., Tononi, G., and Kotter, R. (2005). The human connectome: a structural description of the human brain. *PLoS Comput. Biol.* 1:e42. doi: 10.1371/journal.pcbi.0010042
- Steen, M., Hayasaka, S., Joyce, K., and Laurienti, P. (2011). Assessing the consistency of community structure in complex networks. *Phys. Rev. E Stat. Nonlin. Soft Matter Phys.* 84(1 Pt 2), 016111. doi: 10.1103/PhysRevE.84.016111
- van den Heuvel, M. P., and Sporns, O. (2011). Rich-club organization of the human connectome. *J. Neurosci.* 31, 15775–15786. doi: 10.1523/JNEUROSCI.3539-11.2011
- Varkuti, B., Cavusoglu, M., Kullik, A., Schiffler, B., Veit, R., Yilmaz, O., et al. (2011). Quantifying the link between anatomical connectivity, gray matter volume and regional cerebral blood flow: an integrative MRI Study. *PLoS ONE* 6:e14801. doi: 10.1371/journal.pone.0014801
- Whitfield-Gabrieli, S., and Nieto-Castanon, A. (2012). Conn: a functional connectivity toolbox for correlated and anticorrelated brain networks. *Brain Connect.* 2, 125–141. doi: 10.1089/brain.2012.0073
- Zhang, S., and Li, C. S. (2012). Functional connectivity mapping of the human precuneus by resting state fMRI. *Neuroimage* 59, 3548–3562. doi: 10.1016/j.neuroimage.2011.11.023

**Conflict of Interest Statement:** The authors declare that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.

Received: 12 September 2013; accepted: 01 November 2013; published online: 05 December 2013.

Citation: Ajilore O, Zhan L, GadElkarim J, Zhang A, Feusner JD, Yang S, Thompson PM, Kumar A and Leow A (2013) Constructing the resting state structural connectome. *Front. Neuroinform.* 7:30. doi: 10.3389/fninf.2013.00030

This article was submitted to the journal *Frontiers in Neuroinformatics*.

Copyright © 2013 Ajilore, Zhan, GadElkarim, Zhang, Feusner, Yang, Thompson, Kumar and Leow. This is an open-access article distributed under the terms of the Creative Commons Attribution License (CC BY). The use, distribution or reproduction in other forums is permitted, provided the original author(s) or licensor are credited and that the original publication in this journal is cited, in accordance with accepted academic practice. No use, distribution or reproduction is permitted which does not comply with these terms.



# Electroencephalogram approximate entropy influenced by both age and sleep

Gerick M. H. Lee<sup>1,2</sup>, Sara Fattinger<sup>2</sup>, Anne-Laure Mouthon<sup>2</sup>, Quentin Noirhomme<sup>3</sup> and Reto Huber<sup>2\*</sup>

<sup>1</sup> Institute of Neuroinformatics, University of Zurich and ETH Zurich, Zurich, Switzerland

<sup>2</sup> Child Development Center, University Children's Hospital Zurich, Zurich, Switzerland

<sup>3</sup> Coma Science Group, Neurology Department, Cyclotron Research Centre, University Hospital of Liège, University of Liège, Liège, Belgium

## Edited by:

Jesus M. Cortes, Ikerbasque  
Biocruces Health Research Institute,  
Spain

## Reviewed by:

Adam B. Barrett, University of  
Sussex, UK  
Naoto Burioka, Tottori University,  
Japan

## \*Correspondence:

Reto Huber, University Children's  
Hospital Zurich, Steinwiesstrasse  
75, 8032 Zurich, Switzerland  
e-mail: reto.huber@kispi.uzh.ch

The use of information-based measures to assess changes in conscious state is an increasingly popular topic. Though recent results have seemed to justify the merits of such methods, little has been done to investigate the applicability of such measures to children. For our work, we used the approximate entropy (ApEn), a measure previously shown to correlate with changes in conscious state when applied to the electroencephalogram (EEG), and sought to confirm whether previously reported trends in adult ApEn values across wake and sleep were present in children. Besides validating the prior findings that ApEn decreases from wake to sleep (including wake, rapid eye movement (REM) sleep, and non-REM sleep) in adults, we found that previously reported ApEn decreases across vigilance states in adults were also present in children (ApEn trends for both age groups: wake > REM sleep > non-REM sleep). When comparing ApEn values between age groups, adults had significantly larger ApEn values than children during wakefulness. After the application of an 8 Hz high-pass filter to the EEG signal, ApEn values were recalculated. The number of electrodes with significant vigilance state effects dropped from all 109 electrodes with the original 1 Hz filter to 1 electrode with the 8 Hz filter. The number of electrodes with significant age effects dropped from 10 to 4. Our results support the notion that ApEn can reliably distinguish between vigilance states, with low-frequency sleep-related oscillations implicated as the driver of changes between vigilance states. We suggest that the observed differences between adult and child ApEn values during wake may reflect differences in connectivity between age groups, a factor which may be important in the use of EEG to measure consciousness.

**Keywords: electroencephalogram, development, sleep, consciousness, approximate entropy**

## 1. INTRODUCTION

Recent theoretical work has proposed a link between the ability of the brain to integrate information and its corresponding conscious state (Tononi and Sporns, 2003; Tononi, 2004, 2008, 2012; Balduzzi and Tononi, 2008). Meanwhile, related experimental work has shown a link between changes in informational processing and conscious state (Massimini et al., 2005, 2007; Ferrarelli et al., 2010; Casali et al., 2013). These studies have provided compelling evidence of a causal relationship between the complexity of neural responses to external stimulation, as measured with the electroencephalogram (EEG), and the conscious state of the subject. Nevertheless, the benefits (particularly in the clinical setting) of a metric for conscious state independent of external stimulation are enough to encourage further work toward such a measure.

In this search for an EEG-specific measure of consciousness, many information-based measures have been applied. For this study, we chose the approximate entropy (ApEn), a measure of regularity in the time domain. Originally designed for use on physiological data (Pincus, 1991), ApEn quantifies the predictability of a signal by comparing the number of matching sequences of a given length with the number of matching

sequences one increment (time bin) longer. It has been suggested as an EEG measure of conscious state, and ties into informational theories of consciousness. Theoretical analysis has shown that isolated systems should show decreases in ApEn values (Pincus, 1994). This concurs with findings that non-rapid eye movement (NREM) sleep, associated with decreases in consciousness (Stickgold et al., 2001), tends to feature long-distance connectivity decreases and increases in local clustering (Massimini et al., 2005, 2007; Ferri et al., 2007, 2008; Spoormaker et al., 2010; Uehara et al., 2013). Rapid eye movement (REM) sleep, a state similar to wakefulness in its content of conscious experience, tends to show functional connectivity patterns more similar to those of wake (Massimini et al., 2010).

Prior applications of ApEn as a measure of conscious state have successfully shown correlations with anesthetic depth (Rezek and Roberts, 1998; Bruhn et al., 2000a,b; Zhang et al., 2001; Bruhn et al., 2003; Li et al., 2008; Hayashi et al., 2012), though these findings were contradicted by Jordan et al. (2006), who failed to report certain key parameters. Burioka et al. (2005) applied ApEn to data from adults across wake and sleep, finding a consistent decrease in ApEn from wake to sleep, with the lowest values occurring during deep sleep. Gu et al. (2003) also applied ApEn to data

across multiple stages of sleep, and during epileptic seizure onset, reporting decreases during sleep and during seizure onset, but did not use any statistical testing. Attempts to tie ApEn changes to behavioral changes during wakefulness have found conflicting results: ApEn analysis of subjects driving while sleep deprived found no significant changes in ApEn preceding driving errors (Papadelis et al., 2007a), though Flores Vega et al. (2013) recently showed that ApEn could be used to differentiate between some of the various mental tasks tested. Papadelis et al. (2007b) found no significant changes in ApEn as a function of hypoxia, but ApEn derived metrics did show significant changes. In summary, though its resolution within the wake state is unclear, when analyzing subjects between wakefulness and other conscious states, ApEn values consistently decreased with loss of consciousness. Comparisons of ApEn with other information-based measures typically showed it to be of comparable accuracy and reliability (Rezek and Roberts, 1998; Zhang et al., 2001; Bruhn et al., 2003; Abásolo et al., 2008; Li et al., 2008; Anier et al., 2012).

Past work has documented changes in EEG power across development, during both sleep (Feinberg, 1983; Buchmann et al., 2010; Feinberg and Campbell, 2010; Kurth et al., 2010) and wake (Whitford et al., 2007). To our knowledge, no group has yet applied ApEn to the EEG data of children. Therefore, to further assess the merits of ApEn as a measure of conscious state, we applied ApEn to EEG data recorded across sleep and wake, from both adults and children. Besides replicating the finding that ApEn can mark changes in vigilance state due to sleep in adults, we sought to verify that similar ApEn trends were present across wake and sleep in children, while also assessing any impact of age on ApEn values across both wake and sleep.

## 2. MATERIALS AND METHODS

### 2.1. SUBJECTS

For this study, subjects were pooled into two age groups of six subjects each, hereafter referred to as adults (age range: 19.4–25.1 years; mean age  $\pm$  SD: 23.2  $\pm$  2.06 years; 0 females), and children (age range: 10.6–12.6 years; mean age  $\pm$  SD: 11.4  $\pm$  0.691 years; 2 females). Subjects wore wrist actigraphs and kept sleep diaries to ensure sleep schedule compliance. Napping, alcohol consumption, and taking medication were all forbidden for the 24 h preceding the recording. Informed written consent was obtained from all subjects or their legal guardians. All procedures were performed with approval of the local ethics committee, and in accordance with the Declaration of Helsinki.

### 2.2. DATA ACQUISITION

All data (EEG, electrooculogram, and electromyogram) were gathered previously by our group at the University Children's Hospital Zurich during one evening, night, and morning. All sleep data used were originally published in earlier studies from our group (Kurth et al., 2010, 2012). Of the data recorded previously, subjects within the selected age range and with minimally-artifacted data were used, particularly during wakefulness and REM sleep. Wake data have not yet been used for publication, and were recorded during an auditory oddball task that was performed shortly before and after full night sleep recording. Subjects were awoken at a time allowing for normal school

or work attendance. A 128-electrode high-density EEG array (Electrical Geodesics, Eugene, OR, USA) was used for recording, with a sampling frequency of 500 Hz. Electrodes were referenced to the vertex during recording, which was used for filtering, downsampling, and artifact rejection. Impedances were set below 50 k $\Omega$ . Data were divided into 20 s epochs, the sleep stages of which were categorized using standard criteria (Iber et al., 2007). For the scoring of sleep stages, the recordings were referenced to the mastoid electrodes.

For analysis with the ApEn algorithm, data were then bandpass filtered at frequencies of 1 and 35 Hz, respectively, and downsampled to 250 Hz before being corrected for artifacts. Artifact correction for sleep data involved visual inspection of the power between 0.75–4 Hz, and 20–30 Hz, rejecting individual channels for a given epoch if the power exceeded a mean band power value. Artifact correction for wake data was based on independent component analysis, as presented by Jung et al. (2000). Finally, data were referenced to the average activity of all non-rejected channels above the ears for analysis. To investigate better the role of low-frequency EEG activity on ApEn, we later refiltered our original data with an 8 Hz high-pass filter, and recalculated the ApEn.

ApEn analysis used all 109 electrodes above the ears not rejected during artifact correction. Data preprocessing and all analyses were done using Matlab (The MathWorks, Natick, MA, USA), statistical testing used Matlab, as well as R (R Foundation for Statistical Computing, Vienna, Austria). Data series of 4000 points, corresponding to 16 s of EEG signal, were used for analysis. Because wake epochs were scored in epochs of 4 s duration, analysis used aggregate 16 s epochs comprised of four consecutive artifact-free epochs, taken from the evening recording session preceding sleep. Sleep data was drawn from the first 16 s of unartifacts 20 s epochs. Sleep epochs used were preceded and followed by at least 1 min (three epochs) of sleep all of the same stage, to minimize the influence of stage transitions. For one adult subject, only two epochs (40 s) preceded and followed the epoch for the N3 sleep stage used for all analyses.

### 2.3. APPROXIMATE ENTROPY (ApEn)

The development of ApEn was driven by the need for a distribution-free measure of signal regularity. Unlike the Shannon entropy, the calculation of ApEn is not predicated on the underlying distribution of the data; it is instead based on sequence recurrence. This allows ApEn to be applied to signals of shorter length, and makes model estimation wholly unnecessary, removing the risk for misestimation based on poor model selection.

ApEn can be understood as the logarithmic ratio of component-wise matching sequences from a signal of length  $N$ . The other relevant parameters are  $r$ , a factor based on the standard deviation of the signal being analyzed, and used for comparison. The final parameter is  $m$ , the length of sequences compared. It is measured as an integer count of discrete time bins. The ApEn is computed as follows:

1. The first sequence of length  $m$ , is compared with all other sequences of the same length in a point-wise manner. Those sequences for which all points are within  $r$  of their



corresponding point in the original sequence are counted as a match (including the base sequence with itself). This count is used in step 3.

2. The same comparison is made for sequences of length  $m + 1$ , starting with the first sequence of  $m + 1$  points. This count is used in step 3.
3. The count from step 2 is divided by step 1, and the natural logarithm of this ratio is taken.
4. This process is then repeated for all possible sequences (the final  $m$  points of the signal cannot be used, as there would be no  $m + 1$  sequence for comparison).
5. All logarithm results are then summed, divided by  $N - m$  (the total number of possible base sequences), and multiplied by  $-1$ .

The minimum value for ApEn is 0, suggesting a fully predictable sequence. ApEn values are heavily dependent on parameter choice, and values calculated with different parameter choices cannot be compared. Because the filter factor,  $r$ , typically has its values pegged to the standard deviation of the sequence, the origin of ApEn's robustness to noise and scale invariance can be seen. Our parameters were set per the suggestion of Pincus and Goldberger (1994), as well as other groups applying ApEn to EEG data (Bruhn et al., 2000a; Li et al., 2008; Hayashi et al., 2012), specifically Burioka et al. (2005), to  $m$  and  $r$  values of 2 and  $0.2 \cdot SD$ , respectively. Our  $N$  value, the length of the data series used, was 4000 points.

To confirm the proper functioning of the ApEn algorithm, we computed ApEn values for six regular sine curve sequences of 4000 points, with a simulated sampling rate of 250 Hz. The sine frequencies used were 1, 2, 4, 8, 16, and 32 Hz, frequencies all within the range used in our EEG ApEn analysis. Each sine curve sequence was then randomly shuffled twenty times. ApEn values were calculated for all six sine curve sequences, and all 120 random sequences (twenty random ApEn values per sine curve).

For an excellent appendix detailing the steps in ApEn calculation (including a simple by-hand walkthrough of the steps involved in ApEn calculation, as well as a sample implementation in Visual Basic), please see Bruhn et al. (2000a).

## 2.4. STATISTICAL ANALYSIS

As mentioned above, statistical analyses were performed using Matlab and R. Values were imported into R and log-transformed, to better approximate a normal distribution. A linear mixed model for the subject age groups (independent factor) and vigilance states (repeated-measures factor) was then generated and tested using a repeated-measures ANOVA.

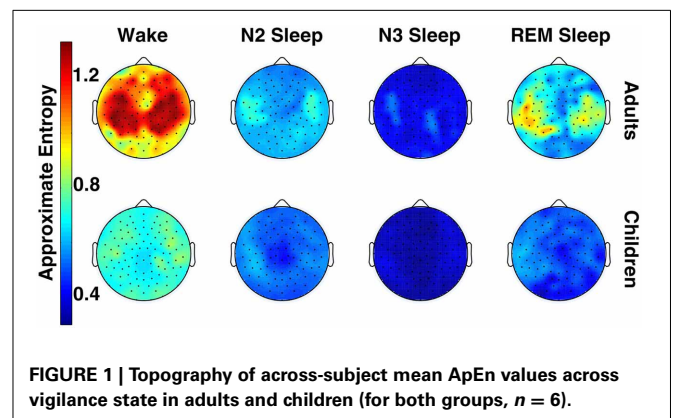
All multiple comparisons corrections were performed using the Holm–Bonferroni method. Because EEG electrodes are not independent, the Holm–Bonferroni correction is overly conservative. For this reason, in order to provide the most informative results,  $p$ -values and significance results from comparisons using all electrodes are reported both with and without correction. To better investigate differences between age groups, unpaired independent-samples  $t$ -tests were performed between each age group within each vigilance state.

## 3. RESULTS

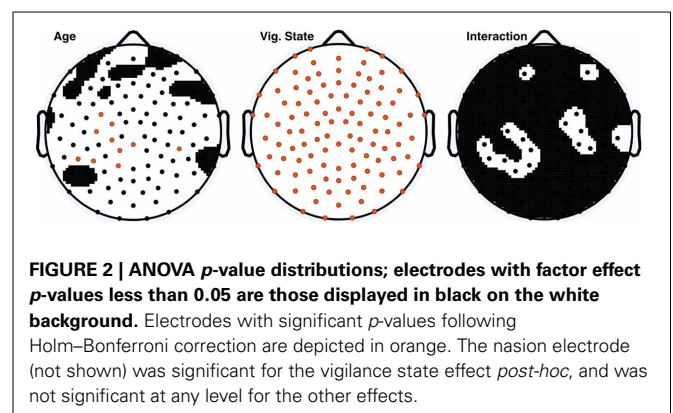
As described above, we analyzed a set of simulated data to validate our ApEn algorithm. ApEn values for the simulated data ranged between 0.07 and 0.29 for the sine curves. Mean ApEn values for the shuffled sequences were all 1.94, with standard deviations of less than 0.01. These results were in line with expectations.

**Figure 1** shows the topographical distribution of mean ApEn per electrode in adults and children. ApEn value trends across vigilance states were similar for both age groups, and were as follows: wake ApEn > REM sleep ApEn > N2 sleep ApEn > N3 sleep ApEn, though REM sleep and N2 sleep were often overlapping, especially in children. **Figure 2** displays the ANOVA results for all factors. All 109 tested electrodes had significant vigilance state effects after *post-hoc* correction. Ninety-three electrodes, widely distributed across the scalp, showed a significant age effect before correction. Ten electrodes had a significant age effect after correction. These ten electrodes were largely clustered over the left parietal and the area between the occipital and temporal lobes, with one isolated over the right temporal lobe.

To better discern the causes of the observed age effects, within-vigilance state pairwise  $t$ -tests were calculated across all electrodes. These results are shown in **Figure 3**, where 66 electrodes had significant age effects during wakefulness before correction, of which 28 electrodes were still significant following correction. N2 sleep and REM sleep had large clusters of significant electrodes before correction; none were significant after correction.



**FIGURE 1 |** Topography of across-subject mean ApEn values across vigilance state in adults and children (for both groups,  $n = 6$ ).



**FIGURE 2 |** ANOVA  $p$ -value distributions; electrodes with factor effect  $p$ -values less than 0.05 are those displayed in black on the white background. Electrodes with significant  $p$ -values following Holm–Bonferroni correction are depicted in orange. The nasion electrode (not shown) was significant for the vigilance state effect *post-hoc*, and was not significant at any level for the other effects.

To fully explore the possibility that sleep regulatory differences between age groups may influence our results (Carskadon et al., 1980; Carskadon and Acebo, 2002), and to verify that ApEn wake values are not influenced by potential changes in overall synaptic weighting during sleep [as proposed by Tononi and Cirelli (2003)], we compared ApEn from both the evening and morning recording sessions, averaged across all 109 electrodes. A Two-Way ANOVA for age and recording session found a significant age effect ( $p < 0.001$ ), as expected from earlier testing, but found no significant effect for the recording session, nor for the interaction of the two ( $p > 0.1$  for both effects). For all other ApEn analysis, evening wakefulness was used to represent wakefulness.

To assess the origin of the observed ApEn differences between children and adults, a high-pass filter of 8 Hz was applied to the data, and ApEn values were again calculated. Two-Way ANOVA results from the high-pass-filtered data of all electrodes are depicted in **Figure 4**. Forty-two electrodes had significant age effects before correction, of which four electrodes were significant following correction. Vigilance state effects were almost entirely abolished; seven electrodes were significant before correction; one electrode was significant after correction.

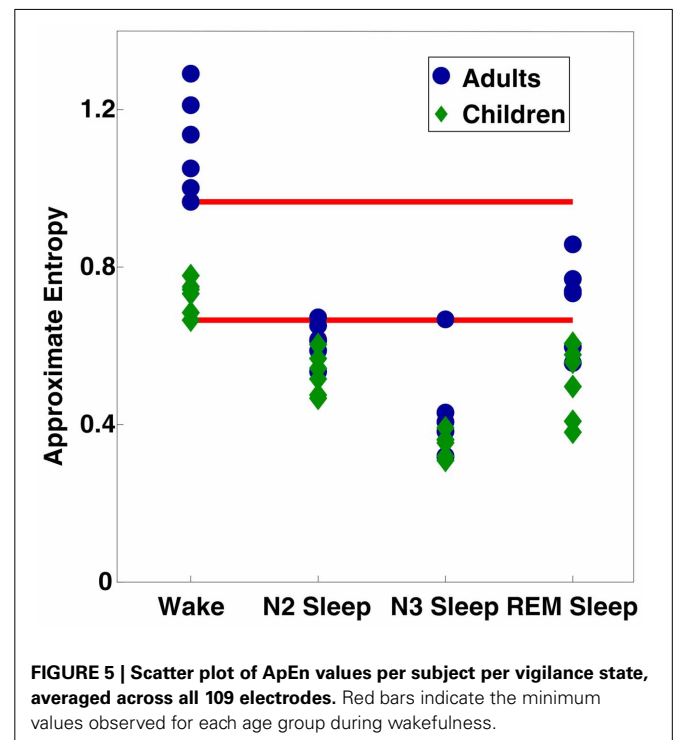
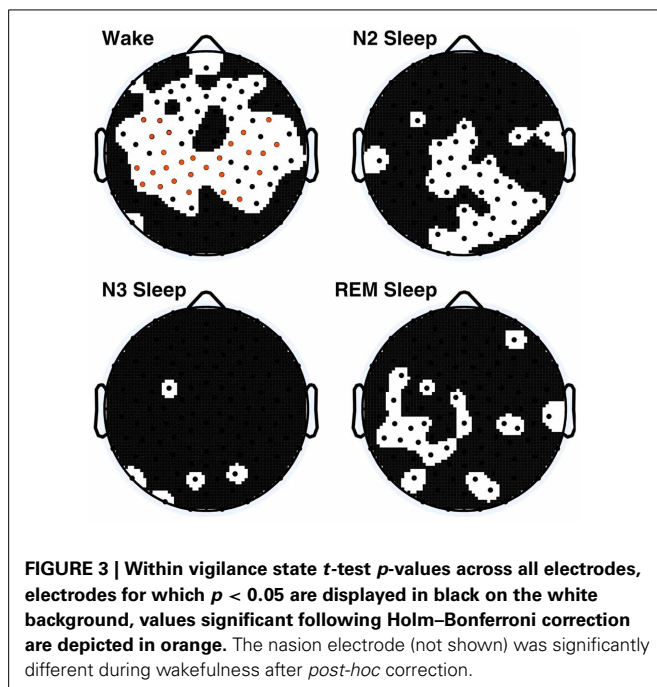
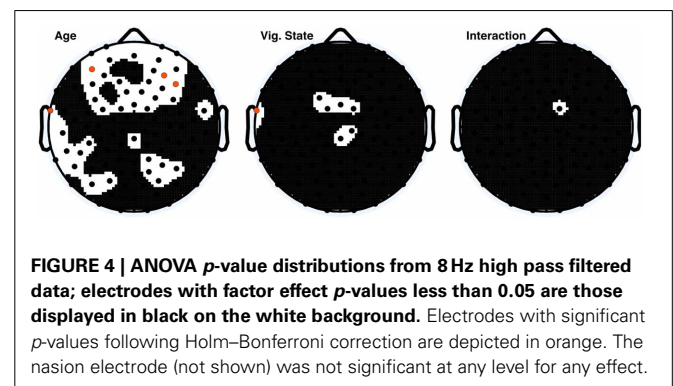
To check for changes in the regional distribution of ApEn, electrode values were normalized to the within-subject-within-vigilance-state mean across all electrodes. One electrode (located near the posterior end of the right frontal area) showed a significant vigilance state effect after correction. No other electrodes were significant for any effect (age, vigilance state, or the interaction of the two), even before *post-hoc* correction.

Finally, to investigate individual differences in ApEn values, we averaged ApEn across all electrodes, and plotted values for each stage as **Figure 5**. The minimum values from wakefulness were invariably higher than the maximum observed ApEn value from

sleep (including both NREM and REM sleep) within the same age group. Comparison of all subjects showed some adult sleep values (especially during REM sleep) greater than some or all wake ApEn values for children.

#### 4. DISCUSSION

Our analysis showed significant ApEn effects due both to vigilance state and age, with age differences being predominantly driven by differences during wakefulness. As a measure of vigilance state, ApEn showed strongly significant results across wake and sleep, with ApEn values in adults following the same trend as those previously reported (Burioka et al., 2005). ApEn results from children followed similar trends between vigilance states, with the only significant age differences occurring during wakefulness. As demonstrated in **Figure 5**, within age group minimum ApEn values for wakefulness were higher than maximum ApEn sleep values for the same age group, supporting the notion that



ApEn can reliably detect changes in vigilance state. The almost complete abolition of significant vigilance state effects observed following application of the 8 Hz high pass filter to our data provide evidence that slow wave activity, the key EEG oscillation of deep (NREM) sleep (Steriade et al., 2001; Buzsaki, 2006), is also the key driver behind the increased regularity observed during sleep.

Pincus (1994) observed that isolated systems have lower ApEn values. If the brain is indeed a more segregated one during NREM sleep, as suggested by experimental work (Massimini et al., 2005, 2007), then one would expect to see decreases in ApEn during NREM sleep, as we did. These findings concur with the proposal presented in Tononi and Massimini (2008), which drew a link between slow wave activity during deep sleep and an interruption in information processing, leading to loss of consciousness. That ApEn differences due to vigilance state mostly disappeared after the removal of the lower frequency bands connects ApEn changes to the presence of sleep oscillations, specifically slow waves. Our results therefore suggest the possibility of a causal relationship between EEG signal changes, as measured via ApEn, and the hyperpolarization phase associated with the slow oscillation (Steriade et al., 2001). This hyperpolarization has been implicated in the induction of loss of consciousness (Massimini et al., 2005).

The almost complete lack of significant vigilance state differences following normalization to the mean value across all electrodes indicates that changes in ApEn values across wake and sleep are not the result of changing topographical distribution. These results were therefore unlike previously observed age-dependent topographical changes in sleep slow wave activity (Kurth et al., 2010), and rather suggest that changes in signal regularity are of a more global nature.

Besides the widely distributed nature of changes due to sleep stage, changes between wake adults and children were also found to be global: Pairwise *t*-tests found a broad distribution of electrodes with significant increases in wake ApEn values across development. These results concur with those of Gasser et al. (1988), who found absolute EEG band power decreases in the delta and theta bands (both of which were below 7.5 Hz), and the overall spectrum, across adolescence when measuring during eyes-closed wake. Our findings also agree with the EEG results of Whitford et al. (2007), who found global power decreases during wakefulness across age, especially in the lower frequency range (0.5–7.5 Hz).

While EEG power changes between adults and children have also been observed during sleep [as reviewed in Feinberg (1983); Feinberg and Campbell (2010), also Buchmann et al. (2010); Kurth et al. (2010)], we only observed age differences in ApEn values during wakefulness. This discrepancy may potentially be explained by the large increase in EEG power during sleep. EEG power differences caused by sleep-related oscillations may be of a large enough scale relative to those due to developmental changes that ApEn age differences are obscured. **Figure 5**, the scatter plot of individual mean ApEn values shows a tendency for ApEn values to be lower in children during sleep (the largest ApEn values for any given stage are invariably from adults; the lowest from children), even though statistical testing reveals no age differences.

Our results from wakefulness may also be in line with this claim; if ApEn age differences during wakefulness reflect anatomical connectivity changes, then the lack of significant differences at occipital and temporal electrodes is in line with what would be expected based on prior developmental research work. The review of Feinberg (1983) drew parallels between their work measuring changes in sleep EEG activity across development, and anatomical work, which showed regional variation in synaptic densities across development [Huttenlocher (1979); Huttenlocher et al. (1982), expanded in Huttenlocher and Dabholkar (1997)]. These works independently demonstrated that primary sensory cortices were first to reach adult-level values, both when measured via EEG power during sleep, and histological synaptic density counts. Coupled MRI and EEG work from our group found correlations between slow-wave activity decreases during sleep and gray matter volume decreases (Buchmann et al., 2010). Similar work during wakefulness from other groups showed correlations between gray matter volume decreases and low-frequency EEG decreases from late childhood through adulthood (subjects ranged between 10 and 30 years of age, Whitford et al., 2007), particularly in the parietal and frontal regions, where our significant differences were focused. Developmental changes in the topographical distribution of low-frequency sleep oscillations followed similar trends; regions converging to adult-level synaptic densities earlier were also the first to converge to adult-level EEG activity (Kurth et al., 2010). Without the use of other tools, such as single-unit recording or transcranial magnetic stimulation, it is difficult to separate EEG slow wave activity from the changes in functional connectivity observed on the neuronal level during slow wave sleep. Nevertheless, the decrease in ApEn observed between wakefulness in children and in adults matches with the increased local anatomical connectivity observed in children. That changes in both vigilance state and sleep result in decreased ApEn values supports the notion that changes in ApEn values may reflect connectivity changes, both anatomical and functional.

Though this claim must be further tested, if true, it would mean that ApEn changes reflect both functional (between wake and sleep) and anatomical (across development) connectivity changes in the brain. As we have shown, ApEn can reliably distinguish between wake and sleep within subject age groups. However, having demonstrated that age has an uneven influence on ApEn values across changes in vigilance state, we highlight the need for future research to fully explore the influence of age on proposed information-based EEG measures of consciousness.

## AUTHOR CONTRIBUTIONS

Gerick M. H. Lee, Anne-Laure Mouchon, Quentin Noirhomme, and Reto Huber designed research; Gerick M. H. Lee and Sara Fattinger performed research; Gerick M. H. Lee, Sara Fattinger, and Reto Huber analyzed data; and Gerick M. H. Lee and Reto Huber wrote the paper.

## ACKNOWLEDGMENTS

The authors would like to thank Salomé Kurth, Maya Ringli, and Anja Geiger for data collection; Caroline Lustenberger, Daniel Heersink, and Mattia Molinaro for statistical advice; and David Balduzzi for helpful input and discussion.



## FUNDING

This work was funded by Swiss National Science Foundation grant PP00P3-135438 to Reto Huber.

## REFERENCES

- Abásolo, D., Escudero, J., Hornero, R., Gómez, C., and Espino, P. (2008). Approximate entropy and auto mutual information analysis of the electroencephalogram in Alzheimer's disease patients. *Med. Biol. Eng. Comput.* 46, 1019–1028. doi: 10.1007/s11517-008-0392-1
- Anier, A., Lipping, T., Ferenets, R., Puumala, P., Sonkajärvi, E., Rätsep, I., et al. (2012). Relationship between approximate entropy and visual inspection of irregularity in the EEG signal, a comparison with spectral entropy. *Br. J. Anaesth.* 109, 928–934. doi: 10.1093/bja/ae312
- Balduzzi, D., and Tononi, G. (2008). Integrated information in discrete dynamical systems: motivation and theoretical framework. *PLoS Comput. Biol.* 4:e1000091. doi: 10.1371/journal.pcbi.1000091
- Bruhn, J., Bouillon, T. W., Radulescu, L., Hoeft, A., Bertaccini, E., and Shafer, S. L. (2003). Correlation of approximate entropy, bispectral index, and spectral edge frequency 95 (SEF95) with clinical signs of anesthetic depth during coadministration of propofol and remifentanyl. *Anesthesiology* 98, 621–627. doi: 10.1097/0000542-200303000-00008
- Bruhn, J., Röpcke, H., and Hoeft, A. (2000a). Approximate entropy as an electroencephalographic measure of anesthetic drug effect during desflurane anesthesia. *Anesthesiology* 92, 715–726. doi: 10.1097/0000542-200003000-00016
- Bruhn, J., Röpcke, H., Rehberg, B., Bouillon, T., and Hoeft, A. (2000b). Electroencephalogram approximate entropy correctly classifies the occurrence of burst suppression pattern as increasing anesthetic drug effect. *Anesthesiology* 93, 981–985. doi: 10.1097/0000542-200010000-00018
- Buchmann, A., Ringli, M., Kurth, S., Schaerer, M., Geiger, A., Jenni, O. G., et al. (2010). EEG sleep slow-wave activity as a mirror of cortical maturation. *Cereb. Cortex* 21, 607–615. doi: 10.1093/cercor/bhq129
- Burioka, N., Miyata, M., Cornélissen, G., Halberg, F., Takeshima, T., Kaplan, D. T., et al. (2005). Approximate entropy in the electroencephalogram during wake and sleep. *Clin. EEG Neurosci.* 36, 21–24. doi: 10.1177/155005940503600106
- Buzsáki, G. (2006). *Rhythms of the Brain*. New York, NY: Oxford University Press. doi: 10.1093/acprof:oso/9780195301069.001.0001
- Carskadon, M. A., and Acebo, C. (2002). Regulation of sleepiness in adolescents: update, insights, and speculation. *Sleep* 25, 606–614.
- Carskadon, M. A., Harvey, K., Duke, P., Anders, T. F., Litt, I. F., and Dement, W. C. (1980). Pubertal changes in daytime sleepiness. *Sleep* 2, 453–460.
- Casali, A. G., Gosseries, O., Rosanova, M., Boly, M., Sarasso, S., Casali, K. R., et al. (2013). A theoretically based index of consciousness independent of sensory processing and behavior. *Sci. Transl. Med.* 5, 198ra105. doi: 10.1126/scitranslmed.3006294
- Feinberg, I. (1982/1983). Schizophrenia: caused by a fault in programmed synaptic elimination during adolescence? *J. Psychiatr. Res.* 17, 319–334. doi: 10.1016/0022-3956(82)90038-3
- Feinberg, I., and Campbell, I. G. (2010). Sleep EEG changes during adolescence: an index of a fundamental brain reorganization. *Brain Cogn.* 72, 56–65. doi: 10.1016/j.bandc.2009.09.008
- Ferrarelli, F., Massimini, M., Sarasso, S., Casali, A., Riedner, B. A., Angelini, G., et al. (2010). Breakdown in cortical effective connectivity during midazolam-induced loss of consciousness. *Proc. Natl. Acad. Sci. U.S.A.* 107, 2681–2686. doi: 10.1073/pnas.0913008107
- Ferri, R., Rundo, F., Bruni, O., Terzano, M. G., and Stam, C. J. (2007). Small-world network organization of functional connectivity of EEG slow-wave activity during sleep. *Clin. Neurophysiol.* 118, 449–456. doi: 10.1016/j.clinph.2006.10.021
- Ferri, R., Rundo, F., Bruni, O., Terzano, M. G., and Stam, C. J. (2008). The functional connectivity of different EEG bands moves towards small-world network organization during sleep. *Clin. Neurophysiol.* 119, 2026–2036. doi: 10.1016/j.clinph.2008.04.294
- Flores Vega, C. H., Noel, J., and Fernández, J. R. (2013). “Cognitive task discrimination using approximate entropy (ApEn) on EEG signals,” in *Biosignals and Biorobotics Conference (BRC), 2013 ISSNIP* (Rio de Janeiro), 1–4.
- Gasser, T., Verleger, R., Bächer, P., and Sroka, L. (1988). Development of the EEG of school-age children and adolescents. I. Analysis of band power. *Electroencephalogr. Clin. Neurophysiol.* 69, 91–99. doi: 10.1016/0013-4694(88)90204-0
- Gu, F., Meng, X., Shen, E., and Cai, Z. (2003). Can we measure consciousness with EEG complexities? *Int. J. Bifurcat. Chaos* 13, 733–742. doi: 10.1142/S0218127403006893
- Hayashi, K., Shigemori, K., and Sawa, T. (2012). Neonatal electroencephalography shows low sensitivity to anesthesia. *Neurosci. Lett.* 517, 87–91. doi: 10.1016/j.neulet.2012.04.028
- Huttenlocher, P. R. (1979). Synaptic density in human frontal cortex—developmental changes and effects of aging. *Brain Res.* 163, 195–205. doi: 10.1016/0006-8993(79)90349-4
- Huttenlocher, P. R., and Dabholkar, A. S. (1997). Regional differences in synaptogenesis in human cerebral cortex. *J. Comp. Neurol.* 387, 167–178. doi: 10.1002/(SICI)1096-9861(19971020)387:2<167::AID-CNE1>3.0.CO;2-Z
- Huttenlocher, P. R., de Courten, C., Garey, L. J., and Van der Loos, H. (1982). Synaptogenesis in human visual cortex - evidence for synapse elimination during normal development. *Neurosci. Lett.* 33, 247–252. doi: 10.1016/0304-3940(82)90379-2
- Iber, C., Ancoli-Israel, S., Chesson, A. L., and Quan, S. F., editors (2007). *The AASM Manual for the Scoring of Sleep and Associated Events: Rules, Terminology and Technical Specification, 1st Edn*. Westchester, IL: American Academy of Sleep Medicine.
- Jordan, D., Schneider, G., Hock, A., Hensel, T., Stockmanns, G., and Kochs, E. F. (2006). EEG parameters and their combination as indicators of depth of anaesthesia. *Biomed. Tech.* 51, 89–94. doi: 10.1515/BMT.2006.016
- Jung, T. P., Makeig, S., Humphries, C., Lee, T. W., McKeown, M. J., Iragui, V., et al. (2000). Removing electroencephalographic artifacts by blind source separation. *Psychophysiology* 37, 163–178. doi: 10.1111/1469-8986.3720163
- Kurth, S., Ringli, M., Geiger, A., LeBourgeois, M., Jenni, O. G., and Huber, R. (2010). Mapping of cortical activity in the first two decades of life: a high-density sleep electroencephalogram study. *J. Neurosci.* 30, 13211–13219. doi: 10.1523/JNEUROSCI.2532-10.2010
- Kurth, S., Ringli, M., LeBourgeois, M. K., Geiger, A., Buchmann, A., Jenni, O. G., et al. (2012). Mapping the electrophysiological marker of sleep depth reveals skill maturation in children and adolescents. *Neuroimage* 63, 959–965. doi: 10.1016/j.neuroimage.2012.03.053
- Li, X., Cui, S., and Voss, L. J. (2008). Using permutation entropy to measure the electroencephalographic effects of sevoflurane. *Anesthesiology* 109, 448–456. doi: 10.1097/ALN.0b013e318182a91b
- Massimini, M., Ferrarelli, F., Esser, S. K., Riedner, B. A., Huber, R., Murphy, M., et al. (2007). Triggering sleep slow waves by transcranial magnetic stimulation. *Proc. Natl. Acad. Sci. U.S.A.* 104, 8496–8501. doi: 10.1073/pnas.0702495104
- Massimini, M., Ferrarelli, F., Huber, R., Esser, S. K., Singh, H., and Tononi, G. (2005). Breakdown of cortical effective connectivity during sleep. *Science* 309, 2228–2232. doi: 10.1126/science.1117256
- Massimini, M., Ferrarelli, F., Murphy, M. J., Huber, R., Riedner, B. A., Casarotto, S., et al. (2010). Cortical reactivity and effective connectivity during REM sleep in humans. *Cogn. Neurosci.* 1, 176–183. doi: 10.1080/17588921003731578
- Papadelis, C., Chen, Z., Kourtidou-Papadelis, C., Bamidis, P. D., Chouvarda, I., Bekiaris, E., et al. (2007a). Monitoring sleepiness with on-board electrophysiological recordings for preventing sleep-deprived traffic accidents. *Clin. Neurophysiol.* 118, 1906–1922. doi: 10.1016/j.clinph.2007.04.031
- Papadelis, C., Kourtidou-Papadelis, C., Bamidis, P. D., Maglaveras, N., and Pappas, K. (2007b). The effect of hypobaric hypoxia on multichannel EEG signal complexity. *Clin. Neurophys.* 118, 31–52. doi: 10.1016/j.clinph.2006.09.008
- Pincus, S. M. (1991). Approximate entropy as a measure of system complexity. *Proc. Natl. Acad. Sci. U.S.A.* 88, 2297–2301. doi: 10.1073/pnas.88.6.2297
- Pincus, S. M. (1994). Greater signal regularity may indicate increased system isolation. *Math. Biosci.* 122, 161–181. doi: 10.1016/0025-5564(94)90056-6
- Pincus, S. M., and Goldberger, A. L. (1994). Physiological time-series analysis; what does regularity quantify? *Am. J. Physiol.* 266, H1643–H1656.
- Rezek, I. A., and Roberts, S. J. (1998). Stochastic complexity measures for physiological signal analysis. *IEEE Trans. Biomed. Eng.* 45, 1186–1191. doi: 10.1109/10.709563
- Spoormaker, V. I., Schröter, M. S., Gleiser, P. M., Andrade, K. C., Dresler, M., Wehrle, R., et al. (2010). Development of a large-scale functional brain network during human non-rapid eye movement sleep. *J. Neurosci.* 30, 11379–11387. doi: 10.1523/JNEUROSCI.2015-10.2010
- Steriade, M., Timofeev, I., and Grenier, F. (2001). Natural waking and sleep states: a view from inside neocortical neurons. *J. Neurophysiol.* 85, 1969–1985.



- Stickgold, R., Malia, A., Fosse, R., Propper, R., and Hobson, J. A. (2001). Brain-mind state: I. longitudinal field study of sleep/wake factors influencing mentation report length. *Sleep* 24, 171–179.
- Tononi, G. (2004). An information integration theory of consciousness. *BMC Neurosci.* 5:42. doi: 10.1186/1471-2202-5-42
- Tononi, G. (2008). Consciousness as integrated information: a provisional manifesto. *Biol. Bull.* 215, 216–242. doi: 10.2307/25470707
- Tononi, G. (2012). Integrated information theory of consciousness: an updated account. *Arch. Ital. Biol.* 150, 293–329. doi: 10.4449/aib.v149i5.1388
- Tononi, G., and Cirelli, C. (2003). Sleep and synaptic homeostasis: a hypothesis. *Brain Res. Bull.* 62, 143–150. doi: 10.1016/j.brainresbull.2003.09.004
- Tononi, G., and Massimini, M. (2008). Why does consciousness fade in early sleep? *Ann. N.Y. Acad. Sci.* 1129, 330–334. doi: 10.1196/annals.1417.024
- Tononi, G., and Sporns, O. (2003). Measuring information integration. *BMC Neurosci.* 4:31. doi: 10.1186/1471-2202-4-31
- Uehara, T., Yamasaki, T., Okamoto, T., Koike, T., Kan, S., Miyauchi, S., et al. (2013). Efficiency of a "small-world" brain network depends on consciousness level: a resting-state fMRI study. *Cereb. Cortex*. doi: 10.1093/cercor/bht004. [Epub ahead of print].
- Whitford, T. J., Rennie, C. J., Grieve, S. M., Clark, C. R., Gordon, E., and Williams, L. M. (2007). Brain maturation in adolescence: concurrent changes in neuroanatomy and neurophysiology. *Hum. Brain Mapp.* 28, 228–237. doi: 10.1002/hbm.20273
- Zhang, X.-S., Roy, R. J., and Jensen, E. W. (2001). EEG complexity as a measure of depth of anesthesia for patients. *IEEE Trans. Biomed. Eng.* 48, 1424–1433. doi: 10.1109/10.966601

**Conflict of Interest Statement:** The authors declare that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.

Received: 01 November 2013; accepted: 19 November 2013; published online: 05 December 2013.

Citation: Lee GMH, Fattinger S, Mouthon A-L, Noirhomme Q and Huber R (2013) Electroencephalogram approximate entropy influenced by both age and sleep. *Front. Neuroinform.* 7:33. doi: 10.3389/fninf.2013.00033

This article was submitted to the journal *Frontiers in Neuroinformatics*.

Copyright © 2013 Lee, Fattinger, Mouthon, Noirhomme and Huber. This is an open-access article distributed under the terms of the Creative Commons Attribution License (CC BY). The use, distribution or reproduction in other forums is permitted, provided the original author(s) or licensor are credited and that the original publication in this journal is cited, in accordance with accepted academic practice. No use, distribution or reproduction is permitted which does not comply with these terms.



# Information gain in the brain's resting state: A new perspective on autism

José L. Pérez Velázquez<sup>1,2</sup> and Roberto F. Galán<sup>3\*</sup>

<sup>1</sup> Neuroscience and Mental Health Programme, Division of Neurology, Hospital for Sick Children, Toronto, ON, Canada

<sup>2</sup> Institute of Medical Science and Department of Paediatrics, Brain and Behaviour Centre, University of Toronto, Toronto, ON, Canada

<sup>3</sup> Department of Neurosciences, School of Medicine, Case Western Reserve University, Cleveland, OH, USA

## Edited by:

Daniele Marinazzo, University of  
Gent, Belgium

## Reviewed by:

Jesus M. Cortes, Ikerbasque,  
Biocruces Health Research Institute,  
Spain  
Filip Van Opstal, Ghent University,  
Belgium

## \*Correspondence:

Roberto F. Galán, Department of  
Neurosciences, School of Medicine,  
Case Western Reserve University,  
10900 Euclid Avenue, Cleveland,  
OH 44106, USA  
e-mail: rfgalan@case.edu

Along with the study of brain activity evoked by external stimuli, an increased interest in the research of background, “noisy” brain activity is fast developing in current neuroscience. It is becoming apparent that this “resting-state” activity is a major factor determining other, more particular, responses to stimuli and hence it can be argued that background activity carries important information used by the nervous systems for adaptive behaviors. In this context, we investigated the generation of information in ongoing brain activity recorded with magnetoencephalography (MEG) in children with autism spectrum disorder (ASD) and non-autistic children. Using a stochastic dynamical model of brain dynamics, we were able to resolve not only the deterministic interactions between brain regions, i.e., the brain's functional connectivity, but also the stochastic inputs to the brain in the resting state; an important component of large-scale neural dynamics that no other method can resolve to date. We then computed the Kullback-Leibler (KLD) divergence, also known as information gain or relative entropy, between the stochastic inputs and the brain activity at different locations (outputs) in children with ASD compared to controls. The divergence between the input noise and the brain's ongoing activity extracted from our stochastic model was significantly higher in autistic relative to non-autistic children. This suggests that brains of subjects with autism create more information at rest. We propose that the excessive production of information in the absence of relevant sensory stimuli or attention to external cues underlies the cognitive differences between individuals with and without autism. We conclude that the information gain in the brain's resting state provides quantitative evidence for perhaps the most typical characteristic in autism: withdrawal into one's inner world.

**Keywords:** brain's resting state, Asperger's syndrome, functional connectivity, stochastic input, relative entropy

## INTRODUCTION

Neuroscience has traditionally focused on the investigation of stimulus-induced activity, whereas spontaneous activity has been considered as noise or background activity of little consequence. However, this view is rapidly changing due in part to empirical evidence indicating the fundamental importance of background, “noisy” activity in the brain for the processing of sensory inputs. Indeed, the brain never rests, for it is constantly receiving inputs, either from the outside or from the body, and even in periods of slow wave sleep the thalamocortical networks display important, coordinated activity. Even when external sensory stimuli are minimized, as in sensory deprivation experiments, the brain responds creating its own world of hallucinations (Sireteanu et al., 2008). Thus, cognitive states in the “idle” brain are not passive and perhaps represent the best opportunity to study the functional connectivity (a term much used these days and perhaps many times abused) of the brain (Galán, 2008; Ringach, 2009; Papo, 2013).

There is a current debate in the autism field about the possible differences in brain connectivity that manifest in the special cognitive style of autistic individuals. In particular, it has been argued

that autistic brains are more “disconnected” than those individuals without autism, notion derived mainly from metabolic brain measures like PET or fMRI (Herbert, 2005; Kennedy and Courchesne, 2008; Monk et al., 2009; Thai et al., 2009). Distinct patterns of synchronization of electroencephalographic or magnetoencephalographic (MEG) signals between individuals with and without autism spectrum disorder (ASD) have also been reported (Murias et al., 2007; Pérez Velázquez et al., 2009; Tsiaras et al., 2011; Teitelbaum et al., 2012). As for anatomical features that could underlie possible differences in functional connectivity and thus brain coordination dynamics, alterations in the frontal cortex have been noted in autism, and particularly, an abnormal spatial organization in the microglial-neuronal components (Morgan et al., 2012). Recent studies with diffusion tensor imaging have also revealed white matter abnormalities in autism, in particular, a possible atypical lateralization in some white matter tracts of the brain and a possible atypical developmental trajectory of white matter microstructure in persons with ASD (Travers et al., 2012).

Recently, based on the notion that brain activity at rest can be accurately described using stochastic linear dynamics, we used a

multivariate Ornstein-Uhlenbeck process (mOUP) to investigate brain dynamics from MEG recordings in ASD and non-ASD individuals (García Domínguez et al., 2013). This method allowed us to estimate not only the functional connections at the sensor level but also the inputs driving the network. Functional connections account for the covariance and lagged correlations between signals recorded from different areas. Inputs reflect contributions to the variance of the recorded signals (outputs) that are not accounted for by the covariance with other signals in the network of sensors. Our results indicated that the dominant connectivity change in ASD relative to controls shows enhanced functional excitation between frontal and parietal/occipital areas. Moreover, the stochastic inputs driving the background activity in the resting state showed a greater spatial homogeneity in ASD than in control individuals, and indeed the spatial complexity of the background noise was significantly lower in ASD subjects. We speculated that higher long-range spatial correlations in the background noise may result from less specificity (or more promiscuity) of thalamo-cortical projections (García Domínguez et al., 2013). All these observations suggest that it may not be a matter of less connectivity in autism, but of changes in connectivity between specific areas as well as in the inputs. As a note of caution, one must bear in mind that in the aforementioned studies with MEG, PET, or fMRI the complex relation between macroscopic recordings and the underlying neuronal activity remains to a certain extent undetermined, so “connectivity” changes are to be understood in a functional rather than an anatomical or physiological sense.

The differences found in previous studies on brain coordination dynamics in ASD suggest that information processing/production could be different as well, for it is the coordinated activity of transiently formed neuronal assemblies that underlie information processing and cognition (Flohry, 1995; Bressler and Kelso, 2001; Kelso, 2008; Pérez Velázquez and Frantseva, 2011). Thus, in this study we investigated whether the production of information in periods of little sensory perturbation (resting state) could differ between individuals with and without ASD. As a measure for information production we used the Kullback-Leibler divergence (KLD) between the brain's inputs and outputs. The KLD is also known as information gain or relative entropy (Ihara, 1993) and quantifies differences between two distributions. In our case, the distributions are the probability density of the stochastic inputs driving the brain's activity and the probability density of the brain's activity itself, as recorded with MEG (outputs). We found an increased divergence in children with ASD compared to controls in the resting conditions in which the MEG recordings were taken, and conjecture that this enhanced information gain could be related to one of the most typical characteristics in autism as described already in the early days of autism research: the withdrawal into one's inner world.

## METHODS

### PARTICIPANTS AND MAGNETOENCEPHALOGRAPHIC RECORDINGS

Data were drawn from a larger sample of children enrolled in previous studies (Pérez Velázquez et al., 2009; Teitelbaum et al., 2012; García Domínguez et al., 2013). In total, MEG data from 19 children, 9 with Asperger's syndrome and 10 age-matched control

children, were analyzed. Age range was between 6 and 14 years for the controls (mean: 11.2 years; standard deviation: 2.6 years) and between 7 and 16 for ASD (mean: 10.8; standard deviation: 3.5). The 9 children with Asperger's syndrome were males while the 10 controls were 6 males and 4 females. We note, however, that boys and girls in the control group were not different from each other in terms of our analysis, as shown in our previous study (García Domínguez et al., 2013). The children's parents provided written consent for the protocol approved by the Hospital for Sick Children Research Ethics Board. Participants, who were evaluated by the psychologists in the Autism Research Unit of the Hospital for Sick Children or were recruited from the Geneva Center for Autism and Autism Ontario, met the criteria for ASD based on DSM-IV. Age-matched control children had no known neurological disorders.

MEG recordings were acquired at 625 Hz sampling rate, DC-100 Hz bandpass, third-order spatial gradient noise cancellation using a CTF Omega 151 channel whole head system (CTF Systems Inc., Port Coquitlam, Canada). Out of the 151 sensors, we discarded 10 that were not comparable across all patients due to artifacts or a very low signal-to-noise ratio. Our analysis thus focused on the recordings from the remaining 141 sensors in all patients. Subjects were tested supine inside the magnetically shielded room. Head movement was tracked by measuring the position of three head coils every 30 ms, located at the nasion, left and right ear, and movements less than 5 mm were considered acceptable. Children were instructed to remain at rest during the recording session that lasted between 30 and 60 s. To facilitate the involvement of the children in the experiment and minimize distraction, they were asked to press a button at will with their right hand a few times during the recording session. For each child, an epoch of 30 s was taken off for analysis of functional brain connectivity. All children were awake and had their eyes open during the experiment. Eye-blinking and muscular artifacts have a much larger amplitude than brain activity and are highly correlated across sensors, so they can be easily identified and removed using a well-established approach based on a principal component analysis (Mitra and Pesaran, 1999). In particular, since the artifacts appear in the first few principal components exclusively, they are efficiently cleaned out by removing those components.

### MODEL OF FUNCTIONAL CONNECTIVITY AND BACKGROUND NOISE

In the resting state, the non-linear dynamics of the brain reduces to noise-driven fluctuations around a state of equilibrium, which corresponds to a stable fixed point in neural-mass models of brain dynamics that include conduction delays, dendritic integration and non-linear firing characteristics of neurons (Robinson et al., 1998, 2001). The presence of background noise does not allow the system to quench at the fixed point but perturbs the system in a continuous manner, so that it fluctuates around the equilibrium (Galán, 2008). Thus, consistent with the approach used by several authors (Tononi et al., 1999; Sporns et al., 2000; Galán, 2008; Barnett et al., 2009; Steinke and Galán, 2011; García Domínguez et al., 2013), large-scale spontaneous brain activity is accurately described as a linear stochastic process that is formally equivalent to a mOUP.

$$x_i(t + dt) = x_i(t) + dt \sum_{j=1}^N W_{ij} x_j(t) + \eta_i(t + dt), \quad (1)$$

where  $W_{ij}$  is the functional connectivity matrix, i.e., the coupling between the  $j$ -th and the  $i$ -th nodes;  $x_i(t)$  is the neural activity of the  $i$ -th node with respect to baseline, measured as the signal from the  $i$ -th MEG channel at time  $t$ ;  $\eta_i$  are the residuals (background, uncorrelated white noise) of the  $i$ -th channel;  $N$  is the number of nodes (sensors) and  $dt$  is the sampling interval (1.6 ms). The sign of  $W_{ij}$  represents *functional* excitation (+) or inhibition (−) and should not be confused with excitatory or inhibitory synaptic connections at the cellular level. At the macroscopic level of MEG recordings, functional excitation and inhibition between nodes result from a combined effect of myriad processes, including multiple synaptic interactions and action potentials, which cannot be resolved. The units of  $W_{ij}$  are reciprocal of time, i.e., frequency units.

The functional connectivity matrix  $W_{ij}$  can be obtained from the empirical data  $x_i(t)$  with the Yule-Walker method for multivariate time series (Priestley, 2001). First, equation (1) is written in vector notation as

$$\vec{x}(t + dt) = \vec{x}(t) + W\vec{x}(t)dt + \vec{\eta}(t + dt). \quad (2)$$

Multiplying from the right by  $\vec{x}(t)^T$  and averaging in time, denoted by brackets  $\langle \dots \rangle$ , one has  $C_+ = (I + Wdt) C$ , with  $C_+ = \langle \vec{x}(t + dt)\vec{x}(t)^T \rangle$ ,  $C = \langle \vec{x}(t)\vec{x}(t)^T \rangle$ , and  $I$  being the identity matrix. After computing  $C$  and  $C_+$  from the recordings, the connectivity matrix is then given by

$$W = (C_+ - C) C^{-1} / dt,$$

where  $C^{-1}$  is the inverse of  $C$ , or its pseudo-inverse if it is rank-deficient. Once  $W$  has been determined, the background noise driving the network  $\eta_i(t)$  can also be obtained from (2), and their covariance is computed as  $Q = \langle \vec{\eta}(t)\vec{\eta}(t)^T \rangle$ . Note that the signals  $x_i(t)$  in the resting state have a stable mean (which is negligible relative to the standard deviation), as shown in **Figures 2A,B** for three arbitrary sensors. For system (2), the covariance matrices of the inputs and outputs are related via (Gardiner, 2004)

$$Q = -dt (W^T C + C W), \quad (3)$$

which allows one to compute  $Q$  directly from  $C$  and  $W$ . This provides a reality check for model (2): the closer the entries in  $Q$  are to the entries in matrix  $\langle \vec{\eta}(t)\vec{\eta}(t)^T \rangle$ , the more accurate is model (2). In our data set, the correlation coefficient between the entries in both matrices is  $r > 0.99$  (García Domínguez et al., 2013).

A multivariate Gaussian distribution of variable  $\vec{u} \in \mathbb{R}^N$  with mean  $\vec{m} = \langle \vec{u}(t) \rangle$ , and covariance  $\Sigma = \langle (\vec{u}(t) - \vec{m})(\vec{u}(t) - \vec{m})^T \rangle \in \mathbb{R}^{N \times N}$  is given by

$$G(\vec{u}; \vec{m}, \Sigma) = \frac{1}{(2\pi)^{N/2} |\Sigma|^{1/2}} \exp \left( -\frac{1}{2} (\vec{u} - \vec{m})^T \Sigma^{-1} (\vec{u} - \vec{m}) \right), \quad (4)$$

where  $|\dots|$  denotes the determinant of the matrix inside, or the pseudo-determinant, if the matrix is rank-deficient. For

mOUP, the stationary distributions of  $\vec{x}$  and  $\vec{\eta}$  are the multivariate Gaussians,  $G(\vec{x}; \vec{0}, C)$  and  $G(\vec{\eta}; \vec{0}, Q)$ , respectively.

## ENTROPY AND INFORMATION GAIN

We computed the entropy of the inputs as the entropy of the distribution of  $\vec{\eta}$  and the entropy of the output, as the entropy of the distribution of  $\vec{x}$ . To this end, we recall that the entropy of a multivariate Gaussian distribution (4) with zero mean is given by

$$\begin{aligned} H(\vec{u}) &= \int_{-\infty}^{\infty} G(\vec{u}; \vec{0}, \Sigma) \ln G(\vec{u}; \vec{0}, \Sigma) d\vec{u}^N = \frac{1}{2} \ln |2\pi e \Sigma| \\ &= \frac{N}{2} (1 + \ln(2\pi)) + \frac{1}{2} \ln |\Sigma|. \end{aligned} \quad (5)$$

So that the entropy of the inputs in (2) is  $H(\vec{\eta}) = 0.5 \cdot \ln |2\pi e Q|$  and the entropy of the outputs is  $H(\vec{x}) = 0.5 \cdot \ln |2\pi e C|$ .

The KLD of two distributions, also known as the relative entropy or information gain, measures how much variability of a stochastic variable  $\vec{u} \in \mathbb{R}^N$  with distribution  $P$  cannot be accounted for by a reference distribution  $Q$ . It is defined as

$$D(P||Q) = \int_{-\infty}^{\infty} P(\vec{u}) \ln \frac{P(\vec{u})}{Q(\vec{u})} d\vec{u}^N.$$

To determine the information gain of a mOUP we computed.

$$\begin{aligned} D(G(\vec{x}; \vec{0}, C) || G(\vec{\eta}; \vec{0}, Q)) &= \int_{-\infty}^{\infty} G(\vec{u}; C) \ln \frac{G(\vec{u}; \vec{0}, C)}{G(\vec{u}; \vec{0}, Q)} d\vec{u}^N \\ &= \frac{1}{2} \left( \text{trace}(Q^{-1}C) - \ln \frac{|C|}{|Q|} - N \right). \end{aligned} \quad (6)$$

The units of the outcome from expressions (5) and (6) are nats. We converted those values to bits by dividing by  $\ln(2)$ , and again by eight to obtain the final result in bytes.

## INVARIANCE OF INFORMATION GAIN

An important property of the information gain is that it is invariant under linear transformations. This implies that the “cross-talk” or mixing of independent source signals does not affect the information gain. In other words, the information gain measured at the sensor level is the same as the information gain at the source level. The mathematical proof is as follows. Recall that model (2) represents the signal model at the sensor level. Let  $U$  denote a linear transformation that “unmixes” the sensor level signals  $\vec{x}(t)$  to obtain the source level signals,  $\vec{y}(t)$ , so that,  $\vec{y}(t) = U\vec{x}(t)$ . In particular, matrix  $U$  can be computed with an independent component analysis. At the source level, model (2) is transformed into

$$\vec{y}(t + dt) = \vec{y}(t) + V\vec{y}(t)dt + \vec{\xi}(t + dt)$$

with  $V = UWU^{-1}$  and  $\vec{\xi}(t) = U\vec{\eta}(t)$ . The covariance matrix of  $\vec{y}(t)$  is then given by  $UCU^{-1}$  and the covariance matrix of  $\vec{\xi}(t)$  by  $UQU^{-1}$ . The information gain at the source level is thus

$$D_{\text{source}} = \frac{1}{2} \left( \text{trace}((UQU^{-1})^{-1} UCU^{-1}) - \ln \frac{|UCU^{-1}|}{|UQU^{-1}|} - N \right).$$



We first note that

$$\text{trace}\left((UQU^{-1})^{-1}UCU^{-1}\right) = \text{trace}(UQ^{-1}CU^{-1}) = \text{trace}(Q^{-1}C),$$

due to the invariance of the trace under similarity transformations. We also note that, since the determinant of the product is the product of the determinants one has

$$\frac{|UCU^{-1}|}{|UQU^{-1}|} = \frac{|U||C||U^{-1}|}{|U||Q||U^{-1}|} = \frac{|C|}{|Q|}.$$

Thus, the information gain at the source level becomes

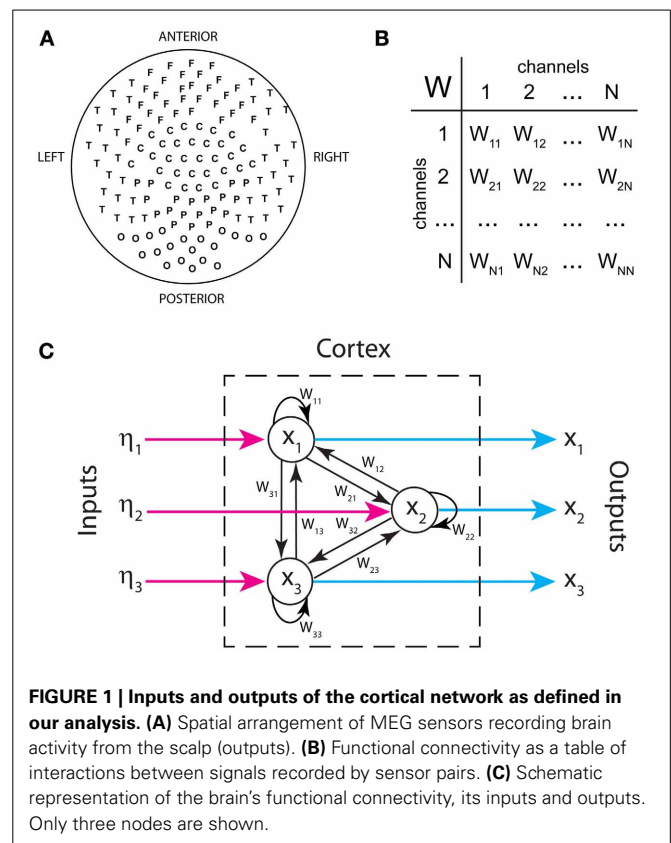
$$D_{\text{source}} = \frac{1}{2} \left( \text{trace}(Q^{-1}C) - \ln \frac{|C|}{|Q|} - N \right),$$

which is identical with the information gain at the sensor level (6).

## RESULTS

**Figure 1A** displays the arrangement of MEG sensors over the scalp. We only show the positions of the 141 out of 151 sensors that were used in all the subjects (as indicated in methods, 10 sensors were left out due to artifacts and/or low signal-to-noise ratios in different patients). Thus, the dimensions of the functional brain connectivity matrix for each subject are  $141 \times 141$ . The sensors cover the occipital (O), frontal (F), central (C), parietal (P), and temporal (T) areas. Each ordered pair of sensors  $(i, j)$  defines an entry in the connectivity matrix  $W_{ij}$  (**Figure 1B**), which is obtained from the data using model (1). Because MEG signals are most sensitive to cortical activity due to the pronounced decay of magnetic fields with distance, matrix  $W_{ij}$  mainly represents functional connections between cortical areas. A thorough analysis of the connectivity matrices and their differences in ASD was presented in our previous study (García Domínguez et al., 2013). Model (1) also allows one to obtain the inputs to the network,  $\eta_i(t)$  as explained in *Methods*. **Figure 1C** schematically shows the black-box interpretation of the brain dynamics described by equation (1), for just three nodes. The stochastic inputs (background noise),  $\eta_i(t)$  impinge on the nodes of the network, which in turn affect each other's activity rate,  $dx_i(t)/dt$  according to the connectivity matrix  $W_{ij}$ . This determines the instantaneous activity fluctuations (outputs) recorded from each node,  $x_i(t)$ . **Figure 2A** shows traces of ongoing activity recorded with three arbitrary sensors from one of the children. Only 3 seconds of the total recording (30 s) are shown. Traces  $x_1$  and  $x_2$  clearly display correlated fluctuations between them but not with  $x_3$ . **Figure 2B** shows the histograms of the fluctuations recorded from each of those three sensors. The fluctuations around the mean were normalized to the standard deviation of the traces so that the normalized amplitude is given by the z-score. Clearly, the fluctuations are normally distributed, as demonstrated by the excellent fit to a Gaussian (red line). The high  $p$ -values confirm the null hypothesis of the chi-square goodness-of-fit test, namely, that the fluctuations have a normal distribution in each sensor.

Model (1) assumes that the noise is additive and hence state independent. In such a case, the mean and variance of small segments of the time series should be independent of each other. To

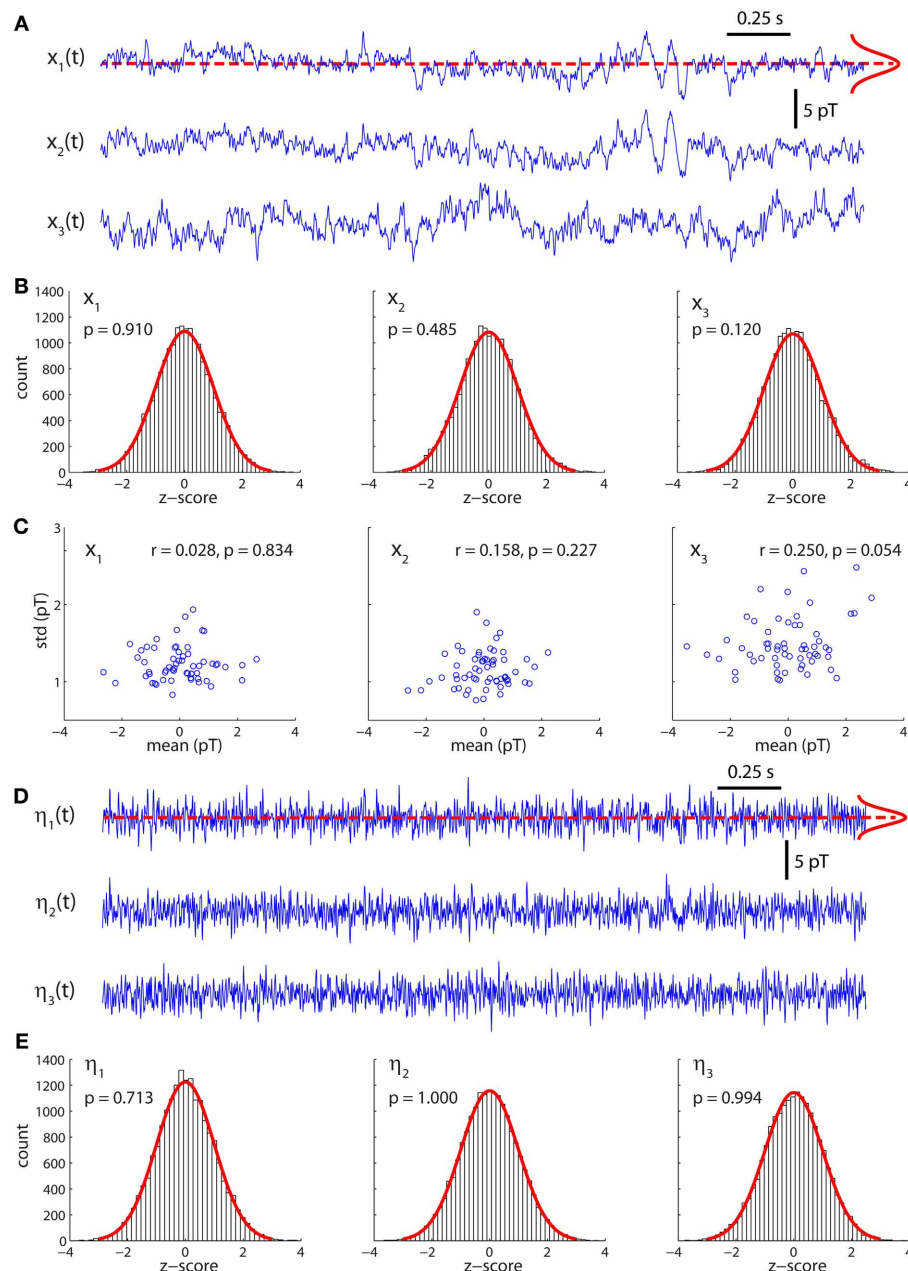


**FIGURE 1 | Inputs and outputs of the cortical network as defined in our analysis. (A)** Spatial arrangement of MEG sensors recording brain activity from the scalp (outputs). **(B)** Functional connectivity as a table of interactions between signals recorded by sensor pairs. **(C)** Schematic representation of the brain's functional connectivity, its inputs and outputs. Only three nodes are shown.

test this, we divided the traces in successive segments of 500 ms and plotted the mean over each segment against its standard deviation (**Figure 2C**). For all traces, the Pearson's correlation coefficient was not statistically significant, suggesting that both quantities are indeed independent of each other.

**Figure 2D** shows the stochastic inputs to the three nodes investigated above,  $\eta_1, \eta_2$  and  $\eta_3$ . Compared to the outputs in **Figure 2A**, the inputs display no significant temporal structure and lower amplitudes, which is what one would expect for the residuals of a parametric model, such as model (1). **Figure 2E** shows that the inputs are also normally distributed.

From the connectivity matrix,  $W_{ij}$  and the covariance matrix of the signals,  $C_{ij}$ , one can readily obtain the covariance matrix of all the inputs,  $Q_{ij}$ , using formula (3) in *Methods*, without having to determine them explicitly. This allows us to efficiently compute information theoretical measures. **Figure 3** shows the entropy of the inputs and outputs for the control and ASD groups. The entropy is larger for the outputs than for the inputs for both groups. However, the differences between both groups for inputs or outputs are not significantly different ( $p > 0.05$ , Wilcoxon rank-sum test). We note that the entropy values are negative. Indeed, while entropy values for discrete signals are non-negative, the entropy of continuous signals (differential entropy) may be negative. Negative entropy values result from expression (5), when  $|2\pi e\Sigma| < 1$ . Note that the value of this determinant depends on the units of the covariance, so our choice of those units affects the value of the entropy. Moreover, the entropy for continuous signals is very sensitive to their variance and because

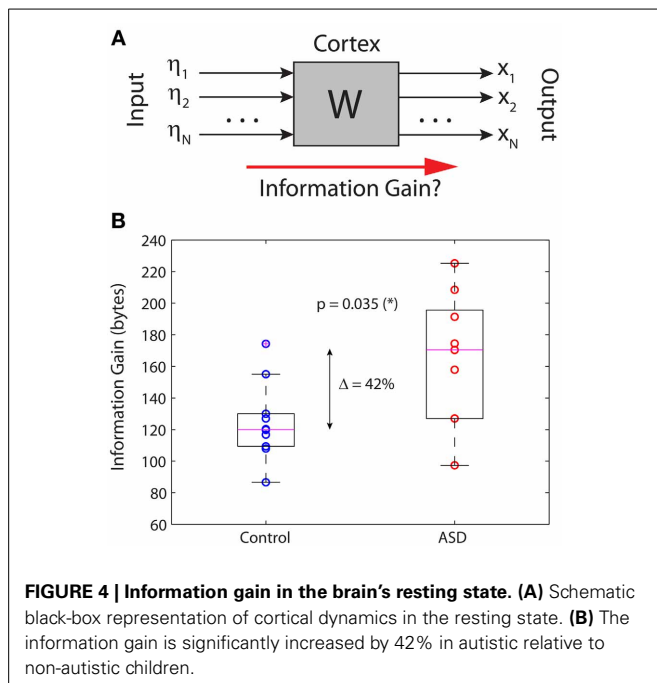
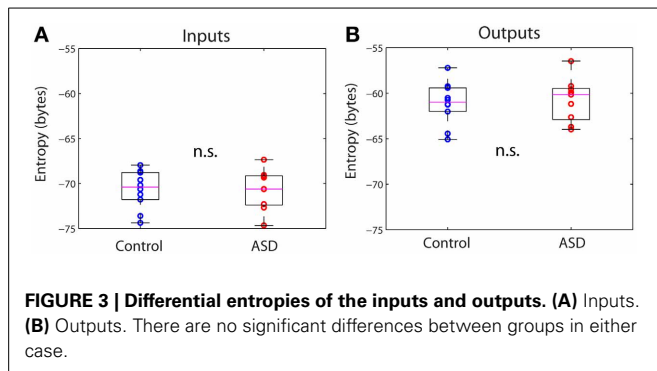


**FIGURE 2 | Activity fluctuations at rest are normally distributed. (A)** Recordings of ongoing activity (3 s long) from three arbitrary sensors in a control subject. **(B)** Activity fluctuations have zero mean and are normally distributed. The histograms were built from segments of 30 s. **(C)** Over

short segments of the time series (500 ms long) the mean and standard deviation are uncorrelated, consistently with the assumption of additive noise. **(D)** Residuals (inputs) of the model for the traces shown in **(A)**. **(E)** The residuals are also normally distributed.

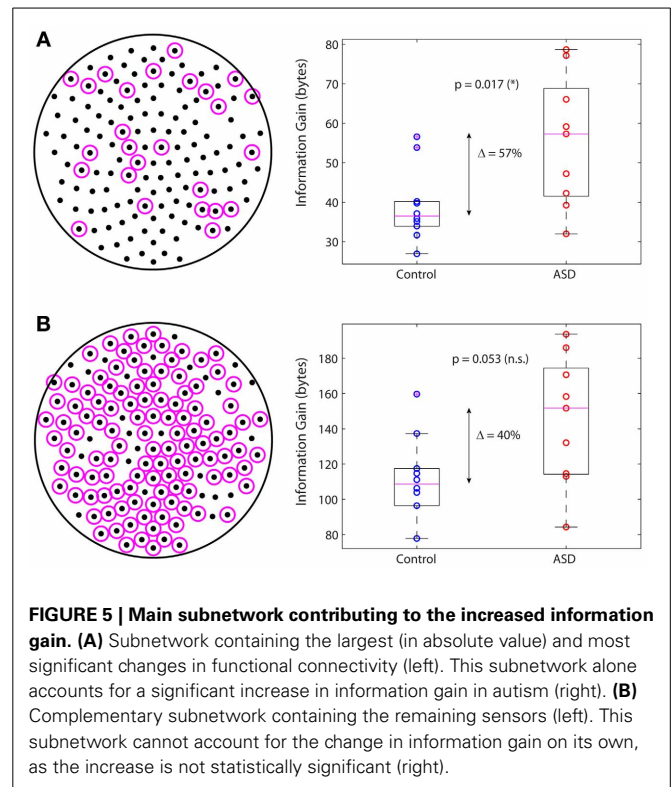
the amplitudes of the activity fluctuations are not significantly different between control and ASD (data not shown), neither are the entropies. These are well-known caveats that preclude the interpretation of entropy (or more accurately, differential entropy) as a measure of information content for continuous signals (Ihara, 1993). This contrasts with the case of discrete signals, for which entropy is legitimately interpreted as the expected value of information contained in a signal (Ihara, 1993).

A more relevant measure of information that has the same interpretation and properties for continuous and discrete signals is the relative entropy, or information gain, defined as the KLD between two distributions (see *Methods*). In lay terms, the KLD measures how much variability of a stochastic variable with distribution  $P$  cannot be accounted for by another stochastic variable with distribution  $P'$ . This interpretation justifies the alternative name of “information gain” about one variable by knowing the



other. In our context, we computed the KLD to *quantify the amount of information of the outputs that cannot be accounted for by the inputs*. In other words, we quantified how much information is “created” by the brain in the resting state. **Figure 4A** shows a simplified black-box interpretation of the brain, in a similar fashion to **Figure 1C** but for an arbitrary number of nodes and without paying attention to the details of the brain's network contained in the box. The key finding of this article is shown in **Figure 4B**, which plots the information gain of the brains in the control and ASD groups. Despite some overlap between the distributions, the medians of both groups are significantly different (Wilcoxon sum-rank test;  $p = 0.035$ ). In particular, the information gain in the ASD group is 42% larger on average, indicating that ASD brains produce more information from the stochastic inputs driving them.

In a previous study we identified the subnetwork of sensors containing the functional connections whose changes in autism are largest in absolute value and most significant relative to control (García Domínguez et al., 2013). We then asked whether this subnetwork on its own can account for the increased information



gain in autism. **Figure 5A** displays the sensors belonging to this subnetwork (left; magenta circles) and the information gain for this subnetwork in the control and ASD groups (right). The difference of the medians is 57% and it is statistically significant ( $p = 0.017$ ; Wilcoxon rank-sum test). In contrast, if one considers the complementary network, i.e., the other nodes in the sensors network (**Figure 5B**, left), the difference of the medians is 40% but not statistically significant, as it falls below the 95% confidence level ( $p = 0.053$ ; Wilcoxon rank-sum test). In conclusion, although all nodes contribute to the information gain, those nodes encompassing the interactions with the largest changes in autism contribute more to the increase in information gain. However, changes in connectivity alone are not sufficient to account for the difference in information gain that we observe in ASD, as the information gain depends not only on  $W$  via  $C$ , but also on matrix  $Q$ , which we know from our previous study that is also significantly different in ASD (García Domínguez et al., 2013). The question then is: do changes in  $W$  compensate for changes in  $Q$  or do these changes act synergistically to increase the information gain? Our analysis suggests the latter may be the correct answer, or at least, that changes in connectivity cannot fully compensate for changes in the inputs.

## DISCUSSION

The term autism (from the Greek *autos*, meaning “self”) was coined in 1911 by Swiss psychiatrist Eugen Bleuler, who used it to describe withdrawal into one's inner world (even though at this time he was referring to schizophrenia patients). Later, other studies defined more precisely the syndrome (Kanner, 1968). The neurophysiological reasons responsible for a certain detachment from

the environment of individuals with ASD remain unknown, and several scholars have proposed ideas mostly centered on the psychological level of description. Whereas much brain structural and genetic studies are being done in autism research, the investigation of the brain dynamics is lagging considerably behind. Here, we have explored what the background brain activity in resting conditions (when individuals are not presented with specific sensory stimuli) may reveal about the inner processing of the brain in terms of information production, quantified as relative entropy. Our analysis of MEG signals recorded at rest indicated that the brains of individuals with ASD, Asperger syndrome in this case, produce more information than the age-matched participants with a 42% increase on average. These significant differences cannot be attributed to the gender-ratio mismatch in our cohort. Although there were 6 males and 4 females in the control group, and no females in the ASD group, the control group was fairly homogeneous: there were no significant differences in the information gain between the boys and girls within the control group ( $p = 0.76$ ; Wilcoxon rank-sum test).

We decided to focus on spontaneous brain activity in resting conditions because the fundamental importance of the ongoing, “noisy” nervous system activity is widely recognized today, and a more in-depth investigation of brain activity in periods of minimal sensory perturbation has been advised by several scholars as it may provide the best opportunity to study the intrinsic connectivity of the brain, in the absence of major sensory perturbations (Galán, 2008; Ringach, 2009; Steinke and Galán, 2011; García Domínguez et al., 2013; Papo, 2013). From an analytical perspective, there are two important reasons for investigating brain activity in the resting state. The first one is that in this case, the brain dynamics are described by stochastic model (1), which implies that functional connections ( $W_{ij}$ ) are constant, in contrast to the stimulated brain, in which interactions between different areas are state-dependent and typically non-linear. The second reason is that in the brain’s resting state the stochastic inputs,  $\eta_i$ , as well as the activity fluctuations,  $x_i$  (outputs) are normally distributed. Thus, the distributions of  $\tilde{\eta}$  and  $\tilde{x}$  are both  $N$ -dimensional Gaussians. This enables an accurate parametric estimation of the entropies and relative entropy, as shown above. If the fluctuations are not normally distributed, as it is frequently the case for stimulus-evoked activity, a parametric estimation of information theoretical measures is in general not possible. To compute entropies and related quantities in such cases, one needs to estimate the probability densities of the data. However, the estimation of high-dimensional probability densities requires very large datasets, which are virtually impossible to collect in current experimental settings.

In our study, functional connections between areas and their inputs are defined operationally from model (1): functional connections account for the covariance and lagged cross-correlations between signals recorded from different areas, whereas the inputs are defined as contributions to the variance that are not accounted for by the covariance with other signals in the network of sensors. Neither the functional connections nor the inputs represent specific neuronal elements, although they obviously emanate from them in a complex, undetermined manner. Certainly, a

multiscale modeling approach, from single cells to neural mass models, is worth attempting. This is, however, a daunting task, as recognized by other authors working on this problem (Deco et al., 2008).

There are two important considerations about the dynamical model used in our study: (1) the suitability of a linear model for large-scale brain dynamics *in the resting state*; and (2) the interpretation of the inputs in the model. As for the first consideration, we note that there is no contradiction between our stochastic linear model and the fact that brain dynamics are strongly non-linear because we do not intend to model neuronal dynamics *per se*. We are rather modeling the recorded signals, which are magnetic fields that do superimpose linearly. An analog dichotomy takes place in weather forecasting: although the dynamics of air masses are turbulent, chaotic and therefore, unpredictable, when considered over a large area the flow of air masses becomes predictable within a time window of a few days. These coarse dynamics of air masses fit very well a linear multivariate stochastic process, which can then be used to accurately forecast variations and co-variations of air pressure and temperature at different locations (Storch and Zwiers, 2001). Similarly, in neural mass models the strong non-linear dynamics of single neurons, when averaged over a fairly large spatial range, display fluctuations around a mean that make a stochastic linear model suitable for the description of large-scale activity. As noted by Nunez and Srinivasan, “the question of brain linearity depends on context and the level [...] addressed [...]. It is only in mathematics that a sharp distinction exists between linear and non-linear system” (Nunez and Srinivasan, 2006). We also note that non-linear neuronal networks, like those based on the celebrated Wilson-Cowan model and the neural-mass models frequently possess hyperbolic fixed points which are linearly stable. The brain’s resting state we record from corresponds to this kind of stable state, as shown in **Figures 2A,B**, in which baseline activity is characterized by fluctuations around a fixed mean. Indeed, linearization of neural-mass dynamics around a hyperbolic fixed point leads to model (2) when stochastic perturbations are included. Several recent papers have taken advantage of this fact to investigate the link between connectivity and spontaneous activity patterns in a neural network model (Galán, 2008; Barnett et al., 2009; Steinke and Galán, 2011; García Domínguez et al., 2013). Outside the resting state, during sensory stimulation, brain activity typically has a moving baseline, or low frequency modulation of the fluctuations, which results from the non-linear regime of the neural-mass dynamics, and therefore, it is inconsistent with model (2). That is the reason why our model should only be applied to brain activity in the resting state.

As for the interpretation of the inputs in our model, we remark that subcortical structures relay inputs to the cortex and probably more (if these could be quantified) than those from the external sensorium, which with the exception of the olfactory system are filtered through the thalamus. Our model considers both sources of fluctuating inputs together: those coming from the external world and those from internal organs are similar for our purposes because the other organs are, after all, external to the brain too, so they are all just inputs. Regarding this matter of



differentiating internal vs. external inputs, we find the thoughts by Nachev and Husain quite appealing; in their words “the contrast between internally and externally-generated actions is empirically intractable” (Nachev and Husain, 2010).

Large-scale recordings, such as MEG traces, have some limitations to keep in mind (Gross et al., 2013). The signals detected by MEG reflect population-scale levels of activity in large neuronal networks. Insights gained from the analysis of MEG data are limited to coarse relationships between large populations of cells rather than the detailed understanding of interactions between individual cells. Moreover, spontaneous activity at any given sensor may contain activity from multiple distributed sources, and conversely, the activity of a single signal source can introduce coordinated changes at multiple sensors. For these reasons, functional connectivity estimated from signals recorded by the sensors does not necessarily reflect the actual connectivity between the brain areas next to where the sensors are located. Thus, a distinction between sensor-level and source-level connectivity is pertinent to MEG but also to all technologies for measuring large-scale brain connectivity that are currently available. Ideally, connectivity analysis should be performed at the source level. However, source reconstruction clearly adds another level of complexity to the analysis and may even yield spurious results, as it is an ill-posed mathematical problem (Gross et al., 2013). This implies that assumptions must be made about the origin and location of the sources in order to properly constrain the solution to the problem. Whereas certain assumptions may be reasonable for stimulus-driven experiments because specific sensory or motor areas are expected to be strongly activated, this is not trivial for ongoing activity where no specific areas are expected to dominate the brain dynamics. Importantly, we have shown here (see *Methods*) that the information gain in the brain's resting state is the same for the source and sensor levels. Thus, the results we report here are unaffected by any possible cross-talk between sensors or mixing of independent source signals at the sensor level.

When addressing queries on information processing in nervous systems, the question of what is meant by “information” always arises. There are several different notions about what information is and represents, and depending on the research field, e.g., thermodynamics, cybernetics, information theory etc., one may come across different definitions. In general terms, however, the concept of information refers to the ability of a given signal to encode a message with a presumed alphabet regardless of its content. That is, the information is agnostic to semantics or meaning. In plain mathematical terms, the information gain used in this study is nothing but a measure of the global differences between the distributions of the input to and output of the brain in its resting state. It therefore quantifies the *degree of transformation of the inputs into the outputs*. Because this transformation is made by the brain's network, the information gain can literally be regarded as the amount of information created by the brain which is not already present in the inputs.

On a more philosophical level, the general expression “brain information processing” is commonly used without specific details as to what this information is, but it serves the purpose

as it relies on certain intuitive knowledge that neuroscientists share and accept. If, as Heinz von Foerster declared, “information is a relative concept that assumes meaning only when related to the cognitive structure of the observer (the recipient)” (Von Foerster, 2003), and the activity of the brain cellular circuits is roughly considered as the production of “novel” associations between stimuli (external or internal), then perhaps an increase in the difference between the stochastic input and output, as found in our work, could conceivably be associated with a more pronounced “mental inner life” that, roughly speaking, may result in the common detachment of individuals with ASD from their environment. Perhaps a bit more specifically, following Davies' recent postulate of two types of information in biological systems (Davies et al., 2013), structural and functional, it could be reasoned that in the nervous system the structural information derived from direct anatomical connections between cells is responsible for the maintenance of memory and other specific aspects that need to be maintained in a stable manner, whereas functional information, which is what we measured in our studies, could be related to the rate of cell assembly formation, to the transient establishment of coordinated activity amongst brain cell networks which is the basis of cognition (Bressler and Kelso, 2001; Kelso, 2008; Pérez Velázquez and Frantseva, 2011). As a predecessor of the current conceptualization, Hans Flohr already proposed almost two decades ago that the rate of cell assembly formation determines cognition (Flohr, 1995). A precise investigation of how cell assemblies form and disappear and the relation between these ephemeral brain functional networks and cognitive/psychological aspects is difficult to achieve with current methods in brain and cognitive science. Nevertheless, these types of research encompassing biophysics and psychological observations, we venture, will be a fundamental part of the immediate future of neuroscience research. In fact, with the current theoretical conceptualization of nervous system dynamics based on dynamical bifurcations that switch brain/cognitive states in a flexible manner, it is not surprising that more investigations on the role of background activity in brain information processing are being conducted at several levels of description (Liljenstrom, 1996; Mcmillen and Kopell, 2003; Pérez Velázquez et al., 2007; Zhou et al., 2010; Luczak et al., 2013).

Combining several empirical observations, the picture that emerges is that a tendency toward enhanced excitatory activity in the cell circuitry in the autistic brain (Rubenstein and Merzenich, 2003; Han et al., 2012) results in hyperactivity in certain brain regions (García Domínguez et al., 2013) that in turn enhances the tendency toward increased synchronous activity in those areas, e.g., parietal cortices (Pérez Velázquez et al., 2009; Teitelbaum et al., 2012), which is reflected in greater spatial correlation in the background activity (García Domínguez et al., 2013) and in a more pronounced information production from the background activity, as found in this study. More generally, these related tendencies toward more than normal excitation and synchronization could underlie most of neurological and psychiatric disorders (Pérez Velázquez and Frantseva, 2011; Yizhar et al., 2011). All these neurophysiological differences between autistic

and non-autistic brains, we propose, could contribute on the behavioral level to the known withdrawal to their inner world of individuals with autism. While, at this stage, this is a conjecture, it is perhaps useful to start the never easy attempt of framing neurophysiological data into psychological aspects. Our study is intended as an initial step in the investigation of how information generation in the brain relates to cognitive/psychological aspects and our results allow us the following speculations. It is noteworthy that the subnetwork of sensors that significantly contributes to the increased information gain in autism (as shown in **Figure 5A**) contains a combination of frontal, temporal and parietal areas which also correspond to the default mode network; the brain areas that reduced their activations during processing of external stimuli and are preferentially active when individuals do not focus on the external world (Buckner et al., 2008). Moreover, this subnetwork contains a number of midline sensors: medial frontal, central and parietal (**Figure 5A**). Remarkably, both the default network and midline brain structures have been proposed to be fundamental regions for self-processing (Northoff and Bermpohl, 2004), and there are numerous studies that reported the association of activation in parietal and medial frontal cortex in self-referential processing (Lou et al., 2004; D'argembeau et al., 2007). Nevertheless, it should be considered that each brain area is "activated" by other connected nets, which means that these regions proposed in the literature, while significantly associated with self-referential processing, receive inputs and integrate their activity with others possibly subcortical areas (Northoff et al., 2011). It is also of interest that distinct patterns of synchronization in the "default areas" have been noted (Fingelkurts, 2011), especially an increase in phase synchrony when subjects attention is internally focused (Kirschner et al., 2012). These previously reported neurophysiological phenomena in those brain areas may contribute to the observed differences between the two groups in the information gain reported in this work, and particularly the higher information gain in the ASD group could therefore be related to the more intense "inner world" that autistic individuals normally have.

Future studies may consider applying our method to other cognitive phenotypes as well. To interpret information gain in other contexts one must bear in mind that it explicitly depends on the inputs and outputs of the resting state network, and implicitly (via the output covariance) on the functional connectivity. Significant changes in information gain must therefore result from changes in at least one of these measures or, as it is the case in our study, in all the three measures. One may then ask whether changes in connectivity tend to compensate for changes in inputs so that the information gain is barely altered, or whether those changes act synergistically to exacerbate alterations in neuronal activity and information gain. Finally, to more explicitly address the relation between information gain and particular psychological traits, it is worth noting that people with schizophrenia are characterized by excessive self-awareness (Frith, 1979), which taken to the limit may lead to hallucinations. We surmise that if our analysis of the brain's resting state were conducted on people with schizophrenia, it would also show a significant increase in information gain that reflects the ability of the brain to generate complex activity on its own, even in the absence of significant stimulation.

## ACKNOWLEDGMENTS

This work was partially supported by a Discovery grant from the Natural Sciences and Engineering Research Council of Canada, NSERC (José L. Pérez Velázquez) and by a scholarship of The Mt. Sinai Health Care Foundation (Roberto F. Galán).

## REFERENCES

- Barnett, L., Buckley, C. L., and Bullock, S. (2009). Neural complexity and structural connectivity. *Phys. Rev. E* 79, 051914. doi: 10.1103/PhysRevE.79.051914
- Bressler, S. L., and Kelso, J. A. (2001). Cortical coordination dynamics and cognition. *Trends Cogn. Sci.* 5, 26–36. doi: 10.1016/S1364-6613(00)01564-3
- Buckner, R. L., Andrews-Hanna, J. R., and Schacter, D. L. (2008). The brain's default network: anatomy, function, and relevance to disease. *Ann. N.Y. Acad. Sci.* 1124, 1–38. doi: 10.1196/annals.1440.011
- D'argembeau, A., Ruby, P., Collette, F., Degueldre, C., Baetens, E., Luxen, A., et al. (2007). Distinct regions of the medial prefrontal cortex are associated with self-referential processing and perspective taking. *J. Cogn. Neurosci.* 19, 935–944. doi: 10.1162/jocn.2007.19.6.935
- Davies, P. C., Rieper, E., and Tuszynski, J. A. (2013). Self-organization and entropy reduction in a living cell. *Biosystems* 111, 1–10. doi: 10.1016/j.biosystems.2012.10.005
- Deco, G., Jirsa, V. K., Robinson, P. A., Breakspear, M., and Friston, K. (2008). The dynamic brain: from spiking neurons to neural masses and cortical fields. *PLoS Comput. Biol.* 4:e1000092. doi: 10.1371/journal.pcbi.1000092
- Fingelkurts, A. A. (2011). Persistent operational synchrony within brain default-mode network and self-processing operations in healthy subjects. *Brain Cogn.* 75, 79–90. doi: 10.1016/j.bandc.2010.11.015
- Flohr, H. (1995). Sensations and brain processes. *Behav. Brain Res.* 71, 157–161. doi: 10.1016/0166-4328(95)00033-X
- Frith, C. D. (1979). Consciousness, information processing and schizophrenia. *Br. J. Psychiatry* 134, 225–235. doi: 10.1192/bjp.134.3.225
- Galán, R. F. (2008). On how network architecture determines the dominant patterns of spontaneous neural activity. *PLoS ONE* 3:e2148. doi: 10.1371/journal.pone.0002148
- García Domínguez, L., Pérez Velázquez, J. L., and Galán, R. F. (2013). A model of functional brain connectivity and background noise as a biomarker for cognitive phenotypes: application to autism. *PLoS ONE* 8:e61493. doi: 10.1371/journal.pone.0061493
- Gardiner, C. W. (2004). *Handbook of Stochastic Methods for Physics, Chemistry, and the Natural Sciences*. Berlin: Springer.
- Gross, J., Baillet, S., Barnes, G. R., Henson, R. N., Hillebrand, A., Jensen, O., et al. (2013). Good practice for conducting and reporting MEG research. *Neuroimage* 65, 349–363. doi: 10.1016/j.neuroimage.2012.10.001
- Han, S., Tai, C., Westenbroek, R. E., Yu, F. H., Cheah, C. S., Potter, G. B., et al. (2012). Autistic-like behaviour in Scn1a<sup>+/−</sup> mice and rescue by enhanced GABA-mediated neurotransmission. *Nature* 489, 385–390. doi: 10.1038/nature11356
- Herbert, M. R. (2005). Large brains in autism: the challenge of pervasive abnormality. *Neuroscientist* 11, 417–440. doi: 10.1177/0091270005278866
- Ihara, S. (1993). *Information Theory for Continuous Systems*. Singapore: River Edge: World Scientific.
- Kanner, L. (1968). Autistic disturbances of affective contact. *Acta Paedopsychiatr.* 35, 100–136.
- Kelso, J. A. (2008). An essay on understanding the mind. *Ecol. Psychol.* 20, 180–208. doi: 10.1080/10407410801949297
- Kennedy, D. P., and Courchesne, E. (2008). The intrinsic functional organization of the brain is altered in autism. *Neuroimage* 39, 1877–1885. doi: 10.1016/j.neuroimage.2007.10.052
- Kirschner, A., Kam, J. W., Handy, T. C., and Ward, L. M. (2012). Differential synchronization in default and task-specific networks of the human brain. *Front. Hum. Neurosci.* 6:139. doi: 10.3389/fnhum.2012.00139
- Liljenstrom, H. (1996). Global effects of fluctuations in neural information processing. *Int. J. Neural Syst.* 7, 497–505. doi: 10.1142/S0129065796000488
- Lou, H. C., Luber, B., Crupain, M., Keenan, J. P., Nowak, M., Kjaer, T. W., et al. (2004). Parietal cortex and representation of the mental Self. *Proc. Natl. Acad. Sci. U.S.A.* 101, 6827–6832. doi: 10.1073/pnas.0400049101
- Luczak, A., Bartho, P., and Harris, K. D. (2013). Gating of sensory input by spontaneous cortical activity. *J. Neurosci.* 33, 1684–1695. doi: 10.1523/JNEUROSCI.2928-12.2013

- Mcmillen, D., and Kopell, N. (2003). Noise-stabilized long-distance synchronization in populations of model neurons. *J. Comput. Neurosci.* 15, 143–157. doi: 10.1023/A:1025860724292
- Mitra, P. P., and Pesaran, B. (1999). Analysis of dynamic brain imaging data. *Biophys. J.* 76, 691–708. doi: 10.1016/S0006-3495(99)77236-X
- Monk, C. S., Peltier, S. J., Wiggins, J. L., Weng, S. J., Carrasco, M., Risi, S., et al. (2009). Abnormalities of intrinsic functional connectivity in autism spectrum disorders. *Neuroimage* 47, 764–772. doi: 10.1016/j.neuroimage.2009.04.069
- Morgan, J. T., Chana, G., Abramson, I., Semendeferi, K., Courchesne, E., and Everall, I. P. (2012). Abnormal microglial-neuronal spatial organization in the dorsolateral prefrontal cortex in autism. *Brain Res.* 1456, 72–81. doi: 10.1016/j.brainres.2012.03.036
- Murias, M., Webb, S. J., Greenson, J., and Dawson, G. (2007). Resting state cortical connectivity reflected in EEG coherence in individuals with autism. *Biol. Psychiatry* 62, 270–273. doi: 10.1016/j.biopsych.2006.11.012
- Nachev, P., and Husain, M. (2010). Action and the fallacy of the ‘internal’: comment on Passingham et al. *Trends Cogn. Sci.* 14, 192–193. author reply: 193–194. doi: 10.1016/j.tics.2010.03.002
- Northoff, G., and Bermpohl, F. (2004). Cortical midline structures and the self. *Trends Cogn. Sci.* 8, 102–107. doi: 10.1016/j.tics.2004.01.004
- Northoff, G., Qin, P., and Feinberg, T. E. (2011). Brain imaging of the self—conceptual, anatomical and methodological issues. *Conscious. Cogn.* 20, 52–63. doi: 10.1016/j.concog.2010.09.011
- Nunez, P. L., and Srinivasan, R. (2006). *Electric Fields of the Brain: the Neurophysics of EEG*. Oxford; New York: Oxford University Press. doi: 10.1093/acprof:oso/9780195050387.001.0001
- Papo, D. (2013). Why should cognitive neuroscientists study the brain’s resting state? *Front. Hum. Neurosci.* 7:45. doi: 10.3389/fnhum.2013.00045
- Pérez Velázquez, J. L., Barceló, F., Hung, Y., Leshchenko, Y., Nenadovic, V., Belkas, J., et al. (2009). Decreased brain coordinated activity in autism spectrum disorders during executive tasks: reduced long-range synchronization in the fronto-parietal networks. *Int. J. Psychophysiol.* 73, 341–349. doi: 10.1016/j.ijpsycho.2009.05.009
- Pérez Velázquez, J. L., and Frantseva, M. V. (2011). *The Brain-Behavior Continuum. The Subtle Transition Between Sanity and Insanity*. Singapore: World Scientific Publishing. doi: 10.1142/8088
- Pérez Velázquez, J. L., García Domínguez, L., and Guevara Erra, R. (2007). Fluctuations in neuronal synchronization in brain activity correlate with the subjective experience of visual recognition. *J. Biol. Phys.* 33, 49–59. doi: 10.1007/s10867-007-9041-4
- Priestley, M. B. (2001). *Spectral Analysis and Time Series*. San Diego, CA: Academic Press.
- Ringach, D. L. (2009). Spontaneous and driven cortical activity: implications for computation. *Curr. Opin. Neurobiol.* 19, 439–444. doi: 10.1016/j.conb.2009.07.005
- Robinson, P. A., Rennie, C. J., Wright, J. J., Bahramali, H., Gordon, E., and Rowe, D. L. (2001). Prediction of electroencephalographic spectra from neurophysiology. *Phys. Rev. E* 63, 021903. doi: 10.1103/PhysRevE.63.021903
- Robinson, P. A., Rennie, C. J., Wright, J. J., and Bourke, P. D. (1998). Steady states and global dynamics of electrical activity in the cerebral cortex. *Phys. Rev. E* 58, 3557–3571. doi: 10.1103/PhysRevE.58.3557
- Rubenstein, J. L., and Merzenich, M. M. (2003). Model of autism: increased ratio of excitation/inhibition in key neural systems. *Genes Brain Behav.* 2, 255–267. doi: 10.1034/j.1601-183X.2003.00037.x
- Sireteanu, R., Oertel, V., Mohr, H., Linden, D., and Singer, W. (2008). Graphical illustration and functional neuroimaging of visual hallucinations during prolonged blindfolding: a comparison to visual imagery. *Perception* 37, 1805–1821. doi: 10.1068/p6034
- Sporns, O., Tononi, G., and Edelman, G. M. (2000). Connectivity and complexity: the relationship between neuroanatomy and brain dynamics. *Neural Netw.* 13, 909–922. doi: 10.1016/S0893-6080(00)00053-8
- Steinke, G. K., and Galán, R. F. (2011). Brain rhythms reveal a hierarchical network organization. *PLoS Comput. Biol.* 7:e1002207. doi: 10.1371/journal.pcbi.1002207
- Storch, H. V., and Zwiers, F. W. (2001). *Statistical Analysis in Climate Research*. Cambridge; New York: Cambridge University Press.
- Teitelbaum, A., Belkas, J., Brian, J., and Pérez Velázquez, J. L. (2012). “Distinct patterns of cortical coordinated activity in autism,” in *Autism Spectrum Disorders: New Research*, eds C. E. Richardson and R. A. Wood (Hauppauge, NY: Nova Publishers), 95–112.
- Thai, N. J., Longe, O., and Rippon, G. (2009). Disconnected brains: what is the role of fMRI in connectivity research? *Int. J. Psychophysiol.* 73, 27–32. doi: 10.1016/j.ijpsycho.2008.12.015
- Tononi, G., Sporns, O., and Edelman, G. M. (1999). Measures of degeneracy and redundancy in biological networks. *Proc. Natl. Acad. Sci. U.S.A.* 96, 3257–3262. doi: 10.1073/pnas.96.6.3257
- Travers, B. G., Adluru, N., Ennis, C., Tromp Do, P. M., Destiche, D., Doran, S., et al. (2012). Diffusion tensor imaging in autism spectrum disorder: a review. *Autism Res.* 5, 289–313. doi: 10.1002/aur.1243
- Tsiaras, V., Simos, P. G., Rezaie, R., Sheth, B. R., Garyfallidis, E., Castillo, E. M., et al. (2011). Extracting biomarkers of autism from MEG resting-state functional connectivity networks. *Comput. Biol. Med.* 41, 1166–1177. doi: 10.1016/j.combiomed.2011.04.004
- Von Foerster, H. (2003). *Understanding Understanding: Essays on Cybernetics and Cognition*. New York, NY: Springer.
- Yizhar, O., Fenno, L. E., Prigge, M., Schneider, F., Davidson, T. J., O’shea, D. J., et al. (2011). Neocortical excitation/inhibition balance in information processing and social dysfunction. *Nature* 477, 171–178. doi: 10.1038/nature10360
- Zhou, Y., Wang, K., Liu, Y., Song, M., Song, S. W., and Jiang, T. (2010). Spontaneous brain activity observed with functional magnetic resonance imaging as a potential biomarker in neuropsychiatric disorders. *Cogn. Neurodyn.* 4, 275–294. doi: 10.1007/s11571-010-9126-9

**Conflict of Interest Statement:** The authors declare that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.

Received: 14 May 2013; accepted: 05 December 2013; published online: 24 December 2013.

Citation: Pérez Velázquez JL and Galán RF (2013) Information gain in the brain’s resting state: A new perspective on autism. *Front. Neuroinform.* 7:37. doi: 10.3389/fninf.2013.00037

This article was submitted to the journal *Frontiers in Neuroinformatics*.

Copyright © 2013 Pérez Velázquez and Galán. This is an open-access article distributed under the terms of the Creative Commons Attribution License (CC BY). The use, distribution or reproduction in other forums is permitted, provided the original author(s) or licensor are credited and that the original publication in this journal is cited, in accordance with accepted academic practice. No use, distribution or reproduction is permitted which does not comply with these terms.



# Mutual information spectrum for selection of event-related spatial components. Application to eloquent motor cortex mapping

Alexei Ossadtchi<sup>1,2,3\*</sup>, Platon Pronko<sup>4</sup>, Sylvain Baillet<sup>5</sup>, Mark E. Pflieger<sup>6</sup> and Tatiana Stroganova<sup>7</sup>

<sup>1</sup> Department of Higher Nervous Activity and Psychophysiology, St. Petersburg State University, St. Petersburg, Russia

<sup>2</sup> Laboratory of Decision Making, National Research University Higher School of Economics, Moscow, Russia

<sup>3</sup> Complex Systems Control Laboratory, Institute for Problems of Mechanical Engineering, RAS, St. Petersburg, Russia

<sup>4</sup> Laboratory of Cognitive Psychophysiology, National Research University Higher School of Economics, Moscow, Russia

<sup>5</sup> Dynamic Neuroimaging Laboratory, McConnell Brain Imaging Centre, Montreal Neurological Institute, McGill University, Montreal, QC, Canada

<sup>6</sup> Source Signal Imaging Inc., San Diego, CA, USA

<sup>7</sup> MEG Centre, Moscow State University of Psychology and Education, Moscow, Russia

## Edited by:

Daniele Marinazzo, University of Gent, Belgium

## Reviewed by:

Alberto Mazzoni, Italian Institute of Technology, Italy

Javier Escudero, University of Edinburgh, UK

## \*Correspondence:

Alexei Ossadtchi, Department of Higher Nervous Activity and Psychophysiology, St. Petersburg State University, 7-9, Universitetskaya nab., St. Petersburg 199034, Russia  
e-mail: a.ossadtchi@spbu.ru

Spatial component analysis is often used to explore multidimensional time series data whose sources cannot be measured directly. Several methods may be used to decompose the data into a set of spatial components with temporal loadings. Component selection is of crucial importance, and should be supported by objective criteria. In some applications, the use of a well defined component selection criterion may provide for automation of the analysis. In this paper we describe a novel approach for ranking of spatial components calculated from the EEG or MEG data recorded within evoked response paradigm. Our method is called Mutual Information (MI) Spectrum and is based on gauging the amount of MI of spatial component temporal loadings with a synthetically created reference signal. We also describe the appropriate randomization based statistical assessment scheme that can be used for selection of components with statistically significant amount of MI. Using simulated data with realistic trial to trial variations and SNR corresponding to the real recordings we demonstrate the superior performance characteristics of the described MI based measure as compared to a more conventionally used power driven gauge. We also demonstrate the application of the MI Spectrum for the selection of task-related independent components from real MEG data. We show that the MI spectrum allows to identify task-related components reliably in a consistent fashion, yielding stable results even from a small number of trials. We conclude that the proposed method fits naturally the information driven nature of ICA and can be used for routine and automatic ranking of independent components calculated from the functional neuroimaging data collected within event-related paradigms.

**Keywords: spatial components, ICA, SVD, components selection, mutual information, eloquent cortex mapping**

## 1. INTRODUCTION

Spatial decomposition is one of the key techniques applied to exploratory analysis of multichannel data in general, and to spatial-temporal electro- and magnetoencephalographic (EMEG) signals in particular. The most commonly used methods to obtain both spatial components and the corresponding temporal loadings are independent component analysis (ICA) (Comon, 1994), principal component analysis (PCA) (Golub and Van Loan, 1996) and factor analysis (FA) (Child, 2006).

The most frequently used approach for analysis of stimulus-locked averaged EMEG data is PCA, which can be performed using the singular value decomposition (SVD) of the stimulus-locked averaged data matrix (Lagerlund et al., 1997). This analysis is followed by thresholding the singular values (SV) spectrum to identify the subspace capturing the largest amount of data variance for a given approximation rank (Vandewalle, 1988). This technique is inherently power driven. Its application to the

identification of the repetitive task-related signal subspace from the averaged ERP/F data relies on the assumption that the individual evoked responses are sufficiently well phase-locked to the stimulus. In that case, the stimulus-locked summation results in an enhanced relative power of the phase-locked component (Misulis, 1994).

SVD is the optimal method for signal subspace detection measured by subspace correlation for a given approximation rank (Vandewalle, 1988). However, the actual value of signal subspace rank,  $R$ , is, in general, unknown. Finding an estimate of  $R$  is not a trivial task. It is often done by visual inspection of the SV spectrum. The method is based on identifying the target index,  $R$ , of a singular component just preceding a sharp drop in power, followed by a slow decaying plateau in the SV spectrum. However, a large disparity of activation amplitudes, spatial proximity of the neuronal sources and powerful noise sources may result in the absence of a clear cut division between task-related and noise



components. In addition, in realistic conditions, the recordings are often contaminated by spatially colored brain activity and/or spatially coherent artifacts. Under these circumstances, component selection based on the SV spectrum may be misleading. As a motivating example, consider the top panel of **Figure 1** that shows the SV spectrum calculated for the averaged data obtained from the simulated MEG timeseries containing the contribution of two non-synchronous dipolar sources. Although the subspace spanned by the first  $\hat{R} = 2$  singular topographies almost exactly matches the true subspace ( $\text{subcorr}([\mathbf{a}_1 \ \mathbf{a}_2], [\mathbf{u}_1 \ \mathbf{u}_2]) = [1, 0.987]$ ) the spectrum of SVs fails to provide evidence that the second component contains task-related signal.

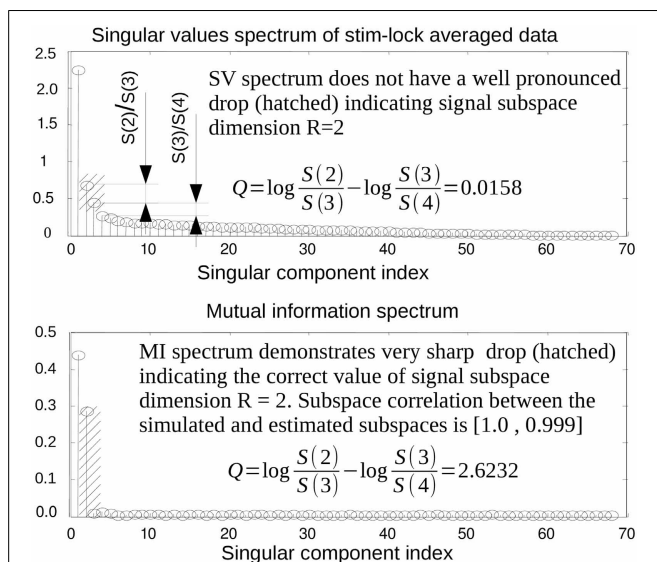
ICA is one of the most widely used approaches to blind source separation, popular for exploratory analysis of multidimensional data. In the analysis of EMEG data from evoked response experiments, this decomposition may be used both to isolate task-related components (Makeig et al., 1996; Vigario et al., 2000) and to remove artifacts (Jung et al., 2000). It can also be used for estimation of source timeseries when proper forward modeling is unavailable and for the estimation of the signal subspace in cases when the experimental paradigm can not guarantee sufficiently

accurate stimulus locking, e.g., in voluntary movement paradigm (Ossadtchi et al., 2000; Delorme, 2010).

The application of ICA to routine analysis of EMEG datasets is limited by the absence of standard approaches for ordering the independent components (Hyvärinen et al., 2001). In the most typical scenario a human observer visually identifies the desired components by exploring their timecourses and topographies. Since the raw EMEG data are often very noisy, it can be difficult to determine which components should be selected. Additionally, such manual selection is often daunting, and selection based on power (Delorme, 2010) or stim-locked averaged power (AP) (Hyvärinen et al., 2001) does not ensure that the components are event-related for reasons essentially similar to those just described for the SV spectrum. It should be also noted that such power-driven ordering methods may be inappropriate, since ICA as an information (rather than power) driven technique. An alternative method, using the correlation metrics between each estimated component and the event trigger, is critically dependent on signal shapes, and is therefore highly unreliable. For completeness we will mention that for some methods of blind source separation, such as AMUSE (Tong et al., 1991), components may have an intrinsic order but such an ordering is not very useful in the context of analysis of EMEG data from ERP studies. These problems hinder efficient utilization of ICA for batch-mode processing of EMEG datasets, and affect the objectivity of the results obtained with manual analysis.

The independent components sorting problem has received considerable attention in the fMRI data analysis literature. Gu et al. (2001) and Esposito et al. (2002) have introduced methods for component ordering based on spatial characteristics. Lu and Rajapakase (2003) suggested ranking based on component timecourse kurtosis. Himberg et al. (2004) used clustering of a succession of ICA realizations to select relevant components. Yang et al. (2008) describes a method for components selection based on the reproducibility principle. In the application of ICA to EEG and MEG, a technique based on measuring the amount of spatial component variance explained by the electromagnetic model was proposed by Grosse-Wentrup and Buss (2008). However, an accurate forward model is required to fully benefit from this approach.

In the current paper, we present a novel mutual information (MI) based approach for ICA components sorting. Moritz et al. (2003) has described a somewhat related method for component ranking, based on the spectral manifestation of stimulus periodicity. However, the periodicity assumption is not always fulfilled, especially in voluntary movement paradigms. In addition, the spectral measure uses only first and second order statistical moments, while our MI-based method implicitly employs higher order moments for estimating the amount of task-related signal present in a component. We report the performance of our new MI based approach and compare it against more conventionally used AP driven technique. Additionally, we demonstrate an application of the MI Spectrum to sorting InfoMax ICA components obtained from real MEG recordings obtained from an experiment designed to non-invasively map primary motor cortex (M1 zone).



**FIGURE 1 | In some practical cases, the first singular topographies remain reasonably good estimators of signal subspace but inspection of the SV spectrum fails to reveal this as illustrated by a model example here.** While the subspace spanned by the first two singular topographies and the actual simulated subspace practically coincide the SV spectrum (**top** panel) fails to reveal the fact that the second singular component also belongs to the signal subspace. On the contrary the MI spectrum (**bottom** panel) computed using raw data projected onto the left singular vectors demonstrates a very clearly cut separation of the task-related and task unrelated subspaces. This figure also introduces the measure of task-related subspace identifiability used in the paper. Since the correct signal subspace rank value is  $R = 2$ , we use discriminating indicator  $q = \log \left( \frac{S(2)}{S(3)} \right) - \log \left( \frac{S(3)}{S(4)} \right)$  that formalizes the strategy employed by the human observers and estimates the amount of drop between the second and the third components referenced to the ratio of the two largest components of the noise range spectrum (with indices 3 and 4) immediately following the two signal components (with indices 1 and 2).

## 2. MATERIALS AND METHODS

### 2.1. EMEG SIGNAL MODEL AND PRELIMINARIES

EMEG data recorded by a  $K$ — sensor array during the  $i$ -th repetition of a neuromotor or cognitive task can be written as the following linear combination

$$\mathbf{x}^i(t) = [\mathbf{a}_1, \dots, \mathbf{a}_R] \begin{bmatrix} f_1^i(t) \\ \vdots \\ f_R^i(t) \end{bmatrix} + [\mathbf{b}_1, \dots, \mathbf{b}_L] \begin{bmatrix} p_1^i(t) \\ \vdots \\ p_L^i(t) \end{bmatrix} + \mathbf{n}(t) \quad (1)$$

For the  $i$ —th epoch, a multichannel signal at each instance of time,  $\mathbf{x}^i(t)$  is a noisy additive mixture of source topographies  $[\mathbf{a}_1, \dots, \mathbf{a}_R]$  weighted by the corresponding stimulus-locked activation timeseries  $[f_1^i(t), \dots, f_R^i(t)]$ , along with a task-unrelated contribution from sources with topographies  $[\mathbf{b}_1, \dots, \mathbf{b}_L]$  activated with task-unrelated timeseries  $[p_1^i(t), \dots, p_L^i(t)]$ , and a random noise vector  $\mathbf{n}(t)$ . Topographies of task related sources form an  $R$ -dimensional signal subspace and topographies of task unrelated sources form an  $L$ -dimensional coherent interference subspace. ERP experiments are usually accompanied by a binary stimulus signal  $s(t)$  marking the task onset. In neuro-motor experiments this binary signal may be derived from the myographic activity record or from the accelerometer signal, using a thresholding procedure. Usually the goal of data analysis is to identify the task related signals and extract the task-related signal subspace to be used subsequently for neuronal source localization. For completeness, we may include induced sources whose activation power is locked to the task-onset moment with random phase. However, since we are interested in analysis of ERP's (which are phase-locked by definition), we do not include the induced component in (1).

Under the ideal and largely unrealistic conditions when activations  $f_r^i(t)$  are exactly reproducible across trials, time locked to the stimulus, and spatially coherent task-unrelated components are absent, the identification of task related components can be done optimally using the SVD of the stimulus locked average data matrix  $\bar{\mathbf{X}} = [\bar{\mathbf{x}}(0), \dots, \bar{\mathbf{x}}(T)]$ , where  $\bar{\mathbf{x}}(t) = \sum_{i=1}^M \mathbf{x}_i(t)$ ,  $M$  is the task repetitions count, and  $T$  is the interval of interest duration (Vandewalle, 1988). The SVD yields the averaged data matrix decomposition  $\bar{\mathbf{X}} = \mathbf{U}\mathbf{S}\mathbf{V}^T$ . Columns of the orthonormal matrix  $\mathbf{U}$  are the singular topographies,  $\mathbf{S}$  is a diagonal matrix of SVs, and columns of the right singular matrix  $\mathbf{V}$  are the singular activations. Task related components are chosen to be the first  $\hat{R}$  components ranked by power.  $\hat{R}$  is determined typically by visual analysis of the SV spectrum. Optionally, the SV spectrum of random matrix may be used as a reference in this task (Golub and Van Loan, 1996).

ICA is usually applied to the raw (unaveraged) spatial-temporal matrix  $\mathbf{X}(t)$  and yields a spatial unmixing matrix  $\mathbf{B}$  and a collection of independent components  $z_i(t)$  obtained as  $\mathbf{z}(t) = \mathbf{B}\mathbf{X}(t)$ ,  $\mathbf{z}(t) = [z_1(t), \dots, z_K(t)]^T$ . We assume that some of these components contain task-related signal and the others do not. In correspondence to  $\mathbf{B}$  we can put matrix  $\mathbf{F} = \mathbf{B}^{-1}$  so that  $\mathbf{X}(t) = \mathbf{F}\mathbf{z}(t)$ . Columns of  $\mathbf{F}$  are called independent topographies and describe the profiles formed by the corresponding independent “sources” on the sensors.

### 2.2. MUTUAL INFORMATION SPECTRUM

We propose to assess the degree to which the  $i$ -th component  $z_i(t)$  is related to the task using the normalized MI spectrum, computed as  $I_i = I(z_i(t), e(t))$ .  $e(t)$  is the expanded stimulus line signal, computed by convolution of the original binary stimulus signal  $s(t)$  with expansion kernel  $k(t)$  as  $e(t) = s(t) * k(t)$  to produce monotonic variations over the interval of interest around each event onset moment. In this work we used a centered (i.e., symmetric around the x-axis) ramp function as the expansion kernel  $k(t)$ .

We used a simple scaled histogram method to compute the MI as the difference between the entropy of an independent component  $z_i(t)$  and its entropy conditioned on the expanded stimulus line signal  $e(t)$ , i.e.,

$$I_0(z_i(t), e(t)) = H(z_i(t)) - H(z_i(t)|e(t)) \quad (2)$$

where  $H(u)$  denotes the entropy of  $u$ .

As suggested by Strehl (2002), we use the geometric mean of the two marginal entropy values to obtain the normalized MI quantities as

$$I_i = I(z_i(t), e(t)) = \frac{I_0(z_i(t), e(t))}{\sqrt{H(z_i(t))H(e(t))}} \quad (3)$$

We then define the MI spectrum as the rank ordered elements  $I_i : \{I_i \geq I_{i+1}\}$ .

As with the more conventional SV spectrum, visual analysis of the MI spectrum can be used to estimate the signal subspace rank. Originally suggested by us in Ossadtchi et al. (2000), this measure of MI with the expanded stimulus signal is a power-invariant way to assess the degree of task-relatedness of raw (unaveraged) timeseries.

We have introduced the notion of MI spectrum in the context of ICA component selection. This method can also be used for ranking singular components obtained via SVD of the stimulus locked average data matrix. To compute the MI spectrum for such singular components, first project the raw unaveraged data matrix  $\mathbf{X}(t)$  onto the left singular column vectors of  $\mathbf{U}$  as  $\mathbf{z}(t) = \mathbf{U}^T\mathbf{X}(t)$  and then apply the MI spectrum calculation procedure as described above.

In neuro-motor tasks, the EMG signal,  $m(t)$ , can be recorded and used instead of the expanded stimulus line. Then the MI spectrum is calculated as  $I_i = I(z_i(t), m(t))$ . Since the EMG signal usually occupies a broader spectrum than that of EEG or MEG signals, it is beneficial to perform a zero-phase-shift band-pass filter to remove excessively low and high frequency components prior to computing the MI.

### 2.3. STATISTICAL TESTING

In automated applications, and for a more informed decisions during the visual analysis of the MI spectrum, we suggest the following randomization testing scheme to estimate the  $p$ -values to reject the null-hypothesis that a component is not task related.

The suggested scheme is based on the observation that for signal components that contain a statistically significant evoked response, the value of the MI is directly related to the consistent

correspondence (not similarity!) between the shape of this component and the expanded stimulus signal. Therefore, when the actual task onset moments are randomized, this correspondence will be destroyed. The MI values for the task-related components will experience a significant drop, while those that pertain to the task-unrelated components will remain in the original range.

We suggest the following simple steps to generate surrogate data and assess statistical significance of the observed MI values. In what follows  $M$  is number of independent components,  $J$ -number of randomization iterations,  $N_t$  is the number of samples in the stimulus signal  $s(t)$  and  $N_r = \sum_{t=1}^{N_t} s(t)$  is the number of task repetitions.

1. for  $j = 1:J$ 
  1. Create a new, surrogate stimulus signal  $s^*(t)$  by randomizing task onset moments:  
 $s^*(t) = 0, \forall t \in [1, N_t];$   
for  $k = 1 : N_r, t \leftarrow U(0, N_t), s^*(t) = 1, \text{end}_k$
  2. Calculate surrogate expanded stimulus signal:  
 $e^*(t) = s^*(t) * k(t)$
  3. Calculate the amount of normalized MI of all the components with this surrogate expanded stimulus:  
for  $i = 1 : M, I_{ij}^* = I(e^*(t), z_i(t)), \text{end}_i$
2. end<sub>j</sub>

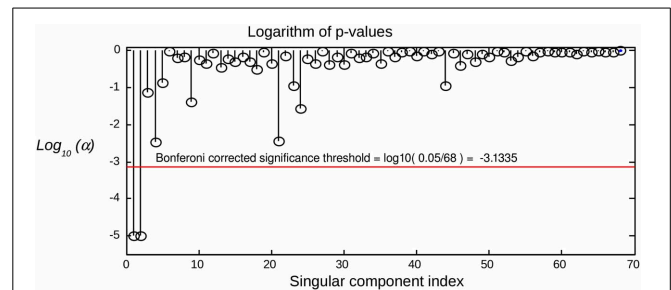
The  $i$ -th row of  $I_{ij}^*$  measures the MI for the  $i$ -th component and the  $j$ -th randomization of the stimulus onset signal. In order to calculate the  $p$ -values of the null-hypothesis that the  $i$ -th component does not contain task-related signal, we compute the fraction of surrogate values  $I_{ij}^*, j \in [1, J]$  that exceed the actual observed value  $I_i$ . If we define a logical function  $L(a, b) = 1$  if  $(a > b)$  and  $L(a, b) = 0$  otherwise, then  $p$ -values for the null-hypothesis that the  $i$ -th component contains no task related signal can be expressed as  $p_i = \frac{1}{M} \sum_{j=1}^J L(I_{ij}^*, I_i)$ .

## 2.4. MULTIPLE COMPARISON CORRECTION

Since we are testing several hypotheses, we need to correct the calculated  $p$ -values for multiple comparisons. Simple Bonferroni correction is appropriate, since the components are independent or at least orthogonal (SVD case). We therefore conclude that a component can be considered as task-related at the significance level of  $\alpha$  if  $p_i < \frac{\alpha}{N_c}$ , where  $N_c$  is the total number of components tested. An example of applying the suggested statistical testing scheme is illustrated in **Figure 2**, where the proposed procedure allows for the correct identification of task related components in a simulated scenario with  $R = 2$  task-related sources with powerful, spatially coherent, interference. For this and subsequent simulations, we used the procedure described in the following section.

## 2.5. SIMULATION PROCEDURE

In order to illustrate the performance of the MI spectrum and compare it with a more standard power driven approach, we performed realistic simulations with the following procedure. To simulate the observed sensor signals, we used Equation (1). We obtained a realistic source configuration from analysis of a



**FIGURE 2 | Raw components'  $\log(p)$ -values computed using the suggested randomization scheme.** The horizontal line corresponds to Bonferroni corrected threshold determined for  $\alpha = 0.05$ . The data were simulated with two dipolar sources and powerful spatially coherent interference. SVD was applied to the stimulus-locked data matrix that yielded first singular topographies. Subspace correlation of the first two singular topographies with the true signal subspace was  $[1, 0.987]$ . We can see that the suggested randomization scheme is able to correctly detect the first two components that span the signal subspace.

somatosensory MEG dataset recorded with a 67 channel CTF MEG system. We applied the RAP-MUSIC localization algorithm to  $[0-200 \text{ ms}]$  range of the stimulus-locked average data and obtained  $R = 2$  dipoles with topographies  $\mathbf{a}_1$  and  $\mathbf{a}_2$  and their corresponding activations  $f_1(t)$  and  $f_2(t)$  [see Equation (1)]. We used these dipoles as sources of task related activity in our simulations. To simulate task related activation timeseries we adapted a kernel-based model of evoked potentials described by Lange et al. (1997). This model includes random trial-to-trial variation in the latency and amplitudes of signal components. It is based on the decomposition of activation timeseries into a superposition of Gaussian kernels with varying amplitudes and delays. The model is justified by the fact that, given relatively poor spatial resolution of MEG, the dipole timeseries may be viewed as the sum of activations of several neuronal assemblies, each with different intensity and activation latency values. A graphical example of such a decomposition is shown in **Figure 3**. The simulated activation of the  $r$ -th dipole during the  $i$ -th epoch can be expressed formally as  $f_r^i(t) = \sum_{k=1}^K \beta_k^r v_k^r(t - \theta_k^r)$  with  $K$  kernels defined as

$$v_k^r(t) = f_r(t) \frac{e^{-\frac{(t-\tau_k^r)^2}{2\sigma_k^2}}}{\sum_{l=1}^K e^{-\frac{(t-\tau_l^r)^2}{2\sigma_l^2}}}, \quad r \in [1, R], \quad k \in [1, K] \quad (4)$$

The model incorporates random variables  $\beta_k, \theta_k, k = \{1, \dots, K\}$  representing amplitude and latency variations. The latency jitter values were independent for all components and were generated using a Gaussian random variable with mean of 50 ms and standard deviation 10 ms. The  $k$ -th kernel amplitude variation  $\beta_k$  was modeled as normally distributed random variable with mean of unity and standard deviation equal to 0.2.

We modeled brain noise with  $L = 1000$  spatially coherent, task-unrelated cerebral sources whose locations and time series varied with each realization. The corresponding topographies  $\mathbf{b}_l$

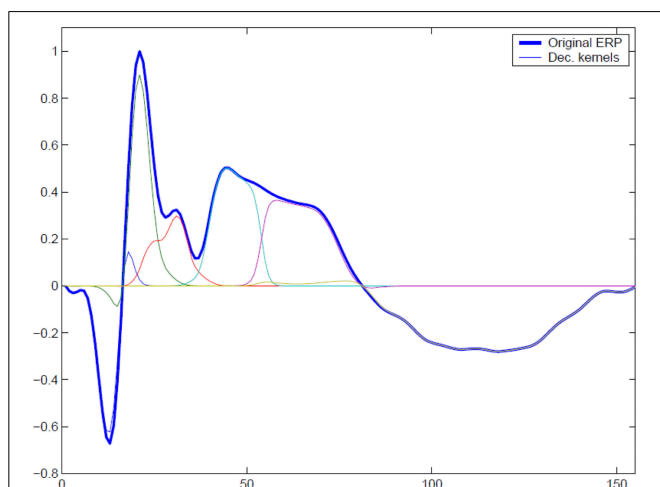
were calculated using locally fitted concentric spheres MEG forward model as implemented in EMSE Software Suite, Source Signal Imaging Inc., San Diego, CA, USA. The activation time series were narrow-band signals obtained via zero-phase filtering of realizations of Gaussian (pseudo)random process by the fifth order band-pass IIR filters in the bands corresponding to theta (4–7 Hz), alpha (8–12 Hz), beta (15–30 Hz) and gamma (30–50 and 50–70 Hz) activity. Their relative contributions were scaled in accordance with  $\frac{1}{f}$  characteristic of the realistic EMEG spectrum. An additional narrow-band alpha-component (9–11 Hz) of occipital origin ( $[-0.05, 0.01, 0.06]$  in EMSE coordinate system) was also included. We scaled the brain noise components to match typical signal-to-noise ratio of real-life recordings.

### 3. RESULTS

#### 3.1. DISCRIMINATIVE POWER OF THE MI SPECTRUM

Consider the situation when a task-related signal is generated by a pair of dipolar sources. When a pair of sources has highly correlated topographies or, in case of a large imbalance in source magnitude, the second singular component may be obscured and may not produce a pronounced SV distinguishable from the baseline. In this case, analysis of the SV spectrum will fail to provide the correct estimate of the signal subspace dimension. An example is illustrated in **Figure 1**, where the SV spectrum of the averaged data matrix obtained from a simulated dataset with  $R = 2$  task-related sources does not have a significant drop between  $R = 2$  and 3. On the contrary, the MI spectrum exhibits a very clear separation between task-related and task-unrelated parts, as it can be seen in the bottom panel of **Figure 1**.

In order to perform a more systematic evaluation of using MI to measure the extent to which a component is task-related, we performed a set of simulations with two dipolar sources in the presence of realistic brain noise. We varied the ratio of activation amplitudes of the two dipoles, performed SVD of the averaged



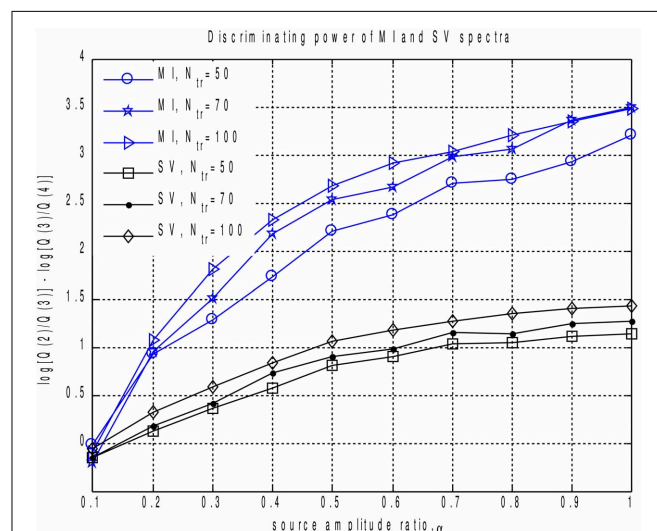
**FIGURE 3 |** To simulate trial-to-trial variation of the responses we used overlapping Gaussian kernel based model. At each trial we varied relative positions of the kernel centers, kernel amplitudes, and global response latency with respect to the binary stimulus signal. A typical response and its representation with a set of overlapping modulated kernels is shown.

data matrix and calculated the MI spectrum for the projections of continuous data onto the left singular vectors. We then compared the discriminating power of the MI and the SV spectra. To do so we introduced the discriminating indicator  $q$ . Since the correct rank value is  $R = 2$  we used  $q = \log \left( \frac{S(2)}{S(3)} \right) - \log \left( \frac{S(3)}{S(4)} \right)$ , see **Figure 1**.  $q$  is sensitive to the drop between the second and the third components, referenced to the ratio of the two largest noise range spectrum values (with indices 3 and 4) immediately following the two signal components (with indices 1 and 2). Results are shown in **Figure 4**, illustrating the discriminating indicator  $q$  as a function of source amplitudes ratio. We performed this numerical experiment for a varying number of trials in a simulated dataset. We found that the proposed MI spectrum outperforms the SV spectrum for all trial counts, and also provides for a clearer separation between the task-related and task-unrelated components. Also note that in most cases the correlation of the subspace spanned by the first two singular topographies and the true signal subspace spanned by  $\mathbf{a}_1$  and  $\mathbf{a}_2$  was sufficiently high to be considered as a correct estimate of the simulated signal subspace.

#### 3.2. RECEIVER OPERATING CHARACTERISTICS OF MI

In this section we describe our experiments on exploring receiver operating characteristics (ROC) of the MI metrics. We consider the task of discriminating between the components that contain task-related signal and those that do not. Spatial components  $z_i(t)$  obtained from signals that can be represented using Equation (1) can be viewed as the superposition of signal and noise, expressed as

$$z(t) = af(t) + \sigma_p p(t) + \sigma_n n(t), \quad (5)$$



**FIGURE 4 |** Discriminating power indicator  $q$  as a function of source amplitudes ratio for different number of trials calculated for MI and SV spectra. Each curve corresponds to a fixed datapoints count. We can see that the proposed MI based measure outperforms power based technique and produces a clearer cut between the task-related and task-unrelated components.



where  $f(t)$  is task-related source activity of amplitude  $a$ .  $p(t)$  and  $n(t)$  are the contributions from spatially coherent and spatially white noise sources, with standard deviation values  $\sigma_p$  and  $\sigma_n$  respectively.

We simulated repetitions of the task-related signal, including the jitter and variations characteristic of realistic brain signals. For each of  $N_{mc} = 1000$  Monte-Carlo realizations, we simulated  $N = 100$  signals according to Equation (5).  $N_1 = 10$  out of 100 signals had  $a = 1$  and the remaining  $N_0 = 90$  signals had  $a = 0$ , i.e., no task-related signal present. We simulated brain noise as described in the Simulations Procedure section. The goal was to detect the components that contain task related signals. We compared the MI values against the more traditionally used stimulus-locked averaged signal power (AP), calculated as

$$P_i = \sum_{t \in W} \bar{z}_i^2(t), \quad (6)$$

where  $\bar{z}_i(t)$  is the stimulus-lock averaged  $i$ -th signal. The summation was performed over a 200 ms window centered on the stimulus. We used the same windows to calculate both MI and AP measures.

To calculate the ROC curves, we applied thresholding to the MI and AP spectra separately, and marked as detected only those components whose corresponding MI or AP values exceeded the threshold. The threshold was originally chosen to be 0.05 of the largest value in the spectrum (AP or MI). In order to obtain the ROC curve we calculated the sensitivity  $p_{sens}(\theta) = \frac{NTP}{N_1}$  and specificity  $p_{spec}(\theta) = 1 - \frac{NFP}{N_0}$  for a succession of evenly spaced threshold values  $\theta = 0.05 k \max_i(I_i)$  or  $\theta = 0.05 k \max_i(P_i)$  for  $k = 1, \dots, 19$ .

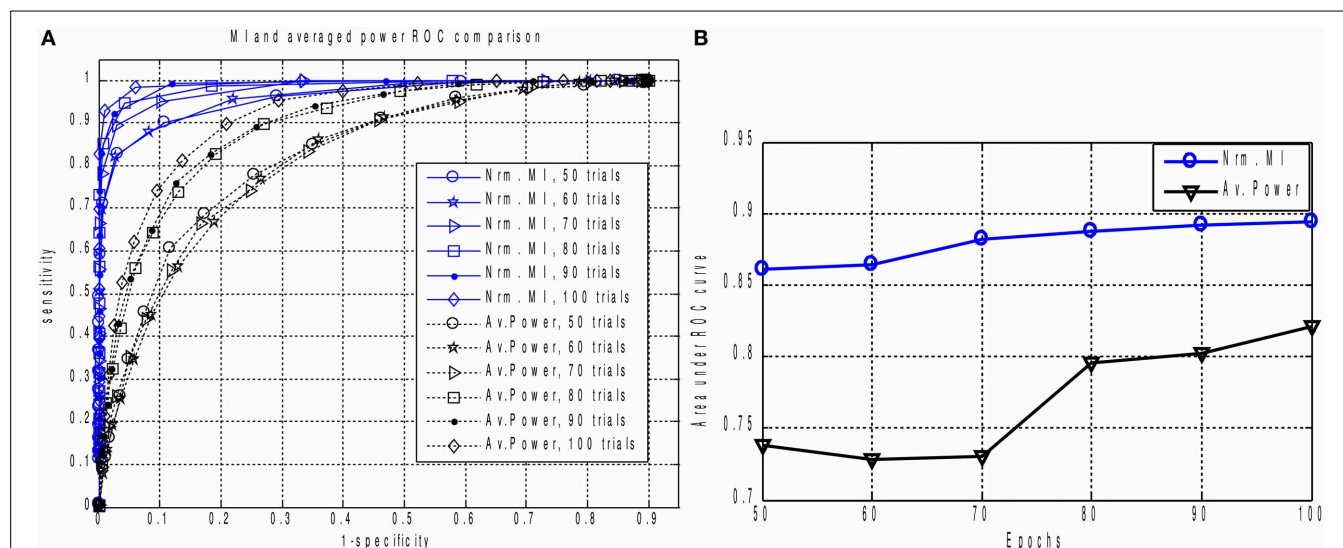
The result is shown in **Figures 5A,B**. For all epoch counts, the MI based measure significantly outperforms the power-based characteristic, and also provides better sensitivity for any selected specificity. We have observed similar behavior when dealing with real MEG data, as described below in the “MI versus AP for small epoch counts in real data” section. The improved ROC may be explained by the fact that the MI based measure implicitly takes into account higher order statistical information as compared to the power based approach, where only the first and second order statistical moments are used.

### 3.3. APPLICATION TO M1 MAPPING

Reliable mapping of the primary motor cortex (M1) based on functional neuroimaging provides an important complement to the use of structural data alone. However, since various zones forming the somatosensor complex appear to be in a coupled interaction even in the motor planning stage, the localization of M1 zone from the functional EEG and MEG data via standard approaches is problematic and often does not yield reliable results (Sanders et al., 1996; Gerloff et al., 1998).

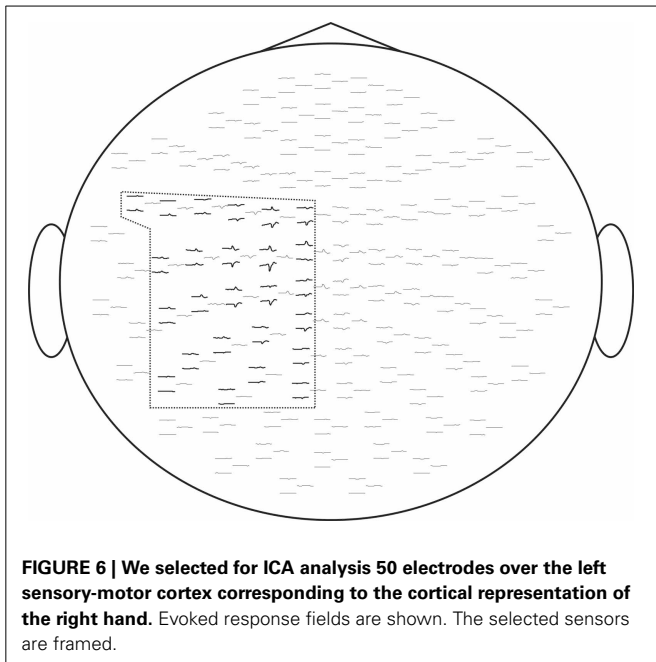
Inspired by the work of Riehle (2005) we explored the possibility of using the information from activation timecourse morphology and looked for spatial properties of activations with sharp non-linear increase just preceding the movement onset. To do so we studied MEG-recorded brain responses during a voluntary index finger movement task performed by 18 healthy right-handed volunteers.

For computational feasibility we used a subset of 50 sensors located over the left sensory-motor region. These were selected based on the grand-average responses, as shown in **Figure 6**. Recordings from all experimental sessions in all subjects were



**FIGURE 5 | (A)** The family of ROC curves for Averaged Response Power and Mutual Information based detection for various number of epochs on the same plot. MI based detection clearly and significantly outperforms the conventional method. Even with 50 trials the MI based criterion (“Nrm. MI, 50 trials” curve) allows to achieve 70 percent of sensitivity with ideal specificity. We can see that for all counts of epochs the MI based measure

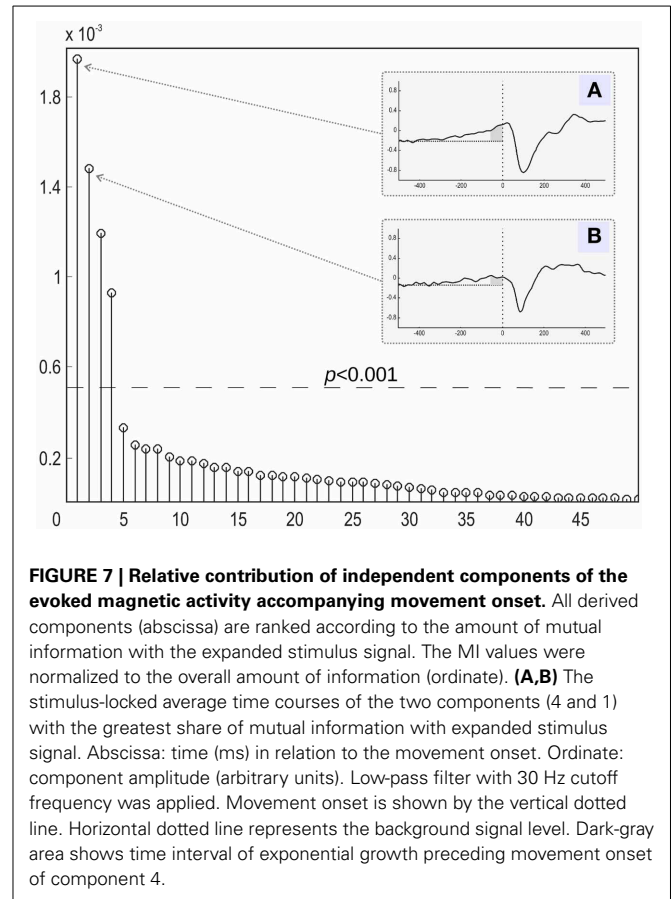
significantly outperforms the power based characteristic. This can be explained by the fact that in calculation of MI we implicitly take into account higher order information as compared to the power based approach where only the first two statistical moments are used. **(B)** Area under ROC curve performance characteristic for Averaged Response Power and Mutual Information.



concatenated into a single sequence and decomposed using the InfoMax-ICA approach. We obtained 50 independent components and ranked them according to the amount of MI with the expanded stimulus signal  $e(t) = s(t) * k(t)$  in the 200 ms window centered around the movement onset moment marked by  $s(t)$ . The choice of the time window is motivated by our interest in the early components reflecting the activity M1 zone.

We then focused on the first two components with the largest mutual information, as shown in **Figure 7**. The temporal dynamics of the first two components showed a slow activation increase starting as early as 400 ms before the actual movement onset. However, as illustrated in **Figures 7A,B**, the two components differed in their behavior during the interval directly preceding the movement onset. The component with the larger value of MI exhibited a sharp quasi-exponential growth starting at around 50–70 ms before the movement onset. The onset dynamics of the second component was smooth, corresponding to a quasi-linear growth. After movement onset both components show a pronounced negative deflection reaching a minimum at around 100 ms after the movement onset.

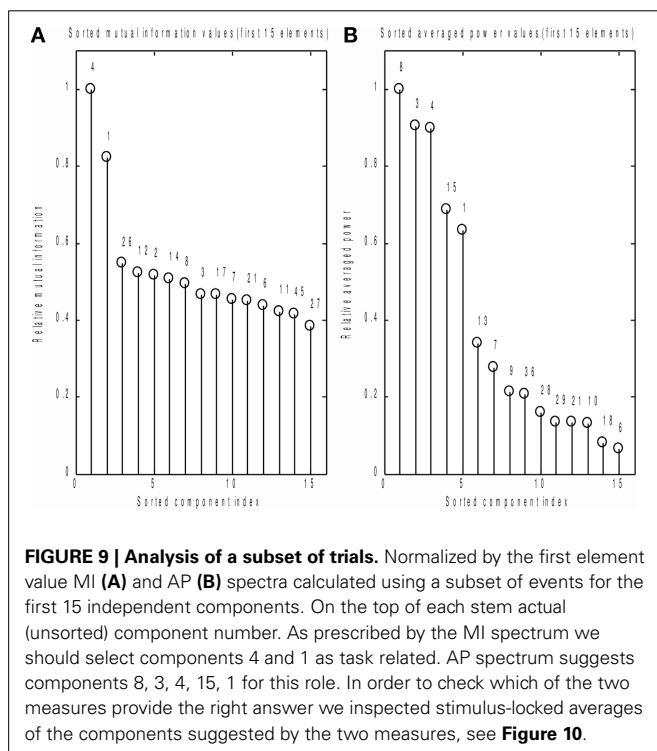
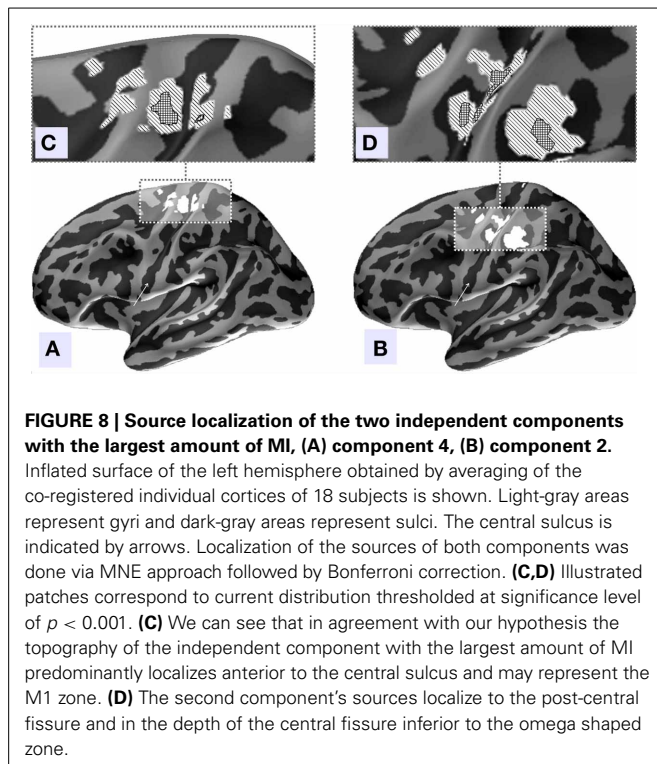
We used MNE distributed source imaging to localize the neuronal generators underlying the topographies of the first two components with the highest values of MI. In both cases we observed activations in the area surrounding the central fissure of the hemisphere contralateral to movement, shown in **Figure 8**. After thresholding at  $p < 0.001$  we observed that cortical areas subserving these two components do not overlap, as shown in **Figure 8**. Cortical sources of the first component localized primarily on the anterior slope of the central sulcus superior to the omega zone, shown in **Figure 8C**. This source most likely lies in M1, based on the anatomy. The cortical sources for the second component were located in the post-central sulcus and in the depth of the central sulcus inferior to the omega shaped zone, shown in **Figure 8D**.



### 3.4. MI VERSUS AP FOR SMALL EPOCH COUNTS IN REAL DATA

We also compared the performance of the MI spectrum with the more conventional AP metric (6) when the number of epochs is limited. We took every 30th event and analyzed the data according to the scheme described above. Independently sorted MI and AP spectra for the first 15 components are shown in **Figures 9A,B** respectively. The MI spectrum clearly shows the presence of task-related signal in the first two components with original indices 4 and 1. The AP spectrum shows five seemingly task-related components (indices 8, 3, 4, 15, 1) standing out from the baseline. Components 1 and 4 are identified as task-related by the both measures of task-relatedness.

In order to check which of the two methods provided the correct answer, we performed stimulus-locked averaging of the first five components obtained by sorting in decreasing order the MI and AP spectra, shown in **Figure 9**. The results are shown in **Figure 10**. The first two components (4 and 1) identified by the MI spectrum (see **Figure 9A**) show a clear task related deflection. The remaining components do not have significant amount of stimulus-locked activity and therefore are most likely unrelated to the task. Three out of five components identified by the AP spectrum (first two and the fourth) do not exhibit any deflection resulting from coherent summation. Note also that components 1 and 4 are among the five components selected by the AP spectrum (the third and the fifth). Visual analysis of



the averages obtained for the other three components suggested by the AP spectrum does not reveal the presence of significant amount of stimulus-locked activity in three out of five components.

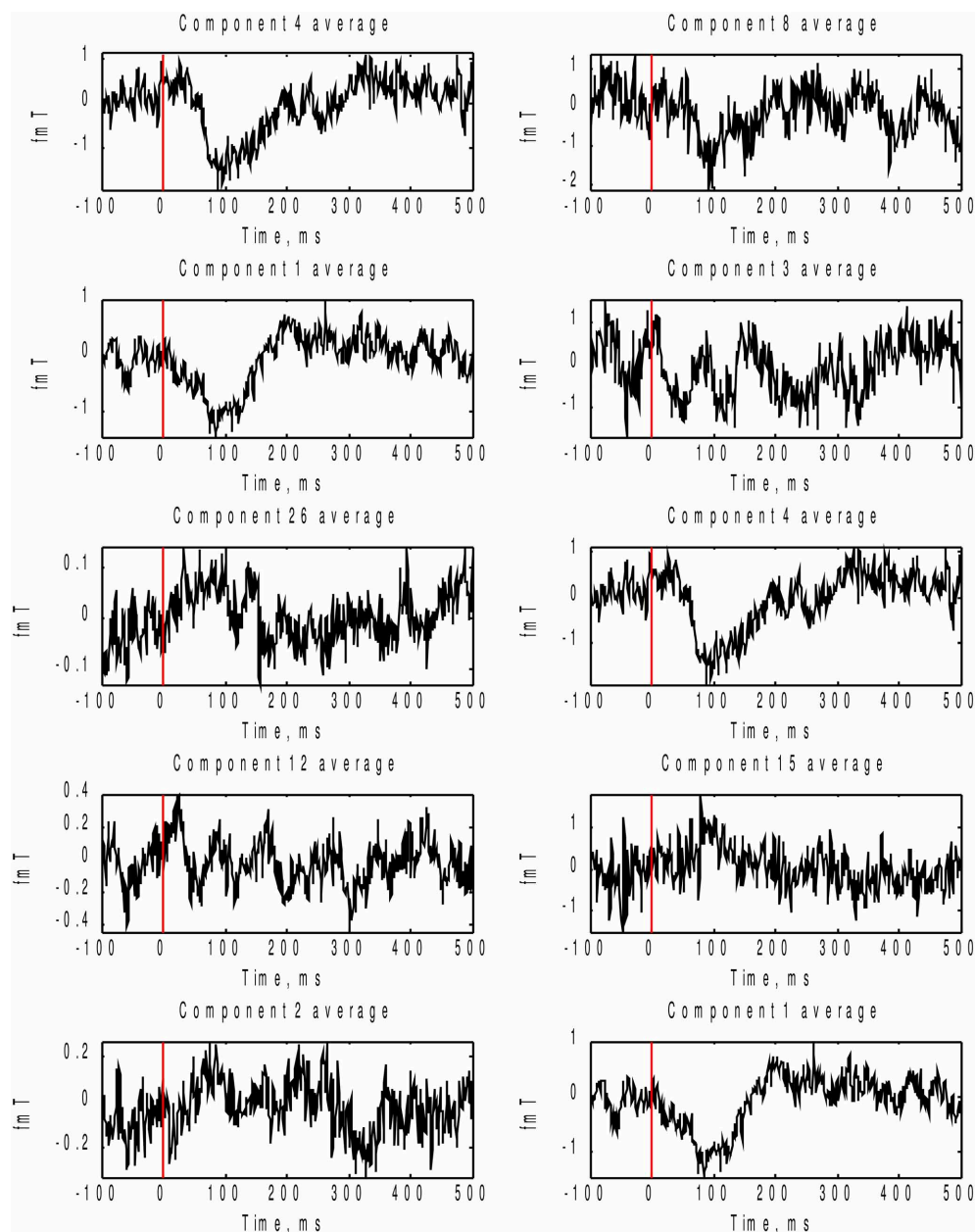
Based on these observations, we conclude that both MI and AP spectra demonstrate identical sensitivity, as both were able to detect two clearly task-related components (1 and 4) that were also found using the full dataset and characterized by inverse modeling (see Figure 8). However, the MI spectrum exhibits optimal specificity, identifying two components (Figure 9A). Both of these components appear to have a task-related deflection in their stimulus-locked averaged profiles (Figure 10A). The specificity of AP based measure  $r$  is poor by comparison with the MI measure, since AP it identified 5 components (Figure 9B), including 3 false positives and 2 correct hits (Figure 10B). The observed behavior is consistent with our simulation studies, illustrated in Figures 5A,B, where the MI spectrum demonstrated significantly higher ROC characteristics, and provided higher specificity for any fixed sensitivity value compared to the AP spectrum.

#### 4. DISCUSSION

We describe a novel information-theoretic approach for spatial components ranking. Our method is based on the MI Spectrum which serves as a power-invariant measure of repetitive task-related signal in the temporal loadings of spatial components. Using realistic simulations we demonstrated that the task-relatedness measure, based on estimating the MI between a component and the expanded binary stimulus signal, allows for significantly higher detector characteristics when compared with conventional alternatives. It also provides a means for more clear-cut separation of task-related and task-unrelated components when compared with the standard power driven approach that is used in SVD, and sometimes used for ranking ICA components as well. The MI measure can be used for sorting the components obtained from any sort of spatial decomposition, as long as it is possible to calculate the quasi-continuous timeseries underlying the components of interest. The demonstrated advantage in performance over the power-driven measure makes the MI spectrum method a candidate for the routine use in ranking both SVD and ICA components in the analysis of ERP data. Since the MI method is insensitive to powerful non-task-related noise sources, it should also facilitate automatic unsupervised analysis of ERP data using ICA.

The method can be easily extended to extract not only the evoked (phase-locked to the stimulus) activity but also band-specific task induced activity that is characterized by random phase but stimulus-locked power fluctuations. Such an extension would require that the band-pass filtered components envelope should be calculated before MI spectrum estimation.

We have also investigated the MI method performance applied to an MEG dataset in a voluntary finger movement task. Such paradigms present special challenges, since they include large amount of random latency jitter when compared with an external stimulus driven paradigms. This increased jitter comes from the inevitable errors in the estimates of motion onset obtained from the accelerometer signal. Nevertheless, the MI measure supported a clear cut separation of four task related components (Figure 7). The component with the largest MI (index 4) demonstrated a



**FIGURE 10 |** In order to check which of the two methods provide the right answer we performed stimulus-locked averaging of the first five components in the order prescribed by sorting of MI and AP spectra (Figure 9). Left panel corresponds to MI prescribed components and the right panel - to AP. As it can be seen the first two components (4 and 1) emphasized by the MI spectrum (Figure 9A)

show a clear task related deflection, the subsequent components do not have significant amount of stimulus-locked activity and therefore are most likely unrelated to the task as correctly indicated by the characteristic drop in the MI spectrum (Figure 9A). Three out of five (8, 3, 15) component averages prescribed by the AP spectrum do not exhibit the expected deflection.

non-linear increase of activation just prior to the motion onset. In agreement with the previous experiments on primates (Riehle, 2005), this component localized primarily to the anterior slope of the central sulcus superior to the omega zone (Figure 8C), and most likely originates in M1. Thus, we demonstrated the potential to localize M1 non-invasively on a group level, using a functional probe.

In order to compare the performance of the MI and power based measures on the experimental dataset, we used a reduced number of trials. As shown by simulations (Figures 4, 5), this reduction should increase the contrast between the performance characteristics of the two methods. It should also mimic more realistic scenarios, when only a single subject dataset is used for ICA analysis. Under these conditions, we demonstrated that the



proposed MI based measure for a fixed sensitivity value yields significantly higher specificity than the more conventional power based measure. This result is in agreement with our simulation studies.

In the current work we used a simple histogram-based approach for calculation of MI, omitting any bias correction. For realizations of independent random processes Treves and Panzeri (1995) have shown that MI estimate bias is quadratically proportional to the number of histogram bins used to approximate the pdf of continuous random processes and is inversely proportional to the number of datapoints. Since we used a large number of datapoints ( $N \approx 6.5 \times 10^5$ ) in our experimental data analysis compared to the  $K = 10$  bins used for approximation of the probability density functions, we do not expect a bias correction procedure to appreciably alter the observed MI. However, it has also been shown (Chrisman, 2013) that the bias decreases as the true MI between the timeseries pairs grow. This means that bias may result in MI values of task-unrelated components being overestimated, yielding a decreased contrast between the task-related and task-unrelated components in the MI spectrum. In our simulation studies we used a relatively small number of datapoints compared to a standard EEG/MEG data recording per single patient. Therefore, we expect that the observed performance (Figures 4, 5) may be further improved with a proper bias correction procedure. The use of a biased estimator in the statistical testing approach we implemented results in less sensitive tests, since the null-hypothesis distribution estimate appears to be “shifted to the right.” Selection of an appropriate bias correction method, however, requires a significant amount of additional numerical experiments and goes beyond the scope of this paper.

## ACKNOWLEDGMENTS

We would like to thank Dr. Richard Greenblatt, Dr. Alexandr Valjamae and the reviewers for the careful reading of the manuscript and their critical comments that helped to improve the readability and added to the clarity of presentation. This study was partly supported by the Russian Targeted Federal Program “Scientific and scientific-pedagogical personnel of innovative Russia” (Agreement 8488), President Grants for Government Support of Young Russian Scientists (5137.2013.4), Russian foundation for basic research (grants 11-06-12037-ofi-m-2011 and 11-06-00449-a), Basic Research Program of the National Research University Higher School of Economics, SPBU grant for priority research 0.37.522.2013 “Neuroeconomics and managerial decision making: interdisciplinary study,” by the funds provided to Alexei Ossadtchi by Source Signal Imaging Inc. as well as Alexei Ossadtchi's personal funds.

## REFERENCES

- Child, D. (2006). *The Essentials of Factor Analysis*. 3rd Edn. London; New York, NY: Continuum.
- Chrisman, L. (2013). Estimation of mutual information. Available online at: <http://blog.lumina.com/2013/estimation-of-mutual-information/>
- Comon, P. (1994). Independent component analysis: a new concept? *Signal Process.* 36, 287–314. doi: 10.1016/0165-1684(94)90029-9
- Delorme, A. (2010). Headit and eeglab: open resources for electrophysiological discovery. *Psychophysiology* 47:S6. [15th Annual Meeting Portland Marriott Downtown Waterfront (Portland, OR)].
- Esposito, F., Formisano, E., Seifritz, E., Goebel, R., Morrone, R., Tedeschi, G., et al. (2002). Spatial independent component analysis of functional MRI time-series: to what extent do results depend on the algorithm used? *Hum. Brain Mapp.* 16, 146–157. doi: 10.1002/hbm.10034
- Gerloff, C., Richard, J., Hadley, J., Schulman, A. E., Honda, M., and Hallett, M. (1998). Functional coupling and regional activation of human cortical motor areas during simple, internally paced and externally paced finger movements. *Brain* 121(Pt 8), 1513–1531. doi: 10.1093/brain/121.8.1513
- Golub, G., and Van Loan, C. (1996). *Matrix Computations*. 3rd Edn. Baltimore; London: Johns Hopkins University Press.
- Grosse-Wentrup, M., and Buss, M. (2008). Multiclass common spatial patterns and information theoretic feature extraction. *IEEE Trans. Biomed. Eng.* 55, 1991–2000. doi: 10.1109/TBME.2008.921154
- Gu, H., Engelen, W., Feng, H., Silbersweig, D. A., Stern, E., and Yang, Y. (2001). Mapping transient, randomly occurring neuropsychological events using independent component analysis. *Neuroimage* 14, 1432–1443. doi: 10.1006/nimg.2001.0914
- Himberg, J., Hyvarinen, A., and Esposito, F. (2004). Validating the independent components of neuroimaging time series via clustering and visualization. *Neuroimage* 22, 1214–1222. doi: 10.1016/j.neuroimage.2004.03.027
- Hyvarinen, A., Hoyer, P., and Inki, M. (2001). Topographic independent component analysis. *Neural Comput.* 13, 1527–1558. doi: 10.1162/089976601750264992
- Jung, T., Makeig, S., Humphries, C., Lee, T., McKeown, M., Iragui, V., et al. (2000). Removing electroencephalographic artifacts by blind source separation. *Psychophysiology* 37, 163–178. doi: 10.1111/1469-8986.3720163
- Lagerlund, T., Sharbrough, F., and Busacker, N. (1997). Spatial filtering of multichannel electroencephalographic recordings through principal component analysis by singular value decomposition. *J. Clin. Neurophysiol.* 14, 73–82.
- Lange, D., Pratt, H., and Inbar, G. (1997). Modelling and estimation of single evoked brain potential components. *IEEE Trans. Biomed. Eng.* 44, 791–799. doi: 10.1109/10.623048
- Lu, W., and Rajapakse, J. (2003). Eliminating indeterminacy in ICA. *Neurocomputing* 50, 271–290. doi: 10.1016/S0925-2312(01)00710-X
- Makeig, S., Bell, A., Jung, T., and Sejnowski, T. (1996). “Independent component analysis of electroencephalographic data,” in *Advances in Neural Information Processing Systems 8: Proceedings of the 1995 Conference*, Vol. 8, eds D. Touretzky, M. Mozer, and M. Hasselmo (Cambridge; London: MIT Press), 145–151.
- Misulis, K. (1994). *Spehlmann's Evoked Potential Primer: Visual, Auditory, and Somatosensory Evoked Potentials in Clinical Diagnosis*. 2nd Edn. Oxford: Butterworth-Heinemann Medical.
- Moritz, C. H., Rogers, B. P., and Meyerand, M. E. (2003). Power spectrum ranked independent component analysis of a periodic fMRI complex motor paradigm. *Hum. Brain Mapp.* 18, 111–122. doi: 10.1002/hbm.10081
- Ossadtchi, A., Baillet, S., Mosher, J., and Leahy, R. (2000). “Using mutual information to select event-related components in ICA,” in *Biomag 2000: Proceedings of the 12-th International Conference on Biomagnetism* (Espoo: Helsinki University of Technology), 199.
- Riehle, A. (2005). “Preparation for action: one of the key functions of motor cortex,” in *Motor Cortex in Voluntary Movements: A Distributed System for Distributed Functions*, eds A. Riehle and E. Vaadia (Boca Raton, FL: CRC Press), 213–240.
- Sanders, J. A., Lewine, J. D., and Orrison, W. W. (1996). Comparison of primary motor cortex localization using functional magnetic resonance imaging and magnetoencephalography. *Hum. Brain Mapp.* 4, 47–57. doi: 10.1002/(SICI)1097-0193(1996)4:1<47::AID-HBM3>3.0.CO;2-P
- Strehl, A. (2002). Cluster ensembles a knowledge reuse framework for combining multiple partitions. *J. Mach. Learn. Res.* 3, 583–617. doi: 10.1162/15324430321897735
- Tong, L., Liu, R.-W., V. C. S., and Huang, Y.-F. (1991). Indeterminacy and identifiability of blind identification. *IEEE Trans. Circuits Syst.* 38, 1–13. doi: 10.1109/31.76486
- Treves, A., and Panzeri, S. (1995). The upward bias in measures of information derived from limited data samples. *Neural Comput.* 7, 399–407. doi: 10.1162/neco.1995.7.2.399
- Vandewalle, J., and De Moor, B. (1988). “A variety of applications of singular value decomposition in identification and signal processing,” in *SVD and*

- Signal Processing, Algorithms, Applications and Architectures*, ed E. Deprettere (Amsterdam: Elsevier), 43–91.
- Vigario, R., Sarela, J., Jousmaki, V., Hamalainen, M., and Oja, E. (2000). Independent component approach to the analysis of eeg and meg recordings. *IEEE Trans. Biomed. Eng.* 47, 589–593. doi: 10.1109/10.841330
- Yang, Z., LaConte, S., Weng, X., and Hu, X. (2008). Ranking and averaging independent component analysis by reproducibility (RAICAR). *Hum. Brain Mapp.* 29, 711–725. doi: 10.1002/hbm.20432

**Conflict of Interest Statement:** The authors declare that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.

Received: 01 November 2013; accepted: 31 December 2013; published online: 20 January 2014.

Citation: Ossadtchi A, Pronko P, Baillet S, Pflieger ME and Stroganova T (2014) Mutual information spectrum for selection of event-related spatial components. Application to eloquent motor cortex mapping. *Front. Neuroinform.* 7:53. doi: 10.3389/fninf.2013.00053

This article was submitted to the journal *Frontiers in Neuroinformatics*.

Copyright © 2014 Ossadtchi, Pronko, Baillet, Pflieger and Stroganova. This is an open-access article distributed under the terms of the Creative Commons Attribution License (CC BY). The use, distribution or reproduction in other forums is permitted, provided the original author(s) or licensor are credited and that the original publication in this journal is cited, in accordance with accepted academic practice. No use, distribution or reproduction is permitted which does not comply with these terms.



# Local active information storage as a tool to understand distributed neural information processing

Michael Wibral<sup>1\*</sup>, Joseph T. Lizier<sup>2</sup>, Sebastian Vögler<sup>3</sup>, Viola Priesemann<sup>4</sup> and Ralf Galuske<sup>3</sup>

<sup>1</sup> MEG Unit, Brain Imaging Center, Goethe University, Frankfurt am Main, Germany

<sup>2</sup> CSIRO Computational Informatics, Marsfield, NSW, Australia

<sup>3</sup> Fakultät für Biologie, Technische Universität, Darmstadt, Germany

<sup>4</sup> Department of Nonlinear Dynamics, Max Planck Institute for Dynamics and Self-Organization, Göttingen, Germany

## Edited by:

Daniele Marinazzo, University of Gent, Belgium

## Reviewed by:

Demian Battaglia, Max Planck Institute for Dynamics and Self-Organization, Germany  
Luca Faes, University of Trento, Italy

## \*Correspondence:

Michael Wibral, MEG Unit, Brain Imaging Center, Goethe University, Heinrich-Hoffmann Strasse 10, Frankfurt am Main, D-602528, Germany  
e-mail: wibral@em.uni-frankfurt.de

Every act of information processing can in principle be decomposed into the component operations of information storage, transfer, and modification. Yet, while this is easily done for today's digital computers, the application of these concepts to neural information processing was hampered by the lack of proper mathematical definitions of these operations on information. Recently, definitions were given for the dynamics of these information processing operations on a local scale in space and time in a distributed system, and the specific concept of local active information storage was successfully applied to the analysis and optimization of artificial neural systems. However, no attempt to measure the space-time dynamics of local active information storage in neural data has been made to date. Here we measure local active information storage on a local scale in time and space in voltage sensitive dye imaging data from area 18 of the cat. We show that storage reflects neural properties such as stimulus preferences and surprise upon unexpected stimulus change, and in area 18 reflects the abstract concept of an ongoing stimulus despite the locally random nature of this stimulus. We suggest that LAIS will be a useful quantity to test theories of cortical function, such as predictive coding.

**Keywords:** visual system, neural dynamics, predictive coding, local information dynamics, voltage sensitive dye imaging, distributed computation, complex systems, information storage

## 1. INTRODUCTION

It is commonplace to state that brains exist to “process information.” Curiously enough, however, it is much more difficult to exactly quantify this putative processing of information. In contrast, we have no difficulties to quantify information processing in a digital computer, e.g., in terms of the information stored on its hard disk, or the amount of information transferred per second from its hard disk to its random access memory, and then on to the CPU. Why then is it so difficult to perform a similar quantification for biological, and especially neural information processing?

One answer to this question is the conceptual difference between a digital computer and a neural system: in a digital computer all components are laid out such that they only perform specific operations on information: a hard disk should store information, and not modify it, while the CPU should quickly modify the incoming information and then immediately forget about it, and system buses exist solely to transfer information. In contrast, in neural systems it is safe to assume that each element of the system (each neuron) *simultaneously* stores, transfers and modifies information in variable amounts, and the component processes are hard to separate quantitatively. Thus, while in digital computers the distinction between information storage, transfer and modification comes practically for free, in neural systems separating the components of distributed information processing requires thorough mathematical definitions of information storage, transfer and modification. Such

definitions, let alone a conceptual understanding of what the terms meant in distributed information processing, were unavailable until very recently (Langton, 1990; Mitchell, 1998; Lizier, 2013).

These necessary mathematical definitions were recently derived building on Turing's old idea that every act of information processing can be decomposed into the component processes of information storage, transfer and modification (Turing, 1936)—very much in line with our everyday view of the subject. Later, Langton and others expanded Turing's concepts to describe the emergence of the capacity to perform arbitrary information processing algorithms, or “universal computation,” in complex systems, such as cellular automata (Langton, 1990; Mitchell et al., 1993), or neural systems. The definitions of information transfer and storage were then given by Schreiber (2000), Crutchfield and Feldman (2003), and Lizier et al. (2012b). However, the definition of information modification is still a matter of debate (Lizier et al., 2013).

Of these three component processes above—information transfer, storage, and modification—information storage in particular has been used with great success to analyze cerebrovascular dynamics (Faes et al., 2013), information processing in swarms (Wang et al., 2012), and most importantly, to evolve (Prokopenko et al., 2006), and optimize (Dasgupta et al., 2013) artificial information processing systems. This suggests that the analysis of information storage could also be very useful for the analysis of neural systems.

Yet, while neuroscientists have given much attention to considering how information is stored structurally in the brain, e.g., via synaptic plasticity, the same attention has not been given to information storage in neural dynamics, and its quantification. As an exception Zipser et al. (1993) clearly contrasted two different ways of storing information: *passive storage*, where information is stored “in modified values of physiological parameters such as synaptic strength,” and *active storage* where “information is preserved by maintaining neural activity throughout the time it must be remembered.” In the same paper, the authors go on to point out that there is evidence for the use of both storage strategies in higher animals, and link the relatively short time scale for active storage (at maximum in the tens of seconds) with short-term or working memory and, therefore, refer to it as “active information storage.”

Despite the importance of information storage for neural information processing, information theoretic measures of active information storage have not yet been used to quantify information processing in neural systems, and in particular not to measure spatiotemporal patterns of information storage dynamics. Therefore, it is the aim of this article to introduce measures of information storage as analysis tools for the investigation of neural systems, and to demonstrate how cortical information storage in visual cortex unfolds in space and time. We will also demonstrate how neural activity may be misinformative about its own future and thereby generates “surprise.”

To this end, we first give a rigorous mathematical definition of information storage in dynamic activity in the form of local active information storage (LAIS). We then show how to apply this measure to voltage sensitive dye imaging data from cat visual cortex. In these data, we found sustained increases in dynamic information storage during visual stimulation, organized in clear spatiotemporal patterns of storage across the cortex, including stimulus-specific spatial patterns, and negative storage, or surprise, upon a change of the stimulus. Finally, we discuss the implications of the LAIS measure for neurophysiological theories of predictive coding [see Bastos et al. (2012), and references therein], that have been suggested to explain general operating principles of the cortex and other hierarchical neural systems.

## 2. MATERIALS AND METHODS

The use of the stored information for information processing inevitably requires its re-expression in neural activity and its interaction with ongoing neural activity and incoming information. Hence, information storage *actively in use for information processing* will inevitably be reflected in the dynamics of neural activity, and is therefore accessible in recordings of neural activity alone. To quantify this stored information that is present in neural time series we will now introduce a measure of information storage called *local active information storage* (Lizier et al., 2012b). In brief, this measure quantifies the amount of information in a sample from a neural time series that is predictable from its past—and thereby has been stored in this past. This is done by simply computing the local mutual information between the past of a neural signal and its next sample at each point in time, and for each channel of a recording. As the

following material is necessarily formal, the reader may consider skipping ahead to section 2.2.3 at first reading to gain an intuitive understanding of mechanisms that serve active information storage.

### 2.1. NOTATION AND INFORMATION THEORETIC PRELIMINARIES

To avoid confusion, we first have to state how we formalize observations from neural systems mathematically. We define that a neural (sub-)system of interest (e.g., a neuron, or brain area)  $\mathcal{X}$  produces an observed time series  $\{x_1, \dots, x_t, \dots, x_N\}$ , sampled at time intervals  $\delta$ . For simplicity we choose our temporal units such that  $\delta = 1$ , and hence index our measurements by  $t \in \{1 \dots N\} \subseteq \mathbb{N}$ , i.e., we index in terms of samples. The full time series is understood as a realization of a *random process*  $X$ . This random processes is nothing but a collection of random variables  $X_t$ , sorted by an integer index ( $t$  in our case). Each random variable  $X_t$ , at a specific time  $t$ , is described by the set of all its  $J$  possible outcomes  $\mathcal{A}_{X_t} = \{a_1, \dots, a_j, \dots, a_J\}$ , and their associated probabilities  $p_t(x_t = a_j)$ . The probabilities of a specific outcome  $p_t(x_t = a)$  may change with  $t$ , i.e., when going from one random variable to the next. In this case, we will indicate the specific random variable  $X_t$  the probability distribution belongs to—hence the subscript in  $p_t(\cdot)$ . For practical estimation of  $p_t(\cdot)$  then, multiple time-series realizations or trials would be required. For stationary processes, where  $p_t(x_t = a)$  does not change with  $t$ , we simply write  $p(x_t)$ , and practical estimation may be done from a single time-series realization. In sum, in this notation the individual random variables  $X_t$  produce realizations  $x_t$ , and the time-point index of a random variable  $X_t$  is necessary when the random process is non-stationary. When using more than one system, the notation is generalized to multiple systems  $\mathcal{X}, \mathcal{Y}, \mathcal{Z}, \dots$

As we will see below, active information storage is nothing but a specific mutual information between collections of random variables in the process in question. We therefore start by giving the definition of *mutual information* (MI)  $I(X; Y)$  as the amount of information held in common by two random variables  $U, V$  on average (Cover and Thomas, 1991):

$$I(U; V) = \sum_{u \in \mathcal{A}_U, v \in \mathcal{A}_V} p(u, v) \log \frac{p(u, v)}{p(u)p(v)}, \quad (1)$$

$$= \sum_{u \in \mathcal{A}_U, v \in \mathcal{A}_V} p(u, v) \log \frac{p(v | u)}{p(v)}, \quad (2)$$

where the log can be taken to an arbitrary base, and choosing base 2 yields the mutual information in bits. Note that the mutual information  $I(U; V)$  is symmetric in  $U$  and  $V$ . As shown more explicitly in Equation (2), the MI  $I(U; V)$  measures the amount of information provided (or the amount that uncertainty is reduced) by an observation of a specific outcome  $u$  of the variable  $U$  about the occurrence of another specific outcome  $v$  of  $V$ —on average over all possible values of  $u$  and  $v$ . As originally pointed out by Fano (1961), the summands  $\log \frac{p(v|u)}{p(v)}$  have a proper interpretation even without the weighted averaging—as the information that observation of a specific  $u$  provides about the occurrence of a specific  $v$ . The *pointwise* or *local mutual*



information is therefore defined as:

$$i(u; v) = \log \frac{p(v | u)}{p(v)}. \quad (3)$$

It is important to note the distinction of the local mutual information measure  $i(x; y)$  considered here from partial localization expressions, i.e., the partial mutual information or specific information  $I(u; V)$  which are better known in neuroscience (DeWeese and Meister, 1999; Butts, 2003; Butts and Goldman, 2006). Partial MI expressions consider information contained in specific values  $u$  of one variable  $U$  about the other (unknown) variable  $V$ . Crucially, there are two valid approaches to measuring partial mutual information, one which preserves the additivity of information and one which retains non-negativity (DeWeese and Meister, 1999). In contrast, the fully local mutual information  $i(x; y)$  that is used here is uniquely defined as shown by Fano (1961).

## 2.2. LOCAL ACTIVE INFORMATION STORAGE

Using the definition in Equation (3), we can immediately quantify how much of the information in the outcome  $x_t$  of the random variable  $X_t$  at time  $t$  was predictable from the observed past state  $\mathbf{x}_{t-1}^{k-}$  of the process at time  $t-1$ :

$$a(x_t) = i(\mathbf{x}_{t-1}^{k-}; x_t) \quad (4)$$

$$= \log \frac{p_t(x_t | \mathbf{x}_{t-1}^{k-})}{p_t(x_t)}. \quad (5)$$

This quantity was introduced by Lizier et al. (2012b) and called *local active information storage* (LAIS). Here,  $\mathbf{x}_{t-1}^{k-}$  is an outcome of the collection of previous random variables  $\mathbf{X}_{t-1}^{k-} = \{X_{t-1}, X_{t-t_1}, \dots, X_{t-t_{k_{\max}}}\}$ , called a *state* (see below). The corresponding expectation value over all possible observations of  $x_t$  and  $\mathbf{x}_{t-1}^{k-}$ ,  $A(X_t) = I(\mathbf{X}_{t-1}^{k-}; X_t)$ , is known simply as the *active information storage*. The naming of this measure aligns well with the concept of active storage in neuroscience by Zipser et al. (1993), but is more general than capturing only sustained firing patterns. In the following subsections, we comment on practical issues involved in estimating the LAIS, and discuss its interpretation.

### 2.2.1. Interpretation and construction of the past state

As indicated above, the joint variable  $\mathbf{x}_{t-1}^{k-}$  in Equation (4) is an outcome of the collection of previous random variables:  $\mathbf{X}_{t-1}^{k-} = \{X_{t-1}, X_{t-t_1}, \dots, X_{t-t_{k_{\max}}}\}$ . This collection should be constructed such, that it captures the *state* of the underlying dynamical system  $\mathcal{X}$ , and can be viewed as a state-space reconstruction of this system. In this sense,  $\mathbf{X}_{t-1}^{k-}$  must be chosen such that  $X_t$  is conditionally independent of all  $X_{t-t_l}$  with  $t_l > t_{k_{\max}}$ , i.e., of all variables that are observed earlier in the process  $\mathbf{X}$  than the variables in the state at  $t-1$ . The choice must be made carefully, since using too few variables  $X_{t-t_l}$  from the history can result in an underestimation of  $a(x_t)$ , while using too many [given the amount of data used to estimate the probability density functions (PDFs) in Equation (4)] will artificially inflate it.

Typically, the state can be captured via Takens delay embedding (Takens, 1981), using  $d$  variables  $X_{t-t_l}$  with the  $t_l$  delays equally spaced by some  $\tau \geq 1$ , with  $d$  and  $\tau$  selected using the Ragwitz criteria (Ragwitz and Kantz, 2002)—as recommended by Vicente et al. (2011) for the related transfer entropy measure (Schreiber, 2000). Alternatively, non-uniform embeddings may be used (e.g., see Faes et al., 2012).

If the process has infinite memory, and  $k_{\max}$  does not exist, then the local active information storage is defined as the limit  $\lim_{k \rightarrow \infty}$  of Equation (4):

$$a(x_t) = \lim_{k \rightarrow \infty} i(\mathbf{x}_{t-1}^{k-}; x_t) \quad (6)$$

$$= \lim_{k \rightarrow \infty} \log \frac{p_t(x_t | \mathbf{x}_{t-1}^{k-})}{p_t(x_t)}. \quad (7)$$

### 2.2.2. Relation to other measures and dynamic state updates

The *average* active information storage (AIS), is related to two measures introduced previously. On the one hand, a similar measure called “regularity” had been introduced by Porta et al. (2000). On the other hand, AIS is closely related to the excess entropy (Crutchfield and Feldman, 2003), as observed in Lizier et al. (2012b). The excess entropy  $E(X_t) = I(\mathbf{X}_{t-1}^{k-}; \mathbf{X}_t^{k+})$ , with  $\mathbf{X}_t^{k+} = \{X_t, X_{t+t_1}, \dots, X_{t+t_{k_{\max}}}\}$  being a similar collection of future random variables from the process, measures the amount of information (on average) in the future outcomes  $\mathbf{x}_t^{k+}$  of the process this is predictable from the observed past state  $\mathbf{x}_{t-1}^{k-}$  at time  $t-1$ . As such, the excess entropy captures all of the information in the future of the process that is predictable from its past. In measuring the subset of that information in only the next outcome of the process, the AIS is focused on the dynamic state updates of the process.

From the point of view of dynamic state updates, the AIS is *complementary* to a well-known measure of uncertainty of the next outcome of the process which cannot be resolved by its past state. Following Crutchfield and Feldman (2003) we refer to this quantity as the “entropy rate,” the conditional entropy of the next outcome given the past state:  $H_\mu(X_t) = H(X_t | \mathbf{x}_{t-1}^{k-}) = \langle -\log_2 p_t(x_t | \mathbf{x}_{t-1}^{k-}) \rangle$ . The complementarity of the entropy rate and AIS was shown by Lizier et al. (2012b):  $H(X_t) = A(X_t) + H_\mu(X_t)$ , where  $H(X_t)$  is the Shannon entropy of the next measurement  $X_t$ .  $H_\mu(X_t)$  is approximated by measures known as the Approximate Entropy (Pincus, 1991), Sample Entropy (Richman and Moorman, 2000), and Corrected Conditional Entropy (Porta et al., 1998), which have been well studied in neuroscience [see e.g., the work by Gómez and Hornero (2010); Vakorin et al. (2011), and references therein]. Many such studies refer to  $H_\mu(X_t)$  as a measure of complexity, however, modern complex systems perspectives focus on complexity as being captured in how much structure can be resolved rather than how much cannot (Crutchfield and Feldman, 2003).

Furthermore, given that the most appropriate measure of complexity of a process is a matter of open debate (Prokopenko et al., 2009), we take the perspective that complexity of a system is best approached as arising out of the interaction of the

component operations of information processing: information storage, transfer and modification (Lizier, 2013), and focus on measuring these quantities since they are rigorously defined and well-understood. Crucially, in comparison to the excess entropy discussed above, the focus of AIS in measuring the information storage in use in dynamic state updates of the process make it directly comparable with measures of information transfer and modification. Of particular importance here is the relationship of AIS to the transfer entropy (Schreiber, 2000), where the two measures together reveal the sources of information (either being the past of that process itself—storage, or of other processes—transfer) which contribute to prediction of the process' next outcome.

The formulation of the transfer entropy specifically eliminates information storage in the past of the target process from being mistakenly considered as having been transferred (Lizier and Prokopenko, 2010; Lizier, 2013; Wibral et al., 2013). An interesting example is where a periodic target process is in fact causally driven by another periodic process—after any initial entrainment period, our information processing view concludes that we have information storage here in the target but no transfer from the driver (Lizier and Prokopenko, 2010). While causally there is a different conclusion, our observational information processing perspective is simply focussed on decomposing apparent information sources of the process, regardless of underlying causality (which in practise cannot often be determined anyway). In this view, a causal interaction can computationally subserve both information storage or transfer (as discussed further in the next section). Information transfer is necessarily linked to a causal interaction, but the reverse is not true. It has previously been demonstrated that the information processing perspective is more relevant to emergent information processing structure in complex systems, e.g., coherent information cascades, in contrast to causal interactions being more relevant to the micro-scale physical structure of a system, e.g., axons in a neural system (Lizier and Prokopenko, 2010).

### 2.2.3. Mechanisms producing active information storage

In contrast to passive storage in terms of modifications to system structure (e.g., synaptic gain changes), the mechanisms underlying active information storage are not immediately obvious. The mechanisms that subserve this task have been formally established, however, and can be grouped as follows:

1. *Physical mechanisms in the system.* This could incorporate some internal memory mechanism in the individual physical element giving rise to the process  $X$  (e.g., some decay function, or the stereotypical processes during the refractory period after a neural spike). More generally, it may involve network structures which offload or distribute the memory function onto edges or other nodes. In particular, Zipser et al. (1993) reported that networks with fixed, *recurrent connections* were sufficient to account for such active storage patterns, which is in line with earlier proposals. Furthermore, Lizier et al. (2012a) quantified the AIS contribution from self-loops, feedback and feedforward loops (as the only network structures contributing to active information storage).

2. *Input-driven storage.* This describes situations where the apparent memory in the process is caused by information storage structure which lies in another element which is driving that process, e.g., a periodically spiking neuron that may cause a downstream neuron to spike with the same period (Obst et al., 2013). As described in section 2.2.2 above, an observer of the process attributes these dynamics to information storage, regardless of the (unobserved) underlying causal mechanism.

Of these mechanisms of active information storage the case of circular causal interactions in a loop motif, and the causal, but repetitive influence from another part of the system may seem counterintuitive at first, as we might think that in these cases there should be information transfer rather than active information storage. To see why these interactions serve storage rather than transfer, it may help to consider that *all* components of information processing, i.e., transfer, active storage and modification, ultimately have to rely on causal interactions in physical systems. Hence, the presence of a causal interaction cannot be linked in a one-to-one fashion to information transfer, as otherwise there would be no possibility for physical causes of active information storage and of information modification left, and no consistent decomposition of information processing would be possible. Therefore, the notion of storage that is measurable in a part of the system but that can be related to external influences onto that part is to be preferred for the sake of mathematical consistency and ultimately, usefulness. We acknowledge that information transfer has often been used as a proxy for a causal influence, dating back to suggestions by Wiener (1956) and Granger (1969). However, now that causal interventional measures and measures of information transfer can be clearly distinguished (Ay and Polani, 2008; Lizier and Prokopenko, 2010) it seems no longer warranted to map causal interactions to information transfer in a one-to-one manner.

### 2.2.4. Interpretation of LAIS values

Measurements of the LAIS tells us the amount to which observing the past state  $\mathbf{x}_{t-1}^{k-}$  reduced our uncertainty about the specific next outcome  $x_t$  that was observed. We can interpret this in terms of *encoding* the outcome  $x_t$  in bits: encoding  $x_t$  using an optimal encoding scheme for the distribution  $p_t(x_t)$  takes  $-\log_2 p_t(x_t)$  bits, whereas encoding  $x_t$  if we know  $\mathbf{x}_{t-1}^{k-}$  using an optimal encoding scheme for the distribution  $p_t(x_t | \mathbf{x}_{t-1}^{k-})$  takes  $-\log_2 p_t(x_t | \mathbf{x}_{t-1}^{k-})$  bits, and the LAIS is the number of bits saved via the latter approach.

At first glance we may assume that the LAIS is a positive quantity. Indeed, as a mutual information, the *average* AIS will always be non-negative. However, the LAIS can be negative as well as positive. It is positive where  $p_t(x_t | \mathbf{x}_{t-1}^{k-}) > p_t(x_t)$ , i.e., where the observed past state  $\mathbf{x}_{t-1}^{k-}$  made the following observation  $x_t$  more likely to occur than we would have guessed without the knowledge of the past state. In this case, we state that  $\mathbf{x}_{t-1}^{k-}$  was informative. In contrast, the LAIS is negative where  $p_t(x_t | \mathbf{x}_{t-1}^{k-}) < p_t(x_t)$ ; i.e., where the observed past state  $\mathbf{x}_{t-1}^{k-}$  made the following observation  $x_t$  less likely to occur than we would have guessed without the knowledge of the past state (but it occurred nevertheless, making

the cue given by  $\mathbf{x}_{t-1}^{k-}$  misleading). In this case, we state that  $\mathbf{x}_{t-1}^{k-}$  was *misinformative* about  $x_t$ . To better understand negative LAIS also see the further discussion in Lizier et al. (2012a), including examples in cellular automata where the past state of a variable was misinformative about the next observation due to the strong influence of an unobserved other source variable at that time point.

### 2.2.5. Choice of the overall time window for constructing probability densities from data

As already pointed out above, active information storage is tightly related to predictability of a given brain area's output as seen by the receiving brain area. This predictability hinges on the ability of the receiver to see the past states in the output of a brain area (see previous section) and to interpret the past states in the received time series in order to make a prediction about the next value. In other words, the receiver needs to guess  $p_t(x_t, \mathbf{x}_{t-1}^{k-})$  correctly in order to exploit the active information storage. If the guess of the receiving neuron ( $n$ ) or brain area, i.e.,  $\tilde{p}_n(x_t, \mathbf{x}_{t-1}^{k-})$ , is incorrect, then only a fraction of the information storage can be used for successfully predicting future events. The losses could be quantified as the extra coding cost for the receiving area, when assuming  $\tilde{p}_n(\cdot)$  instead of  $p_t(\cdot)$ . This loss would simply be the Kullback–Leibler divergence  $D_{\text{KL}}(p_t || \tilde{p}_n)$ . This scenario sees the receiving brain area mostly as an optimal encoder or compressor. In contrast, the cost occurring in the framework of predictive coding theories would arise because the receiving brain area could not predict the incoming signal well, and thereby inhibit it via feedback to the sending brain area (Rao and Ballard, 1999). In this scenario, the cost of imperfect predictions resulting from using  $\tilde{p}_n$  instead of  $p_t$ , would be reduced inhibition and a more frequent signaling of prediction errors by the sending system, leading to a metabolic cost.

To see the storage that the receiving brain area can exploit, the time interval used for the practical estimation of the probability density functions (PDFs) from neural recordings should best match the expected sampling strategy of the receiving brain area. For example, if we think that probabilities are evaluated over long time frames, then it might make sense to pool all available data in the experiment, as even a mis-estimation of the true probability densities  $p_t(\cdot)$  (due to potential non-stationarities) then will better reflect the internal estimate  $\tilde{p}_n(x_t, \mathbf{x}_{t-1}^{k-})$ , and thus the *internally* predictable information. However, if we think that probabilities are only estimated instantaneously by pooling over all available inputs to a brain area at any time point, then we should construct the necessary PDFs only from all simultaneously acquired data from all measurement channels, but not pool over time. The latter view could also be described as assuming that the brain area receiving the signals in question computes the PDF instantaneously by pooling over all its inputs, without keeping any longer term memory of the observed probabilities. This construction of a PDF would be linked closely to an instantaneous physical ensemble approach, considering that all incoming channels are physically equivalent, but are only assessed at a single instant in time. In contrast, if we assume that learning of

the relevant PDFs takes place on a lifelong timescale, then PDFs should be acquired from very long recordings of a freely behaving subject or animal in a natural environment, and the outcomes of a specific experiment should be interpreted using this “life-long” PDF. Here we lean toward this latter approach and pool all available data to estimate the internally available  $\tilde{p}_n$ .

Note that while we indeed pool over all the available data to obtain the distribution  $\tilde{p}_n$ , the interpretation of the data in terms of the active information storage is *local per agent and time step*. This is exactly the meaning of “local” in local active information storage as introduced in Lizier et al. (2012b) (this is also akin to the relation of the local mutual information introduced by Fano (1961) and the corresponding global PDF). The local active information storage values are thus obtained by interpreting realizations for a single agent and a single time step in the light of a probability distribution that is obtained over a more global view of the system in space and time. This is also indicated by the use of  $\tilde{p}_n$  instead of  $p_t$ . Also see the discussion section for potential other choices of obtaining  $p$ .

## 2.3. ACQUISITION OF NEURAL DATA

### 2.3.1. Animal preparation

Data were obtained from an anesthetized cat. The animal had been anesthetized and artificially ventilated with a mixture of O<sub>2</sub> and N<sub>2</sub>O (30/70%) supplemented with Halothane (0.7%). All procedures were along the guidelines of the Society for Neuroscience, in accordance with the German law for the protection of laboratory animals, permitted by the local authorities and overseen by a designated veterinarian.

### 2.3.2. Voltage sensitive dye imaging

For optical imaging the visual cortex (area 18) was exposed and an imaging chamber was implanted over the craniotomy. The chamber was filled with silicone oil and sealed with a glass plate. A voltage sensitive dye (RH1691, Optical Imaging Ltd, Rehovot, Israel) was applied to the cortex for about 2 h and subsequently the excess of the dye was washed out. For imaging we used a CMOS camera system (Imager 3001, Optical Imaging Ltd, Rehovot, Israel, Camera: Photon Focus MV1 D1312, chip size 1312 × 1082 pixel) fitted with a lens system consisting of two 50 mm Nikon objectives providing a field of view of 8.7 × 10.5 mm and an epifluorescence illumination system (excitation: 630 ± 10 nm, emission high pass 665 nm). In order to optimize the signal-to-noise ratio raw camera signals were spatially binned to 32 × 32 camera pixels allowing for a spatial resolution of 30 × 32 μm<sup>2</sup> per data pixel. Camera frames were collected at a rate 150 Hz, resulting in a temporal resolution of 6.7 ms.

### 2.3.3. Visual stimulation

Stimuli were presented triggered to the heartbeat of the animal for 2 s and camera frames were collected during the entire stimulation period. We will denote such a single stimulation period and the corresponding data acquisition as a trial here. Each trial consisted of 1 s stimulation with an isoluminant gray screen followed by stimulation with fields of randomly positioned dots (dot size: 0.23° visual angle; 384 dots distributed over an area of 30° (vertical) by 40° (horizontal) visual angle) moving coherently in

one of eight different directions at 16 degree/s. Stimuli were presented in blocks of 16 trials, consisting of eight trials using the stimuli described before and an additional eight trials which consisted only of the presentation of the isoluminant gray screen for 2 s (“blank trials”). Each motion direction condition was presented eight times in total (eight trials), resulting in the presentation on 64 stimulus trials and 64 blank trials in total. Of the presented set of eight stimulus types, seven were used for the final analysis, as the computational process for one condition did not finish on time before local compute clusters were taken down for service.

### 2.3.4. VSD data post-processing

After spatial binning of  $32 \times 32$  camera pixels into one data pixel, VSD data were averaged over all presentations of blank trials and this average was subtracted from the raw data to remove the effects of dye-bleaching and heartbeat. Finally, the data were denoised using a median filter of  $3 \times 3$  data pixels.

### 2.4. MEASUREMENT OF LAIS ON VSD NEURAL DATA

Estimation of LAIS was performed using the open source *Java information dynamics toolkit* (JIDT) (Lizier, 2012), with a history parameter  $k_{\max}$  of ten time points, spaced 2 samples, or  $(2/150 \text{ Hz}) = 13.3 \text{ ms}$ , apart. The total history length thus covered 133 ms, or roughly one cycle of a neural theta oscillation, which seems to be a reasonable time horizon for a downstream neural population that ultimately must assess these states. To enable LAIS estimation from a sufficient amount of samples, we considered the data pixels as homogeneous variables executing comparable state transitions, such that the pixels form a physical ensemble in terms of information storage dynamics. Pooling data over pixels thus enables an ensemble estimate of the PDFs in question. This approach seems justified as all pixels reported activity from a single brain area (area 18 of cat visual cortex, see below). Mutual information was estimated using a box kernel-estimator (Kantz and Schreiber, 2003) with a kernel width of 0.5 standard deviations of the data.

Here we assume that the neural system is at least capable of exploiting the statistics arising from the stimulation given throughout the experiment and thus construct PDFs from all data (time points and pixels) for a given condition. Therefore, we pool data over the full time course from  $-1$  to  $1$  s of the

experiment. Thus, each image of the VSD data had a spatial configuration of  $67 \times 137$  spatial data pixels after removal of the two rows/columns on each side of an image because of the median filter that was applied. Each trial (of a total of eight trials per condition) resulted in 288 LAIS values, based on an original data length of 298 samples and a history length (state dimension) of 10 pixels. The product of final image size and LAIS samples resulted in  $2.64 \cdot 10^6$  data points per trial for the estimation of the PDF for each of the eight motion direction conditions. Due to computational limitations, LAIS estimates were performed on two blocks of four trials separately, resulting in  $1.06 \cdot 10^7$  data points entering the estimation in JIDT.

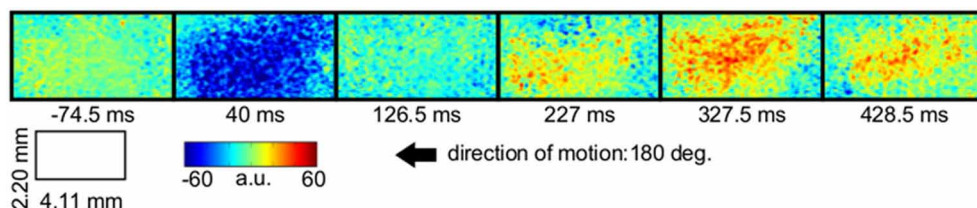
### 2.5. CORRELATION ANALYSIS OF LAIS AND VSD DATA

For each of the seven analyzed motion direction conditions, VSD data and LAIS were initially organized separately per condition into 5 dimensional data structures, with dimensions: blocks (1,2), trials (1–4), time ( $-1$  to  $1$  s), and pixel row (67) and columns (137). For correlation analysis, these arrays were linearized and entered into a Spearman rank correlation analysis to obtain correlation coefficients  $\rho(\text{VSD}, \text{LAIS})$  and significance values.

## 3. RESULTS

LAIS values exhibited a clear spatial and temporal pattern. The temporal pattern exhibited higher LAIS values during stimulation with a moving random dot pattern than under baseline stimulation with an isoluminant gray screen, with effects being largest in spatially clearly segregated regions (Figures 1–3). The spatial pattern of LAIS under stimulation was dependent on the motion direction of the drifting random dots in the stimulus (Figure 2).

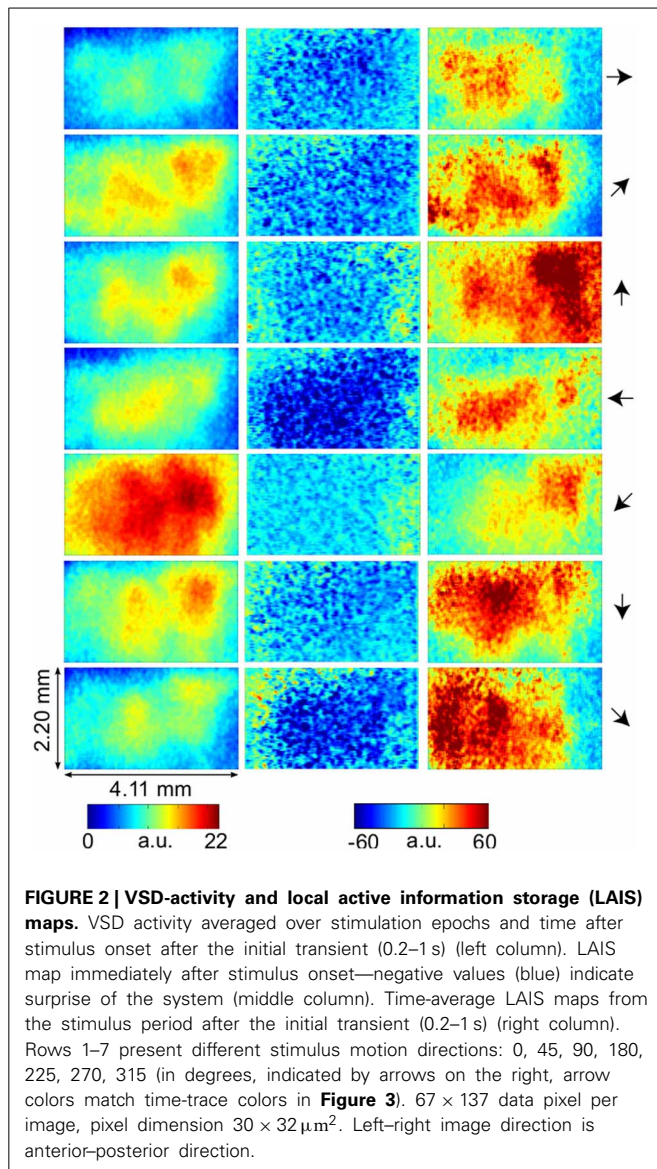
In contrast to this spatially highly selective elevation of LAIS values under stimulation, there was a sharp drop in LAIS values at approximately 40 ms after stimulus onset, with negative LAIS values measured at many pixels (Figure 1, 40 ms window; Figure 2, middle column; Figure 3, lower row). This indicates that the baseline activity was misinformative about the following stimulus related activity (since an observer would expect the baseline activity to continue). This transient, stimulus induced drop in LAIS was more evenly distributed throughout the imaging window than the elevated LAIS in the later stimulus period post 200 ms (Figure 2, middle column). The transient drop in



**FIGURE 1 | Local active information storage (LAIS) allows to trace neural information processing in space and time.** Spatio-temporal structure of LAIS in cat area 18—seven frames from the spatio-temporal LAIS data, taken at the times indicated below each frame. Stimulation onset was at time 0. Baseline activity ( $-74.5$  ms) is around zero and

mostly uniform. At 40 ms after stimulus onset, LAIS is negative in a region that correlates to the region that later exhibits high LAIS. Around 227 ms increased LAIS sets in and lasts until the end of the data epoch, albeit with slow fluctuations (up to 1 s, see Figure 3). Also see the post-stimulus time-average in Figure 2.





LAIS had a recovery time of approximately 34 ms, also giving an estimate of the dominant intrinsic storage duration of the neural processes.

In all conditions we observed a positive, but weak correlation between the local VSD activity values and LAIS values over time and space (**Table 1**). Looking at individual time intervals, we found stronger, and negative, correlation coefficients both, for the baseline interval (−1 to 0 s), and for the initial interval after the onset of the moving dot stimulus (0.04–0.14 s). In contrast, we observed a strong positive correlation at the late stimulus interval (0.2–1 s). This means that the increased dynamic range observed in the VSD signals during stimulation with the moving stimuli led to an increased amount of predictable information, rather than to a decrease. This correlation also means that storage was generally higher in neurons that were preferentially activated by the respective moving stimulus (also compare left and right columns in **Figure 2** for each motion direction).

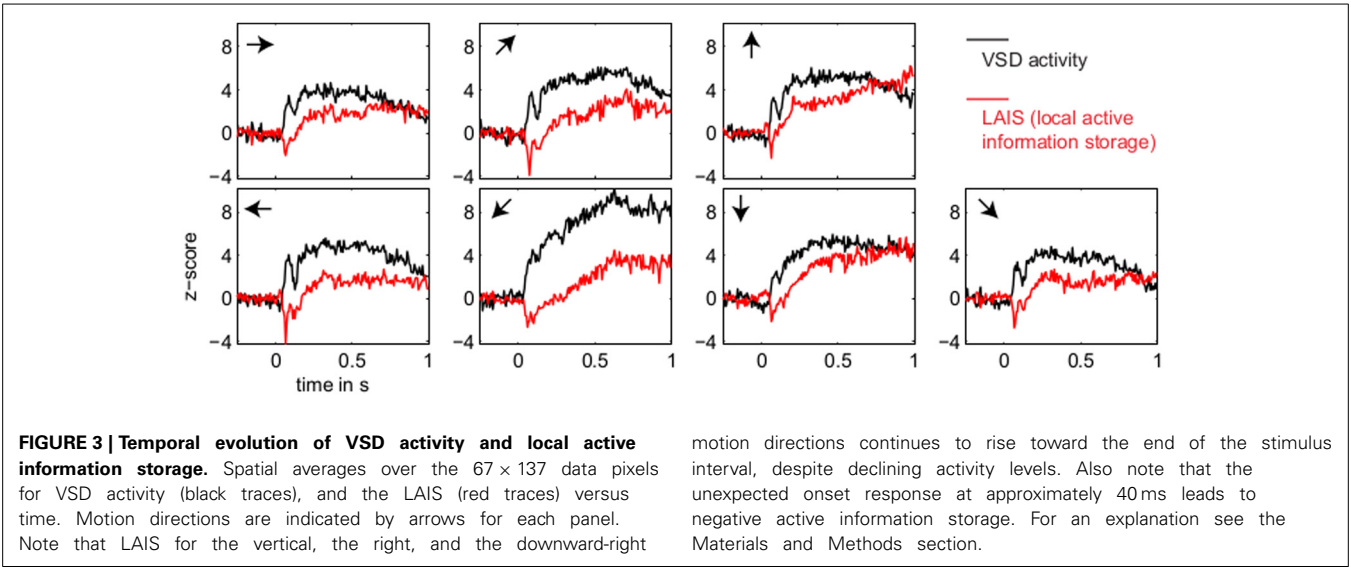
## 4. DISCUSSION

Our results demonstrate increased local active information storage in the primary visual cortex of the cat under sustained stimulation, compared to baseline. The spatial pattern of the LAIS increase was clustered spatially and stimulus-specific (**Figure 2**). The temporal pattern of LAIS consisted of a first sharp drop in LAIS from 0.04 to 0.14 s after onset of the moving stimulus and a sustained rise in LAIS up to the end of the stimulation epoch (**Figure 3**). The sharp drop at stimulus onset for many pixels is important because it indicates the past activity of the pixels was surprising or misinformative about the next outcomes near that onset. This has the potential to be used in detecting changes of processing regimes directly from neural activity.

The subsequent sustained rise in LAIS is particularly notable because of the *random spatial* structure of each stimulus on a local scale; this random spatial structure translates into a random temporal stimulation sequence in the receptive field of each neuron because of the stimulus motion. The increased LAIS despite random stimulation of the neurons suggests that our observation is not due to input-driven storage, i.e., memory or storage contained already in the spatio-temporal stimulus features that drive the observed LAIS [as discussed in section 2.2.3 and by Obst et al. (2013)]. Nevertheless, as revealed by correlation analysis, storage was highest in regions preferentially activated by the stimulus, suggesting a representational nature of LAIS in these data with respect to the motion features of the stimulus. In sum, the changes of LAIS with stimulation onset, stimulation duration, and stimulus type clearly demonstrate that LAIS reflects neural processing, rather than mere physiological or instrumentation-dependent noise regularities. This leads us to believe that LAIS is a promising tool for the analysis of neural data in general, and of VSD data in particular.

### 4.1. LOCAL ACTIVE INFORMATION STORAGE AND NEURAL ACTIVITY LEVELS

Any increase in LAIS may in principle arise from two sources: first, a richer dynamics with a larger amplitude range—increasing overall information content, while maintaining the predictability of the time series (e.g., quantified as the inverse of the signal prediction error, or the entropy-normalized LAIS), may increase LAIS. Alternatively, increased LAIS may be based on increased predictability under essentially unchanged dynamics. The significant positive correlation between LAIS and VSD activity after stimulus onset suggests that a richer, but still predictable, dynamics of VSD activity is at the core of the stimulus-dependent effects observed here. As a caveat we have to note that the use of a kernel estimator for LAIS measurement, coupled with pooling of observations over the whole ensemble of pixels and time points may also have introduced a slight bias in favor of a positive correlation between high VSD activity and LAIS, as it allows storage to be more easily measured in pixels with larger amplitude here. The negative correlation observed in the baseline interval, however, demonstrates that this bias is not a dominant effect in our data. This is because a dominant effect of the kernel-based bias would also assign higher storage values to high amplitude data in the baseline interval, and thereby result in a positive correlation in the baseline. This was not the case. The relatively low correlation



**FIGURE 3 | Temporal evolution of VSD activity and local active information storage.** Spatial averages over the  $67 \times 137$  data pixels for VSD activity (black traces), and the LAIS (red traces) versus time. Motion directions are indicated by arrows for each panel. Note that LAIS for the vertical, the right, and the downward-right

motion directions continues to rise toward the end of the stimulus interval, despite declining activity levels. Also note that the unexpected onset response at approximately 40 ms leads to negative active information storage. For an explanation see the Materials and Methods section.

**Table 1 | Correlation of LAIS and local VSD activity.**

Motion direction	Correlation coefficient			
	Full epoch	−1 to 0 s	0.04–0.14 s	0.2–1 s
0°	0.05*	−0.33*	−0.09*	0.45*
4 °	0.09*	−0.50*	−0.20*	0.65*
90°	0.12*	−0.30*	−0.13*	0.48*
180°	0.07*	−0.27*	−0.22*	0.44*
225°	0.07*	−0.58*	−0.22*	0.71 *
270°	0.17*	−0.39*	−0.33*	0.68*
315°	0.03*	−0.37*	−0.17*	0.40*

Correlation coefficients are Spearman rank correlations.

\* $p < 0.05/7$ .

coefficients across the complete time-interval, which are between 0.02 and 0.13, further suggest that LAIS increases due not follow higher VSD signals tightly. Therefore, LAIS extracts additional useful information about neural processing. This point is further supported by the stimulus-dependent changes that seem more pronounced in LAIS maps than in the VSD activity maps (compare left and right columns in **Figure 2**).

For future studies the amplitude-bias problem introduced by the fixed-width kernel estimator should easily be overcome using a Kraskov-type variable width kernel estimator—see the original work of Kraskov et al. (2004), and Lindner et al. (2011); Vicente et al. (2011); Wibral et al. (2011, 2013); Lizier (2012) for implementation details of Kraskov-type estimators. Another possibility would be to condition the analysis on the activity level, as for example done for the transfer entropy measure by Stetter et al. (2012).

4.2. TIMESCALES OF LAIS

The recovery time of the stimulus-induced, transient drop in LAIS was 34 ms. A drop of this kind means that the activity before the drop (baseline activity) was not useful to predict the activity

during the drop (the onset response). This is expected as the stimulus is presented in an unpredictable way to the neural system. However, the recovery time of this drop of approximately 34 ms yields an insight into the intrinsic storage time scales of the neural processes. We note that the observed time-scale corresponds to the high beta frequency band around 29 Hz (1/34 ms). In how far this is an incidental finding or bears significance must be clarified in future studies.

4.3. ON THE INTERPRETATION OF LOCAL ACTIVE INFORMATION STORAGE MEASURES IN NEUROSCIENCE

When working with measures from information theory, it is important to keep in mind that the basic definition of information as given by Shannon revolves around the probabilities of events and the possibility to encode something using these events. To separate Shannon information content from information about something (new) in a more colloquial sense, one often also speaks about *potential* or *syntactic* information, when referring to Shannon information content, of *semantic* information when referring to human interpretable information, and last of pragmatic information for our everyday notion of information as in “news” [for details see for example the treatment of this topic by Deacon (2010)]. In the same way, LAIS does not directly describe information that the neural system stores about things in the outside world—rather, it quantifies how much of the future (Shannon) information in the activity can be predicted from its past.

In fact, information in the neural system *about* something in the outside world would have to be quantified by some kind of mutual information between aspects of the outside world and neural activity, while information in the classic sense of semantic information represented symbolically (e.g., in books, and other media) would be even more complicated: theoretically it should be quantified as a mutual information between the medium containing the symbols and activity in the neural system, while additionally satisfying the constraint that this mutual information should vanish when conditioning

on the states of the world variables represented by the symbols.

While this lack of a more semantic interpretation of LAIS may seem disappointing at first, the quantification of the predictable amount of information makes this measure highly useful in understanding information processing at a more abstract level. This is important wherever we have not yet gained insights into what (if anything) may be explicitly represented by a neural system. Moreover, the focus on predictability provides a non-trivial link between LAIS and current theories of brain function as pointed out below. Nevertheless, a use of the concept in neuroscience may have to take the properties of the receiving neuron or brain area into account to consider how much of the mathematical storage in a signal is accessible to neural information processing. To address this concern, we used a pooling over all available data in space and time here as it seems to represent a way by which a receiving brain area could construct its (implicit) guesses of the underlying probability densities. However, also other strategies are possible and need to be explored in the future. As one example for another strategy of probability-density estimation, we have investigated a construction of probability densities via pooling over all data pixels but separately for each point in time. This approach avoids any potential issues with non-stationarities, but obscures the view of the “typical transitions” in the system over time to a point that no interpretable results were obtained (data not shown).

#### 4.4. LOCAL ACTIVE INFORMATION STORAGE AND PREDICTIVE CODING THEORIES

Information storage in neural activity means that information from the past of a neural process will predict some non-zero fraction of information in the future of this process. It is via this predictability improvement that information storage is also tightly connected with predictive coding, an important family of theories of cortical function. Predictive coding theories propose that a neural system is constantly generating predictions about the incoming sensory input (Rao and Ballard, 1999; Knill and Pouget, 2004; Friston, 2005; Bastos et al., 2012) to adapt internal behavior and processing accordingly. These predictions of incoming information must be implemented in neural activity, and they typically need to be maintained for a certain duration—as it will typically be unknown to the system when the predictive information will be needed. Hence, the neural activity subserving prediction must itself have a predictable character, i.e., non-zero information storage *in activity*. Analysis of active information storage may thereby enable us to test central assumptions of predictive coding theories rather directly. This is important because tests of predictive coding theories so far mostly relied on the predictions being explicitly known and then violated—a condition not given for most brain areas beyond early sensory cortices, and for most situations beyond simple experimental designs. Here, the quantification of the predictability of brain signals themselves via LAIS may open a second approach to testing these important theories. To this end we may scan brain signals for negative LAIS, as negative LAIS values indicate the past states of the neural signals in question were not informative about the future, i.e., negative LAIS signals a breakdown of predictions. In our example dataset this

was brought about by the sudden, unexpected onset of the stimulus. However, the same analyses may be applied in situations that are not under external control—for example to analyze internally driven changes in information processing regimes.

In relation to predictive coding theories it is also encouraging that the predictive information was found on timescales related to the beta band. This is because this frequency band has been implied in the intra-cortical transfer of predictions (Bastos et al., 2012).

#### 4.5. SUB-SAMPLING AND COARSE GRAINING, AND NON-LOCALITY OF PDF ESTIMATION

When interpreting LAIS values it should be kept in mind that in neural recordings we typically do not observe the system fully or at the relevant scales—in contrast to artificial systems, such as cellular automata and robots, where the full system is accessible. More precisely, in neural data one of two types of sub-sampling is typically present—either coarse graining with local averaging of activity indices (as in VSD) or sub-sampling proper, where neural activity is recorded faithfully (e.g., via intracellular recordings) but with incomplete coverage of the full system. This sub-sampling may have non-trivial effects on the probability distributions of neural events [see for example Priesemann et al. (2009, 2013)]. Hence, LAIS values obtained under sub-sampling should be interpreted as *relative* rather than absolute measures and should only be compared to other experiments, or experimental conditions, when obtained under identical sampling conditions.

In addition there is necessarily temporal subsampling in the form of finite data; we therefore note again the potential for bias in the actual MI values returned via the use of kernel estimation here, particularly for large embedding dimensions and small kernel widths. Alternatives to kernel estimator are known to be more effective in bias compensation [e.g., Kraskov-Grassberger-Stögbauer estimation (Kraskov et al., 2004)]; or use of use kernel estimation is solely motivated by practical computational reasons. Effects of temporal subsampling also mandates to focus on relative rather than absolute values within this experiment.

Even within the experiment though, the bias may not be evenly distributed amongst the local MI values, which tend to exhibit larger bias for low frequency events. With that said, our experiment did use a large amount of data (by pooling observations over pixels and time), which counteracts such concerns to a large degree, and many of the key results (e.g., **Figure 3**) involve averaging or correlating over many local values, which further ameliorates this. There are techniques suggested to alleviate bias in local or pointwise MI, e.g., Turney and Pantel (2010), and while none were applied here, we do not believe this alters the general conclusions of our experiment for the aforementioned reasons. As a particular example, the surprise caused by the onset of stimulus is still clearly visible as negative LAIS, despite any propensity for such low frequency events to have been biased strongly toward positive values.

#### 4.6. ON THE LOCALITY OF INFORMATION VALUES

As a concluding remark, we would like to point out again that various “levels of locality” have to be carefully chosen in the analysis



of neural data. One important level is the spatial extent (ensemble of agents) and the time span over which data are pooled to obtain the PDF. However, even pooling over a large spatial extent, i.e., many agents and a long time span, may still allow to interpret the information value of the data agent-by-agent and time step-by-time step, if agents  $i$  are *identical* and samples at subsequent time points  $t$  come from a *stationary* random process [see the book of Lizier (2013) for several examples]. This is because one may pool data to estimate a PDF as long as these data can be considered “replications,” i.e., as coming from the same random variable. Pooling data under these conditions will obviously not bias the PDF estimate away from the ground truth for any agent or time step. Irrespective of how many data points are pooled this way, it is then still possible to interpret each data point  $(x_{i,t}, \mathbf{x}_{i,t-1}^{k-})$  individually in terms of its LAIS,  $a(x_t, \mathbf{x}_{t-1}^{k-})$ . This locality of information values is identical to the local interpretation of the (Shannon) information terms  $h(x_i) = -\log(p(x_i))$  that together, as a weighted average over all possible outcomes  $x_i$ , yield the (Shannon) entropy  $H(X) = \sum_i p(x_i)h(x_i)$  of a random variable  $X$ . As explained for example by MacKay (2003, chapter 4), each and every outcome  $x_i$  of a random variable  $X$  has its own meaningful Shannon information value  $h(x_i)$ , that may be very different from that of another outcome  $x_j$ , although repeated draws from this random variable can be considered stationary. It is this sense of “local” that gives *local* active information storage its name. In contrast, how locally in space and time we obtain the PDF is more important for the precision of the LAIS estimates.

In the analysis of LAIS from neural data three issues will necessarily blur locality, and impair the precision of the LAIS estimate to some extent:

1. If a pool of identical agents  $i$ , all running identical stationary random processes  $X_i$ , is available, the only blurring of locality arises due to the intrinsic temporal extent of the state variables. However, while the stored information may be encoded in a temporally non-local state  $\mathbf{x}_{t-1}^{k-}$ , this information is used to predict the next value of the process  $x_t$  at a *single* point in time.
2. If agents are non-identical, but their data are pooled nonetheless, then the overall empirical PDF obtained across these agents is no longer fully representative of each single agent and the local information storage values per agent are biased due to the use of this non-optimal PDF. This effect may be present to some extent in our analysis, as we cannot guarantee that all parts of area 18 behave strictly identical.
3. If the random process in question is not stationary, then a PDF obtained via pooling samples across time is also not representative of what happens at single points in time, and again a bias in the LAIS values for each agent and time step arises. This bias is potentially more severe. Nevertheless, we pooled data across all available time samples here, as this seems to be closer to the strategy available to a neuron in a downstream brain area (also see section 2.2.5), when trying to estimate, or adapt to, its input distribution. This is because a neuron may more easily estimate approximate PDFs of its inputs across time than across all possible neurons in an upstream brain area, to most of which it simply doesn't interface.

## 5. CONCLUSION

Distributed information processing in neural systems can be decomposed into component processes of information transfer, storage and modification. Information storage can be quantified locally in space and time using an information theoretic measure termed local active information storage (LAIS). Here we present for the first time the application of this measure to neural data. We show that storage reflects neural properties such as stimulus preferences and surprise, and reflects the abstract concept of an ongoing stimulus despite the locally random nature of this stimulus. We suggest that LAIS will be a useful quantity to test theories of cortical function, such as predictive coding.

## ACKNOWLEDGMENTS

The authors thank Matthias Kaschube from the Frankfurt Institute for Advanced Studies (FIAS) for fruitful discussions on active information storage. Viola Priesemann received funding from the Federal Ministry of Education and Research (BMBF) Germany under grant number 01GQ0811 (Bernstein).

## FUNDING

Michael Wibral and Viola Priesemann were supported by LOEWE Grant “Neuronale Koordination Forschungsschwerpunkt Frankfurt (NeFF).” Michael Wibral thanks the Commonwealth Scientific and Industrial Research Organisation (CSIRO) for supporting a visit in Sydney which contributed to this work. Sebastian Vögler was supported by the Bernstein Focus: Neurotechnology (BFNT) Frankfurt/M.

## REFERENCES

- Ay, N., and Polani, D. (2008). Information flows in causal networks. *Adv. Complex Syst.* 11, 17. doi: 10.1142/S0219525908001465
- Bastos, A. M., Usrey, W. M., Adams, R. A., Mangun, G. R., Fries, P., and Friston, K. J. (2012). Canonical microcircuits for predictive coding. *Neuron* 76, 695–711. doi: 10.1016/j.neuron.2012.10.038
- Butts, D. A. (2003). How much information is associated with a particular stimulus? *Network* 14, 177–187. doi: 10.1088/0954-898X/14/2/301
- Butts, D. A., and Goldman, M. S. (2006). Tuning curves, neuronal variability, and sensory coding. *PLoS Biol.* 4:e92. doi: 10.1371/journal.pbio.0040092
- Cover, T. M., and Thomas, J. A. (1991). *Elements of Information Theory*. New York, NY: Wiley-Interscience. doi: 10.1002/0471200611
- Crutchfield, J. P., and Feldman, D. P. (2003). Regularities unseen, randomness observed: levels of entropy convergence. *Chaos* 13, 25–54. doi: 10.1063/1.1530990
- Dasgupta, S., Wörgötter, F., and Manoonpong, P. (2013). Information dynamics based self-adaptive reservoir for delay temporal memory tasks. *Evol. Syst.* 4, 235–249. doi: 10.1007/s12530-013-9080-y
- Deacon, T. W. (2010). “What is missing from theories of information?” (chapter 8) in *Information and the Nature of Reality*, eds P. Davies and N. H. Gregersen (Cambridge: Cambridge University Press), 146.
- DeWeese, M. R., and Meister, M. (1999). How to measure the information gained from one symbol. *Network* 10, 325–340. doi: 10.1088/0954-898X/10/4/303
- Faes, L., Nollo, G., and Porta, A. (2012). Non-uniform multivariate embedding to assess the information transfer in cardiovascular and cardiorespiratory variability series. *Comput. Biol. Med.* 42, 290–297. doi: 10.1016/j.combiomed.2011.02.007
- Faes, L., Porta, A., Rossato, G., Adami, A., Tonon, D., Corica, A., et al. (2013). Investigating the mechanisms of cardiovascular and cerebrovascular regulation in orthostatic syncope through an information decomposition strategy. *Auton. Neurosci.* 178, 76–82. doi: 10.1016/j.autneu.2013.02.013
- Fano, R. (1961). *Transmission of Information: A Statistical Theory of Communications*. Cambridge, MA: The MIT Press.



- Friston, K. (2005). A theory of cortical responses. *Philos. Trans. R. Soc. Lond. B Biol. Sci.* 360, 815–836. doi: 10.1098/rstb.2005.1622
- Gómez, C., and Hornero, R. (2010). Entropy and complexity analyses in alzheimer's disease: an MEG study. *Open Biomed. Eng. J.* 4, 223. doi: 10.2174/1874120701004010223
- Granger, C. W. J. (1969). Investigating causal relations by econometric models and cross-spectral methods. *Econometrica* 37, 424–438. doi: 10.2307/1912791
- Kantz, H., and Schreiber, T. (2003). *Nonlinear Time Series Analysis*. 2nd Edn. Cambridge: Cambridge University Press. doi: 10.1017/CBO9780511755798
- Knill, D. C., and Pouget, A. (2004). The bayesian brain: the role of uncertainty in neural coding and computation. *Trends Neurosci.* 27, 712–719. doi: 10.1016/j.tins.2004.10.007
- Kraskov, A., Stogbauer, H., and Grassberger, P. (2004). Estimating mutual information. *Phys. Rev. E Stat. Nonlin. Soft Matter Phys.* 69, 066138. doi: 10.1103/PhysRevE.69.066138
- Langton, C. G. (1990). Computation at the edge of chaos: phase transitions and emergent computation. *Physica D* 42, 12–37. doi: 10.1016/0167-2789(90)90064-V
- Lindner, M., Vicente, R., Priesemann, V., and Wibral, M. (2011). Trentool: a Matlab open source toolbox to analyse information flow in time series data with transfer entropy. *BMC Neurosci.* 12:1–22. doi: 10.1186/1471-2202-12-119
- Lizier, J. T. (2012). JIDT: an information-theoretic toolkit for studying the dynamics of complex systems. Available online at: <http://code.google.com/p/information-dynamics-toolkit/>
- Lizier, J. T. (2013). *The Local Information Dynamics of Distributed Computation in Complex Systems*. Springer theses. Springer. doi: 10.1007/978-3-642-32952-4\_2
- Lizier, J. T., Atay, F. M., and Jost, J. (2012a). Information storage, loop motifs, and clustered structure in complex networks. *Phys. Rev. E Stat. Nonlin. Soft Matter Phys.* 86(2 Pt 2), 026110. doi: 10.1103/PhysRevE.86.026110
- Lizier, J. T., Prokopenko, M., and Zomaya, A. Y. (2012b). Local measures of information storage in complex distributed computation. *Inform. Sci.* 208, 39–54. doi: 10.1016/j.ins.2012.04.016
- Lizier, J. T., Flecker, B., and Williams, P. L. (2013). “Towards a synergy-based approach to measuring information modification,” in *Proceedings of the 2013 IEEE Symposium on Artificial Life (ALIFE)* (Singapore), 43–51. doi: 10.1109/ALIFE.2013.6602430
- Lizier, J. T., and Prokopenko, M. (2010). Differentiating information transfer and causal effect. *Eur. Phys. J. B* 73, 605–615. doi: 10.1140/epjb/e2010-00034-5
- MacKay, D. J. (2003). *Information Theory, Inference and Learning Algorithms*. Cambridge: Cambridge University Press.
- Mitchell, M. (1998). “Computation in cellular automata: a selected review,” in *Non-Standard Computation*, eds T. Gramß, S. Bornholdt, M. Groß, M. Mitchell, and T. Pellizzari (Weinheim: Wiley-VCH Verlag GmbH & Co. KGaA), 95–140.
- Mitchell, M., Hraber, P., and Crutchfield, J. P. (1993). Revisiting the edge of chaos: evolving cellular automata to perform computations. *Complex Systems* 7, 89–130.
- Obst, O., Boedecker, J., Schmidt, B., and Asada, M. (2013). On active information storage in input-driven systems. arXiv: 1303.5526.
- Pincus, S. M. (1991). Approximate entropy as a measure of system complexity. *Proc. Natl. Acad. Sci. U.S.A.* 88, 2297–2301. doi: 10.1073/pnas.88.6.2297
- Porta, A., Baselli, G., Liberati, D., Montano, N., Cogliati, C., Gnechi-Ruscione, T., et al. (1998). Measuring regularity by means of a corrected conditional entropy in sympathetic outflow. *Biol. Cybernet.* 78, 71–78. doi: 10.1007/s004220050414
- Porta, A., Guzzetti, S., Montano, N., Pagani, M., Somers, V., Malliani, A., et al. (2000). Information domain analysis of cardiovascular variability signals: evaluation of regularity, synchronisation and co-ordination. *Med. Biol. Eng. Comput.* 38, 180–188. doi: 10.1007/BF02344774
- Priesemann, V., Munk, M. H. J., and Wibral, M. (2009). Subsampling effects in neuronal avalanche distributions recorded *in vivo*. *BMC Neurosci.* 10:40. doi: 10.1186/1471-2202-10-40
- Priesemann, V., Valderrama, M., Wibral, M., and Le Van Quyen, M. (2013). Neuronal avalanches differ from wakefulness to deep sleep—evidence from intracranial depth recordings in humans. *PLoS Comput. Biol.* 9:e1002985. doi: 10.1371/journal.pcbi.1002985
- Prokopenko, M., Boschiatti, F., and Ryan, A. J. (2009). An information-theoretic primer on complexity, self-organization, and emergence. *Complexity* 15, 11–28. doi: 10.1002/cplx.20249
- Prokopenko, M., Gerasimov, V., and Tanev, I. (2006). “Evolving spatiotemporal coordination in a modular robotic system,” in *From Animals to Animats 9: Proceedings of the Ninth International Conference on the Simulation of Adaptive Behavior (SAB'06)*. Lecture notes in computer science. Vol. 4095, eds S. Nolfi, G. Baldassarre, R. Calabretta, J. C. T. Hallam, D. Marocco, J.-A. Meyer, et al. (Berlin: Springer), 558–569.
- Ragwitz, M., and Kantz, H. (2002). Markov models from data by simple nonlinear time series predictors in delay embedding spaces. *Phys. Rev. E Stat. Nonlin. Soft Matter Phys.* 65(5 Pt 2), 056201. doi: 10.1103/PhysRevE.65.056201
- Rao, R. P., and Ballard, D. H. (1999). Predictive coding in the visual cortex: a functional interpretation of some extra-classical receptive-field effects. *Nat. Neurosci.* 2, 79–87. doi: 10.1038/4580
- Richman, J. S., and Moorman, J. R. (2000). Physiological time-series analysis using approximate entropy and sample entropy. *Am. J. Physiol. Heart. Circ. Physiol.* 278, H2039–H2049.
- Schreiber, T. (2000). Measuring information transfer. *Phys. Rev. Lett.* 85, 461–464. doi: 10.1103/PhysRevLett.85.461
- Stetter, O., Battaglia, D., Soriano, J., and Geisel, T. (2012). Model-free reconstruction of excitatory neuronal connectivity from calcium imaging signals. *PLoS Comput. Biol.* 8:e1002653. doi: 10.1371/journal.pcbi.1002653
- Takens, F. (1981). “Detecting strange attractors in turbulence” (chapter 21) in *Dynamical Systems and Turbulence, Warwick 1980*. Lecture notes in mathematics. Vol. 898, eds D. Rand and L.-S. Young (Berlin: Springer), 366–381.
- Turing, A. M. (1936). On computable numbers, with an application to the Entscheidungsproblem. *Proc. Lond. Math. Soc.* 42, 230–265.
- Turney, P. D., and Pantel, P. (2010). From frequency to meaning: vector space models of semantics. *J. Artif. Intell. Res.* 37, 141–188.
- Vakorin, V. A., Mišić, B., Krakovska, O., and McIntosh, A. R. (2011). Empirical and theoretical aspects of generation and transfer of information in a neuromagnetic source network. *Front. Syst. Neurosci.* 5:96. doi: 10.3389/fnsys.2011.00096
- Vicente, R., Wibral, M., Lindner, M., and Pipa, G. (2011). Transfer entropy—a model-free measure of effective connectivity for the neurosciences. *J. Comput. Neurosci.* 30, 45–67. doi: 10.1007/s10827-010-0262-3
- Wang, X. R., Miller, J. M., Lizier, J. T., Prokopenko, M., and Rossi, L. F. (2012). Quantifying and tracing information cascades in swarms. *PLoS ONE* 7:e40084. doi: 10.1371/journal.pone.0040084
- Wibral, M., Pampu, N., Priesemann, V., Siebenhühner, F., Seiwert, H., Lindner, M., et al. (2013). Measuring information-transfer delays. *PLoS ONE* 8:e55809. doi: 10.1371/journal.pone.0055809
- Wibral, M., Rahm, B., Rieder, M., Lindner, M., Vicente, R., and Kaiser, J. (2011). Transfer entropy in magnetoencephalographic data: quantifying information flow in cortical and cerebellar networks. *Prog. Biophys. Mol. Biol.* 105, 80–97. doi: 10.1016/j.pbiomolbio.2010.11.006
- Wiener, N. (1956). “The theory of prediction,” in *Modern Mathematics for the Engineer*, ed E. F. Beckmann (New York, NY: McGraw-Hill).
- Zipser, D., Kehoe, B., Littlewort, G., and Fuster, J. (1993). A spiking network model of short-term active memory. *J. Neurosci.* 13, 3406–3420.

**Conflict of Interest Statement:** The authors declare that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.

Received: 09 November 2013; paper pending published: 02 December 2013; accepted: 09 January 2014; published online: 28 January 2014.

Citation: Wibral M, Lizier JT, Vögler S, Priesemann V and Galuske R (2014) Local active information storage as a tool to understand distributed neural information processing. *Front. Neuroinform.* 8:1. doi: 10.3389/fninf.2014.00001

This article was submitted to the journal *Frontiers in Neuroinformatics*.

Copyright © 2014 Wibral, Lizier, Vögler, Priesemann and Galuske. This is an open-access article distributed under the terms of the Creative Commons Attribution License (CC BY). The use, distribution or reproduction in other forums is permitted, provided the original author(s) or licensor are credited and that the original publication in this journal is cited, in accordance with accepted academic practice. No use, distribution or reproduction is permitted which does not comply with these terms.



# Reduced predictable information in brain signals in autism spectrum disorder

Carlos Gómez<sup>1</sup>, Joseph T. Lizier<sup>2</sup>, Michael Schaum<sup>3</sup>, Patricia Wollstadt<sup>3</sup>, Christine Grützner<sup>4</sup>, Peter Uhlhaas<sup>5</sup>, Christine M. Freitag<sup>6</sup>, Sabine Schlitt<sup>6</sup>, Sven Bölte<sup>6</sup>, Roberto Hornero<sup>1</sup> and Michael Wibral<sup>3\*</sup>

<sup>1</sup> Biomedical Engineering Group, E. T. S. Ingenieros de Telecomunicación, University of Valladolid, Valladolid, Spain

<sup>2</sup> Commonwealth Scientific and Industrial Research Organisation, Computational Informatics, Marsfield, NSW, Australia

<sup>3</sup> MEG Unit, Brain Imaging Center, Johann Wolfgang Goethe University, Frankfurt am Main, Germany

<sup>4</sup> Department of Neurophysiology, Max-Planck Institute for Brain Research, Frankfurt am Main, Germany

<sup>5</sup> Institute of Neuroscience and Psychology, University of Glasgow, Glasgow, UK

<sup>6</sup> Department of Child and Adolescent Psychiatry, Psychosomatics and Psychotherapy, Johann Wolfgang Goethe University, Frankfurt am Main, Germany

## Edited by:

Daniele Marinazzo, University of Gent, Belgium

## Reviewed by:

Karl Friston, University College London, UK

Rikkert Hindriks, Universitat Pompeu Fabra, Spain

## \*Correspondence:

Michael Wibral, MEG Unit, Brain Imaging Center, Johann Wolfgang Goethe University, Heinrich-Hoffmann Strasse 10, Frankfurt am Main, D-602528, Germany  
e-mail: wibral@em.uni-frankfurt.de

Autism spectrum disorder (ASD) is a common developmental disorder characterized by communication difficulties and impaired social interaction. Recent results suggest altered brain dynamics as a potential cause of symptoms in ASD. Here, we aim to describe potential information-processing consequences of these alterations by measuring active information storage (AIS)—a key quantity in the theory of distributed computation in biological networks. AIS is defined as the mutual information between the past state of a process and its next measurement. It measures the amount of stored information that is used for computation of the next time step of a process. AIS is high for rich but predictable dynamics. We recorded magnetoencephalography (MEG) signals in 10 ASD patients and 14 matched control subjects in a visual task. After a beamformer source analysis, 12 task-relevant sources were obtained. For these sources, stationary baseline activity was analyzed using AIS. Our results showed a decrease of AIS values in the hippocampus of ASD patients in comparison with controls, meaning that brain signals in ASD were either less predictable, reduced in their dynamic richness or both. Our study suggests the usefulness of AIS to detect an abnormal type of dynamics in ASD. The observed changes in AIS are compatible with Bayesian theories of reduced use or precision of priors in ASD.

**Keywords:** autism spectrum disorder, information theory, active information storage, complex systems, magnetoencephalography, hippocampus, predictive coding

## 1. INTRODUCTION

It has been 70 years since Kanner (1943) and Asperger (1944) first described an intriguing disorder characterized by the children's inability to relate themselves in the ordinary way to people and situations from the beginning of life. The symptom cluster described by Kanner has been called autism spectrum disorder (ASD), and it is clinically defined by a triad of deficits comprising impairments in communication, social interaction, and behavioral flexibility (Wing and Gould, 1979). Prevalence studies estimate that ASD affects 2–10 children per 1000 births (Yeargin-Allsopp et al., 2003; Baird et al., 2006). It is characterized by an early onset, since these typical behaviors show up before the age of 36 months. Nevertheless, ASD is a permanent developmental disorder that will continue into adulthood. Great heterogeneity in development has been reported, with some individuals losing skills over time, others reaching a plateau in adolescence, and still others manifesting a pattern of continued development in adulthood (Seltzer et al., 2003). Due to the complexity and variety of the symptoms with which autistic individuals present to clinicians, it has been difficult to conceptualize a defining neurological mechanism that might underlie the core features of this disorder (Bauman and Kemper,

2005). Therefore, new techniques are necessary to achieve a more detailed understanding of this disorder, and ultimately an earlier identification and more effective interventions and treatment (Bauman and Kemper, 2005).

ASD symptoms, but also self-reports of ASD patients (Williams and Bishop, 1994) and the phenomenon of “savants” (Treffert, 2009) point to fundamentally altered modes of information processing in the brain of patients with autism. While autism has most likely genetic roots, the final disease outcome is the result of a developmental trajectory and of an interaction with the environment. It seems safe to assume that in the autistic brain information processing is also optimized or adapted during development in some way that genetics and environment allow, and the complex developmental trajectory of neuroanatomical changes in ASD supports this view (Bauman and Kemper, 2005). However, even describing this adapted, but altered information processing in ASD at a *neurophysiological* level—beyond behavioral outcomes—has remained difficult. Results at the neurophysiological level have so far mostly dealt with descriptors of the dynamics, such as time-frequency analysis (Sun et al., 2012), connectivity methods (Belmonte et al., 2004a,b) or entropy measures (Bosl et al., 2011). However, it has been difficult to address

information processing more directly. These difficulties were foremost conceptual because what we actually mean when using the term *information processing* in biological systems has been unclear. Only recently formal, operational descriptions of information processing and its components have become available (Langton, 1990; Mitchell, 2011; Lizier, 2013). These descriptions can be traced back to Turing's finding that every act of information processing can be decomposed into the component processes of information storage, transfer and modification (Turing, 1936). Later, Langton and others expanded these concepts to describe the emergence of the capacity to perform arbitrary information processing algorithms, or "universal computation," in complex systems, such as cellular automata (Langton, 1990; Mitchell, 2011).

Of the three component processes above—information transfer, storage, and modification—information storage in particular has been used with great success to evolve (Prokopenko et al., 2006), and optimize (Dasgupta et al., 2013) self-organizing information processing systems. This success was enabled by the introduction of quantitative measures of information storage in the form of excess entropy by Grassberger (1986) (introduced as "effective measure complexity," and later reintroduced as excess entropy by Crutchfield and Feldman, 2003), and in the form of the *active information storage* (AIS) by Lizier et al. (2012). Despite their success in artificial systems, however, these measures have not been applied yet to biological neural systems.

One reason for this slow adoption may be that we face an apparent complication in biological neural systems, as in these systems information storage may at first sight take various forms, e.g., as reverberant neural activity or as synaptic changes (Zipser et al., 1993). However, a use of the stored information for information processing inevitably requires its re-expression in neural activity and its interaction with ongoing neural activity and incoming information. Hence, information storage actively in use for a computation will be reflected in the dynamics of neural activity, and is therefore accessible based on recordings of neural activity. Information storage in neural activity will be reflected by the fact that information from the past of a neural process will serve to predict a certain fraction of information in the future of this process, by virtue of the very definition of storage. To measure information-theoretically this amount of information in the future of a process that is predicted by its past state, we use AIS (Lizier et al., 2012), described in section 2.1.

It is via this predictable information that information storage is also tightly connected with predictive coding, an important family of theories of cortical function. Predictive coding theories propose that a neural system is constantly generating predictions about the incoming sensory input (Rao and Ballard, 1999; Friston et al., 2006; George and Hawkins, 2009; Bastos et al., 2012; Grossberg, 2013) to adapt internal behavior and processing accordingly. The prediction of incoming information that forms the central idea of predictive coding theory must happen via neural activity. These predictions typically need to be maintained for a short interval—as it is not known precisely *a priori* when the predictive information

will be needed. Hence, the neural activity subserving prediction must itself have a predictable character, i.e., non-zero information storage. Analysis of AIS thereby enables us to test central assumptions of predictive coding theories rather directly.

The close link between information storage and general theories of cortical function makes AIS also a promising candidate measure to investigate altered information processing in ASD. Influential accounts of altered perception in ASD hold that there is either some form of reduced top-down control (Happé and Frith, 2006; Pellicano and Burr, 2012; Friston et al., 2013), or a reduced noise in the ascending sensory systems (e.g., Mottron et al., 2006). Both views can be formalized using a Bayesian formalism, i.e., a predictive coding theory of perception in ASD (Pellicano and Burr, 2012). Despite this semi-quantitative formalism, many aspects of altered perception in ASD can be explained in a Bayesian framework in one of two opposing ways—either by less dominant (top-down) expectations or more precise sensory inputs (Brock, 2012). Here, quantities such as the amount of predictable information in a neural signal—as measured by AIS—may play a crucial role in distinguishing these theoretical accounts based on experimental evidence, as they quantify the amount of information that is reliably obtainable from a brain area.

To explore the potential of AIS for ASD research, we here apply this measure to magnetoencephalographic (MEG) data obtained from a group of patients with high functioning ASD and matched healthy controls. We focus our study on the visual system, as atypical perception is particularly well documented in this system (see Williams and Bishop, 1994; Plaisted, 2001; Ropar and Mitchell, 2002; Mitchell and Ropar, 2004; Bertone et al., 2005; Rogers and Ozonoff, 2005; Happé and Frith, 2006; Mottron et al., 2006; Sheppard et al., 2007; Baron-Cohen et al., 2009; David et al., 2010, but also see, Ropar and Mitchell, 1999, 2001).

## 2. MATERIALS AND METHODS

### 2.1. ACTIVE INFORMATION STORAGE—DEFINITION AND PRACTICAL ESTIMATION

We assume that the neural signals we record from a system  $\mathcal{X}$  can be treated as realizations  $x_t$  of random variables  $X_t$  that together form the random process  $\mathbf{X}$ , describing the system's dynamics.

AIS is then simply defined as the mutual information  $I(\mathbf{X}_{t-1}^{k-}; X_t)$ —see Cover and Thomas (1991)—between the past state random variable  $\mathbf{X}_{t-1}^{k-} = \{X_{t-1}, \dots, X_{t-1-k}\}$  of a process and its next random variable  $X_t$  (Lizier et al., 2012):

$$A_{X_t} = \lim_{k \rightarrow \infty} I(\mathbf{X}_{t-1}^{k-}; X_t) = \lim_{k \rightarrow \infty} \left\langle \log \frac{p(\mathbf{x}_{t-1}^{k-}, x_t)}{p(\mathbf{x}_{t-1}^{k-}) p(x_t)} \right\rangle. \quad (1)$$

Here the averaging  $\langle \cdot \rangle$  via  $p(\mathbf{x}_{t-1}^{k-}, x_t)$  in principle has to be taken over an ensemble of realizations of the process at time point  $t$  (e.g., via physical replications of the system  $\mathcal{X}$ ). For stationary processes, however, where all random variables that form the process  $\mathbf{X}$  have identical probability distribution, we can use

time-averaging instead of the ensemble average and Equation (1) simplifies to:

$$A_X = \lim_{k \rightarrow \infty} \left\langle \log \frac{p(\mathbf{x}_{t-1}^{k-}, x_t)}{p(\mathbf{x}_{t-1}^{k-}) p(x_t)} \right\rangle_t \quad (2)$$

where the averaging can now be taken with respect to time  $t$ .

The use of the multivariate collection  $\mathbf{X}_{t-1}^{k-}$  is particularly important here—it is intended to capture the *state* of the underlying dynamical system  $\mathcal{X}$ , and can be viewed as a state-space reconstruction of it. In this fashion, AIS brings together aspects of both dynamical systems theory and information theory in its analysis. The AIS tells us how much information could be *predicted* about the next measurement of a process by examining its past state. For a linear perspective, this is akin to building a classical autoregressive model of order  $k$  and measuring how well that model predicts the next measurements of the process. Importantly though, the use of information theory here is a more general approach which captures non-linear auto-dependencies in the process, and does so in a model-free way. As such, we refer to this component of the prediction of the next measurement as *information storage*, capturing the information-theoretic basis of the self-prediction. This also highlights that AIS quantifies how much information from the past state is involved in generating or computing the next value of a process, in contrast to other information sources (i.e., information transferred from other processes as quantified by the transfer entropy Schreiber, 2000, a non-linear analogy of the Granger causality) as discussed by Lizier et al. (2010). We call this the *active* component of information storage since it quantifies the stored information actively *in use* in this generation of the next value, as opposed to that passively stored for later use, e.g., in synaptic weights. Zipser et al. (1993) discuss this contrast in active and passive storage, though our perspective generalizes the active storage beyond merely “maintaining neural activity” (as described by Zipser et al., 1993) to more complex non-linear auto-correlations, and may additionally capture contributions of passive storage when they are re-expressed in dynamics.

Now, if the history before a certain time point  $t - k_{\max}$  does not help to improve the prediction of  $X_t$  we can further simplify Equation (2). Technically speaking,  $X_t$  then is conditionally independent of all  $X_{t-k_i}$  with  $k_i > k_{\max}$ :

$$\lim_{k \rightarrow \infty} p(x_t | \mathbf{x}_{t-1}^{k-}) = p(x_t | \mathbf{x}_{t-1}^{k_{\max}-}), \quad (3)$$

and Equation (2) becomes:

$$A_X = \left\langle \log \frac{p(\mathbf{x}_{t-1}^{k_{\max}-}, x_t)}{p(\mathbf{x}_{t-1}^{k_{\max}-}) p(x_t)} \right\rangle_t. \quad (4)$$

The parameter  $k_{\max}$  can be determined using Ragwitz’ criterion (Ragwitz and Kantz, 2002), as suggested for example in Vicente et al. (2011), and implemented in the TRENTOOL toolbox (Lindner et al., 2011). For the analyses presented here, we used  $k_{\max} = 10$  on data with a sampling rate of 300 Hz.

For the practical estimation of Equation (4) for continuous data, as analyzed here, various estimation techniques exist, such as binning and kernel approaches. Here, we used a kernel-based estimator (see Kantz and Schreiber, 2003 for more information on kernel-based estimators) as implemented in the open source JAVA Information Dynamics Toolkit (Lizier, 2012), with a kernel width  $\epsilon$  of 0.5 standard deviations of the data.

## 2.2. DATA ACQUISITION

We recorded magnetoencephalography (MEG) signals in 10 ASD patients and 14 matched healthy control (HC) subjects in a visual task. More details of this study can be found in Sun et al. (2012). Its most important aspects are summarized in the following paragraphs.

### 2.2.1. Participants and task

All ASD patients (mean age:  $30.3 \pm 9.6$ ) were clinically diagnosed and suffered from Asperger’s disorder, or pervasive developmental disorder not otherwise specified (PDD-NOS) according to DSM-IV (American Psychiatric Association, 2000). The clinical diagnosis was corroborated using the German form of the Autism Diagnostic Interview-Revised (Schmötzer et al., 1993; Lord et al., 1994) and the Autism Diagnostic Observation Schedule (Lord et al., 2000). The patients were recruited from the Department of Child and Adolescent Psychiatry, Psychosomatics, and Psychotherapy of the Goethe University at Frankfurt/M. The healthy controls (mean age:  $29.7 \pm 6.9$ ) were screened for psychopathology with the German version of Structured Clinical Interview for DSM-IV-R Non-Patient Edition (Saß et al., 2003). Both groups showed no significant differences in age, sex distribution and IQ. The study was performed according to the Declaration of Helsinki and approved by the ethics committee of the Goethe University (Frankfurt, Germany).

All subjects performed a perceptual closure task where stimuli consisted of degraded pictures of human faces in which all shades of gray had been converted into black or white (Mooney and Ferguson, 1951). In addition, scrambled and vertically mirrored versions of these stimuli were created, for which face perception was not possible. One hundred and sixty different stimuli for each stimulus category were presented in a random sequence, where each stimulus was shown for 200 ms, separated by a random inter-stimulus intervals between 3500 and 4500 ms. Participants had to indicate with a button press whether they saw a face or not. Response hands were counterbalanced across participants in each group. This set of stimuli allowed us to identify the visual system and higher cortices related to object and face perception for further investigation using AIS.

### 2.2.2. MR and MEG data acquisition

Individual structural MR images were acquired with a Siemens Allegra scanner (Siemens Medical Solutions, Erlangen, Germany), using a 3D MPRAGE sequence. MEG signals were recorded with a 275-channel system (Omega 2005; VSM MedTech, Coquitlam, BC, Canada) with 600 Hz sampling rate, third-order gradiometers. The acquired data were bandpass filtered between 0.5 and 150 Hz (fourth order Butterworth filter). Before and after each run, the head position was localized using



localization coils. Recordings with movements larger than 5 mm were discarded.

### 2.2.3. MEG data preprocessing

MEG data were analyzed using the FieldTrip open source MATLAB Toolbox (Oostenveld et al., 2011). The continuously recorded data were segmented into trials from  $-1000$  to  $1000$  ms with respect to the onset of the visual stimulus. Eye blinks, signal jumps caused by the SQUID sensors, and muscle artefacts were detected automatically (in this sequence) using the preprocessing functions of FieldTrip, followed by visual inspection for residual artefacts. Affected trials were rejected completely as suggested in Gross et al. (2012). The remaining trials were linearly detrended and baseline corrected.

### 2.2.4. Analysis of sensor-level spectral power changes

Time-frequency representations (TFRs) were computed from sensor data using a multi-taper method [frequency range from 25 to 140 Hz in 2 Hz steps over a time range of  $-500$  to  $1000$  ms in 10 ms steps, discrete prolate spheroidal sequences (DPSS), length of sliding time window,  $5/\text{frequency}$ , width of frequency smoothing,  $0.4 * \text{frequency}$ ]. The power of the time-frequency-transformed trial data was averaged over all sensors and trials and, subsequently, all subjects. The optimal beamformer bandwidth (Brookes et al., 2008) was then estimated based on the observed power changes induced by the visual stimulus (analysis interval 75–375 ms) relative to baseline (analysis interval  $-350$  to  $-50$  ms).

### 2.2.5. Source reconstruction and selection

As the estimation of AIS is computationally very demanding, we were not able to compute this measure on a source grid covering the whole brain. We therefore chose to investigate a selection of source locations showing differences between baseline and the perceptual closure task. This decision was based on previous reports of changes in visual perception in ASD (see Pellicano and Burr, 2012 and references therein); this goal for a selection of sources substantially differs from the goal of detecting sources with spectral power differences between ASD subjects and controls that was pursued in the study by Sun et al. (2012), and the analysis strategies differ accordingly. Note that preselecting sources with power differences between ASD and healthy controls would potentially bias a subsequent analysis of ASD (“double dipping”), whereas selecting areas that represent the visual system does not entail such a bias *a priori*. To identify visually responsive areas, we first performed a beamformer source analysis in the high gamma frequency range (60–120 Hz), as the initial TFR-analysis indicated sustained responses triggered by the visual stimulus in this frequency range. Note that even though differences between ASD patients and healthy controls have been demonstrated previously in this band, the choice of this band does not unduly bias the analysis as such differences have been shown in all major frequency bands (see Figure 3 in Sun et al., 2012).

After identifying locations with significant differences in the high-frequency gamma band, we then recomputed broadband beamformer filters for these locations and extracted the individual source time courses for each subject and source location for further analysis. Note that we only used baseline intervals (time

interval from  $-1000$  to  $0$  ms with respect to stimulus onset) in our analysis of AIS to ensure the stationarity of the underlying processes. Note that differences in brain dynamics are also expected in this baseline interval, as there is of course ongoing visual experience.

In more detail, we constructed head models of each individual subject from anatomical MRI for beamformer source reconstruction. To this end, first a regular source grid with a spacing of 1 cm was constructed in MNI space. After computing the (linear) transformations from the MNI template head to each individual subject's anatomical MRI, these transformations were applied to the source grid to obtain individual source grids in physical space for each subject. After segmentation of the MRI to find the inner boundary of the skull, the lead fields for the individual source grid locations were then computed using a realistic single shell model introduced by Nolte (2003).

Next, the cross spectral density (CSD) matrices were computed for all trials for both patients and controls, separately for baseline ( $-350$  to  $-50$  ms) and task intervals (75–375 ms) in the high (60–120 Hz) gamma-band frequency range, using a multi-taper method (center frequency 90 Hz, smoothing bandwidth  $\pm 30$  Hz, DPSS, 17 tapers). Based on the lead fields and the computed CSD matrices, spatial filters were computed for each grid point using a frequency domain beamformer (Gross et al., 2001) as provided by FieldTrip, using real valued filter coefficients. To compensate for the rather short time intervals underlying the computation of CSD matrices, a matrix regularization of  $\lambda = 5\%$  of the trace of the CSD matrix was used. In order to avoid that statistical differences arise because of different filters for the two intervals, we computed common filters which are based on the combined CSD matrices from both, task segments (face and no-faces) and the corresponding baseline segments (Nieuwenhuis et al., 2008; Gross et al., 2012). The source power estimate at each grid point was computed by applying the corresponding common filter of this grid point to the filtered trial data. This was done separately for task and baseline segments of each subject.

To obtain common sources that responded to the perceptual closure task for all subjects and independent of their group affiliation, a non-parametric randomization test (test-statistic: cluster sum of dependent samples *t*-metric, Monte Carlo estimate with 5000 randomizations, Maris and Oostenveld, 2007) was computed based on source power data of all subjects. Using a within-subject design on subject-wise source power for task and baseline, activation-versus-baseline effects were identified; *t*-metrics within a cluster were used to identify local extrema of source power changes inside the significant clusters.

## 2.3. ANALYSIS OF ACTIVE INFORMATION STORAGE IN SOURCE TIME COURSES

For further analysis of AIS based on source time courses, we obtained these time courses at the identified source locations that responded to the perceptual closure task, using a broadband beamformer, so that AIS computation could draw on a signal bandwidth of 10–150 Hz—the analysis of AIS at even lower frequencies was not possible due to the finite length (1 s) of the baseline data. On the three source time courses extracted for the three cardinal spatial directions (*x*, *y*, *z*) at each location we then performed a principal component analysis in order to

determine the dominant dipole orientation (direction with the largest variance), and kept only the signal for this direction. As indicated above, AIS was computed using the Java Information Dynamics Toolkit (Lizier, 2012), with a box kernel of a width of 0.5 standard deviations of the data, and a history length  $k$  of 10 time steps. Per subject and source, approximately 40,000 samples entered the AIS analysis, composed of 1 s of baseline data per trial, sampled at 300 Hz, repeated across approximately 66 correct and artefact-free trials per condition, and two conditions. For statistical comparison between ASD patients and controls we used a randomization test, and corrected the significance threshold for multiple comparisons across 12 sources using the false discovery rate (FDR) with a threshold of  $q < 0.1$ , as suggested in Genovese et al. (2002).

#### 2.4. CORRELATION OF BAND-LIMITED SPECTRAL POWER, AUTOCORRELATION DECAY TIME, AND ACTIVE INFORMATION STORAGE

To investigate whether the obtained AIS values were driven by spectral power changes, or contained information not accessible

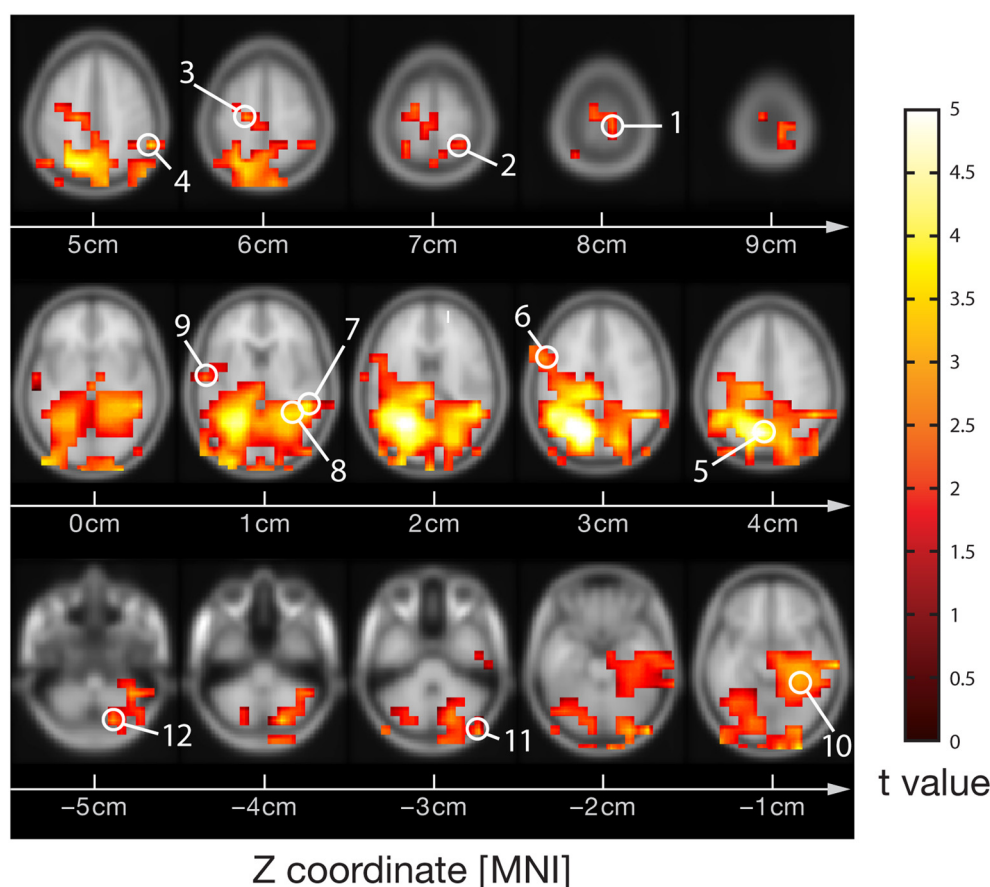
by an analysis of spectral power, we computed the correlation coefficients between spectral power in the 10–12 Hz  $\alpha$ -, the 13–15 Hz  $\beta$ -, the 25–60 Hz low frequency  $\gamma$ -, and the 60–120 Hz high-frequency  $\gamma$ -bands. Spectral power within these bands was computed per subject using a Hanning window on the full baseline data followed by a fast Fourier transform for frequencies up to 25 Hz. Above 25 Hz spectral power was determined by a multitaper approach using 34, and 59 DPSS-tapers for the bands from 25 to 60 Hz and from 60 to 120 Hz, respectively.

In addition, we determined the autocorrelation decay time (ACT) as a measure of linear memory time scales in the data. The ACT was obtained by computing the autocorrelation function and determining the lag at which the autocorrelation function had dropped to a fraction of  $1/e$  of its center peak.

### 3. RESULTS

#### 3.1. SOURCES PARTICIPATING IN THE PERCEPTUAL CLOSURE TASK

Visual stimulation increased neural activity in the high-frequency gamma band in several occipital, parietal, temporal and central



**FIGURE 1 | MEG-beamformer source locations used for the analysis of active information storage.** MEG sources with enhanced power in the high-frequency gamma band (60–120 Hz) upon visual stimulation with Mooney face images (see Grützner et al., 2010 for stimulus details); permutation test on  $t$ -metrics  $p < 0.05$ , cluster-based correction for multiple comparisons. Source locations: 1—Primary motor cortex BA4a R (10, -30, 80), 2—Superior parietal lobule 7PC R (30, -50, 70),

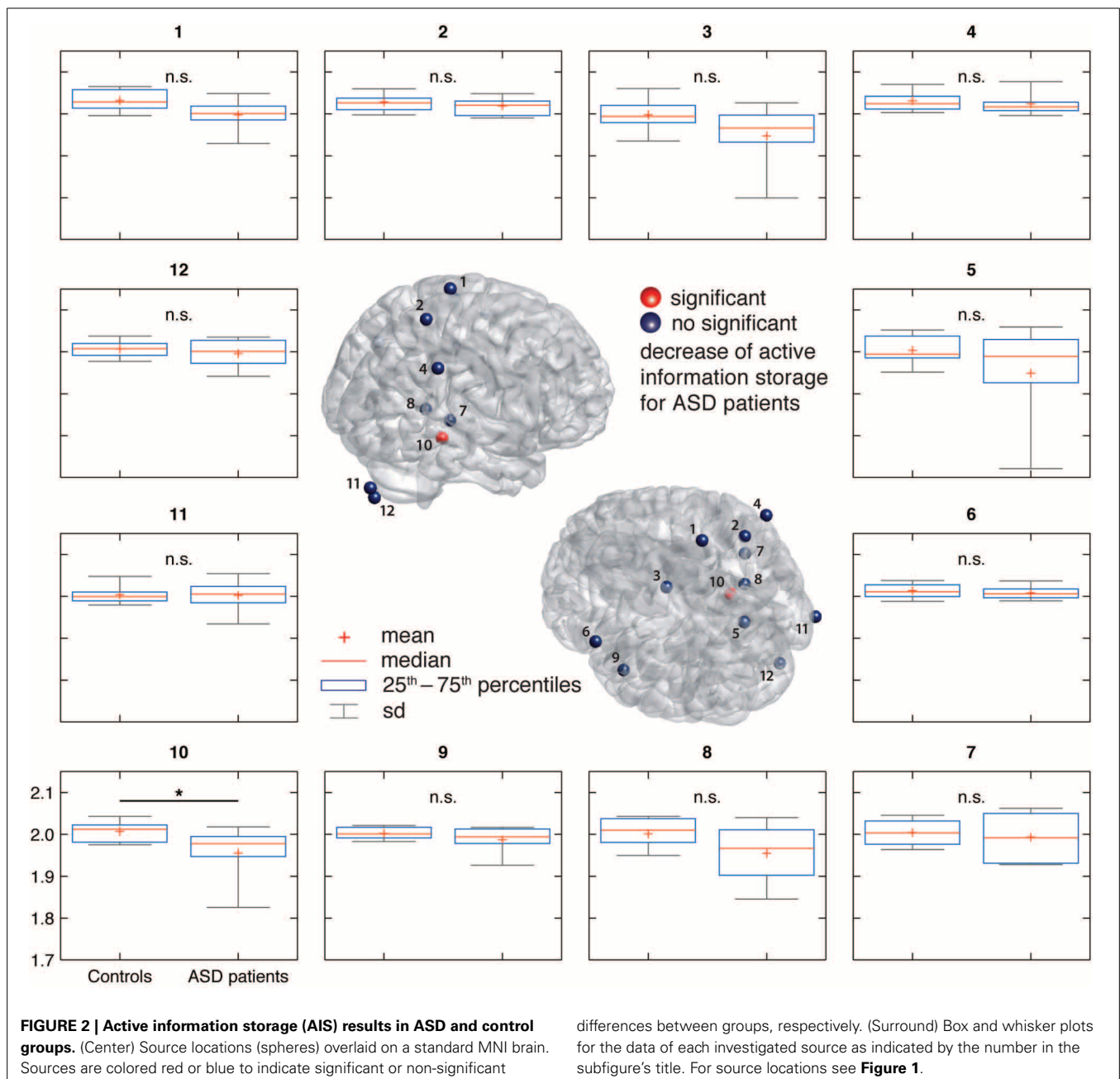
3—Premotor cortex BA6 L (-20, -20, 60), 4—Parietal lobe (60, -50, 50), 5—Precuneus/Superior parietal lobule 7P L (-10, -70, 40), 6—Broca's area BA44 L (-60, 10, 30), 7—Temporal lobe (50, -40, 10), 8—Visual cortex V2 BA18 R/V1 BA17 R (30, -50, 10), 9—Secondary somatosensory cortex/ Parietal operculum OP4 L (-60, -10, 10), 10—Hippocampus/Subiculum R (30, -40, -10), 11—Right cerebellum (50, -90, -30), 12—Cerebellum (20, -80, -50)

cortical regions and in the cerebellum ( $p < 0.05$ , corrected) (**Figure 1**), as expected from previous studies (Grützner et al., 2010; Sun et al., 2012). Source locations of significantly increased gamma-band power were: 1—Primary motor cortex BA4a R (10, -30, 80), 2—Superior parietal lobule 7PC R (30, -50, 70), 3—Premotor cortex BA6 L (-20, -20, 60), 4—Parietal lobe (60, -50, 50), 5—Precuneus/Superior parietal lobule 7P L (-10, -70, 40), 6—Broca's area BA44 L (-60, 10, 30), 7—Temporal lobe (50, -40, 10), 8—Visual cortex V2 BA18 R/V1 BA17 R (30, -50, 10), 9—Secondary somatosensory cortex/Parietal operculum OP4 L (-60, -10, 10), 10—Hippocampus/Subiculum R (30, -40, -10), 11—Right cerebellum (50, -90, -30), 12—Cerebellum (20, -80, -50). For

these locations broadband beamformer source time-courses for the baseline interval were extracted and subjected to AIS-analysis (recall that only baseline interval values were analyzed to ensure stationarity of the data).

### 3.2. ACTIVE INFORMATION STORAGE

AIS-analysis revealed significantly reduced AIS in ASD in the hippocampus ( $q < 0.1$ , FDR corrected) (**Figure 2**). At an uncorrected significance level ( $p < 0.05$ ), we observed additional differences in visual cortex, primary motor cortex and premotor cortex. At a purely descriptive level, in 11 out of 12 sources the observed median AIS values were lower in the ASD group compared to controls.



### 3.3. CORRELATION OF SPECTRAL POWER, AUTOCORRELATION DECAY TIME, AND AIS

Spectral power in none of the investigated bands (10–12, 13–15, 25–60, and 60–120 Hz) was significantly correlated with AIS values after correction for multiple comparisons (Table 1) (minimum  $p$ -value reached by any correlation:  $p = 0.03$ , uncorrected), indicating that AIS provides information that is at least partially independent of spectral power indices (Figure 3). Moreover, correlation coefficients were mostly negative—in contrast to what would be expected of the bias properties of the AIS estimator (see the companion paper on *local* AIS in this *Frontiers special topic* Wibral et al., 2014 for details), further supporting the independence of the two measures for our data.

Similarly, ACT showed a negative Pearson correlation with the AIS ( $p < 0.034$ ) (Figure 4).

## 4. DISCUSSION

In line with our initial hypothesis, we found reduced AIS in individuals with ASD. More specifically, AIS was reduced in the hippocampus (subiculum). As our study is the first of its kind, extra care has to be taken to ensure both, an understanding meaning of the results at the conceptual level, as well as a clear view of the limitations of the current study. Therefore, we start by discussing two technical points related to a proper interpretation of AIS; first, we clarify the relation between AIS and signal prediction errors; second, we discuss the relationship of AIS and more high level concepts of memory in a neural system. Next, we detail the limitations of the current study in terms of small sample size and region of interest analysis. After these technical points, we discuss our findings in relation to known anatomical and cellular changes in ASD. We close by discussing our finding in relation to predictive coding theories of this disorder.

### 4.1. ACTIVE INFORMATION STORAGE AND SIGNAL PREDICTION ERRORS

In the introduction, we pointed out how measures of information storage may be useful tools to investigate predictive-coding type of theories of cortical function (Rao and Ballard, 1999; Friston et al., 2006; George and Hawkins, 2009; Bastos et al., 2012; Grossberg, 2013). In this respect, it seems important to stress the difference between the amount of *predicted information*—as measured by information storage—and the *signal* prediction error, i.e., the amount of information not predicted in a signal (not

to be confused with a neural prediction error in predictive coding theories). While the sum of these is the total information in a process, this total information is not necessarily constant. In fact, in most task-related studies we expect the neural processes to be non-stationary, i.e., to have probability distributions changing across time, leading to changing total information. This, in turn, results in predicted information (information storage) and unpredicted information (prediction error) describing complementary aspects of the information processing system—and one cannot be obtained from the other.

### 4.2. ACTIVE INFORMATION STORAGE, MEMORY, AND NEURAL PROCESSING

While seemingly similar, AIS as an information theoretic measure should not be confused with high-level concepts of memory, or the storage of information about the external world. Rather, it describes the predictability and complexity of a neural process. AIS is low for processes that produce little information, such as a constant process, but also for unpredictable processes, such as chaotic ones (Lizier et al., 2011). Only when sufficiently rich dynamics and predictability meet, a high AIS value is obtained. In the context of our data, high AIS values are linked to transitions in the dynamics that are repeatedly seen across the multiple trials used for analysis—albeit not necessarily at the same time. Therefore, one source for reduced AIS values in the ASD group could be a more erratic signal behavior across trials in baseline dynamics between the stimuli. With respect to this baseline dynamics it is important to note that the baseline activity is not necessarily independent of stimuli and task. In this respect, the reduced AIS in the baseline epochs in ASD could still be linked to the specific stimulus material used here (faces, a social stimulus) and the detection task. In how far our results can be generalized for other experimental designs is an open question.

Correlation analysis between spectral power and AIS values revealed that AIS provides additional information, not immediately accessible using an analysis of spectral power. In contrast, the significant Pearson correlation between ACT and AIS indicates that the AIS reflects also the linear memory in the process, as would be expected. However, the correlation coefficient was below 0.5, indicating that also for the comparison of ACT and AIS, AIS yields additional useful information.

### 4.3. REDUCED ACTIVE INFORMATION STORAGE IN ASD

Our results showed that AIS values were reduced in the neural signals obtained from the hippocampus/subiculum. Other sources showed reduction at least at an uncorrected significance level (visual cortex, primary motor cortex, premotor cortex). Before we proceed to the potential implications of these findings, we will briefly discuss several reasons that warrant a cautious interpretation of our results.

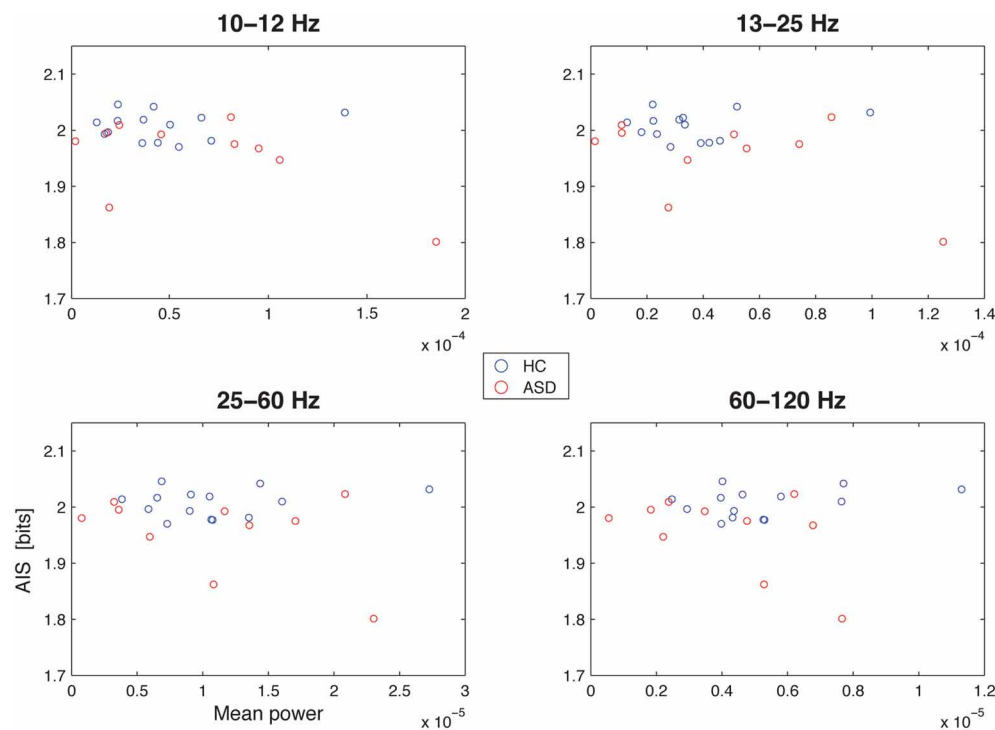
#### 4.3.1. Limited sample size

Perhaps the most important reason for caution is the relatively low number of patients in this study ( $n = 10$ ), limiting statistical power. Therefore, our study should be understood as a pilot study for a larger, normative study of AIS values in ASD. Such a larger scale study seems highly promising as a close inspection

**Table 1 | Correlation coefficients between AIS and spectral power, and autocorrelation decay time (ACT) in the hippocampal source.**

Frequency (Hz)	Spearman		Pearson	
	$\rho$	$p$	$\rho$	$p$
10–12	−0.202	0.343	−0.444	0.030
13–25	−0.143	0.502	−0.364	0.080
25–60	−0.034	0.876	−0.194	0.364
60–120	0.117	0.586	−0.048	0.825
ACT	−0.318	0.130	−0.434	0.034





**FIGURE 3 | Correlation between spectral power and AIS in the hippocampal source.** Correlation between spectral power in the 10–12 Hz  $\alpha$ -, the 13–15 Hz  $\beta$ -, the 25–60 Hz low frequency  $\gamma$ -, and

the 60–120 Hz high-frequency  $\gamma$ -bands (x-axes) and the active information storage (y-axes). See **Table 1** for details on correlation coefficients.

of **Figure 2** reveals that the mean and median AIS values in the ASD group are lower in *all* investigated sources, except one (right cerebellum). Despite the fact that none of these effects reaches statistical significance, the relatively uniform sign of the effect may point to a more pervasive reduction of AIS in ASD. This, however, can only be tested in a study with improved statistical power.

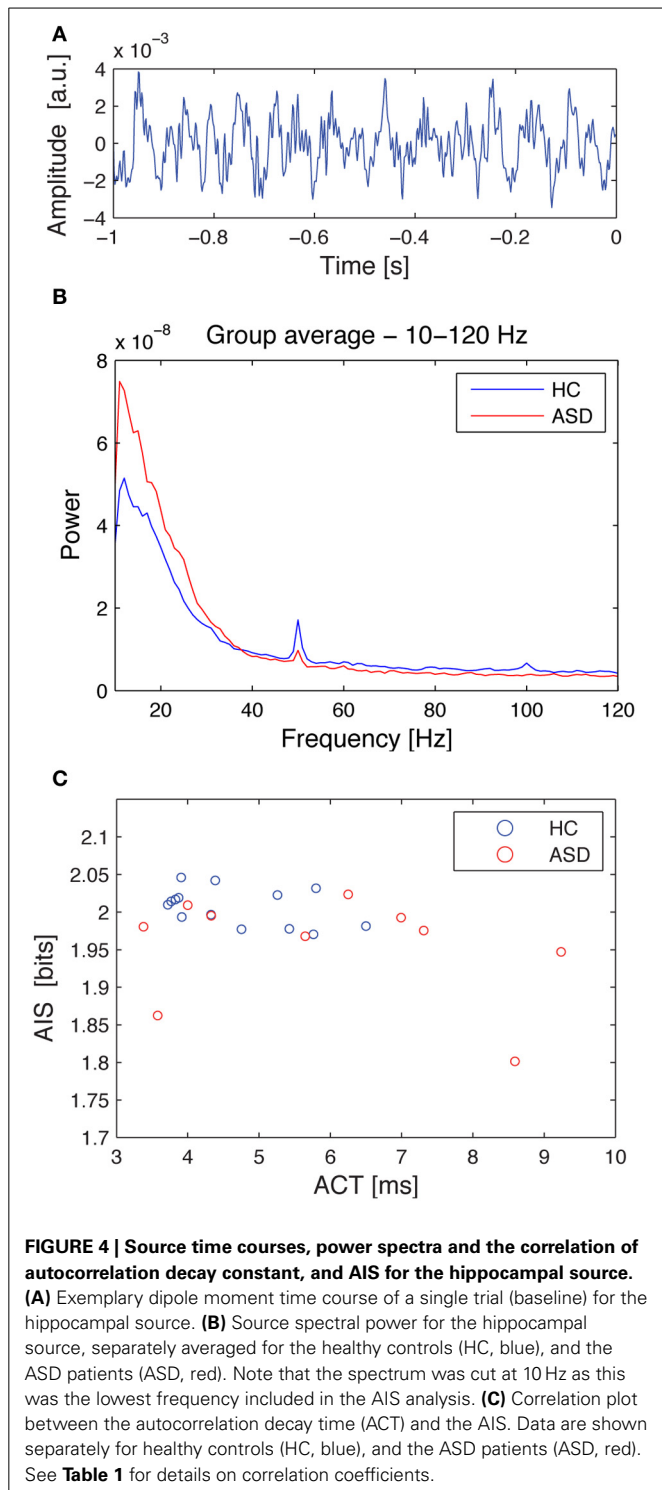
#### 4.3.2. Region of interest analysis and magnetoencephalographic detection of deep sources

Our focus on selected brain areas limits any statements on the ubiquity of reduced AIS in ASD. The fact that we focus on preselected brain areas (for purely practical reasons), clearly forbids any statements of the type “AIS is most strongly reduced in brain area A” or “AIS is only reduced in brain area A.” Thus, we can only link the current findings to literature on ASD-related changes in the specific brain areas that were investigated here. Furthermore, MEG source reconstruction is of limited spatial precision. We therefore discuss our findings of reduced AIS in the hippocampus/subiculum—where the analyzed source was located—more broadly as reduced AIS in the hippocampal region, and note the possibility of signal leakage from the nearby amygdala (note, however, that the cellular organization of the amygdala makes detectable MEG signals less likely to be picked up). With respect to the hippocampus it is often questioned whether the sensitivity of MEG recordings is high enough to capture this relatively deep source. However, a large body

of evidence has accumulated in recent years that confirms that hippocampal activity can be reconstructed with modern MEG devices, artefact suppression techniques, and beamformer source reconstruction (Tesche et al., 1996; De Araújo et al., 2002; Hanlon et al., 2005; Cornwell et al., 2008; Riggs et al., 2009; Taylor et al., 2012). Moreover, nearby inferior temporal brain areas are routinely localized using MEG, for example the fusiform face area (Grützner et al., 2010); in addition, traces of thalamic activity have been recently revealed, using a combination of MEG cross-frequency analyses and transfer entropy techniques (Roux et al., 2013), and even auditory brain stem responses have been localized using MEG (Parkkonen et al., 2009). We therefore think it is safe to assume that our results indeed derive from changes in activity patterns in the hippocampal region.

#### 4.4. THE HIPPOCAMPUS IN ASD

Interestingly, there is a number of anatomical findings of atypical hippocampal structure in patients with ASD. At the cellular level, increased cell packing density and reduced cell size was reported by Bauman and Kemper (1994). Raymond et al. (1995) further showed reduced dendritic branching of CA4 and CA1 cells. Blatt et al. (2001) reported reduced binding of GABA<sub>A</sub> receptors in the hippocampus. Furthermore, some rare Autism-linked point mutations coding for to Neuroligins seem to selectively target AMPA receptor-mediated neurotransmission in Hippocampus



and dramatically change synaptic function in a mouse model of autism (Eherton et al., 2011).

At the macroscopic level, an enlargement of the right hippocampus was found across all studied age groups by Schumann et al. (2004), and an enlargement of the left hippocampus by Groen et al. (2010), Dager et al. (2007), and Nicolson et al. (2006) also reported shape abnormalities of the hippocampus in ASD. In

addition, metabolic abnormalities in the hippocampus-amygdala region (Otsuka et al., 1999) have been reported as well.

Taken together, these structural and cellular findings suggest an involvement of hippocampal changes in the pathophysiology of ASD. Our findings are compatible with this idea and add a computational perspective to the neuroanatomical and cellular evidence by indicating that brain signals from hippocampus have less predictable information in patients with ASD. Next, we will discuss how these findings tie in with predictive coding accounts of ASD.

#### 4.5. REDUCED AIS AND PREDICTIVE CODING THEORIES OF ASD

Given that the hippocampus is a plausible locus for changes in information processing in ASD from an anatomical perspective—how does the observation of reduced AIS in this brain area fit the various theoretical accounts of information processing in ASD?

For a more detailed understanding of the meaning of a reduction in AIS in relation to ASD, we have to consider first that reductions in AIS may indicate various changes in cortical dynamics—either a reduced dynamic richness of the neural process captured in the measurement, a decrease of predictability, or a combination of the two. Irrespective of the exact underlying change in the dynamics, however, a brain area receiving signals from another one with reduced AIS will inevitably be faced with a signal that has less predictable information. We can therefore state that information from cortical signals from one brain area will be harder to predict internally by another brain area in ASD patients.

If we look at this reduction in predictable information in ASD from the perspective of predictive coding theory, we may speculate that this reduction will result in difficulties in learning internal predictive models and in a less accurate model of the external world. Taking into account additionally that internal models should be organized hierarchically with internal models in sensory areas lowest in the hierarchy, as suggested for example by the hierarchical temporal model of George and Hawkins (2009), and by the model of Kiebel et al. (2008), the fact that we observe the most significant differences in hippocampus is particularly interesting. This is because the hippocampus resides at a high level in these hierarchies (George and Hawkins, 2009), where it would be most vulnerable to difficulties in learning of internal models. Moreover, many anatomical, physiological and computational reasons suggest that deeper or more central models at higher levels of the hierarchy entail dynamics that have greater temporal depth (Kiebel et al., 2008), and therefore should have more predictable information. One may speculate that this adds to the visibility of changes in AIS deep into the hierarchy, e.g., in hippocampus.

The fact that we obtained significantly reduced AIS values specifically in the *hippocampus* is also remarkable in relation to previous whole brain analyses of neuronal responses to explicit manipulations of predictability—using temporal dependencies in sequences of stimuli. For example, Strange et al. (2005) report hippocampal selectivity for the predictability of stimuli, consistent with the notion that the hippocampus is of central importance in the processing of temporal succession (MacDonald et al., 2011). This processing of a temporal aspect of predictability

in hippocampus fits comfortably with the hierarchical Bayesian inference and predictive coding formulations of autism, when combined with the changes in hippocampal processing reported here. In other words, if altered hippocampal processing lead to a loss of hierarchically deep encoding of hidden causes in the world, this would necessarily entail a loss of deep temporal structure and a failure to encode temporal regularities over extended periods of time, and thereby global temporal context.

Such a loss of central or deep coherence in time and space has also been proposed previously as a psychological mechanism that explains many of the symptoms in ASD (“weak central coherence theory,” Frith, 1989, but see Bernardino et al., 2012 for conflicting data). This view indeed pre-dates modern perspectives from the point of view of hierarchical inference and predictive coding (Pellicano and Burr, 2012), but is fully compatible with its successors.

In sum, our data are fully compatible with predictive coding accounts of ASD (e.g., Pellicano and Burr, 2012). In contrast, our data are not compatible with theories that suggest enhanced sensory representations, and/or lower physiological noise, as both of these should lead to increased rather than reduced AIS values, which was not observed. Thus, our data favor accounts of autism in terms of compromised top-down processing.

#### 4.6. AIS AND PREVIOUS STUDIES ON SIGNAL-ENTROPY AND COMPLEXITY IN ASD

Several previous studies have analyzed brain activity in ASD by means of complexity and/or entropy measures. For instance, Catarino et al. (2011) analyzed EEG data using multi-scale entropy, which quantifies the complexity of a physiological signal by measuring entropy across multiple time scales. Their results demonstrated a complexity reduction in autism group in comparison with controls, especially over tempo-parietal and occipital sensors. Using a modified version of multi-scale entropy, a decrease in resting-state EEG complexity in children at high risk for ASD has been reported (Bosl et al., 2011). In another study, Ahmadi et al. (2010) reported significant differences at several EEG locations using the fractal dimension algorithms proposed by Higuchi and Katz. Finally, a statistically significant reduction in Lempel-Ziv complexity was found in ASD group in comparison with controls, at EEG electrodes F7, F3, and T5 (Sheikhan et al., 2007). All of these previous results are fully compatible with our findings of reduced AIS, as a reduced entropy limits also the maximally possible AIS. In addition, the findings of reduced Lempel-Ziv complexity align well with decreased AIS, as the Lempel-Ziv algorithm entails cataloguing recurrent events and this is tightly linked to predictable recurrence inside the repeated sequences in the signals. Our results extend these previous findings by localizing the dominant changes to the hippocampus. Moreover, the use of AIS, rather than more generic measures of entropy or complexity allows a straightforward interpretation in terms of component processes of information processing, i.e., information storage.

## 5. CONCLUSION

In this study, we present the first application of information theoretic measures of information storage to experimental neural

data. Using MEG and source signal reconstruction from 12 selected brain areas, we show that AIS is reduced in the hippocampus of individuals with ASD. Future studies on larger samples of patients, combined with whole brain analyses, will have to show in how far our results generalize across brain areas, and to broader populations of ASD patients. The relatively uniform sign of the observed AIS differences across all investigated brain areas suggests that reduced AIS may be a pervasive change of information processing in ASD.

## ACKNOWLEDGMENTS

The authors would like to thank Viola Priesemann from the Max Planck Institute for Dynamics and Self-Organization, Göttingen, Germany, for stimulating discussions on the nature of active information storage in neural signals.

## FUNDING

Michael Wibral was supported by LOEWE Grant “Neuronale Koordination Forschungsschwerpunkt Frankfurt (NeFF).” Michael Wibral thanks the Commonwealth Scientific and Industrial Research Organisation (CSIRO) for supporting a visit which contributed to this work. Carlos Gómez received a travel grant from LOEWE Grant “Neuronale Koordination Forschungsschwerpunkt Frankfurt (NeFF).” Carlos Gómez and Roberto Hornero were supported in part by the “Ministerio de Economía y Competitividad” and FEDER under project TEC2011-22987.

## REFERENCES

- Ahmadi, M., Adeli, H., and Adeli, A. (2010). Fractality and a wavelet-chaos-neural network methodology for EEG-based diagnosis of autistic spectrum disorder. *J. Clin. Neurophysiol.* 27, 328–333. doi: 10.1097/WNP.0b013e3181f40dc8
- American Psychiatric Association. (2000). *Diagnostic and Statistical Manual of Mental Disorders: DSM-IV-TR®*. Washington, DC: American Psychiatric Association.
- Asperger, H. (1944). Die autistischen psychopathen im kindesalter. *Eur. Arch. Psychiatry Clin. Neurosci.* 117, 76–136.
- Baird, G., Simonoff, E., Pickles, A., Chandler, S., Loucas, T., Meldrum, D., et al. (2006). Prevalence of disorders of the autism spectrum in a population cohort of children in south thames: the special needs and autism project (snap). *Lancet* 368, 210–215. doi: 10.1016/S0140-6736(06)69041-7
- Baron-Cohen, S., Ashwin, E., Ashwin, C., Tavassoli, T., and Chakrabarti, B. (2009). Talent in autism: hyper-systemizing, hyper-attention to detail and sensory hypersensitivity. *Philos. Trans. R. Soc. B Biol. Sci.* 364, 1377–1383. doi: 10.1098/rstb.2008.0337
- Bastos, A. M., Usrey, W. M., Adams, R. A., Mangun, G. R., Fries, P., and Friston, K. J. (2012). Canonical microcircuits for predictive coding. *Neuron* 76, 695–711. doi: 10.1016/j.neuron.2012.10.038
- Bauman, M. L., and Kemper, T. L. (1994). “Neuroanatomic observations of the brain in autism,” in *The Neurobiology of Autism*, eds M. L. Bauman and T. L. Kemper (Baltimore, MD: Johns Hopkins University Press), 119–145.
- Bauman, M. L., and Kemper, T. L. (2005). Neuroanatomic observations of the brain in autism: a review and future directions. *Int. J. Dev. Neurosci.* 23, 183–187. doi: 10.1016/j.ijdevneu.2004.09.006
- Belmonte, M. K., Allen, G., Beckel-Mitchener, A., Boulanger, L. M., Carper, R. A., and Webb, S. J. (2004a). Autism and abnormal development of brain connectivity. *J. Neurosci.* 24, 9228–9231. doi: 10.1523/JNEUROSCI.3340-04.2004
- Belmonte, M. K., Cook, E. H., Anderson, G. M., Rubenstein, J. L., Greenough, W. T., Beckel-Mitchener, A., et al. (2004b). Autism as a disorder of neural information processing: directions for research and targets for therapy. *Mol. Psychiatry* 9, 646–663. doi: 10.1038/sj.mp.4001499
- Bernardino, I., Mouga, S., Almeida, J., van Asselen, M., Oliveira, G., and Castelo-Branco, M. (2012). A direct comparison of local-global integration in autism

- and other developmental disorders: implications for the central coherence hypothesis. *PLoS ONE* 7:e39351. doi: 10.1371/journal.pone.0039351
- Bertone, A., Mottron, L., Jelenic, P., and Faubert, J. (2005). Enhanced and diminished visuo-spatial information processing in autism depends on stimulus complexity. *Brain* 128, 2430–2441. doi: 10.1093/brain/awh561
- Blatt, G. J., Fitzgerald, C. M., Guptill, J. T., Booker, A. B., Kemper, T. L., and Bauman, M. L. (2001). Density and distribution of hippocampal neurotransmitter receptors in autism: an autoradiographic study. *J. Autism Dev. Disord.* 31, 537–543. doi: 10.1023/A:1013238809666
- Bosl, W., Tierney, A., Tager-Flusberg, H., and Nelson, C. (2011). EEG complexity as a biomarker for autism spectrum disorder risk. *BMC Med.* 9:18. doi: 10.1186/1741-7015-9-18
- Brock, J. (2012). Alternative bayesian accounts of autistic perception: comment on pellicano and burr. *Trends Cogn. Sci.* 16, 573–574. doi: 10.1016/j.tics.2012.10.005
- Brookes, M. J., Vrba, J., Robinson, S. E., Stevenson, C. M., Peters, A. M., Barnes, G. R., et al. (2008). Optimising experimental design for MEG beamformer imaging. *Neuroimage*, 39, 1788–1802. doi: 10.1016/j.neuroimage.2007.09.050
- Catarino, A., Churches, O., Baron-Cohen, S., Andrade, A., and Ring, H. (2011). Atypical EEG complexity in autism spectrum conditions: a multiscale entropy analysis. *Clin. Neurophysiol.* 122, 2375–2383. doi: 10.1016/j.clinph.2011.05.004
- Cornwell, B. R., Johnson, L. L., Holroyd, T., Carver, F. W., and Grillon, C. (2008). Human hippocampal and parahippocampal theta during goal-directed spatial navigation predicts performance on a virtual morris water maze. *J. Neurosci.* 28, 5983–5990. doi: 10.1523/JNEUROSCI.5001-07.2008
- Cover, T. M., and Thomas, J. A. (1991). *Elements of Information Theory*. New York, NY: Wiley-Interscience. doi: 10.1002/0471200611
- Crutchfield, J. P., and Feldman, D. P. (2003). Regularities unseen, randomness observed: levels of entropy convergence. *Chaos* 13, 25–54. doi: 10.1063/1.1530990
- Dager, S., Wang, L., Friedman, S., Shaw, D., Constantino, J., Artru, A., et al. (2007). Shape mapping of the hippocampus in young children with autism spectrum disorder. *Am. J. Neuroradiol.* 28, 672–677. Available online at: <http://www.ajnr.org/content/28/4/672.full>
- Dasgupta, S., Wörgötter, F., and Manoonpong, P. (2013). Information dynamics based self-adaptive reservoir for delay temporal memory tasks. *Evol. Syst.* 4, 235–249. doi: 10.1007/s12530-013-9080-y
- David, N., Rose, M., Schneider, T. R., Vogele, K., and Engel, A. K. (2010). Brief report: altered horizontal binding of single dots to coherent motion in autism. *J. Autism Dev. Disord.* 40, 1549–1551. doi: 10.1007/s10803-010-1008-9
- De Araújo, D. B., Baffa, O., and Wakai, R. T. (2002). Theta oscillations and human navigation: a magnetoencephalography study. *J. Cogn. Neurosci.* 14, 70–78. doi: 10.1162/08992902317205339
- Etherton, M. R., Tabuchi, K., Sharma, M., Ko, J., and Südhof, T. C. (2011). An autism-associated point mutation in the neuroligin cytoplasmic tail selectively impairs ampa receptor-mediated synaptic transmission in hippocampus. *EMBO J.* 30, 2908–2919. doi: 10.1038/emboj.2011.182
- Friston, K., Kilner, J., and Harrison, L. (2006). A free energy principle for the brain. *J. Physiol. Paris* 100, 70–87. doi: 10.1016/j.jphysparis.2006.10.001
- Friston, K. J., Lawson, R., and Frith, C. D. (2013). On hyperpriors and hypopriors: comment on pellicano and burr. *Trends Cogn. Sci.* 17, 1. doi: 10.1016/j.tics.2012.11.003
- Frith, U. (1989). *Autism: Explaining the Enigma*. Oxford: Wiley Online Library.
- Genovese, C. R., Lazar, N. A., and Nichols, T. (2002). Thresholding of statistical maps in functional neuroimaging using the false discovery rate. *Neuroimage* 15, 870–878. doi: 10.1006/nimg.2001.1037
- George, D., and Hawkins, J. (2009). Towards a mathematical theory of cortical micro-circuits. *PLoS Comput. Biol.* 5:e1000532. doi: 10.1371/journal.pcbi.1000532
- Grassberger, P. (1986). Toward a quantitative theory of self-generated complexity. *Int. J. Theor. Phys.* 25, 907–938. doi: 10.1007/BF00668821
- Groen, W., Teluij, M., Buitelaar, J., and Tendolkar, I. (2010). Amygdala and hippocampus enlargement during adolescence in autism. *J. Am. Acad. Child Adolesc. Psychiatry* 49, 552–560. doi: 10.1016/j.jaac.2009.12.023
- Gross, J., Baillet, S., Barnes, G. R., Henson, R. N., Hillebrand, A., Jensen, O., et al. (2012). Good-practice for conducting and reporting meg research. *Neuroimage* 65, 349–363. doi: 10.1016/j.neuroimage.2012.10.001
- Gross, J., Kujala, J., Hamalainen, M., Timmermann, L., Schnitzler, A., and Salmelin, R. (2001). Dynamic imaging of coherent sources: studying neural interactions in the human brain. *Proc. Natl. Acad. Sci. U.S.A.* 98, 694–699. doi: 10.1073/pnas.98.2.694
- Grossberg, S. (2013). Adaptive resonance theory: how a brain learns to consciously attend, learn, and recognize a changing world. *Neural Netw.* 37, 1–47. doi: 10.1016/j.neunet.2012.09.017
- Grützner, C., Uhlhaas, P. J., Genc, E., Kohler, A., Singer, W., and Wibral, M. (2010). Neuroelectromagnetic correlates of perceptual closure processes. *J. Neurosci.* 30, 8342–8352. doi: 10.1523/JNEUROSCI.5434-09.2010
- Hanlon, F. M., Weisend, M. P., Yeo, R. A., Huang, M., Lee, R. R., Thoma, R. J., et al. (2005). A specific test of hippocampal deficit in schizophrenia. *Behav. Neurosci.* 119, 863. doi: 10.1037/0735-7044.119.4.863
- Happé, F., and Frith, U. (2006). The weak coherence account: detail-focused cognitive style in autism spectrum disorders. *J. Autism Dev. Disord.* 36, 5–25. doi: 10.1007/s10803-005-0039-0
- Kanner, L. (1943). Autistic disturbances of affective contact. *Nervous Child* 2, 217–250.
- Kantz, H., and Schreiber, T. (2003). *Nonlinear Time Series Analysis, 2 Edn*. Cambridge: Cambridge University Press. doi: 10.1017/CBO9780511755798
- Kiebel, S. J., Daunizeau, J., and Friston, K. J. (2008). A hierarchy of time-scales and the brain. *PLoS Comput. Biol.* 4:e1000209. doi: 10.1371/journal.pcbi.1000209
- Langton, C. G. (1990). Computation at the edge of chaos: phase transitions and emergent computation. *Physica D* 42, 12–37. doi: 10.1016/0167-2789(90)90064-V
- Lindner, M., Vicente, R., Priesemann, V., and Wibral, M. (2011). Trentool: a Matlab open source toolbox to analyse information flow in time series data with transfer entropy. *BMC Neurosci.* 12:119. doi: 10.1186/1471-2202-12-119
- Lizier, J. T. (2012). JIDT: an information-theoretic toolkit for studying the dynamics of complex systems. Available online at: <http://code.google.com/p/information-dynamics-toolkit/>.
- Lizier, J. T. (2013). *The Local Information Dynamics of Distributed Computation in Complex Systems*. Springer theses. Berlin: Springer. doi: 10.1007/978-3-642-32952-4
- Lizier, J. T., Pritam, S., and Prokopenko, M. (2011). Information dynamics in small-world Boolean networks. *Artif. Life* 17, 293–314. doi: 10.1162/artl\_a\_00040
- Lizier, J. T., Prokopenko, M., and Zomaya, A. Y. (2010). Information modification and particle collisions in distributed computation. *Chaos* 20, 037109. doi: 10.1063/1.3486801
- Lizier, J. T., Prokopenko, M., and Zomaya, A. Y. (2012). Local measures of information storage in complex distributed computation. *Inform. Sci.* 208, 39–54. doi: 10.1016/j.ins.2012.04.016
- Lord, C., Risi, S., Lambrecht, L., Cook, E. H. Jr., Leventhal, B. L., DiLavore, P. C., et al. (2000). The autism diagnostic observation schedule generic: a standard measure of social and communication deficits associated with the spectrum of autism. *J. Autism Dev. Disord.* 30, 205–223. doi: 10.1023/A:1005592401947
- Lord, C., Rutter, M., and Le Couteur, A. (1994). Autism diagnostic interview-revised: a revised version of a diagnostic interview for caregivers of individuals with possible pervasive developmental disorders. *J. Autism Dev. Disord.* 24, 659–685. doi: 10.1007/BF02172145
- MacDonald, C. J., Lepage, K. Q., Eden, U. T., and Eichenbaum, H. (2011). Hippocampal time cells bridge the gap in memory for discontinuous events. *Neuron* 71, 737–749. doi: 10.1016/j.neuron.2011.07.012
- Maris, E., and Oostenveld, R. (2007). Nonparametric statistical testing of EEG- and MEG-data. *J. Neurosci. Methods* 164, 177–190. doi: 10.1016/j.jneumeth.2007.03.024
- Mitchell, M. (2011). Ubiquity symposium: biological computation. *Ubiquity* 2011, 3. doi: 10.1145/1940721.1944826
- Mitchell, P., and Ropar, D. (2004). Visuo-spatial abilities in autism: a review. *Infant Child Dev.* 13, 185–198. doi: 10.1002/icd.348
- Mooney, C., and Ferguson, G. A. (1951). A new closure test. *Can. J. Psychol.* 5, 129. doi: 10.1037/h0083540
- Mottron, L., Dawson, M., Soulières, I., Hubert, B., and Burack, J. (2006). Enhanced perceptual functioning in autism: an update, and eight principles of autistic perception. *J. Autism Dev. Disord.* 36, 27–43. doi: 10.1007/s10803-005-0040-7
- Nicolson, R., DeVito, T. J., Vidal, C. N., Sui, Y., Hayashi, K. M., Drost, D. J., et al. (2006). Detection and mapping of hippocampal abnormalities in autism. *Psychiatry Res.* 148, 11–21. doi: 10.1016/j.psychres.2006.02.005
- Nieuwenhuis, I. L., Takashima, A., Oostenveld, R., Fernández, G., and Jensen, O. (2008). Visual areas become less engaged in associative



- recall following memory stabilization. *Neuroimage* 40, 1319–1327. doi: 10.1016/j.neuroimage.2007.12.052
- Nolte, G. (2003). The magnetic lead field theorem in the quasi-static approximation and its use for magnetoencephalography forward calculation in realistic volume conductors. *Phys. Med. Biol.* 48, 3637–3652. doi: 10.1088/0031-9155/48/22/002
- Oostenveld, R., Fries, P., Maris, E., and Schoffelen, J.-M. (2011). Fieldtrip: open source software for advanced analysis of MEG, EEG, and invasive electrophysiological data. *Comput. Intell. Neurosci.* 2011, 156869. doi: 10.1155/2011/156869
- Otsuka, H., Harada, M., Mori, K., Hisaoka, S., and Nishitani, H. (1999). Brain metabolites in the hippocampus-amygdala region and cerebellum in autism: an 1h-mr spectroscopy study. *Neuroradiology* 41, 517–519. doi: 10.1007/s002340050795
- Parkkonen, L., Fujiki, N., and Mäkelä, J. P. (2009). Sources of auditory brainstem responses revisited: contribution by magnetoencephalography. *Hum. Brain Mapp.* 30, 1772–1782. doi: 10.1002/hbm.20788
- Pellicano, E., and Burr, D. (2012). When the world becomes too real: a bayesian explanation of autistic perception. *Trends Cogn. Sci.* 16, 504–510. doi: 10.1016/j.tics.2012.08.009
- Plaisted, K. C. (2001). “Reduced generalization in autism: an alternative to weak central coherence,” in *The Development of Autism: Perspectives from Theory and Research*, eds J. A. Burack, T. Charman, N. Yirmiya, and P. R. Zelazo (Mahwah, NJ: Lawrence Erlbaum Associates Publishers), xvii, 149–169.
- Prokopenko, M., Gerasimov, V., and Tanev, I. (2006). “Evolving spatiotemporal coordination in a modular robotic system,” in *From Animals to Animats 9: Proceedings of the Ninth International Conference on the Simulation of Adaptive Behavior (SAB'06)*. Lecture notes in computer science, Vol. 4095, eds S. Nolfi, G. Baldassarre, R. Calabretta, J. C. T. Hallam, D. Marocco, J.-A. Meyer, et al. (Berlin: Springer), 558–569.
- Ragwitz, M., and Kantz, H. (2002). Markov models from data by simple nonlinear time series predictors in delay embedding spaces. *Phys. Rev. E Stat. Nonlin. Soft Matter Phys.* 65, 056201. doi: 10.1103/PhysRevE.65.056201
- Rao, R. P., and Ballard, D. H. (1999). Predictive coding in the visual cortex: a functional interpretation of some extra-classical receptive-field effects. *Nat. Neurosci.* 2, 79–87. doi: 10.1038/4580
- Raymond, G. V., Bauman, M. L., and Kemper, T. L. (1995). Hippocampus in autism: a golgi analysis. *Acta Neuropathol.* 91, 117–119. doi: 10.1007/s004010050401
- Riggs, L., Moses, S. N., Bardouille, T., Herdman, A. T., Ross, B., and Ryan, J. D. (2009). A complementary analytic approach to examining medial temporal lobe sources using magnetoencephalography. *Neuroimage* 45, 627–642. doi: 10.1016/j.neuroimage.2008.11.018
- Rogers, S. J., and Ozonoff, S. (2005). Annotation: what do we know about sensory dysfunction in autism? a critical review of the empirical evidence. *J. Child Psychol. Psychiatry* 46, 1255–1268. doi: 10.1111/j.1469-7610.2005.01431.x
- Ropar, D., and Mitchell, P. (1999). Are individuals with autism and asperger's syndrome susceptible to visual illusions? *J. Child Psychol. Psychiatry* 40, 1283–1293. doi: 10.1111/1469-7610.00544
- Ropar, D., and Mitchell, P. (2001). Susceptibility to illusions and performance on visuospatial tasks in individuals with autism. *J. Child Psychol. Psychiatry* 42, 539–549. doi: 10.1111/1469-7610.00748
- Ropar, D., and Mitchell, P. (2002). Shape constancy in autism: the role of prior knowledge and perspective cues. *J. Child Psychol. Psychiatry* 43, 647–653. doi: 10.1111/1469-7610.00053
- Roux, F., Wibral, M., Singer, W., Aru, J., and Uhlhaas, P. J. (2013). The phase of thalamic alpha activity modulates cortical gamma-band activity: evidence from resting-state meg recordings. *J. Neurosci.* 33, 17827–17835. doi: 10.1523/JNEUROSCI.5778-12.2013
- Saß, H., Wittchen, H.-U., Zaudig, M., and Houben, I. (2003). *Dsm-iv-tr*. Göttingen: Diagnostische Kriterien, Hogrefe.
- Schmötzer, G., Rühl, D., Thies, G., and Poustka, F. (1993). *Autismus diagnostisches interview-revision*. Frankfurt/Main: Deutsche Übersetzung. Universität Frankfurt (Eigendruck).
- Schreiber (2000). Measuring information transfer. *Phys. Rev. Lett.* 85, 461–464. doi: 10.1103/PhysRevLett.85.461
- Schumann, C. M., Hamstra, J., Goodlin-Jones, B. L., Lotspeich, L. J., Kwon, H., Buonocore, M. H., et al. (2004). The amygdala is enlarged in children but not adolescents with autism; the hippocampus is enlarged at all ages. *J. Neurosci.* 24, 6392–6401. doi: 10.1523/JNEUROSCI.1297-04.2004
- Seltzer, M. M., Krauss, M. W., Shattuck, P. T., Orsmond, G., Swe, A., and Lord, C. (2003). The symptoms of autism spectrum disorders in adolescence and adulthood. *J. Autism Dev. Disord.* 33, 565–581. doi: 10.1023/B:JADD.0000005995.02453.0b
- Sheikhan, A., Behnam, H., Mohammadi, M. R., Noroozian, M., and Golabi, P. (2007). “Analysis of quantitative electroencephalogram background activity in autism disease patients with Lempel-Ziv complexity and Short Time Fourier Transform measure,” in *Conference Proceedings of 4th IEEE-EMBS International Summer School and Symposium on Medical Devices and Biosensors* (Cambridge). doi: 10.1109/ISSMDBS.2007.4338305
- Sheppard, E., Ropar, D., and Mitchell, P. (2007). The impact of meaning and dimensionality on copying accuracy in individuals with autism. *J. Autism Dev. Disord.* 37, 1913–1924. doi: 10.1007/s10803-006-0321-9
- Strange, B. A., Duggins, A., Penny, W., Dolan, R. J., and Friston, K. J. (2005). Information theory, novelty and hippocampal responses: unpredicted or unpredictable? *Neural Netw.* 18, 225–230. doi: 10.1016/j.neunet.2004.12.004
- Sun, L., Grützner, C., Bölte, S., Wibral, M., Tozman, T., Schlitt, S., et al. (2012). Impaired gamma-band activity during perceptual organization in adults with autism spectrum disorders: evidence for dysfunctional network activity in frontal-posterior cortices. *J. Neurosci.* 32, 9563–9573. doi: 10.1523/JNEUROSCI.1073-12.2012
- Taylor, M., Donner, E., and Pang, E. (2012). fmri and meg in the study of typical and atypical cognitive development. *Neurophysiol. Clin.* 42, 19–25. doi: 10.1016/j.neucli.2011.08.002
- Tesche, C., Karhu, J., and Tissari, S. (1996). Non-invasive detection of neuronal population activity in human hippocampus. *Cogn. Brain Res.* 4, 39–47. doi: 10.1016/0926-6410(95)00044-5
- Treffert, D. A. (2009). The savant syndrome: an extraordinary condition. A synopsis: past, present, future. *Philos. Trans. R. Soc. B Biol. Sci.* 364, 1351–1357. doi: 10.1098/rstb.2008.0326
- Turing, A. M. (1936). On computable numbers, with an application to the Entscheidungsproblem. *Proc. Lond. Math. Soc.* 42, 230–265.
- Vicente, R., Wibral, M., Lindner, M., and Pipa, G. (2011). Transfer entropy – a model-free measure of effective connectivity for the neurosciences. *J. Comput. Neurosci.* 30, 45–67. doi: 10.1007/s10827-010-0262-3
- Wibral, M., Lizier, J., Vögler, S., Priesemann, V., and Galuske, R. (2014). Local active information storage as a tool to understand distributed neural information processing. *Front. Neuroinform.* 8:1. doi: 10.3389/fninf.2014.00001
- Williams, D., and Bishop, J. (1994). *Somebody Somewhere: Breaking Free from the World of Autism*. New York, NY: Times Book.
- Wing, L., and Gould, J. (1979). Severe impairments of social interaction and associated abnormalities in children: epidemiology and classification. *J. Autism Dev. Disord.* 9, 11–29. doi: 10.1007/BF01531288
- Yeargin-Allsopp, M., Rice, C., Karapurkar, T., Doernberg, N., Boyle, C., and Murphy, C. (2003). Prevalence of autism in a us metropolitan area. *JAMA* 289, 49–55. doi: 10.1001/jama.289.1.49
- Zipser, D., Kehoe, B., Littlewort, G., and Fuster, J. (1993). A spiking network model of short-term active memory. *J. Neurosci.* 13, 3406–3420.

**Conflict of Interest Statement:** The authors declare that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.

Received: 15 November 2013; accepted: 23 January 2014; published online: 14 February 2014.

Citation: Gómez C, Lizier JT, Schaum M, Wollstadt P, Grützner C, Uhlhaas P, Freitag CM, Schlitt S, Bölte S, Hornero R and Wibral M (2014) Reduced predictable information in brain signals in autism spectrum disorder. *Front. Neuroinform.* 8:9. doi: 10.3389/fninf.2014.00009

This article was submitted to the journal *Frontiers in Neuroinformatics*.

Copyright © 2014 Gómez, Lizier, Schaum, Wollstadt, Grützner, Uhlhaas, Freitag, Schlitt, Bölte, Hornero and Wibral. This is an open-access article distributed under the terms of the Creative Commons Attribution License (CC BY). The use, distribution or reproduction in other forums is permitted, provided the original author(s) or licensor are credited and that the original publication in this journal is cited, in accordance with accepted academic practice. No use, distribution or reproduction is permitted which does not comply with these terms.



# Energy landscapes of resting-state brain networks

Takamitsu Watanabe<sup>1,2</sup>, Satoshi Hirose<sup>1</sup>, Hiroyuki Wada<sup>3</sup>, Yoshio Imai<sup>3</sup>, Toru Machida<sup>3</sup>, Ichiro Shirouzu<sup>3</sup>, Seiki Konishi<sup>1</sup>, Yasushi Miyashita<sup>1</sup> and Naoki Masuda<sup>4,5\*</sup>

<sup>1</sup> Department of Physiology, The University of Tokyo School of Medicine, Tokyo, Japan

<sup>2</sup> Awareness Group, Institute of Cognitive Neuroscience, University College London, London, UK

<sup>3</sup> Department of Radiology, NTT Medical Center Tokyo, Tokyo, Japan

<sup>4</sup> Department of Mathematical Informatics, The University of Tokyo, Tokyo, Japan

<sup>5</sup> CREST, Japan Science and Technology Agency, Saitama, Japan

## Edited by:

Daniele Marinazzo, University of Gent, Belgium

## Reviewed by:

John A. Hertz, Niels Bohr Institute, Denmark

Shan Yu, National Institute of Mental Health, USA

## \*Correspondence:

Naoki Masuda, Department of Mathematical Informatics, The University of Tokyo, 7-3-1 Hongo, Bunkyo-ku, Tokyo 113-8655, Japan  
e-mail: masuda@mist.i.u-tokyo.ac.jp

During rest, the human brain performs essential functions such as memory maintenance, which are associated with resting-state brain networks (RSNs) including the default-mode network (DMN) and frontoparietal network (FPN). Previous studies based on spiking-neuron network models and their reduced models, as well as those based on imaging data, suggest that resting-state network activity can be captured as attractor dynamics, i.e., dynamics of the brain state toward an attractive state and transitions between different attractors. Here, we analyze the energy landscapes of the RSNs by applying the maximum entropy model, or equivalently the Ising spin model, to human RSN data. We use the previously estimated parameter values to define the energy landscape, and the disconnectivity graph method to estimate the number of local energy minima (equivalent to attractors in attractor dynamics), the basin size, and hierarchical relationships among the different local minima. In both of the DMN and FPN, low-energy local minima tended to have large basins. A majority of the network states belonged to a basin of one of a few local minima. Therefore, a small number of local minima constituted the backbone of each RSN. In the DMN, the energy landscape consisted of two groups of low-energy local minima that are separated by a relatively high energy barrier. Within each group, the activity patterns of the local minima were similar, and different minima were connected by relatively low energy barriers. In the FPN, all dominant local minima were separated by relatively low energy barriers such that they formed a single coarse-grained global minimum. Our results indicate that multistable attractor dynamics may underlie the DMN, but not the FPN, and assist memory maintenance with different memory states.

**Keywords: resting-state network, maximum entropy model, Ising model, attractor dynamics, functional connectivity**

## INTRODUCTION

In the last few decades, a line of neuroimaging studies have accumulated evidence supporting that spontaneous brain activity during rest is not random enough to be averaged out in statistical analysis (Biswal et al., 1995; Raichle et al., 2001; Greicius et al., 2003). The brain activity in resting states shows consistent spatial patterns called the resting-state networks (RSNs) (Raichle et al., 2001; Greicius et al., 2003; Fox et al., 2005; Dosenbach et al., 2006; Fair et al., 2009). Connections between the RSNs and cognitive functions have been revealed in previous studies. In particular, the default-mode network (DMN), one of the representative RSNs, is suggested to be engaged in self-referential mental processes and maintenance of long-term memory (Raichle et al., 2001; Greicius et al., 2003; Buckner et al., 2008; Uddin et al., 2009). The frontoparietal network (FPN), another RSN, is known to be recruited during cognitive tasks with relatively high loads that require continuous attention (Dosenbach et al., 2006; Corbetta et al., 2008; Fair et al., 2009).

Most of these results on the RSNs were derived from correlations between slow oscillations of brain activity (0.01–0.1 Hz) in different brain regions. However, the neural activity as observed

in the RSNs at a macroscopic spatial scale is dynamic on a much shorter time scale. Experimental and computational studies indicate that within a RSN, a group of brain regions is specifically activated within a specific time window, and that different groups of regions are activated during different time windows (Honey et al., 2007; Chang and Glover, 2010; Kiviniemi et al., 2011; Allen et al., 2012; Hutchison et al., 2013). Such spatio-temporal dynamics of the RSNs may facilitate, for example, the flexibility of human cognitive functions (Allen et al., 2012).

These results pertaining to the dynamics of the resting-state brain activity suggest that the activity of the RSNs may be captured in terms of transitions among locally stable states, i.e., attractor states (Deco et al., 2012, 2013; Nakagawa et al., 2013). In fact, beyond the description of RSNs, attractor network models of spiking neurons and firing-rate models derived by the reduction of spiking-neuron models have been used to model cortical dynamics (for reviews, see Barbieri and Brunel, 2008; Wang, 2009; Braun and Mattia, 2010; Knierim and Zhang, 2012). In particular, the role of attractor dynamics has been implicated in brain activity during various cognitive functions such as associative long-term memory, non-spatial working memory, spatial

working memory, place field recognition, decision making, and attention. The aforementioned models are particularly successful in describing persistent activity recorded during these cognitive tasks. Although attractor network models may be too simple to describe fast transients of brain activity accurately (Rabinovich et al., 2008; Rabinovich and Varona, 2011), these experimental and numerical results are consistent with the notion that brain dynamics are multistable and that the brain's state travels from one state to another depending on, for example, external input and endogenous cognitive processes.

In associative memory models, an energy function often exists such that each state possesses a corresponding energy value and a state with a low energy is taken with a large probability (Hopfield, 1982; Hertz et al., 1991). In this case, the attractor dynamics can be described by a trajectory that represents a dynamical state of the brain in an energy landscape. Therefore, estimating the energy landscapes of brain activity contributes to understanding of brain dynamics from the perspective of attractor dynamics. In the present study, we investigate the energy landscapes of resting-state brain activity using the functional magnetic resonance imaging (fMRI) data previously collected by our group (Watanabe et al., 2013). In the previous work based on these data, we demonstrated that the so-called pairwise maximum entropy model (MEM) (Schneidman et al., 2006; Shlens et al., 2006; Tang et al., 2008; Yu et al., 2008; Ohiorhenuan et al., 2010; Santos et al., 2010; Ganmor et al., 2011) described the activities of the DMN and FPN with high accuracy (Watanabe et al., 2013). For the fitted models from that study and randomized RSNs, here we calculated the energy of all the brain states and identified local minima of energy that would correspond to the attractors in attractor dynamics. Then, we applied the so-called disconnectivity graph method (Becker and Karplus, 1997) to the empirical and artificial energy landscapes of the RSNs. We found that the energy landscapes of the DMN and FPN are qualitatively different.

## MATERIALS AND METHODS

### DATA ACQUISITION AND FITTING OF THE PAIRWISE MEM

To examine the energy landscape of the RSNs, we used the parameter values estimated in our previous study in which we fitted the so-called pairwise MEM to the resting-state fMRI data (Watanabe et al., 2013) (Figure 1A). The fMRI data were recorded while six healthy right-handed subjects (aged 20–23 years; three males) were resting inside a 3T MRI scanner (Philips Achieva X 3T Rel. 2.6, Best, The Netherlands; gradient-echo echo-planar sequences:  $TR = 9.045$  s,  $TE = 35$  ms, flip angle =  $90^\circ$ , resolution =  $2 \times 2 \times 2$  mm<sup>3</sup>, 75 slices). In total, 17,820 volumes of resting-state fMRI images were obtained. The entire procedure for the MRI scanning was approved by the institutional review board of The University of Tokyo, School of Medicine.

The pairwise MEM and the fitting procedure are outlined as follows. Readers interested in the detailed procedures should refer to our previous article (Watanabe et al., 2013). First, we conducted a conventional preprocessing procedure that consisted of slice-timing correction, spatial normalization, spatial smoothing, motion correction, and temporal band-pass filtering. Second, to normalize the fMRI data, we subtracted the average from the signals and divided the obtained values by their standard deviation

for each brain region. Third, we binarized the normalized signals with a threshold of 0.1. The binarized activity at brain region  $i$  and discrete time  $t$ , denoted by  $\sigma_i^t$ , is either active (+1) or inactive (0). The network state at time  $t$  is described by

$$V^t = [\sigma_1^t, \sigma_2^t, \dots, \sigma_N^t], \quad (1)$$

where  $N$  is the number of the brain regions in a RSN. It should be noted that there are  $2^N$  network states. The empirical activation probability of region  $i$ , denoted by  $\langle \sigma_i \rangle$ , is equal to  $(1/T) \sum_{t=1}^T \sigma_i^t$ , where  $T$  is the number of images. The empirical joint activation probability of regions  $i$  and  $j$ , denoted by  $\langle \sigma_i \sigma_j \rangle$ , is given by  $(1/T) \sum_{t=1}^T \sigma_i^t \sigma_j^t$ .

Fourth, we adopted the distribution of the network state that maximized the entropy under the restriction that  $\langle \sigma_i \rangle$  and  $\langle \sigma_i \sigma_j \rangle$  ( $1 \leq i \leq N$ ,  $1 \leq j \leq N$ ,  $i \neq j$ ) for the inferred model were equal to the empirical values. Such a distribution is known to have the form

$$P(V_k) = e^{-E(V_k)} / \sum_{\ell=1}^{2^N} e^{-E(V_\ell)}, \quad (2)$$

where  $P(V_k)$  is the probability of the  $k$ th network state  $V_k$ , and

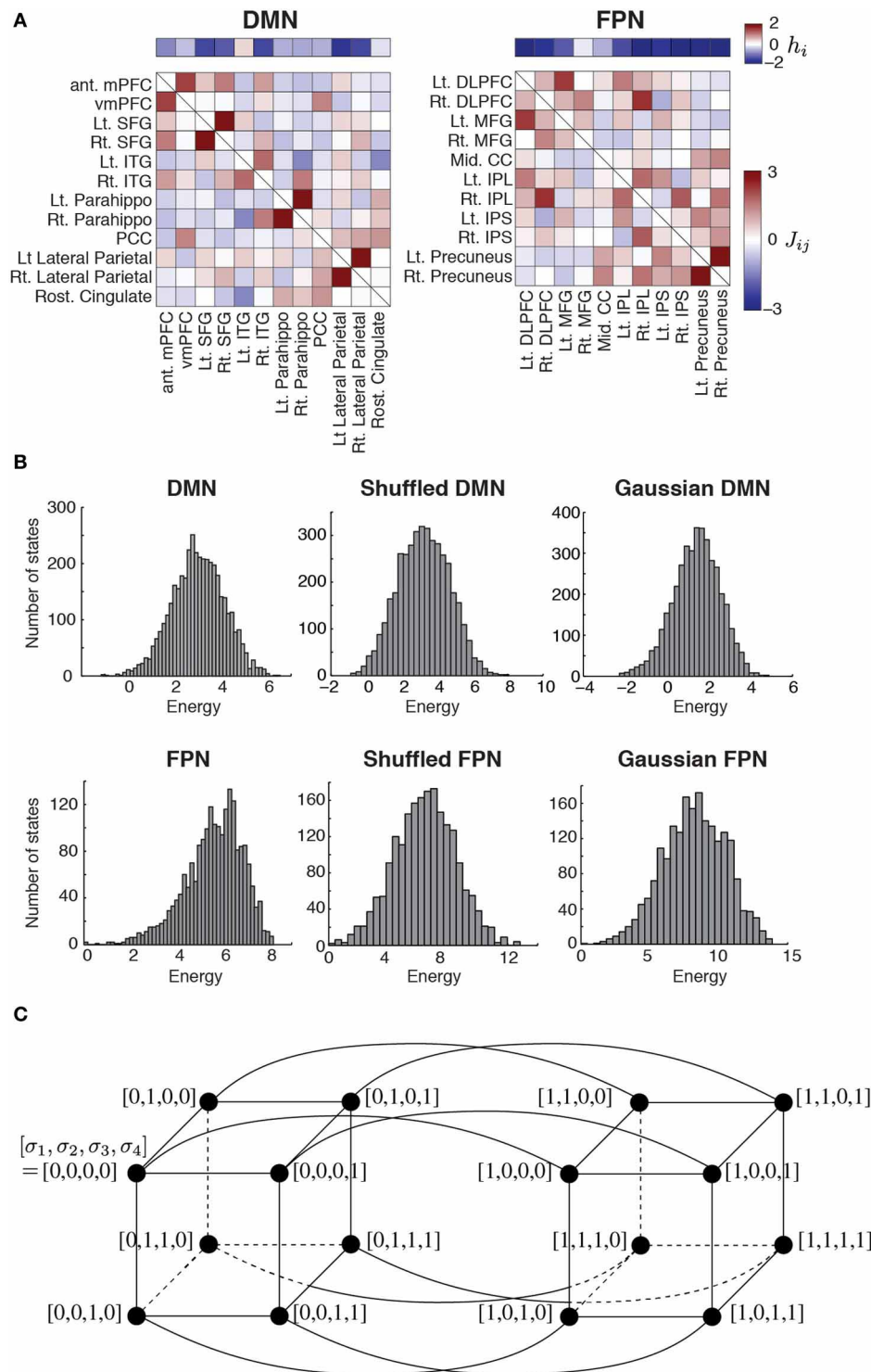
$$E(V_k) = - \sum_{i=1}^N h_i \sigma_i(V_k) - \frac{1}{2} \sum_{i=1}^N \sum_{j=1, j \neq i}^N J_{ij} \sigma_i(V_k) \sigma_j(V_k) \quad (3)$$

is the energy of network state  $V_k$ . Variable  $\sigma_i(V_k)$  indicates the value of  $\sigma_i$  (i.e., 1 or 0) under network state  $V_k$ . For the inferred model, the expected activation probability,  $\langle \sigma_i \rangle_m$ , and the expected pairwise joint activation probability,  $\langle \sigma_i \sigma_j \rangle_m$ , are given by  $\langle \sigma_i \rangle_m = \sum_{\ell=1}^{2^N} \sigma_i(V_\ell) P(V_\ell)$  and  $\langle \sigma_i \sigma_j \rangle_m = \sum_{\ell=1}^{2^N} \sigma_i(V_\ell) \sigma_j(V_\ell) P(V_\ell)$ , respectively. We determined  $h_i$  and  $J_{ij}$  by iteratively adjusting  $\langle \sigma_i \rangle_m$  and  $\langle \sigma_i \sigma_j \rangle_m$  toward  $\langle \sigma_i \rangle$  and  $\langle \sigma_i \sigma_j \rangle$ , respectively, with a gradient descent algorithm. As a result, we obtained  $h_i$  ( $1 \leq i \leq N$ ) and  $J_{ij}$  ( $= J_{ji}$ ;  $1 \leq i \leq N$ ,  $1 \leq j \leq N$ ,  $i \neq j$ ) for a RSN (DMN or FPN) (Figure 1A). Here,  $h_i$  is considered to represent the basal brain activity of region  $i$ , i.e., the expected brain activity when the region is isolated.  $J_{ij}$  represents the functional interaction between regions  $i$  and  $j$ . The brain regions constituting each RSN, with the labels being indicated in Figure 1A, were determined on the basis of previous studies (Dosenbach et al., 2006; Fair et al., 2009).

### DISCONNECTIVITY GRAPH

The energy landscape of a RSN is specified by two factors: the energy  $E(V_k)$  of the  $2^N$  network states  $V_k$ , which are regarded as nodes in a network of network states; and the connectivity between different nodes (i.e., network states). One RSN inferred by the pairwise MEM defines an energy landscape. Two nodes are defined to be adjacent by a link if and only if they take the opposite binary activity at just one brain region (i.e., one  $\sigma_i$ ; see Figure 1C for the case of  $N = 4$ ).

We analyzed the energy landscape for each RSN using disconnectivity graphs (Becker and Karplus, 1997; Wales, 2010). In



**FIGURE 1 | (A)** Parameter values estimated for the two RSNs. The horizontal bars show the basal brain activity ( $h_i$ ). The square matrices show the functional connectivity between pairs of regions ( $J_{ij}$ ) as determined by the fitting of the pairwise MEM. The obtained parameter values were identical to those obtained in our previous study (Watanabe et al., 2013). DMN, default mode network; FPN, fronto-parietal network; ant. mPFC, anterior medial prefrontal cortex; vmPFC, ventro-medial prefrontal cortex; Lt, left; Rt, right; SFG, superior frontal gyrus; ITG, inferior temporal gyrus; Parahippo, parahippocampal gyrus; PCC, posterior cingulate cortex; DLPFC, dorso-lateral

prefrontal cortex; MFG, middle prefrontal cortex; Mid, middle; CC, cingulate cortex; IPL, inferior parietal lobule; IPS, inferior parietal sulcus. **(B)** Distribution of energy for each network. To generate the histograms, we weighted each state equally, i.e., not with the probability that the state is realized. The results for the shuffled and Gaussian networks are based on a single realization of the network. **(C)** Concept of neighbors in a network of network states. For illustration, we set  $N = 4$ . The circles represent nodes, i.e., network states. A link between a pair of nodes indicates that the two nodes are adjacent.



short, a disconnectivity graph represents the (dis)connectivity between local minima of the energy. It has also been used to study the Ising spin model, which is equivalent to the pairwise MEM, and its variants (Garstecki et al., 1999; Zhou and Wong, 2009; Zhou, 2011). In the context of the spin systems, a disconnectivity graph with a continuous energy threshold, where the energy threshold is defined in the following, is also referred to as a barrier tree (Fontanari and Stadler, 2002; Hordijk et al., 2003).

We constructed disconnectivity graphs in the following way: (1) A local minimum is a node whose energy is smaller than those of all the  $N$  neighboring nodes. We exhaustively examined whether each of the  $2^N$  nodes is a local minimum. (2) We set a threshold energy level, denoted by  $E_{th}$ , to the largest energy level realized by (at least) one of the  $2^N$  nodes. (3) We removed the nodes whose energy level was higher than  $E_{th}$ . We also removed all links incident to a removed node. In fact, no node or link was removed when the threshold was equal to the largest possible energy level. Some nodes and links were removed when we revisited this step after lowering the  $E_{th}$  value. (4) We judged whether each pair of local minima was connected by a path in the reduced network. In general, the local minima are classified into some connected components. (5) We repeated steps (3) and (4) after moving  $E_{th}$  down to the next largest energy level realized by a node. Finally, we obtained a reduced network of the local minima in which each local minimum was isolated. (6) On the basis of these results, we built a disconnectivity graph, i.e., a hierarchical tree whose leaves (i.e., terminal nodes down in the tree) were the local minima. The vertical position of the leaves and internal nodes of the disconnectivity graph represents an energy value. An internal node represents the point at which the branching of different groups of local minima takes place. In other words, local minima that are contained in different branches belong to distinct connected components for an  $E_{th}$  larger than the value at the common internal root node. Local minima in the different branches belong to the same connected component for  $E_{th}$  smaller than this value.

### **Basin Size of Local Minimum**

We then calculated the size of the basin of each local minimum as follows (Stillinger and Weber, 1982, 1984; Becker and Karplus, 1997; Zhou, 2011). We first selected a starting node  $i$ , which was one of the  $2^N$  nodes in the network of network states. Then, we identified the neighbor of node  $i$  possessing the smallest energy level and denoted it by  $j$ . If  $E(V_j) < E(V_i)$ , we moved to node  $j$ . This move is in accordance with the steepest descent at node  $i$ . If such a node  $j$  did not exist, we remained at node  $i$ . In the latter case,  $i$  is a local minimum. If we moved to node  $j$ , we looked for the steepest descent from node  $j$  and continued to travel until we arrived at a local minimum. The starting node  $i$  belongs to the basin of the local minimum that is finally reached. We performed the same procedure for all  $i$ . The basin size of a local minimum is the fraction of nodes that belong to the basin of the local minimum.

### **Energy Barrier**

For a given disconnectivity graph, we estimated the energy barrier opposing transitions between two local minima denoted by

$i$  and  $j$ . Specifically, we defined the energy barrier between  $i$  and  $j$  as  $\min [E^b(V_i, V_j) - V_i, E^b(V_i, V_j) - V_j]$ , where  $E^b(V_i, V_j)$  is the threshold energy level at which the disconnectivity graph branches into a group of nodes that includes  $i$  and a group that includes  $j$ . Any path connecting  $i$  and  $j$  in the network of network states contains a node whose energy is at least  $E^b(V_i, V_j)$ . If the energy barrier is high, the transition of network states between  $i$  and  $j$  occurs at a small rate at least in one direction. In fact, the transition occurs at different rates in the two directions if  $V_i$  and  $V_j$  are different (Becker and Karplus, 1997). However, for simplicity, we used the symmetric definition given above (Zhou, 2011).

### **Hierarchical Clustering**

We carried out hierarchical clustering of the brain regions and local minima as follows by using MATLAB. First, we set a distance threshold,  $d_{th}$  to the smallest Hamming distance realized by a pair of nodes. If the distance between a node pair was equal to or less than the current  $d_{th}$  value, we bridged the two nodes through a parent node, which is located at  $d_{th}$  along the axis in the dendrogram. We repeated this procedure by gradually elevating  $d_{th}$  until all nodes were connected as a single dendrogram.

### **Randomized RSNs**

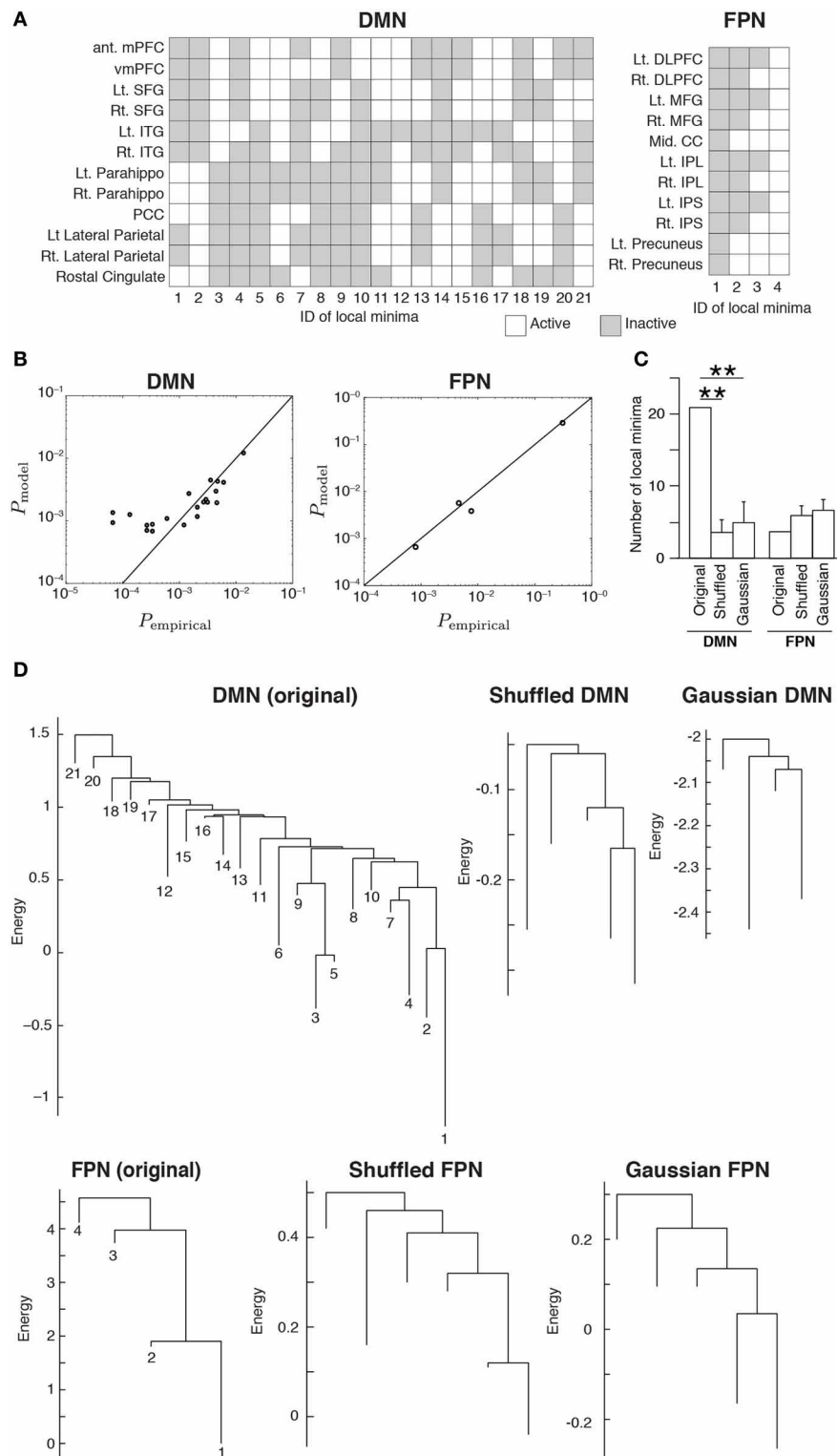
As controls, we calculated the disconnectivity graph and other properties of the energy landscape for two types of randomized MEMs. We generated the first type of network by randomly permuting  $h_i$  ( $1 \leq i \leq N$ ) of the original MEM and doing the same for  $J_{ij}$  ( $= J_{ji}$ ;  $1 \leq i \leq j \leq N$ ). We refer to the generated network as a shuffled network. We also generate a second type of randomized network by independently drawing the values of  $h_i$  ( $1 \leq i \leq N$ ) from a normal distribution with the same mean and standard deviation as those of the original MEM and doing the same for  $J_{ij}$  ( $= J_{ji}$ ;  $1 \leq i \leq j \leq N$ ). We refer to the generated network as a Gaussian network.

## **RESULTS**

### **Local Minima and the Disconnectivity Graph**

The parameter values of the pairwise MEM inferred for the DMN and FPN are shown in **Figure 1A**. The distribution of the energy on the basis of all the  $2^N$  network states is shown in **Figure 1B** for the two RSNs. The distribution of the energy was unimodal for both RSNs. The shape of the distribution did not significantly differ from that obtained from either of the randomized networks (for both shuffled and Gaussian networks,  $P > 0.6$  in the Kolmogorov–Smirnov test; **Figure 1B**).

The inferred MEMs for the DMN and FPN had 21 and 4 local minima, respectively. The activity pattern of each local minimum is shown in **Figure 2A**. In both RSNs, the probabilities that different local minima were visited were similar between the empirical data and the pairwise MEM (**Figure 2B**). The similarity is particularly evident for the local minima with a low energy (i.e., large probability of the visit) in the DMN. In fact, the error averaged over all 21 local minima in the DMN was 260%. Here, we defined the error for a local minimum as the absolute difference between the empirical and estimated probabilities that the local minimum is realized, divided by the empirical probability. However, the



**FIGURE 2 | (A)** Activation patterns of the local minima. The IDs of the local minima are shown on the horizontal axis. The local minima are sorted in order of ascending energy. Each local minimum is specified by an activation pattern, which is an  $N$ -dimensional binary vector. The white and gray elements indicate active and inactive brain regions, respectively. **(B)** Comparison of the probability that the local minima are realized between the empirical data and the model. Each circle represents a local minimum. **(C)** The number of local minima for the

original RSNs and the average number of local minima for the randomized RSNs, where the average is taken over 100 realizations of each type of the randomized networks. Error bars show the standard deviation. \*\* $P < 0.01$ , Bonferroni-corrected. **(D)** Disconnectivity graphs. The vertical axis represents the energy. The numbers immediately under the leaves (i.e., end nodes) represent the IDs of the local minima as defined in panel (A). The energy value at the bottom end of a leaf is equal to that of the corresponding local minimum.

large error was due to three outliers with small probabilities (ID 12, 13, and 16). If the three minima were excluded, the averaged error was 33.2%. Moreover, the error averaged over the 11 local minima with the lowest energy values (i.e., largest probabilities) was 26.2%. In the FPN, the error averaged over all four local minima was 18%. Together with these error values, the results shown in **Figure 2B** justify the use of local minima of the pairwise MEM in the following analysis as stochastic footprints of the network state.

We calculated the number of local minima for 100 realizations of the two types of randomized RSNs. For the DMN, the number of local minima was significantly larger for the original network than for either type of randomized network ( $P < 0.01$ , Bonferroni-corrected; **Figure 2C**). For the FPN, there was no significant difference in the number of local minima between the original and randomized networks.

We then constructed disconnectivity graphs to illustrate relationships between the local minima. The disconnectivity graphs for the original RSN and one realization for each type of randomized network are shown in **Figure 2D**, separately for the DMN and FPN. In the DMN, the structure of the empirical disconnectivity graph was apparently more complex than that of the randomized networks, partly because the former had more local minima than the latter (**Figure 2C**). The disconnectivity graph of the DMN has a complex and forked structure relative to that of the FPN. In contrast, the disconnectivity graph of the FPN seems not as complex as the randomized networks and is composed of a single dominant minimum with weak fluctuations, which is one of the main subtypes of the disconnectivity graph (Becker and Karplus, 1997; Wales et al., 1998).

### CLUSTERING OF BRAIN REGIONS AND LOCAL MINIMA

To probe the relationships between different local minima, we performed hierarchical clustering on the basis of similarity between local minima. The (dis)similarity between two local minima was defined by the Hamming distance between the activity patterns of the local minima, i.e., the number of brain regions at which the two local minima possess the opposite binary activity. We constructed a dendrogram for each RSN (see Materials and Methods for the algorithm).

The dendrogram shown in **Figure 3A** suggests that, in the DMN, bilateral brain regions show similar activation patterns in most of the local minima. In particular, in the parahippocampal gyri, superior frontal gyri, and lateral parietal region, the bilateral regions had exactly the same activation patterns in all the local minima. In contrast, the resemblance of bilateral regions is uncommon in the FPN. According to the dendrogram, a region in a bilateral region pair was not the nearest to its counterpart, except in the case of the precuneus.

We also quantified the similarity among the local minima by the same hierarchical clustering algorithm (**Figure 3B**). In the DMN, local minima with the lowest energies (e.g., local min #1 to #6) were relatively dissimilar. The energy landscape of the DMN is composed of relatively distinct local minima that yield multistability. In contrast, in the FPN, the local minima with the lowest energies (e.g., #1 and #2) were more similar to each other than in the case of the DMN. Therefore, we consider that the

energy landscape of the FPN is essentially composed of a single global minimum. We provide support of this interpretation in the following sections.

### SIZE OF BASIN

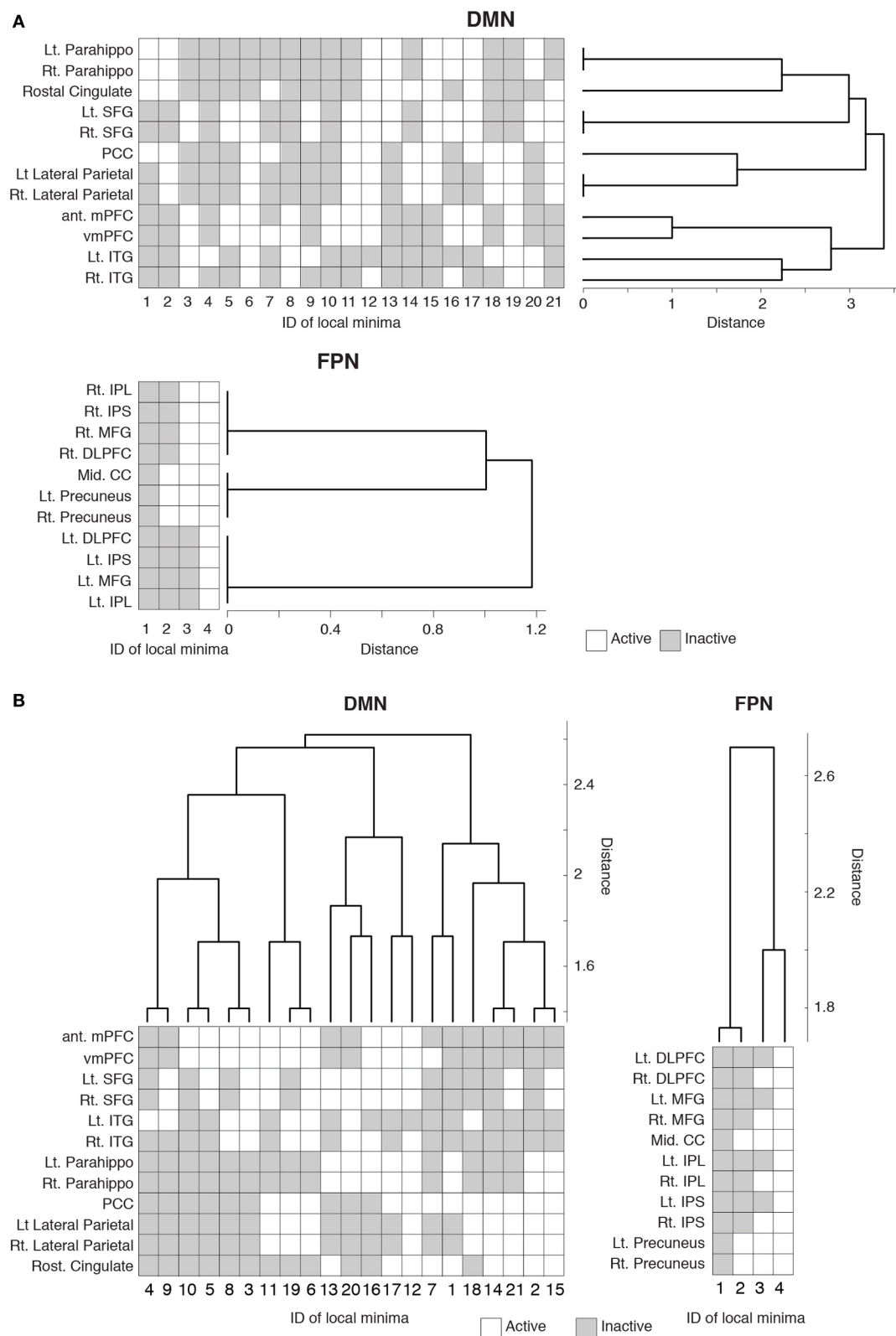
To further characterize the energy landscape of the two RSNs, we calculated the size of the basin of the local minima. The relationship between the size of the basin and the energy value is shown in **Figure 4A**. In the figure, an open circle represents a local minimum. In both RSNs, a local minimum with a small energy value tended to have a large basin. This tendency was even stronger in the randomized networks. In both empirical and randomized RSNs, a small number of the local minima with the lowest energy values attracts a majority of the network states (in the sense of the steepest descent walk in the energy landscape). The fraction of network states attracted to one of the local minima with the lowest energies is shown in **Figure 4B**. For example, when the value at the fraction of local minima is equal to 0.5, the accumulated size of basins is over 0.8; that is, when the half of the local minima with the lowest energies is considered, over 80% of the network states belong to the basin of one of these local minima. In fact, only the six local minima with the lowest energies (ca. 28% of the local minima) attracted more than 80% of the network states in the DMN (solid line in **Figure 4B**). In the FPN, the local minimum with the lowest energy (25% of the local minima) attracted approximately 60% of the network states (dashed line in **Figure 4B**).

These results suggest that the lower part of the disconnectivity graph comprising the local minima with the smallest energies, i.e., a connected tree that contains leaves near the bottom in **Figure 2D**, reflects the backbone of the energy landscape. A visual inspection of **Figure 2D** reveals that the lower part of the disconnectivity graph for the DMN comprises two main branches, one consisting of local minima labeled 6, 9, 3, and 5, and the other consisting of local minima labeled 8, 10, 7, 4, 2, and 1. In contrast, the lower part of the disconnectivity graph for the FPN is composed of a single main branch.

### ENERGY BARRIER

To further quantify the difference between the DMN and FPN, we evaluated the transition rates between local minima by calculating the energy barrier between each pair of local minima. If the barrier is high relative to unity, transitions between the two local minima are rare, at least in one direction. The energy barriers for all the pairs of local minima are shown in **Figure 5A** for each RSN. In the figure, the local minima are sorted according to the energy value. **Figure 5A** suggests that, in the DMN, transitions among the major local minima accompany high energy barriers such that they occur at small rates. In contrast, in the FPN, transitions between local minima occur relatively easily at least in one direction because of the low barriers that separate them.

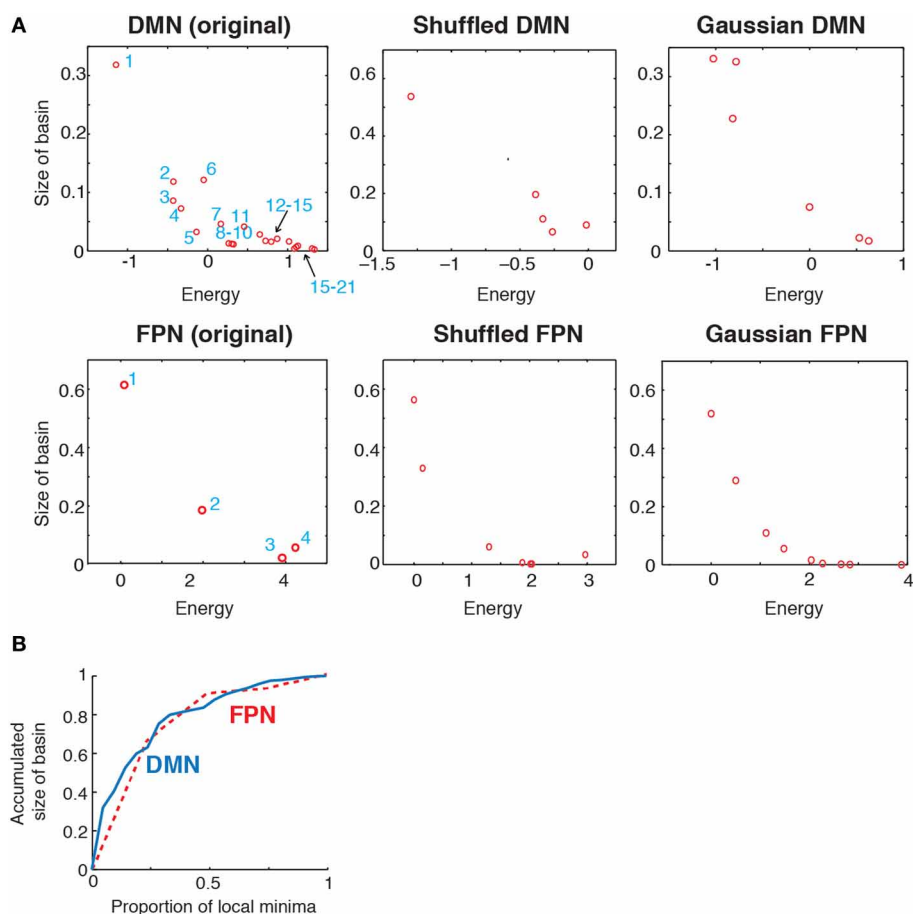
Subsequently, we calculated the average of energy barrier between pairs of local minima with the lowest energies. This amounts to averaging the energy barrier values contained in the leading principal minor of the matrix shown in **Figure 5A** (i.e., top left square submatrix) excluding the diagonal elements. The results are shown in **Figure 5B** as a function of the size



**FIGURE 3 | Hierarchical clustering of the brain regions and local minima.** Each row represents the activity pattern of a brain region in different local minima. Each column represents the activity pattern of a local minimum in different brain regions. **(A)** Dendrogram showing the

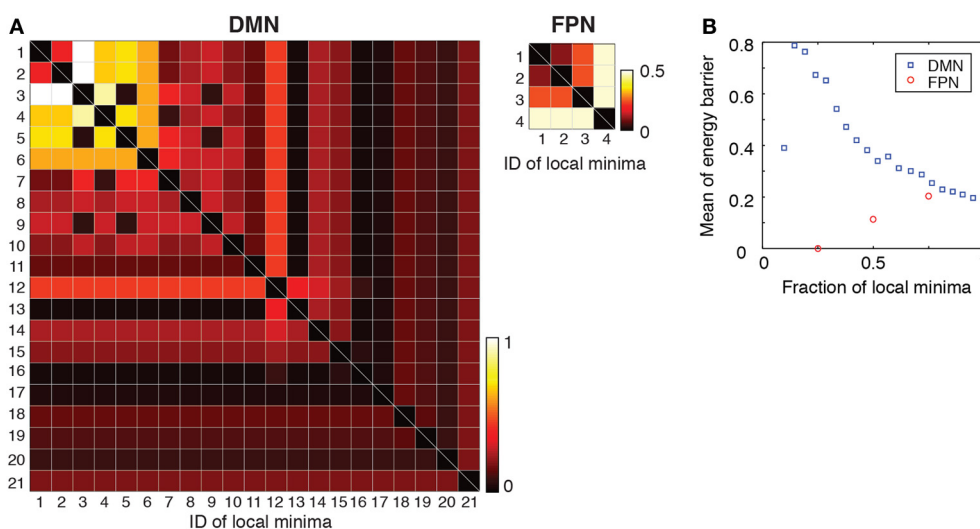
similarity among the brain regions in a hierarchical fashion. The similarity is measured by the Hamming distance between the activity patterns of two local minima. **(B)** Dendrogram showing the similarity among the local minima.





**FIGURE 4 | (A)** Relationship between the size of basin and the energy of local minima. In the two panels on the left, the numbers indicate the IDs of local minima used in **Figure 2**. **(B)** Accumulated size of the basin for the local minima. The vertical axis shows the fraction of the network states that

belong to the basin of one of the local minima with the lowest energies. This quantity is plotted against the fraction of local minima with the lowest energies. The solid and dashed curves indicate the results for the DMN and FPN, respectively.



**FIGURE 5 | (A)** Energy barrier between pairs of local minima. The local minima are sorted in order of ascending energy. **(B)** Average of the energy barrier between pairs of local minima with the lowest energies. For example,

the values at a fraction 0.5 of local minima indicate the average when we consider only pairs of the local minima whose energies are among the lowest 50%.

of the minor (i.e., the number of local minima with the lowest energies included in the analysis). As shown in **Figure 5B**, the averaged energy barrier was much larger in the DMN than in the FPN ( $P < 0.01$  in one-sample  $t$ -tests when the linear size of the minor is 25 and 50% of the entire DMN, respectively. The mean value of the FPN was used as a baseline in the  $t$ -tests). The difference between the two RSNs was larger when fewer local minima with the lowest energies were considered.

The results shown in **Figure 5** imply that, in the DMN, the brain activity may linger in the neighborhood of one of the several local minima for some time and wander from one to another. This interpretation is consistent with the result that the major local minima exhibit distinct activation patterns in the DMN (**Figure 3B**). In contrast, the brain activity in the FPN may tend to stay near the global minimum albeit with some fluctuations.

## DISCUSSION

We found that the DMN has dominant local minima that are relatively distinct in terms of the activation pattern and an energy barrier of the order of unity separating them from one another. The observed energy barrier is not large enough for the local minima to be justified as metastable states. However, if the brain state gradually changes, it may linger near a major local minimum for some time before transiting to another minimum. Therefore, roughly speaking, the present result is consistent with the concept of the multistable attractor dynamics for the RSN, in which the brain state is considered to travel from one relatively stable state to another, either in a spontaneous manner or triggered by external input (Deco et al., 2012, 2013). Such attractor dynamics may facilitate, for example, large capabilities of computation (Deco et al., 2013). It should be noted that we did not consider dynamics in the present study. Hence, dynamical variants of the present study warrant future work. Accounting for the dynamics will require better temporal resolution in imaging experiments.

In the DMN, the major local minima with small energies and large basins can be roughly classified into two groups separated by a relatively high energy barrier (**Figures 3B, 5A**). One group consists of the local minima in which the posterior brain regions are activated (local minima #1 and #2) and accounts for approximately 50% of the network states. The other main group consists of those in which the medial prefrontal regions are mainly activated (local minima #3, #4, and #5) and dominates approximately 30% of the network states. Therefore, the DMN is suggested to have two major coarse-grained states marked by posterior-centric activation (the first group) and frontal-centric activation (the second group). Previous studies suggested that the RSNs could be described by attractor dynamics (Deco et al., 2012, 2013). Our empirical evidence indicating the existence of two major coarse-grained states adds to these previous arguments. At a cellular level, multistable neural activity in the hippocampus represents multiple memory items (Leutgeb et al., 2005; Wills et al., 2005; Knierim and Zhang, 2012). The present macroscopic results lend a support to the possibility that multistable activity patterns in the DMN,

which includes the parahippocampal cortex, underlie various cognitive functions such as memory maintenance and self-referential thought (Raichle et al., 2001; Buckner et al., 2008; Uddin et al., 2009).

In contrast, the energy landscape of the FPN appears to be roughly monostable. In fact, the local minima were separated by low energy barriers (**Figure 5A**). A possible reason for the absence of multistability is that the activity pattern of the FPN during rest may be different from that when subjects are performing cognitively demanding tasks. The FPN was originally determined as a brain network for attentional cognition (Dosenbach et al., 2006; Corbetta et al., 2008). We should investigate the activity of the FPN during tasks to better understand the FPN.

An obvious limitation of the present study is that we have not directly examined attractor dynamics. Instead, we focused on the energy landscape of the network states constructed from the probability distribution of network states. There are several implicit assumptions underlying our energy landscape analysis. First, the network state was assumed to change gradually. Otherwise, a network state could jump from one local minimum to another by simultaneously flipping the binary states of multiple regions without passing through a network state that realizes the energy barrier. Our analysis, which exploits the concept of the energy barrier, would then be invalidated. The time window for constructing snapshots of brain activity should be small to track possibly step-by-step transitions of the network state. We followed our previous study (Watanabe et al., 2013) and used a time window of approximately 9 s because it was effective at decorrelating different snapshots. It should be noted that the energy landscape does not depend on the temporal resolution if we have sufficiently long data. Analyzing data with improved time resolution may be of interest. We should keep it in mind that the dynamics may not be gradual in fact; non-gradual transition can occur when fMRI signals at different brain regions are tightly synchronized.

Second, an energy barrier analysis implicitly assumes that state transitions depend on the difference between the energy values in the current and subsequent network states. Therefore, we implicitly ignored the effect of past network states on state transitions. To assess the extent of the history dependence of the trajectory is a relevant question. Addressing this question calls for a large amount of data; hence, it should be investigated in tandem with the effect of time window size because correlated snapshots would lead to stronger history dependence under the discrete time frame whose unit is defined by the size of the time window of the measurement.

## AUTHOR CONTRIBUTIONS

Takamitsu Watanabe and Naoki Masuda designed the research. Satoshi Hirose, Hiroyuki Wada, Yoshio Imai, Toru Machida, Ichiro Shirouzu, and Seiki Konishi conducted imaging experiments. Takamitsu Watanabe and Naoki Masuda analyzed the data and wrote the manuscript. Yasushi Miyashita discussed the results and commented on the manuscripts.

## ACKNOWLEDGMENTS

This work was supported by Grants-in-Aid for Scientific Research (23681033) from MEXT Japan to Naoki Masuda, a support from a JSPS fellowship for research abroad to Takamitsu Watanabe, a Grant-in-Aid for Specially Promoted Research (19002010) to Yasushi Miyashita, a Grant-in-Aid for Scientific Research B (22300134) to Seiki Konishi, and a research grant from Takeda Science Foundation to Yasushi Miyashita.

## REFERENCES

- Allen, E. A., Damaraju, E., Plis, S. M., Erhardt, E. B., Eichele, T., and Calhoun, V. D. (2012). Tracking whole-brain connectivity dynamics in the resting state. *Cereb. Cortex*. doi: 10.1093/cercor/bhs352. [Epub ahead of print].
- Barbieri, F., and Brunel, N. (2008). Can attractor network models account for the statistics of firing during persistent activity in prefrontal cortex? *Front. Neurosci.* 2, 114–122. doi: 10.3389/neuro.01.003.2008
- Becker, O. M., and Karplus, M. (1997). The topology of multidimensional potential energy surfaces: theory and application to peptide structure and kinetics. *J. Chem. Phys.* 106, 1495. doi: 10.1063/1.473299
- Biswal, B., Yetkin, F. Z., Haughton, V. M., and Hyde, J. S. (1995). Functional connectivity in the motor cortex of resting human brain using echo-planar MRI. *Magn. Reson. Med.* 34, 537–541. doi: 10.1002/mrm.1910340409
- Braun, J., and Mattia, M. (2010). Attractors and noise: twin drivers of decisions and multistability. *Neuroimage* 52, 740–751. doi: 10.1016/j.neuroimage.2009.12.126
- Buckner, R. L., Andrews-Hanna, J. R., and Schacter, D. L. (2008). The brain's default network: anatomy, function, and relevance to disease. *Ann. N.Y. Acad. Sci.* 1124, 1–38. doi: 10.1196/annals.1440.011
- Chang, C., and Glover, G. H. (2010). Time-frequency dynamics of resting-state brain connectivity measured with fMRI. *Neuroimage* 50, 81–98. doi: 10.1016/j.neuroimage.2009.12.011
- Corbetta, M., Patel, G., and Shulman, G. L. (2008). The reorienting system of the human brain: from environment to theory of mind. *Neuron* 58, 306–324. doi: 10.1016/j.neuron.2008.04.017
- Deco, G., Jirsa, V., and Friston, K. J. (2012). *Principles of Brain Dynamics*. Cambridge, MA: MIT Press.
- Deco, G., Jirsa, V. K., and McIntosh, A. R. (2013). Resting brains never rest: computational insights into potential cognitive architectures. *Trends Neurosci.* 36, 268–274. doi: 10.1016/j.tins.2013.03.001
- Dosenbach, N. U. F., Visscher, K. M., Palmer, E. D., Miezin, F. M., Wenger, K. K., Kang, H. C., et al. (2006). A core system for the implementation of task sets. *Neuron* 50, 799–812. doi: 10.1016/j.neuron.2006.04.031
- Fair, D. A., Cohen, A. L., Power, J. D., Dosenbach, N. U. F., Church, J. A., Miezin, F. M., et al. (2009). Functional brain networks develop from a “local to distributed” organization. *PLoS Comput. Biol.* 5:e1000381. doi: 10.1371/journal.pcbi.1000381
- Fontanari, J. F., and Stadler, P. F. (2002). Fractal geometry of spin-glass models. *J. Phys. A Math. Gen.* 35, 1509. doi: 10.1088/0305-4470/35/7/303
- Fox, M. D., Snyder, A. Z., Vincent, J. L., Corbetta, M., van Essen, D. C., and Raichle, M. E. (2005). The human brain is intrinsically organized into dynamic, anticorrelated functional networks. *Proc. Natl. Acad. Sci. U.S.A.* 102, 9673–9678. doi: 10.1073/pnas.0504136102
- Ganmor, E., Segev, R., and Schneidman, E. (2011). Sparse low-order interaction network underlies a highly correlated and learnable neural population code. *Proc. Natl. Acad. Sci. U.S.A.* 108, 9679–9684. doi: 10.1073/pnas.1019641108
- Garstecki, P., Hoang, T. X., and Cieplak, M. (1999). Energy landscapes, supergraphs, and “folding funnels” in spin systems. *Phys. Rev. E Stat. Phys. Plasmas Fluids Relat. Interdiscip. Topics* 60, 3219–3226. doi: 10.1103/PhysRevE.60.3219
- Grecius, M. D., Krasnow, B., Reiss, A. L., and Menon, V. (2003). Functional connectivity in the resting brain: a network analysis of the default mode hypothesis. *Proc. Natl. Acad. Sci. U.S.A.* 100, 253–258. doi: 10.1073/pnas.0135058100
- Hertz, J., Krogh, A., and Palmer, R. G. (1991). *Introduction to the Theory of Neural Computation*. New York, NY: Addison Wesley Publishing Company.
- Honey, C. J., Kötter, R., Breakspear, M., and Sporns, O. (2007). Network structure of cerebral cortex shapes functional connectivity on multiple time scales. *Proc. Natl. Acad. Sci. U.S.A.* 104, 10240–10245. doi: 10.1073/pnas.0701519104
- Hopfield, J. J. (1982). Neural networks and physical systems with emergent collective computational abilities. *Proc. Natl. Acad. Sci. U.S.A.* 79, 2554–2558. doi: 10.1073/pnas.79.8.2554
- Hordijk, W., Fontanari, J. F., and Stadler, P. F. (2003). Shapes of tree representations of spin-glass landscapes. *J. Phys. A Math. Gen.* 36, 3671–3681. doi: 10.1088/0305-4470/36/13/302
- Hutchison, R. M., Womelsdorf, T., Gati, J. S., Everling, S., and Menon, R. S. (2013). Resting-state networks show dynamic functional connectivity in awake humans and anesthetized macaques. *Hum. Brain Mapp.* 34, 2154–2177. doi: 10.1002/hbm.22058
- Kiviniemi, V., Vire, T., Remes, J., Elseoud, A. A., Starck, T., Tervonen, O., et al. (2011). A sliding time-window ICA reveals spatial variability of the default mode network in time. *Brain Connect.* 1, 339–347. doi: 10.1089/brain.2011.0036
- Knierim, J. J., and Zhang, K. (2012). Attractor dynamics of spatially correlated neural activity in the limbic system. *Annu. Rev. Neurosci.* 35, 267–285. doi: 10.1146/annurev-neuro-062111-150351
- Leutgeb, J. K., Leutgeb, S., Treves, A., Meyer, R., Barnes, C. A., McNaughton, B. L., et al. (2005). Progressive transformation of hippocampal neuronal representations in “morphed” environments. *Neuron* 48, 345–358. doi: 10.1016/j.neuron.2005.09.007
- Nakagawa, T. T., Jirsa, V. K., Spiegler, A., McIntosh, A. R., and Deco, G. (2013). Bottom up modeling of the connectome: linking structure and function in the resting brain and their changes in aging. *Neuroimage* 80, 318–329. doi: 10.1016/j.neuroimage.2013.04.055
- Ohiorhenuan, I. E., Mechler, F., Purpura, K. P., Schmid, A. M., Hu, Q., and Victor, J. D. (2010). Sparse coding and high-order correlations in fine-scale cortical networks. *Nature* 466, 617–621. doi: 10.1038/nature09178
- Rabinovich, M., Huerta, R., and Laurent, G. (2008). Neuroscience: transient dynamics for neural processing. *Science* 321, 48–50. doi: 10.1126/science.1155564
- Rabinovich, M. I., and Varona, P. (2011). Robust transient dynamics and brain functions. *Front. Comput. Neurosci.* 5:24. doi: 10.3389/fncom.2011.00024
- Raichle, M. E., MacLeod, A. M., Snyder, A. Z., Powers, W. J., Gusnard, D. A., and Shulman, G. L. (2001). A default mode of brain function. *Proc. Natl. Acad. Sci. U.S.A.* 98, 676–682. doi: 10.1073/pnas.98.2.676
- Santos, G. S., Gireesh, E. D., Plenz, D., and Nakahara, H. (2010). Hierarchical interaction structure of neural activities in cortical slice cultures. *J. Neurosci.* 30, 8720–8733. doi: 10.1523/JNEUROSCI.6141-09.2010
- Schneidman, E., Berry, M. J., Segev, R., and Bialek, W. (2006). Weak pairwise correlations imply strongly correlated network states in a neural population. *Nature* 440, 1007–1012. doi: 10.1038/nature04701
- Shlens, J., Field, G. D., Gauthier, J. L., Grivich, M. I., Petrusca, D., Sher, A., et al. (2006). The structure of multi-neuron firing patterns in primate retina. *J. Neurosci.* 26, 8254–8266. doi: 10.1523/JNEUROSCI.1282-06.2006
- Stillinger, F. H., and Weber, T. A. (1982). Hidden structure in liquids. *Phys. Rev. A* 25, 978–989. doi: 10.1103/PhysRevA.25.978
- Stillinger, F. H., and Weber, T. A. (1984). Packing structures and transitions in liquids and solids. *Science* 225, 983–989. doi: 10.1126/science.225.4666.983
- Tang, A., Jackson, D., Hobbs, J., Chen, W., Smith, J. L., Patel, H., et al. (2008). A maximum entropy model applied to spatial and temporal correlations from cortical networks in vitro. *J. Neurosci.* 28, 505–518. doi: 10.1523/JNEUROSCI.3359-07.2008
- Uddin, L. Q., Kelly, A. M., Biswal, B. B., Xavier Castellanos, F., and Milham, M. P. (2009). Functional connectivity of default mode network components: correlation, anticorrelation, and causality. *Hum. Brain Mapp.* 30, 625–637. doi: 10.1002/hbm.20531
- Wales, D. J. (2010). Energy landscapes: some new horizons. *Curr. Opin. Struct. Biol.* 20, 3–10. doi: 10.1016/j.sbi.2009.12.011
- Wales, D. J., Miller, M. A., and Walsh, T. R. (1998). Archetypal energy landscapes. *Nature* 394, 758–760. doi: 10.1038/29487
- Wang, X. J. (2009). Attractor network models. *Encycl. Neurosci.* 1, 667–679. doi: 10.1016/B978-008045046-9.01397-8
- Watanabe, T., Hirose, S., Wada, H., Imai, Y., Machida, T., Shirouzu, I., et al. (2013). A pairwise maximum entropy model accurately describes resting-state human brain networks. *Nat. Commun.* 4, 1370. doi: 10.1038/ncomms2388
- Wills, T. J., Lever, C., Cacucci, F., Burgess, N., and O'Keefe, J. (2005). Attractor dynamics in the hippocampal representation of the local environment. *Science* 308, 873–876. doi: 10.1126/science.1108905

- Yu, S., Huang, D., Singer, W., and Nikolic, D. (2008). A small world of neuronal synchrony. *Cereb. Cortex* 18, 2891–2901. doi: 10.1093/cercor/bhn047
- Zhou, Q. (2011). Random walk over basins of attraction to construct Ising energy landscapes. *Phys. Rev. Lett.* 106:180602. doi: 10.1103/PhysRevLett.106.180602
- Zhou, Q., and Wong, W. H. (2009). Energy landscape of a spin-glass model: exploration and characterization. *Phys. Rev. E* 79:051117. doi: 10.1103/PhysRevE.79.051117

**Conflict of Interest Statement:** The authors declare that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.

Received: 20 November 2013; accepted: 28 January 2014; published online: 25 February 2014.

Citation: Watanabe T, Hirose S, Wada H, Imai Y, Machida T, Shirouzu I, Konishi S, Miyashita Y and Masuda N (2014) Energy landscapes of resting-state brain networks. *Front. Neuroinform.* 8:12. doi: 10.3389/fninf.2014.00012

This article was submitted to the journal *Frontiers in Neuroinformatics*.

Copyright © 2014 Watanabe, Hirose, Wada, Imai, Machida, Shirouzu, Konishi, Miyashita and Masuda. This is an open-access article distributed under the terms of the Creative Commons Attribution License (CC BY). The use, distribution or reproduction in other forums is permitted, provided the original author(s) or licensor are credited and that the original publication in this journal is cited, in accordance with accepted academic practice. No use, distribution or reproduction is permitted which does not comply with these terms.





# Isotropic non-white matter partial volume effects in constrained spherical deconvolution

Timo Roine<sup>1\*</sup>, Ben Jeurissen<sup>1</sup>, Daniele Perrone<sup>2</sup>, Jan Aelterman<sup>2</sup>, Alexander Leemans<sup>3</sup>, Wilfried Philips<sup>2</sup> and Jan Sijbers<sup>1</sup>

<sup>1</sup> iMinds-Vision Lab, Department of Physics, University of Antwerp, Antwerp, Belgium

<sup>2</sup> Ghent University-iMinds/Image Processing and Interpretation, Ghent, Belgium

<sup>3</sup> Image Sciences Institute, University Medical Center Utrecht, Utrecht, Netherlands

## Edited by:

Daniele Marinazzo, University of Gent, Belgium

## Reviewed by:

Fang-Cheng Yeh, Carnegie Mellon University, USA

Shawna Farquharson, The Florey Institute of Neuroscience and Mental Health, Australia

## \*Correspondence:

Timo Roine, iMinds-Vision Lab, Department of Physics, University of Antwerp, Universiteitsplein 1, Building N, 2610 Wilrijk, Antwerp, Belgium  
e-mail: timo.roine@uantwerpen.be

Diffusion-weighted (DW) magnetic resonance imaging (MRI) is a non-invasive imaging method, which can be used to investigate neural tracts in the white matter (WM) of the brain. Significant partial volume effects (PVEs) are present in the DW signal due to relatively large voxel sizes. These PVEs can be caused by both non-WM tissue, such as gray matter (GM) and cerebrospinal fluid (CSF), and by multiple non-parallel WM fiber populations. High angular resolution diffusion imaging (HARDI) methods have been developed to correctly characterize complex WM fiber configurations, but to date, many of the HARDI methods do not account for non-WM PVEs. In this work, we investigated the isotropic PVEs caused by non-WM tissue in WM voxels on fiber orientations extracted with constrained spherical deconvolution (CSD). Experiments were performed on simulated and real DW-MRI data. In particular, simulations were performed to demonstrate the effects of varying the diffusion weightings, signal-to-noise ratios (SNRs), fiber configurations, and tissue fractions. Our results show that the presence of non-WM tissue signal causes a decrease in the precision of the detected fiber orientations and an increase in the detection of false peaks in CSD. We estimated 35–50% of WM voxels to be affected by non-WM PVEs. For HARDI sequences, which typically have a relatively high degree of diffusion weighting, these adverse effects are most pronounced in voxels with GM PVEs. The non-WM PVEs become severe with 50% GM volume for maximum spherical harmonics orders of 8 and below, and already with 25% GM volume for higher orders. In addition, a low diffusion weighting or SNR increases the effects. The non-WM PVEs may cause problems in connectomics, where reliable fiber tracking at the WM–GM interface is especially important. We suggest acquiring data with high diffusion-weighting 2500–3000 s/mm<sup>2</sup>, reasonable SNR (~30) and using lower SH orders in GM contaminated regions to minimize the non-WM PVEs in CSD.

**Keywords:** diffusion MRI, fiber orientation, partial volume effect, constrained spherical deconvolution, gray matter

## INTRODUCTION

Diffusion-weighted (DW) magnetic resonance imaging (MRI) is a non-invasive imaging method to investigate tissue microstructure via the measurement of the displacement of water molecules (Stejskal and Tanner, 1965; Jones, 2010). Diffusion in white matter (WM) neural tracts is anisotropic: it is larger parallel to the tract than in the perpendicular direction. In liquid, such as cerebrospinal fluid (CSF), diffusion is isotropic, i.e., equal in all directions. This diffusion property can be exploited to extract fiber orientations from DW data and investigate neural tracts in the brain WM using fiber tractography algorithms (Conturo et al., 1999; Basser et al., 2000; Mori and van Zijl, 2002; Jones, 2008; Tournier et al., 2010; Jeurissen et al., 2011).

The image resolution in DW-MRI is typically about 2–3 mm in all directions. Thus, significant partial volume effects (PVEs) are present in the measured signal (Alexander et al., 2001; Vos et al., 2011). These may be caused by multiple non-parallel neural tracts passing through a voxel (Vos et al., 2011; Jeurissen et al., 2013),

or several tissue types present in a voxel (Pasternak et al., 2009; Metzler-Baddeley et al., 2012a).

Currently, the most common method in the analysis of DW-MRI data is diffusion tensor imaging (DTI; Basser et al., 1994a,b; Jones and Leemans, 2011; Tournier et al., 2011). The shortcoming of DTI is the inability to identify complex fiber configurations consisting of multiple fiber orientations (Alexander et al., 2001; Frank, 2001, 2002), present in 60–90% of WM voxels (Jeurissen et al., 2013). To overcome this, high angular resolution diffusion imaging (HARDI) methods (Tuch et al., 2002; Jansons and Alexander, 2003; Tournier et al., 2004, 2007; Tuch, 2004; Dell'Acqua et al., 2007; Descoteaux et al., 2007; Behrens et al., 2007) and methods based on diffusion spectrum imaging (DSI; Wedeen et al., 2005, 2008) have been developed. However, although able to identify complex fiber configurations, most of the HARDI methods do not account for PVEs caused by non-WM tissue, such as gray matter (GM) and CSF (Dell'Acqua et al., 2010; Metzler-Baddeley et al., 2012a).

The presence of non-WM PVEs is known in DW-MRI (Alexander et al., 2001; Pasternak et al., 2009; Dell'Acqua et al., 2010; Metzler-Baddeley et al., 2012a), but their effects in HARDI methods have not been widely studied. Diffusion in non-WM tissue is mostly isotropic within the resolution of DW-MRI (Dell'Acqua et al., 2010). Isotropic non-WM PVEs have been shown to affect DTI (Alexander et al., 2001; Pasternak et al., 2009) and tensor-derived measures (Metzler-Baddeley et al., 2012a). Pasternak and coworkers used constrained optimization of a bi-tensor model for “free water elimination” (FWE) in DTI (Pasternak et al., 2009), but they did not investigate GM PVEs. Metzler-Baddeley and coworkers used FWE to correct for CSF-contamination in tensor-derived measures in constrained spherical deconvolution (CSD) based tractography (Metzler-Baddeley et al., 2012a). Both fractional anisotropy (FA) and mean diffusivity (MD) were shown to increase in the presence of CSF-contamination (Pasternak et al., 2009). Moreover, diffusivity metrics were shown to be more sensitive to PVEs than anisotropy metrics (Metzler-Baddeley et al., 2012a). However, FWE-based approaches are not suitable for GM-contaminated regions.

In HARDI methods, very few studies account for the non-WM PVEs. The “ball and stick” model is the only method, which initially included an isotropic compartment and could be extended into multiple fiber orientations (Behrens et al., 2003, 2007; Jbabdi et al., 2012). In another study involving HARDI methods, isotropic PVEs were dampened by using adaptive regularization in the iterative Richardson–Lucy deconvolution algorithm (Dell'Acqua et al., 2010). Other methods that also account for isotropic compartments include diffusion basis spectrum imaging (Wang et al., 2011) and diffusion decomposition (Yeh et al., 2011; Yeh and Tseng, 2013). However, the non-WM PVEs are not taken into account and have not been studied earlier in CSD, one of the most popular, clinically feasible and readily available HARDI methods (Leemans et al., 2009; Tournier et al., 2012).

In this work, we perform simulations to assess non-WM PVEs in CSD (Tournier et al., 2004, 2007). This kind of comprehensive analysis has not been performed before, although the method is widely used and the consequences may be significant when studying the connectivity between GM regions. In addition, we analyze the proportion of voxels affected by isotropic PVEs, and present the fiber orientation distribution functions (fODFs) estimated with CSD in real data affected by non-WM PVEs.

## MATERIALS AND METHODS

We investigated the isotropic PVEs caused by non-WM tissue on fODFs estimated with CSD. DW signals were simulated with varying diffusion weightings, signal-to-noise ratios, fiber configurations, and tissue fractions. In addition, experiments with real data were performed.

### ESTIMATION OF FIBER ORIENTATIONS WITH CONSTRAINED SPHERICAL DECONVOLUTION

In CSD, the full fODF is deconvolved from the DW signal using a kernel constructed from the single-fiber response function (RF), which can be estimated from the data (Tournier et al., 2004; Tax et al., 2014). During the deconvolution procedure, constraints are

imposed to suppress the negative peaks in the fODF (Tournier et al., 2007, 2008). The number of distinct gradient directions limits the maximum order of the spherical harmonics (SH) decomposition, which can be used in the estimation in the fODF. However, the constraints used to suppress the negative peaks in the fODF can be exploited to estimate higher order solutions and thus, describe more complex fODFs. This is called super-resolved CSD (Tournier et al., 2007).

To find the peaks of the fODF estimated with CSD, a Newton optimization algorithm was used to extract the local maxima of the fODF directly based on the SH decompositions (Jeurissen et al., 2013). Optimization was initialized on a dense set of uniformly distributed spherical sample points. A threshold of 33% of the maximum amplitude of the fODF was used to discard small peaks. A maximum of six of the highest peaks were identified. The peaks were clustered around the peaks of the average fODF calculated over all simulation repetitions performed with the same parameter configuration. Peaks further away than half of the crossing angle (with an upper limit of 35°) from any of the peaks of the average fODF were not included in the clusters. A mean dyadic tensor was then used to derive the mean orientation for each of the identified fiber clusters (Basser and Pajevic, 2000; Jones, 2003). This orientation was then compared to the true orientations of the fiber bundles. Peaks in clusters that were less than half of the crossing angle (with an upper limit of 35°) from the true orientations were defined as true, and rest of the peaks, also if they were not assigned to a cluster, as false. From the true clusters, accuracy and precision (95th percentile confidence interval, CI) with respect to the orientation of the mean dyadic tensor were calculated.

### SIMULATION OF THE DW SIGNAL WITH PVES

Two crossing WM fiber configurations were simulated with equal weights. The orientation of the first fiber bundle was randomly selected, after which the orientation of the second fiber bundle was calculated in spherical coordinates with the defined crossing angle. The resulting angle was verified to be correct in each case.

Then, the DW signal was simulated separately for different tissue types, and the resulting signals were combined assuming no exchange between the compartments (Leemans et al., 2005). The number of gradient directions uniformly distributed on the unit hemisphere was 64 (Jones et al., 1999). To eliminate any bias caused by the gradient orientations, a different gradient set was used for each simulated DW signal. Signal from the specific WM fiber configurations was combined with isotropic CSF and GM compartments. In addition, air compartments were simulated to investigate only the effect of reduced signal of the WM compartment without any isotropic diffusion. Derived based on Basser and Jones (2002), the combined simulated DW signal  $S$  is:

$$S = (1 - f_{\text{isot}}) \left( f_{\text{fiber}} e^{-\text{Trace}(\mathbf{bD}_{\text{fiber}1})} + (1 - f_{\text{fiber}}) e^{-\text{Trace}(\mathbf{bD}_{\text{fiber}2})} \right) + f_{\text{isot}} e^{-\text{Trace}(\mathbf{bD}_{\text{isot}})}, \quad (1)$$

where  $f_{\text{isot}}$  is the fraction of isotropic volume,  $f_{\text{fiber}}$  is the fraction of the first fiber compartment with respect to the WM compartment,  $\mathbf{b}$  is the b-matrix summarizing the attenuation in all three

directions of the diffusion tensor (including information of the diffusion-weighting and the gradient orientations; Mattiello et al., 1997), and  $\mathbf{D}_{\text{fiber1}}$ ,  $\mathbf{D}_{\text{fiber2}}$  and  $\mathbf{D}_{\text{isot}}$  are the diffusion tensors of the two WM fibers and the isotropic compartment respectively (Basser et al., 1994b). The diffusion tensors were created with the following values. The MD for the simulation of different tissue types was  $0.002 \text{ mm}^2/\text{s}$  for CSF,  $0.0007 \text{ mm}^2/\text{s}$  for WM and GM (Dell'Acqua et al., 2010), and for air the signal was assumed to be zero. The FA was 0.8 for the WM signal and 0 for other tissue types. Rician noise was added to the combined DW signal. Finally, the DW signals were decomposed into an eighth-order series of SH (maximum possible order with the number of gradient orientations used).

## SIMULATION EXPERIMENTS

We performed simulation experiments to investigate the PVEs with different tissue compartments. Simulations and analyses of the simulation experiments were performed in Matlab (The MathWorks, Inc., Natick, MA, USA), by using dedicated software programmed by the authors.

The fraction of isotropic GM, CSF or air volume was varied from 0.00 to 0.95 with intervals of 0.05. We analyzed angles between the fiber populations in configurations ranging from  $40^\circ$  to  $90^\circ$  and with diffusion weightings ( $b$ -values) from 1000 to  $3500 \text{ s/mm}^2$ . Signal-to-noise ratio (SNR) was calculated with respect to the non-diffusion weighted signal and simulated from 10 to 60, also generating a noiseless version of the DW signal. We performed 1000 repetitions with different noise realizations (resulting in Rician distributed data) for each parameter configuration. The fODFs were estimated from the simulated DW signals with CSD or super-resolved CSD using maximum orders of the SH from 4 to 14.

In addition to the isotropic volume fraction (VF) and PVE type (GM, CSF, or air) only one parameter at a time was investigated. The default values for the non-varying parameters were:  $b$ -value:  $3000 \text{ s/mm}^2$ ; angle between the crossing fiber configurations:  $70^\circ$ ; SNR: 30. The default maximum order of the SH was 8 for CSD and 12 for super-resolved CSD.

## ACQUISITION AND ANALYSIS OF REAL DATA

High angular resolution DW data were acquired on a 3T MRI system with a 32-channel head coil. The subject gave written informed consent to participate in this study under a protocol approved by the local ethics committee. A single-shot echo-planar imaging (EPI) sequence was used with  $\text{TR} = 8100 \text{ ms}$ ,  $\text{TE} = 116 \text{ ms}$  and  $2.5 \text{ mm} \times 2.5 \text{ mm} \times 2.5 \text{ mm}$  voxel size. The field of view (FOV) was  $240 \times 240 \text{ mm}^2$  with a  $96 \times 96$  acquisition matrix and the number of excitations (NEX) was 1. Fifty-four axial slices were imaged with 2.5 mm thickness and no gap. Diffusion sensitizing gradients with a  $b$ -value of  $2800 \text{ s/mm}^2$  were applied along 75 non-collinear directions. Ten images without diffusion weighting ( $b = 0 \text{ s/mm}^2$ ) were acquired, of which one was acquired with reverse phase-encoding, for the purpose of EPI distortion correction. High-resolution anatomical T1-weighted images were acquired using a 3D magnetization-prepared rapid gradient-echo (MPRAGE) sequence (Mugler and Brookeman, 1990) with  $\text{TR} = 1900 \text{ ms}$ ,  $\text{TE} = 2.52 \text{ ms}$ ,  $\text{TI} = 900 \text{ ms}$  and

$1 \text{ mm} \times 1 \text{ mm} \times 1 \text{ mm}$  voxel size (flip angle =  $9^\circ$  and NEX = 1). FOV was  $250 \text{ mm} \times 250 \text{ mm} \times 176 \text{ mm}$  with a  $256 \times 256 \times 176$  acquisition matrix.

The DW data were corrected for subject motion and eddy current induced distortions (Leemans and Jones, 2009; Andersson et al., 2012), and TOPUP was used to correct for EPI distortions (Andersson et al., 2003). The MRtrix package (J-D Tournier, Brain Research Institute, Melbourne, Australia, <http://www.brain.org.au/software/>; Tournier et al., 2012) was used for visualization of the real data. Tissue VFs for the DW data were estimated from the T1-weighted images, using a similar approach as presented by Smith et al. (2012).

The percentage of WM voxels affected by significant non-WM PVEs was estimated from real data. WM voxels were defined using a threshold of 25% WM tissue. The voxels with PVEs were estimated by using two threshold values: 25 and 10% non-WM volume.

## RESULTS

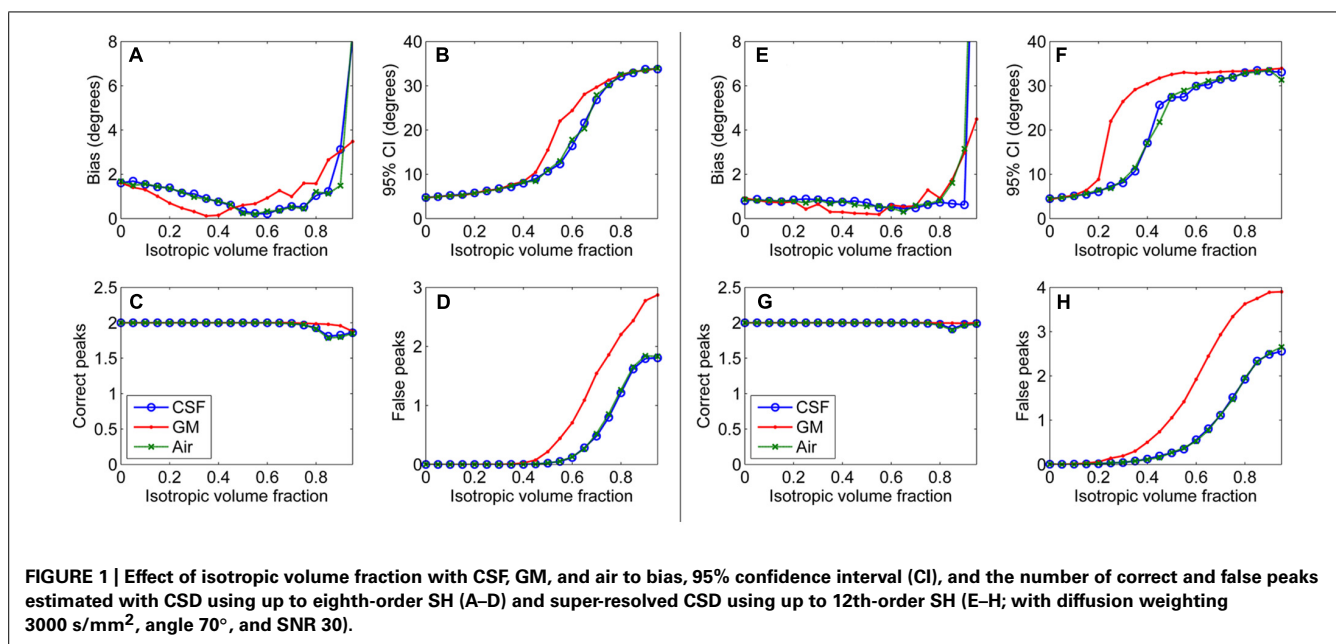
First, results of the simulation experiments are presented. **Figure 1** shows the effects of isotropic non-WM VF in CSD (**Figures 1A–D**) and super-resolved CSD (**Figures 1E,H**). The bias and the 95% CI of the fiber orientations extracted with CSD are presented in **Figures 1A,B**. We also studied the effects on the number of correctly and falsely identified peaks (**Figures 1C,D**). The number of falsely identified peaks increased and the precision of the identified fiber orientations decreased, when the isotropic VF increased. The effects were stronger in GM than in CSF or air. However, the accuracy of the identified fiber orientations and the number of true peaks identified did not change until very high non-WM fractions. The similar performance with CSF and air PVEs using a high  $b$ -value indicates that the effect in CSF is mostly an SNR effect, which is clearly not the case in GM. Default values were used for the other parameters as specified in the methods section. The non-WM PVEs in super-resolved CSD, using up to 12th-order SH, started to affect the precision and the number of false peaks (**Figures 1F,H**) with lower fractions than when using up to eighth-order SH. Accuracy remained high and was similar to the results when using SH up to eighth-order. However, the ability to detect the two correct fiber orientations stayed higher with high isotropic fractions than while using SH up to eighth order.

An illustration of the estimated fODFs based on only one noise realization per fraction is shown in **Figure 2**. The false peaks became more numerous and the correct peaks lost precision, when the isotropic VF increased. Next, the effects of varying maximum SH orders, diffusion weightings, SNRs, and angles between the two crossing fiber configurations were analyzed, while keeping the isotropic non-WM fraction constant at 0.5.

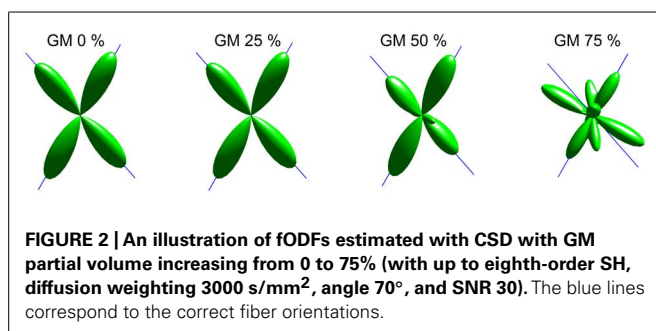
In **Figure 3**, the effects of maximum SH order on the non-WM PVEs are shown. The 95% CI and the number of false peaks increased when higher maximum SH orders were used. Bias was low with all orders except for the lowest maximum order 4 with GM PVEs. However, the correct peaks could be found properly with GM PVEs even with the lowest order, but not with CSF or air PVEs.

In **Figure 4**, the effects of varying diffusion weightings to the non-WM PVEs are shown. The 95% CI and the number of false





**FIGURE 1 |** Effect of isotropic volume fraction with CSF, GM, and air to bias, 95% confidence interval (CI), and the number of correct and false peaks estimated with CSD using up to eighth-order SH (A–D) and super-resolved CSD using up to 12th-order SH (E–H; with diffusion weighting 3000 s/mm<sup>2</sup>, angle 70°, and SNR 30).



**FIGURE 2 |** An illustration of fODFs estimated with CSD with GM partial volume increasing from 0 to 75% (with up to eighth-order SH, diffusion weighting 3000 s/mm<sup>2</sup>, angle 70°, and SNR 30). The blue lines correspond to the correct fiber orientations.

peaks increased when diffusion weighting decreased. The number of false peaks was high and the precision of the correct peaks was low under GM PVEs compared to CSF, air, or 100% WM regardless of the diffusion weighting. The difference between CSF and air PVEs was visible only with very low diffusion weightings of 1000–1500 s/mm<sup>2</sup>.

The effects of SNR on the non-WM PVEs are presented in **Figure 5**. **Figures 5A–D** show effects with 50% non-WM fractions and **Figures 5E–H** with 75% non-WM fractions. With 50% PVEs, increasing SNR improved the precision and reduced the number of false peaks identified. However, with 75% non-WM fractions, increasing SNR could not improve the situation with GM PVEs, and there were problems with precision also with high SNRs under CSF PVEs.

**Figure 6** shows the effects of varying angle between the two crossing fiber configurations. With an angle of 40° between the two fiber configurations, the correct peaks could not be properly identified. However, with an angle of 50°, they could still be reliably detected without isotropic PVEs, but any type of non-WM volume caused a decrease in the fraction of the correct peaks identified (**Figure 6C**). With higher angles, the correct peaks were identified correctly and without more bias than in pure WM (**Figure 6A**).

The precision of the identified fiber orientations and the number of false peaks identified improved when the angle between the fiber configurations increased (**Figures 6B,D**).

From real data, we estimated that 35.7% of WM voxels, defined to have at least 25% WM volume, had significant PVEs with non-WM tissue, also defined to be more than 25% VF. Lowering the non-WM tissue threshold to 10%, the proportion of WM voxels affected by PVEs increased to 46.8%. Of these voxels with non-WM PVEs, 96.0% were affected by PVEs with GM and 5.3% with CSF.

**Figure 7** shows the fODFs estimated with CSD, using up to eight order SH, from real data overlaid on the WM tissue probability map of corona radiata extending towards cortical GM. The areas where WM interfaces with GM were affected both with CSD and super-resolved CSD. A large amount of voxels in the area had significant PVEs (gray-colored voxels), and perpendicular or spurious peaks appeared in the voxels with no apparent anatomical origin.

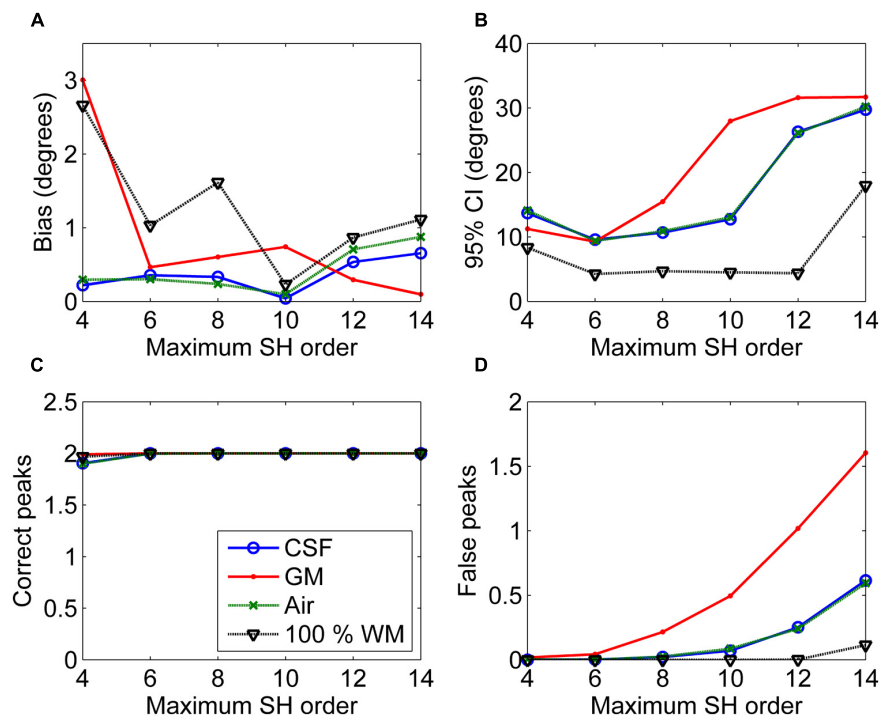
**Figure 8** shows the effect of CSF PVEs on the estimation of fODFs at the interface between the corpus callosum and CSF. Spurious orientations can be noticed, but they are much smaller in amplitude and the principal fiber orientation can still be clearly distinguished.

## DISCUSSION

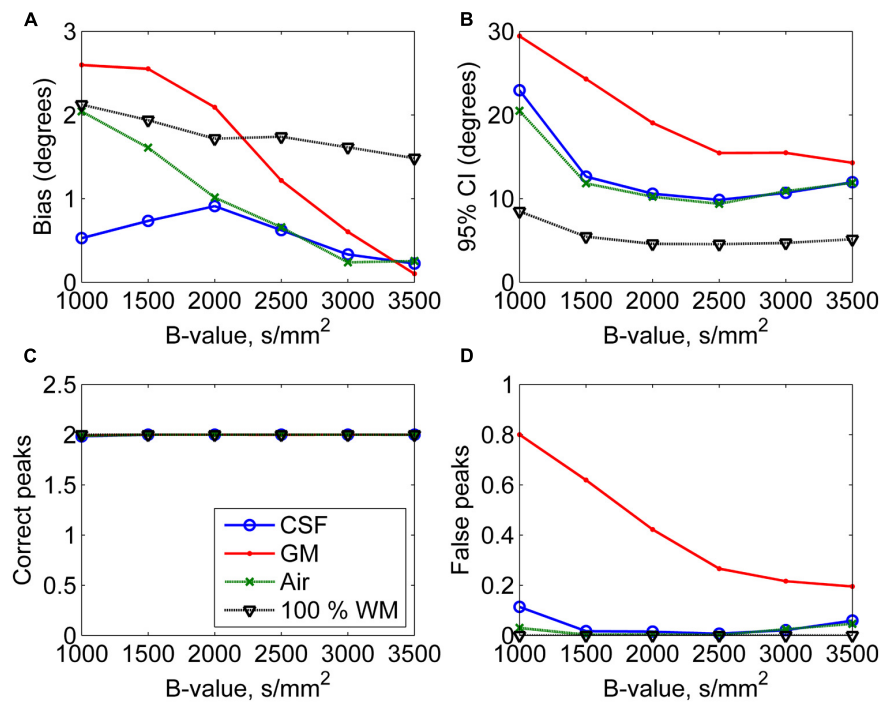
We studied the effects of isotropic non-WM partial volume on the fiber orientations estimated with CSD and super-resolved CSD by performing extensive simulations and real data experiments. CSD is a widely used method and knowledge about the implications of non-WM PVEs should be augmented.

Our results demonstrate that although CSD is efficient in the detection of PVEs caused by complex fiber configurations within a voxel, problems arise in the detection of the fODFs in the case of non-WM PVEs, which we estimated to be present in 35–50% of

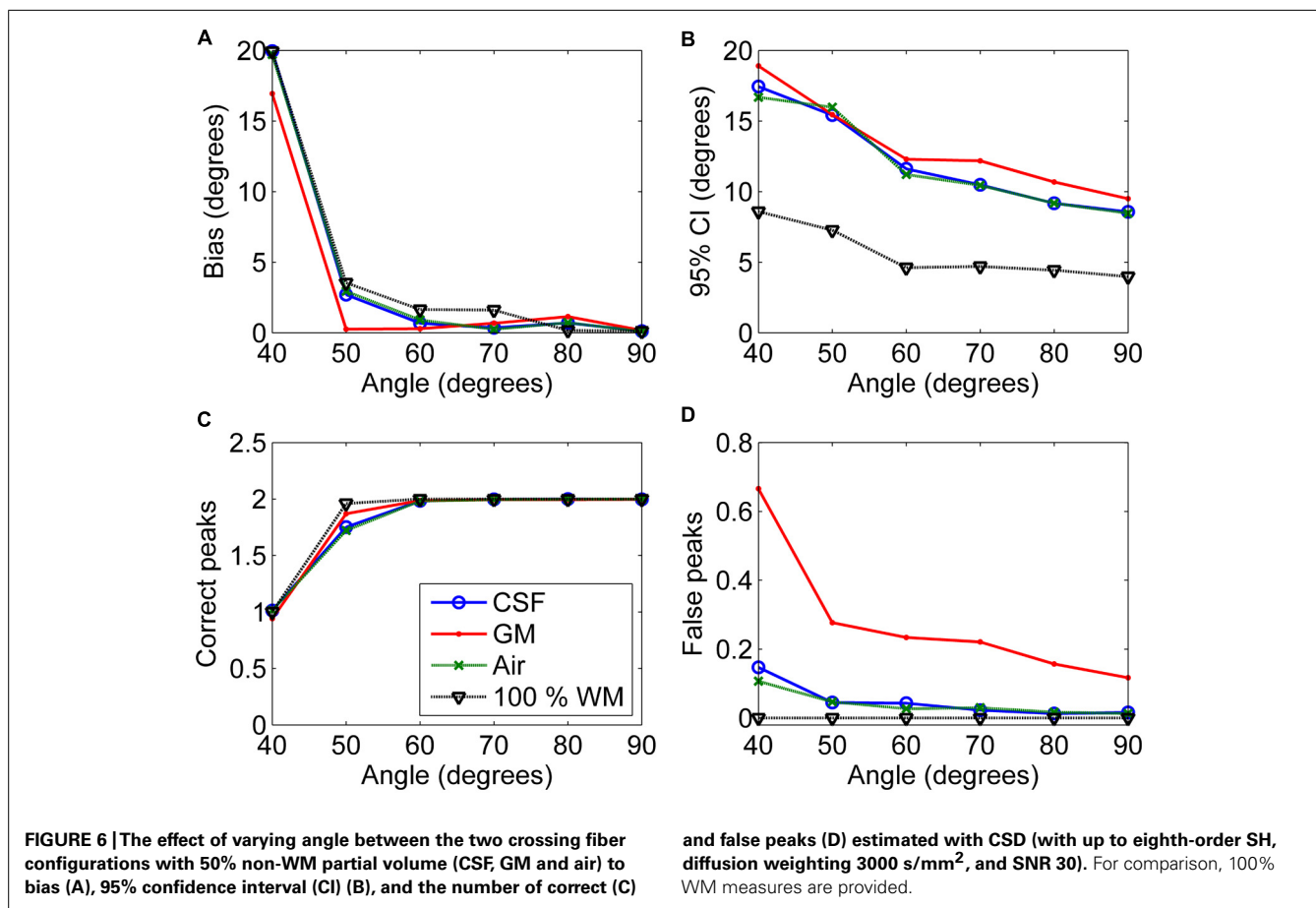
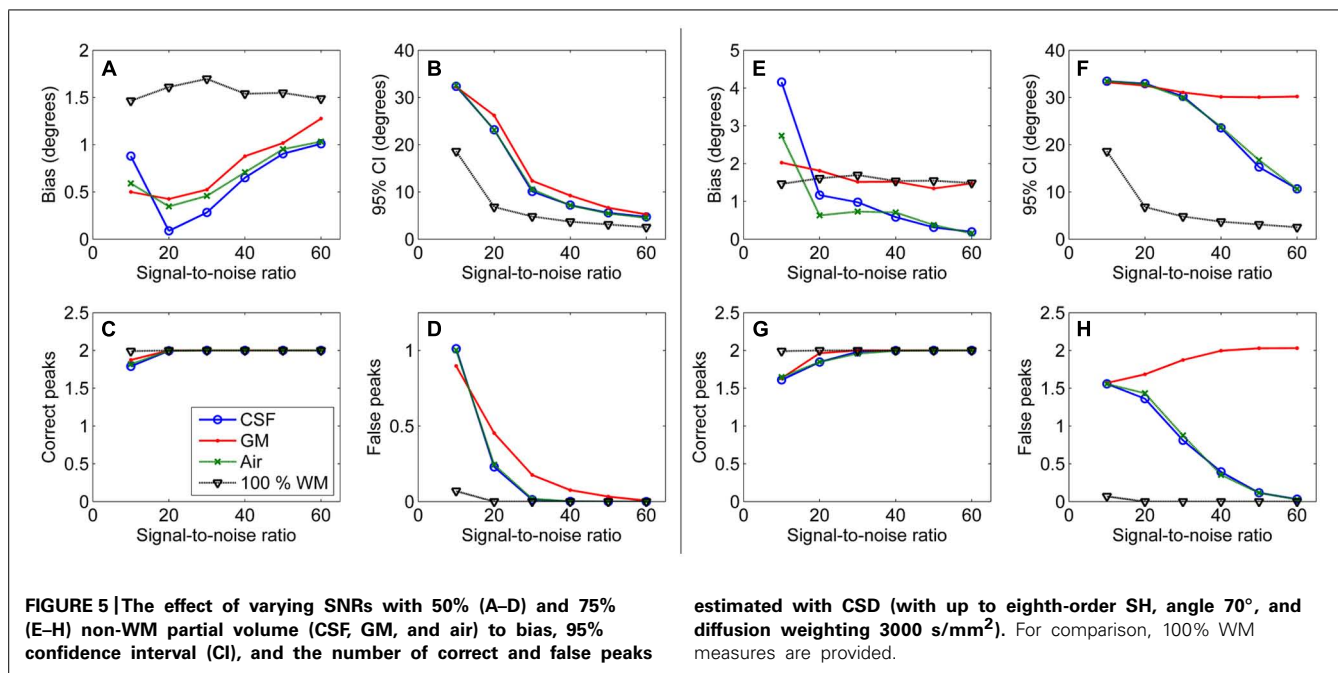


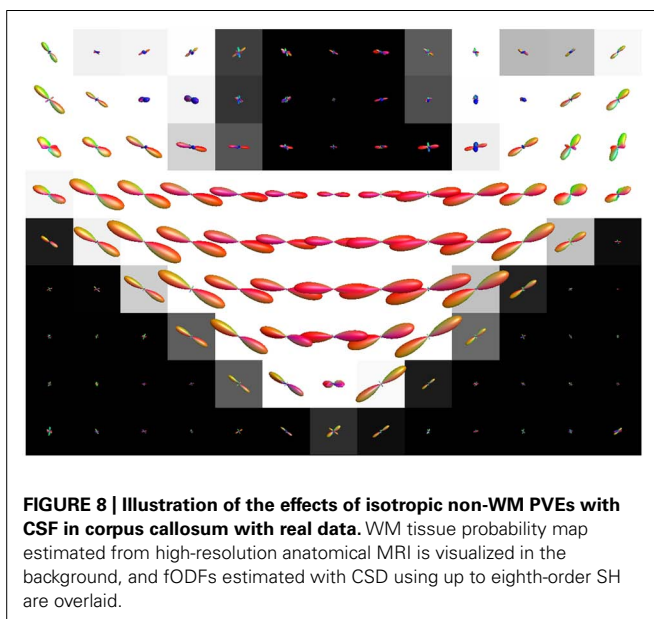
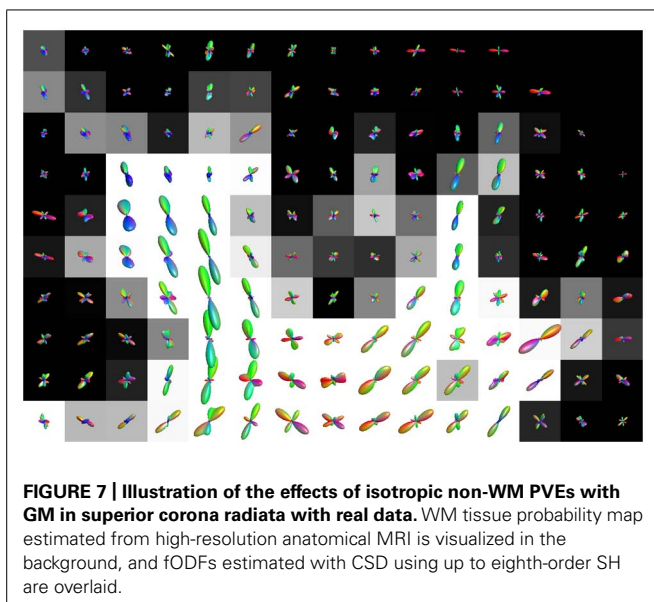


**FIGURE 3 |** The effect of varying maximum SH orders with 50% non-WM partial volume (CSF, GM, and air) to bias (A), 95% confidence interval (CI) (B), and the number of correct (C) and false peaks (D) estimated with CSD (with diffusion weighting 3000 s/mm<sup>2</sup>, angle 70°, and SNR 30). For comparison, 100% WM measures are provided.



**FIGURE 4 |** The effect of varying diffusion weightings with 50% non-WM partial volume (CSF, GM, and air) to bias (A), 95% confidence interval (CI) (B), and the number of correct (C) and false peaks (D) estimated with CSD (with up to eighth-order SH, angle 70°, and SNR 30). For comparison, 100% WM measures are provided.





the WM voxels. As shown in **Figure 1**, the precision of the detected fiber orientations decreases and false peaks appear in the fODFs. This effect is most prominent for GM PVEs. The increase in bias with very high isotropic fractions may be at least partly caused by the inability to distinguish reliably both of the correct fiber orientations.

Part of these effects is due to the reduction of relative SNR in the WM tissue, which is caused by the decreased WM volume in the voxel, and not the isotropic diffusion properties of the non-WM tissue. PVEs with CSF tissue are mostly due to this effect, as shown by the similarity to air PVEs (**Figure 1**). Another part of the effects is caused by the isotropic diffusion, which invalidates the single fiber RF originally designed for pure WM. The more prominent PVEs in GM than in CSF or air are caused by this effect.

In addition, we showed that the PVEs increased when the maximum SH order increased. Therefore, the high maximum SH orders, although able to improve the angular resolution (Tournier et al., 2007), should be used with caution in the estimation of the fODFs under significant non-WM PVEs. Although the maximum angular frequency in the DW data is relatively low (Tournier et al., 2013), the fODFs contain higher angular frequencies, so a higher maximum SH order could still be useful in the estimation of the fODFs within pure WM regions. The use of lower diffusion weighting than generally used in HARDI sequences (i.e., less than 3000 s/mm<sup>2</sup>) increased the PVEs. Larger crossing angles could be detected with higher precision. With higher SNRs, moderate PVEs could be handled better, but high PVEs continued to decrease precision and increase the number of false peaks especially in GM.

Based on these results, we provide the following advice on how to operate CSD to maintain reasonable precision and number of false peaks under non-WM PVEs. Conditions with 95% CI lower than 20° and less than one identified false peak were considered reasonable. Thus, we suggest acquiring data with a high diffusion-weighting 2500–3000 s/mm<sup>2</sup>, and a reasonable SNR (~25–30). To extract the fiber orientations with CSD in regions with GM PVEs, we suggest using relatively low, from 6 to 8, maximum SH orders to minimize the loss in precision and the increase in the number of identified false fiber orientations. Nevertheless, the identified fiber orientations should be considered unreliable with higher than 60% GM and higher than 80% CSF VFs.

The isotropic PVEs, present in a significant proportion of WM voxels, lead to decreased precision and a high number of false peaks in the fODFs estimated with CSD, which in turn affects subsequent tractography algorithms, and may introduce false positives and hinder tract propagation into the cortex or near subcortical GM tissue. An algorithm already exists to discard tracts based on their anatomical feasibility and thus, only accept tracts that correctly propagate to the cortex (Smith et al., 2012). However, enabling the tracts to propagate properly into the cortex or adjacent to subcortical GM tissue would reduce the time needed for tracking and improve the precision of anatomically feasible tracts. Especially in connectomics, where reliable reconstruction of the fiber orientations profiles at the GM–WM interface is required to compute connectivity matrices, taking isotropic PVEs into account will be valuable.

Limitations of this study include the restriction to only one HARDI method, although it is one of the most commonly used ones (Metzler-Baddeley et al., 2012b; Emsell et al., 2013; Forde et al., 2013; Kristo et al., 2013; McGrath et al., 2013a,b; Reijmer et al., 2013a,b; Thompson et al., 2014). Previous studies indicate that the non-WM PVEs are present in DW-MRI in general (Alexander et al., 2001; Pasternak et al., 2009; Dell'Acqua et al., 2010; Metzler-Baddeley et al., 2012a). While some of the analysis methods already acknowledge or account for these PVEs (Behrens et al., 2003, 2007; Pasternak et al., 2009; Dell'Acqua et al., 2010; Wang et al., 2011; Yeh et al., 2011; Jbabdi et al., 2012; Yeh and Tseng, 2013), many of the currently used methods do not. For example, in CSD they have not yet been taken into account, and no detailed investigation about these effects had been performed previously. It is likely that also other methods which do

not appropriately account for the non-WM PVEs will suffer from similar consequences. An additional limitation of this study is that there is no ground truth available in real data. Considering the clear effects demonstrated in the simulations, it is reasonable to assume that the spurious fiber orientations visible at the tissue interfaces are in fact false peaks also in real data. However, further experiments with real data are still necessary to completely understand the phenomenon and its effects in tractography. This would in turn help in the development of improvements for the fODF estimation with CSD, applicable also in real data, and thus allow improved tracking especially in the WM–GM interface.

In conclusion, we studied the effects of isotropic non-WM PVEs in CSD and found decreased precision and increased number of false peaks in the estimated fODFs. The effect was more pronounced with GM tissue. Considering the clear effects present in real and simulated data and the large proportion of WM voxels affected, it is important to take the non-WM PVEs into account in the extraction of fiber orientations with CSD. Therefore, we provide simple recommendations for the parameters used in the acquisition and the analysis, but acknowledge the need for more sophisticated methods to account for non-WM tissue in CSD.

## ACKNOWLEDGMENTS

This work was supported by the Fund for Scientific Research–Flanders (FWO), and by the Interuniversity Attraction Poles Program (P7/11) initiated by the Belgian Science Policy Office.

## REFERENCES

- Alexander, A. L., Hasan, K. M., Lazar, M., Tsuruda, J. S., and Parker, D. L. (2001). Analysis of partial volume effects in diffusion-tensor MRI. *Magn. Reson. Med.* 45, 770–780. doi: 10.1002/mrm.1105
- Andersson, J. L., Skare, S., and Ashburner, J. (2003). How to correct susceptibility distortions in spin-echo echo-planar images: application to diffusion tensor imaging. *Neuroimage* 20, 870–888. doi: 10.1016/S1053-8119(03)00336-7
- Andersson, J. L. R., Xu, J., Yacoub, E., Auerbach, E., Moeller, S., and Ugurbil, K. (2012). “A comprehensive Gaussian process framework for correcting distortions and movements in diffusion images,” in *Proceedings of the 20th Annual Meeting of ISMRM*, Melbourne, 2426.
- Basser, P. J., and Jones, D. K. (2002). Diffusion-tensor MRI: theory, experimental design and data analysis: a technical review. *NMR Biomed.* 15, 456–467. doi: 10.1002/nbm.783
- Basser, P. J., Mattiello, J., and LeBihan, D. (1994a). MR diffusion tensor spectroscopy and imaging. *Biophys. J.* 66, 259–267. doi: 10.1016/S0006-3495(94)80775-1
- Basser, P. J., Mattiello, J., and LeBihan, D. (1994b). Estimation of the effective self-diffusion tensor from the NMR spin echo. *J. Magn. Res. B* 103, 247–254. doi: 10.1006/jmrb.1994.1037
- Basser, P. J., and Pajevic, S. (2000). Statistical artifacts in diffusion tensor MRI (DT-MRI) caused by background noise. *Magn. Reson. Med.* 44, 41–50. doi: 10.1002/1522-2594(200007)44:1<41::AID-MRM8>3.0.CO;2-O
- Basser, P. J., Pajevic, S., Pierpaoli, C., Duda, J., and Aldroubi, A. (2000). In vivo fiber tractography using DT-MRI data. *Magn. Reson. Med.* 44, 625–632. doi: 10.1002/1522-2594(200010)44:4<625::AID-MRM17>3.0.CO;2-O
- Behrens, T. E. J., Berg, H. J., Jbabdi, S., Rushworth, M. F. S., and Woolrich, M. W. (2007). Probabilistic diffusion tractography with multiple fibre orientations: what can we gain? *Neuroimage* 34, 144–155. doi: 10.1016/j.neuroimage.2006.09.018
- Behrens, T. E. J., Woolrich, M. W., Jenkinson, M., Johansen-Berg, H., Nunes, R. G., Clare, S., et al. (2003). Characterization and propagation of uncertainty in diffusion-weighted MR imaging. *Magn. Reson. Med.* 50, 1077–1088. doi: 10.1002/mrm.10609
- Conturo, T. E., Lori, N. F., Cull, T. S., Akbudak, E., Snyder, A. Z., Shimony, J. S., et al. (1999). Tracking neuronal fiber pathways in the living human brain. *Proc. Natl. Acad. Sci. U.S.A.* 96, 10422–10427. doi: 10.1073/pnas.96.18.10422
- Dell’Acqua, F., Rizzo, G., Scifo, P., Clarke, R. A., Scotti, G., and Fazio, F. (2007). A model-based deconvolution approach to solve fiber crossing in diffusion-weighted MR imaging. *IEEE Trans. Biomed. Eng.* 54, 462–472. doi: 10.1109/TBME.2006.888830
- Dell’Acqua, F., Scifo, P., Rizzo, G., Catani, M., Simmons, A., Scotti, G., et al. (2010). A modified damped Richardson–Lucy algorithm to reduce isotropic background effects in spherical deconvolution. *Neuroimage* 49, 1446–1458. doi: 10.1016/j.neuroimage.2009.09.033
- Descoteaux, M., Angelino, E., Fitzgibbons, S., and Deriche, R. (2007). Regularized, fast, and robust analytical Q-ball imaging. *Magn. Reson. Med.* 58, 497–510. doi: 10.1002/mrm.21277
- Emsell, L., Leemans, A., Langan, C., Van Hecke, W., Barker, G. J., McCarthy, P., et al. (2013). Limbic and callosal white matter changes in euthymic bipolar I disorder: an advanced diffusion magnetic resonance imaging tractography study. *Biol. Psychiatry* 73, 194–201. doi: 10.1016/j.biopsych.2012.09.023
- Forde, N. J., Ronan, L., Suckling, J., Scanlon, C., Neary, S., Holleran, L., et al. (2013). Structural neuroimaging correlates of allelic variation of the BDNF Val66met polymorphism. *Neuroimage* 90, 280–289. doi: 10.1016/j.neuroimage.2013.12.050
- Frank, L. R. (2001). Anisotropy in high angular resolution diffusion-weighted MRI. *Magn. Reson. Med.* 45, 935–939. doi: 10.1002/mrm.1125
- Frank, L. R. (2002). Characterization of anisotropy in high angular resolution diffusion-weighted MRI. *Magn. Reson. Med.* 47, 1083–1099. doi: 10.1002/mrm.10156
- Jansons, K. M., and Alexander, D. C. (2003). Persistent angular structure: new insights from diffusion magnetic resonance imaging data. *Inverse Problems* 19, 1031. doi: 10.1088/0266-5611/19/5/303
- Jbabdi, S., Sotiropoulos, S. N., Savio, A. M., Graña, M., and Behrens, T. E. (2012). Model-based analysis of multishell diffusion MR data for tractography: how to get over fitting problems. *Magn. Reson. Med.* 68, 1846–1855. doi: 10.1002/mrm.24204
- Jeurissen, B., Leemans, A., Jones, D. K., Tournier, J. D., and Sijbers, J. (2011). Probabilistic fiber tracking using the residual bootstrap with constrained spherical deconvolution. *Hum. Brain Mapp.* 32, 461–479. doi: 10.1002/hbm.21032
- Jeurissen, B., Leemans, A., Tournier, J. D., Jones, D. K., and Sijbers, J. (2013). Investigating the prevalence of complex fiber configurations in white matter tissue with diffusion magnetic resonance imaging. *Hum. Brain Mapp.* 34, 2747–2766. doi: 10.1002/hbm.22099
- Jones, D. K. (2003). Determining and visualizing uncertainty in estimates of fiber orientation from diffusion tensor MRI. *Magn. Reson. Med.* 49, 7–12. doi: 10.1002/mrm.10331
- Jones, D. K. (2008). Studying connections in the living human brain with diffusion MRI. *Cortex* 44, 936–952. doi: 10.1016/j.cortex.2008.05.002
- Jones, D. K. (ed). (2010). *Diffusion MRI: Theory, Methods, and Applications*. Oxford: Oxford University Press. doi: 10.1093/med/9780195369779.001.0001
- Jones, D. K., Horsfield, M. A., and Simmons, A. (1999). Optimal strategies for measuring diffusion in anisotropic systems by magnetic resonance imaging. *Magn. Reson. Med.* 42, 515–525. doi: 10.1002/(SICI)1522-2594(199909)42:3<515::AID-MRM14>3.0.CO;2-Q
- Jones, D. K., and Leemans, A. (2011). “Diffusion tensor imaging,” in *Magnetic Resonance Neuroimaging*, eds M. Modo and J. Bulte (New York: Humana Press), 127–144.
- Kristo, G., Leemans, A., Raemaekers, M., Rutten, G. J., Gelder, B., and Ramsey, N. E. (2013). Reliability of two clinically relevant fiber pathways reconstructed with constrained spherical deconvolution. *Magn. Reson. Med.* 70, 1544–1556. doi: 10.1002/mrm.24602
- Leemans, A., Jeurissen, B., Sijbers, J., and Jones, D. K. (2009). ExploreDTI: a graphical toolbox for processing, analyzing, and visualizing diffusion MR data. *Proc. Int. Soc. Mag. Reson. Med.* 17, 3536.
- Leemans, A., and Jones, D. K. (2009). The B-matrix must be rotated when correcting for subject motion in DTI data. *Magn. Reson. Med.* 61, 1336–1349. doi: 10.1002/mrm.21890
- Leemans, A., Sijbers, J., Verhoye, M., Van der Linden, A., and Van Dyck, D. (2005). Mathematical framework for simulating diffusion tensor MR neural fiber bundles. *Magn. Reson. Med.* 53, 944–953. doi: 10.1002/mrm.20418
- Mattiello, J., Basser, P. J., and Le Bihan, D. (1997). The b matrix in diffusion tensor echo-planar imaging. *Magn. Reson. Med.* 37, 292–300. doi: 10.1002/mrm.1910370226



- McGrath, J., Johnson, K., O'Hanlon, E., Garavan, H., Gallagher, L., and Leemans, A. (2013a). White matter and visuospatial processing in autism: a constrained spherical deconvolution tractography study. *Autism Res.* 6, 307–319. doi: 10.1002/aur.1290
- McGrath, J., Johnson, K., O'Hanlon, E., Garavan, H., Leemans, A., and Gallagher, L. (2013b). Abnormal functional connectivity during visuospatial processing is associated with disrupted organisation of white matter in autism. *Front. Hum. Neurosci.* 7:434. doi: 10.3389/fnhum.2013.00434
- Metzler-Baddeley, C., O'Sullivan, M. J., Bells, S., Pasternak, O., and Jones, D. K. (2012a). How and how not to correct for CSF-contamination in diffusion MRI. *Neuroimage* 59, 1394–1403. doi: 10.1016/j.neuroimage.2011.08.043
- Metzler-Baddeley, C., Hunt, S., Jones, D. K., Leemans, A., Aggleton, J. P., and O'Sullivan, M. J. (2012b). Temporal association tracts and the breakdown of episodic memory in mild cognitive impairment. *Neurology* 79, 2233–2240. doi: 10.1212/WNL.0b013e31827689e8
- Mori, S., and van Zijl, P. (2002). Fiber tracking: principles and strategies: a technical review. *NMR Biomed.* 15, 468–480. doi: 10.1002/nbm.781
- Mugler, J. P., and Brookeman, J. R. (1990). Three-dimensional magnetization-prepared rapid gradient-echo imaging (3D MP RAGE). *Magn. Reson. Med.* 15, 152–157. doi: 10.1002/mrm.1910150117
- Pasternak, O., Sochen, N., Gur, Y., Intrator, N., and Assaf, Y. (2009). Free water elimination and mapping from diffusion MRI. *Magn. Reson. Med.* 62, 717–730. doi: 10.1002/mrm.22055
- Reijmer, Y. D., Leemans, A., Brundel, M., Kappelle, L. J., and Biessels, G. J. (2013a). Disruption of the cerebral white matter network is related to slowing of information processing speed in patients with type 2 diabetes. *Diabetes Metab. Res. Rev.* 62, 2112–2115. doi: 10.2337/db12-1644
- Reijmer, Y. D., Freeze, W. M., Leemans, A., and Biessels, G. J. (2013b). The effect of lacunar infarcts on white matter tract integrity. *Stroke* doi: 10.1161/strokeaha.113.001321
- Smith, R. E., Tournier, J. D., Calamante, F., and Connelly, A. (2012). Anatomically-constrained tractography: improved diffusion MRI streamlines tractography through effective use of anatomical information. *Neuroimage* 62, 1924–1938. doi: 10.1016/j.neuroimage.2012.06.005
- Stejskal, E. O., and Tanner, J. E. (1965). Spin diffusion measurements: spin echoes in the presence of a time-dependent field gradient. *J. Chem. Phys.* 42, 288. doi: 10.1063/1.1695690
- Tax, C. M., Jeurissen, B., Vos, S. B., Viergever, M. A., and Leemans, A. (2014). Recursive calibration of the fiber response function for spherical deconvolution of diffusion MRI data. *Neuroimage* 86, 67–80. doi: 10.1016/j.neuroimage.2013.07.067
- Thompson, D. K., Thai, D., Kelly, C. E., Leemans, A., Tournier, J. D., Kean, M. J., et al. (2014). Alterations in the optic radiations of very preterm children: perinatal predictors and relationships with visual outcomes. *Neuroimage Clin.* 4, 145–153. doi: 10.1016/j.nicl.2013.11.007
- Tournier, J., Calamante, F., and Connelly, A. (2007). Robust determination of the fibre orientation distribution in diffusion MRI: non-negativity constrained super-resolved spherical deconvolution. *Neuroimage* 35, 1459–1472. doi: 10.1016/j.neuroimage.2007.02.016
- Tournier, J., Calamante, F., and Connelly, A. (2012). MRtrix: diffusion tractography in crossing fiber regions. *Int. J. Imaging Syst. Technol.* 22, 53–66. doi: 10.1002/ima.22005
- Tournier, J., Calamante, F., and Connelly, A. (2013). Determination of the appropriate b value and number of gradient directions for high-angular-resolution diffusion-weighted imaging. *NMR Biomed.* 26, 1775–1786. doi: 10.1002/nbm.3017
- Tournier, J., Calamante, F., Gadian, D. G., and Connelly, A. (2004). Direct estimation of the fiber orientation density function from diffusion-weighted MRI data using spherical deconvolution. *Neuroimage* 23, 1176–1185. doi: 10.1016/j.neuroimage.2004.07.037
- Tournier, J., Yeh, C. H., Calamante, F., Cho, K. H., Connelly, A., and Lin, C. P. (2008). Resolving crossing fibres using constrained spherical deconvolution: validation using diffusion-weighted imaging phantom data. *Neuroimage* 42, 617–625. doi: 10.1016/j.neuroimage.2008.05.002
- Tournier, J. D., Calamante, F., and Connelly, A. (2010). “Improved probabilistic streamlines tractography by 2nd order integration over fibre orientation distributions,” in *Proceedings of the International Society for Magnetic Resonance Medicine*, Stockholm, Sweden. Abstract 1670.
- Tournier, J. D., Mori, S., and Leemans, A. (2011). Diffusion tensor imaging and beyond. *Magn. Reson. Med.* 65, 1532–1556. doi: 10.1002/mrm.22924
- Tuch, D. S. (2004). Q-ball imaging. *Magn. Reson. Med.* 52, 1358–1372. doi: 10.1002/mrm.20279
- Tuch, D. S., Reese, T. G., Wiegell, M. R., Makris, N., Belliveau, J. W., and Wedeen, V. J. (2002). High angular resolution diffusion imaging reveals intravoxel white matter fiber heterogeneity. *Magn. Reson. Med.* 48, 577–582. doi: 10.1002/mrm.10268
- Vos, S. B., Jones, D. K., Viergever, M. A., and Leemans, A. (2011). Partial volume effect as a hidden covariate in DTI analyses. *Neuroimage* 55, 1566–1576. doi: 10.1016/j.neuroimage.2011.01.048
- Wang, Y., Wang, Q., Halder, J. P., Yeh, F. C., Xie, M., Sun, P., et al. (2011). Quantification of increased cellularity during inflammatory demyelination. *Brain* 134, 3590–3601. doi: 10.1093/brain/awr307
- Wedeen, V. J., Hagmann, P., Tseng, W. Y. I., Reese, T. G., and Weisskoff, R. M. (2005). Mapping complex tissue architecture with diffusion spectrum magnetic resonance imaging. *Magn. Reson. Med.* 54, 1377–1386. doi: 10.1002/mrm.20642
- Wedeen, V. J., Wang, R. P., Schmahmann, J. D., Benner, T., Tseng, W. Y. I., Dai, G., et al. (2008). Diffusion spectrum magnetic resonance imaging (DSI) tractography of crossing fibers. *Neuroimage* 41, 1267–1277. doi: 10.1016/j.neuroimage.2008.03.036
- Yeh, F. C., and Tseng, W. Y. I. (2013). Sparse solution of fiber orientation distribution function by diffusion decomposition. *PLoS ONE* 8:e75747. doi: 10.1371/journal.pone.0075747
- Yeh, F. C., Wedeen, V. J., and Tseng, W. Y. I. (2011). Estimation of fiber orientation and spin density distribution by diffusion deconvolution. *Neuroimage* 55, 1054–1062. doi: 10.1016/j.neuroimage.2010.11.087

**Conflict of Interest Statement:** The authors declare that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.

Received: 06 December 2013; accepted: 02 March 2014; published online: 28 March 2014.

Citation: Roine T, Jeurissen B, Perrone D, Aelterman J, Leemans A, Philips W and Sijbers J (2014) Isotropic non-white matter partial volume effects in constrained spherical deconvolution. *Front. Neuroinform.* 8:28. doi: 10.3389/fninf.2014.00028

This article was submitted to the journal *Frontiers in Neuroinformatics*. Copyright © 2014 Roine, Jeurissen, Perrone, Aelterman, Leemans, Philips and Sijbers. This is an open-access article distributed under the terms of the Creative Commons Attribution License (CC BY). The use, distribution or reproduction in other forums is permitted, provided the original author(s) or licensor are credited and that the original publication in this journal is cited, in accordance with accepted academic practice. No use, distribution or reproduction is permitted which does not comply with these terms.



# Variational Bayesian causal connectivity analysis for fMRI

Martin Luessi<sup>1,2\*</sup>, S. Derin Babacan<sup>3</sup>, Rafael Molina<sup>4</sup>, James R. Booth<sup>5</sup> and Aggelos K. Katsaggelos<sup>2</sup>

<sup>1</sup> Athinoula A. Martinos Center for Biomedical Imaging, Harvard Medical School, Massachusetts General Hospital, Charlestown, MA, USA

<sup>2</sup> Department of Electrical Engineering and Computer Science, Northwestern University, Evanston, IL, USA

<sup>3</sup> Google Inc., Mountain View, CA, USA

<sup>4</sup> Departamento de Ciencias de la Computación e I.A., Universidad de Granada, Granada, Spain

<sup>5</sup> Department of Communication Sciences and Disorders, Northwestern University, Evanston, IL, USA

## Edited by:

Jesus M. Cortes, Ikerbasque,  
Biocruces Health Research Institute,  
Spain

## Reviewed by:

Sebastiano Stramaglia, Università  
degli Studi di Bari, Italy  
Lotfi Chaari, IRI-TENSEIHT, France

## \*Correspondence:

Martin Luessi, Athinoula A. Martinos  
Center for Biomedical Imaging,  
Harvard Medical School,  
Massachusetts General Hospital,  
Building 149, Room 2301, 13th  
Street, Charlestown, MA 02129,  
USA  
e-mail: mluessi@  
nmr.mgh.harvard.edu

The ability to accurately estimate effective connectivity among brain regions from neuroimaging data could help answering many open questions in neuroscience. We propose a method which uses causality to obtain a measure of effective connectivity from fMRI data. The method uses a vector autoregressive model for the latent variables describing neuronal activity in combination with a linear observation model based on a convolution with a hemodynamic response function. Due to the employed modeling, it is possible to efficiently estimate all latent variables of the model using a variational Bayesian inference algorithm. The computational efficiency of the method enables us to apply it to large scale problems with high sampling rates and several hundred regions of interest. We use a comprehensive empirical evaluation with synthetic and real fMRI data to evaluate the performance of our method under various conditions.

**Keywords: fMRI, causality, connectivity, variational Bayesian method, Granger causality**

## 1. INTRODUCTION

Traditionally, functional neuroimaging has been used to obtain spatial maps of brain activation, e.g., using functional magnetic resonance imaging (fMRI) or positron emission tomography (PET), or to study the spatio-temporal progression of activity using magneto- or electroencephalography (M/EEG). Due to the increasing availability of MRI scanners to researchers and due to their high spatial resolution, the question of how fMRI can be used to obtain measures of *effective connectivity*, describing directed influence and causality in brain networks (Friston, 1994), has recently received significant attention.

An idea that forms the basis of several methods is that causality can be used to infer effective connectivity, i.e., if activity in one region can be used to accurately predict future activity in another region, it is likely that a directed connection between the regions exists. An exhaustive review of causality based methods for fMRI is beyond the scope of this work; we only provide a short introduction and refer to Roebroeck et al. (2011) for a recent review of related methods. Effective connectivity methods for fMRI can be divided into two groups. Methods in the first group are referred to as *dynamic causal modeling* (DCM) methods (Friston et al., 2003). In DCM, the relationship between neuronal activity in different regions of interest (ROIs) is described by bilinear ordinary differential equations (ODEs) and the fMRI observation process is modeled by a biophysical model based on the Balloon model (Buxton et al., 1998, 2004). While providing an accurate model of the hemodynamic process underlying fMRI, the non-linearity of the observation model poses difficulties when estimating the latent variables describing the neuronal activity from the fMRI observations. Due to this, DCM is typically used for small numbers of ROIs (less than 10) and DCM methods

typically are confirmatory approaches, i.e., the user provides a number of different candidate models describing the connectivity, which are then ranked based on an approximation to the model evidence.

The second class of methods attempts to estimate effective connectivity between ROIs from causal interactions that exist in the observed fMRI time series. In the widely used *Wiener–Granger causality* (WGC) measure (Wiener, 1956; Granger, 1969) (refer to Bressler and Seth, 2010 for a recent review of related methods), a linear prediction model is employed to predict the future of one time series using either only its past or its past and the past of the time series from a different ROI. If the latter leads to a significantly lower prediction error, the other time series is considered to exert a causal influence on the time series being evaluated, which is indicative of directed connectivity between the underlying ROIs. Related methods estimate the causal connectivity between all time series simultaneously by employing a vector autoregressive (VAR) model. The magnitudes of the estimated VAR coefficients are considered a measure of connectivity between regions. In Valdés-Sosa et al. (2005), a first order VAR model is employed and the connectivity graph is assumed to be sparse, i.e., only few regions are connected. The sparsity assumption is formalized by using  $\ell_1$ -norm regularization of the VAR coefficients. It has been shown in Haufe et al. (2008) that the use of higher order VAR models in combination with  $\ell_1\ell_2$ -norm (group-lasso) (Yuan and Lin, 2006; Meier et al., 2008) regularization of the VAR coefficients across lags leads to a more accurate estimation of the connectivity structure.

There are two main concerns when estimating effective connectivity from causal relations in the observed fMRI time series. First, the processing times at the neuronal level are in the order of

milliseconds, which is several orders of magnitude shorter than the sampling interval (time to repeat, TR) of the MRI scanner. Second, fMRI measures neuronal activity indirectly through the so-called blood oxygen level dependent (BOLD) contrast (Ogawa et al., 1990; Frahm et al., 1992), which depends on slow hemodynamic processes. The observation process can be modeled as a convolution of the time series describing the neuronal activity with a hemodynamic response function (HRF). As there is variability in the shape of the HRF among brain regions and individuals (Handwerker et al., 2004) and the sampling rate of the MRI scanner is low, detecting effective connectivity from causal interactions that exist in the observed fMRI data is a challenging problem. There has recently been some controversy if this is indeed the case. In David et al. (2008), a study using simultaneous fMRI, EEG, and intra-cerebral EEG recordings from rats was performed and it was found that the performance of WGC for fMRI is indeed poor, unless the fMRI time series of each region is first deconvolved with the measured HRF of the same region. Using simulations with synthetic fMRI data generated using the biophysical model underlying DCM, it was also found in Smith et al. (2011) that WGC methods perform poorly relative to the other evaluated connectivity methods. On the other hand, another recent study (Deshpande et al., 2010) found that WGC methods provide a high accuracy for the detection of causal interactions at the neuronal level with interaction lengths of hundreds of milliseconds, i.e., much shorter than the TR of the MRI scanner, even when HRF variations are present. The minor influence of HRF variations may be explained by the property that typical HRF variations do not simply correspond to temporal shifts of an HRF with the same shape, which would change the causality of interactions present in the fMRI data. Instead, as pointed out in Deshpande et al. (2010), the HRF variability among brain regions is mostly apparent in the shape of the peak of the HRF and the time-to-peak (Handwerker et al., 2004), which may explain why causal interactions at the neuronal level can still be present after convolution with varying HRFs. This is in agreement with recent results. It has been shown that WGC is invariant to filtering with invertible filters (Barnett and Seth, 2011) and in Seth et al. (2013) simulations were performed that confirm that the invariance typically holds for HRF convolution. However, at the same time it was found that WGC can be severely confounded when HRF convolution is combined with downsampling and measurement noise is added to the data.

Several methods have been proposed that account for HRF variability when analyzing WGC from fMRI data. In David et al. (2008) a noise-regularized HRF deconvolution was employed, and in Smith et al. (2010) a switching linear dynamical system (SLDS) model is proposed to describe the interaction between latent variables representing the neuronal activity together with a linear observation model based on a convolution with a (unknown) HRF for each region. The method employs a Bayesian formulation and obtains estimates of the latent variables using the maximum-likelihood approach. In contrast to WGC methods, the SLDS model can also account for modulatory inputs which change the effective connectivity of the network and introduce non-stationarity in the observed fMRI data. The method in Smith et al. (2010) can be seen as a convergence of DCM methods and

WGC-type methods (Roebroeck et al., 2011). A similar method is proposed in Ryali et al. (2011), which can be considered a multi-variate extension of methods which perform deconvolution of the neuronal activity for a single fMRI time series (Penny et al., 2005; Makni et al., 2008). Joint estimation of the HRF and detection of neuronal activity is also an important problem for event-related fMRI, we refer to Cassidy et al. (2012) and Chaari et al. (2013) for recently proposed methods addressing this problem.

In this paper, we propose a causal connectivity method for fMRI which employs a VAR model of arbitrary order for the time series of neuronal activity in combination with a linear hemodynamic convolution model for the fMRI observation process. We use a Bayesian formulation of the problem and draw inference based on an approximation to the posterior distribution which we obtain using the variational Bayesian (VB) method (Jordan et al., 1999; Attias, 2000). In contrast to previous methods (Smith et al., 2010; Ryali et al., 2011), our method is designed to be computationally efficient, enabling application to large scale problems with large numbers of regions and high temporal sampling rates. Computational efficiency is achieved by the introduction of an approximation to the neuronal time series in the Bayesian modeling. When drawing inference, introducing this approximation has the effect that the hemodynamic deconvolution can be separated from the estimation of the neuronal time series, leading to a reduction of the state-space dimension of the variational Kalman smoother (Beal and Ghahramani, 2001; Ghahramani and Beal, 2001), which forms a part of the VB inference algorithm. The lower state-space dimension drastically reduces the processing and memory requirements. Another key difference to previous Bayesian methods is that we assume that the VAR coefficient matrices are sparse and that the coefficient matrices at different lags have non-zero entries at mostly the same locations, i.e., the matrices have similar sparsity profiles. In Haufe et al. (2008) this assumption is formalized using an  $\ell_1\ell_2$ -norm regularization term for the VAR coefficient matrices. In our work, we employ Gaussian priors with shared precision hyperparameters for the VAR coefficient matrices, which is a Bayesian alternative to  $\ell_1\ell_2$ -norm regularization and results in a higher estimation performance of the method.

Our results show that the proposed method offers a higher detection performance than WGC when the number of nodes is large or when the SNR is low. In addition, our method is less affected when the VAR model order assumed in the method is higher than the order present in the data. We also perform simulations using a modified version of our method, which is similar to the method in Ryali et al. (2011), and show that the approximation to the neuronal time series used in our method has a negligible effect on the estimation performance while allowing the application of the proposed method to large problems with hundreds of ROIs. We perform an extensive series of simulations where we vary both the downsampling ratio and the neuronal delay. The results show that the proposed method offers some benefits over WGC, especially in low SNR situations and when HRF variations are present. However, both the proposed method and WGC can at times detect a causal influence with the opposite direction of the true influence, which is a known problem for WGC methods (David et al., 2008; Deshpande et al., 2010; Seth

et al., 2013). Finally, we apply the proposed method to resting-state fMRI data from the Human Connectome Project (Van Essen et al., 2012), where it successfully detects connections between regions that belong to known resting-state networks.

This paper is outlined as follows. First, we introduce a hierarchical Bayesian formulation for the generative model underlying the fMRI connectivity estimation problem. Next, we present the Bayesian inference scheme which estimates the latent variables of the model using a variational approximation to the posterior distribution. We then perform extensive simulations with synthetic fMRI data. Finally, we apply the method to real fMRI data and conclude the paper.

## 1.1. NOTATION

We use the following notation throughout this work: Matrices are denoted by uppercase bold letters, e.g.,  $\mathbf{A}$ , while vectors are denoted by lowercase bold letters, e.g.,  $\mathbf{a}$ . The element at the  $i$ -th row and  $j$ -th column of matrix  $\mathbf{A}$  is denoted by  $a_{ij}$ , while  $\mathbf{a}_i$  and  $\mathbf{a}_j$  denote column vectors with the elements from the  $i$ -th row and the  $j$ -th column of  $\mathbf{A}$ , respectively. The operator  $\text{diag}(\mathbf{A})$  extracts the main diagonal of  $\mathbf{A}$  as a column vector, whereas  $\text{Diag}(\mathbf{a})$  is a diagonal matrix with  $\mathbf{a}$  as its diagonal. The operator  $\text{vec}(\mathbf{A})$  vectorizes  $\mathbf{A}$  by stacking its columns,  $\text{tr}(\mathbf{A})$  denotes the trace of matrix  $\mathbf{A}$ , and  $\otimes$  denotes the Kronecker product. The identity matrix of size  $N \times N$  is denoted by  $\mathbf{I}_N$ . Similarly,  $\mathbf{0}_N$  and  $\mathbf{0}_{N \times M}$  denote  $N \times N$  and  $N \times M$  all-zero matrices, respectively.

## 2. BAYESIAN MODELING

The goal of this work is to infer effective connectivity implied by the causal relations between  $N$  time series of neuronal activity from  $N$  different regions in the brain. To this end, we employ a vector autoregressive (VAR) model of order  $P$  to model the time series as follows

$$\mathbf{s}(t) = \sum_{p=1}^P \mathbf{A}^{(p)} \mathbf{s}(t-p) + \boldsymbol{\eta}(t), \quad (1)$$

where  $\mathbf{s}(t) \in \mathbb{R}^N$  denotes the neuronal activity of all regions at time  $t$ ,  $\mathbf{A}^{(p)} \in \mathbb{R}^{N \times N}$  is a matrix with VAR coefficients for lag  $p$ , and  $\boldsymbol{\eta}(t) \sim \mathcal{N}(0, \boldsymbol{\Lambda}^{-1})$  denotes the innovation. In this model, the activity at any time point is predicted from the activity at  $P$  previous time points. More specifically, the activity of the  $i$ -th time series at time  $t$ , denoted by  $s_i(t)$ , is predicted from the past of the  $j$ -th time series using the coefficients  $\{a_{ij}^{(p)}\}_{p=1}^P$ . Hence, if any of these coefficients is significantly larger than zero, we can conclude that the  $j$ -th time series exerts a causal influence on the  $i$ -th time series, implying connectivity between the regions. This is the idea underlying Wiener–Granger causality (Wiener, 1956; Granger, 1969) and related methods using vector autoregressive models (Valdés-Sosa et al., 2005; Haufe et al., 2008).

We can now introduce an embedding process (Weigend and Gershenfeld, 1994; Penny et al., 2005)  $\mathbf{x}(t)$  defined by

$$\mathbf{x}(t) = [\mathbf{s}(t)^T \mathbf{s}(t-1)^T \dots \mathbf{s}(t-P+1)^T]^T, \quad (2)$$

which allows us to express (Equation (1)) by a first order VAR model as follows

$$\mathbf{x}(t) = \tilde{\mathbf{A}} \mathbf{x}(t-1) + \tilde{\boldsymbol{\eta}}(t), \quad (3)$$

where  $\tilde{\mathbf{A}} \in \mathbb{R}^{PN \times PN}$  is given by

$$\tilde{\mathbf{A}} = \begin{bmatrix} \mathbf{A}^{(1)} & \mathbf{A}^{(2)} & \dots & \mathbf{A}^{(P-1)} & \mathbf{A}^{(P)} \\ \mathbf{I}_N & \mathbf{0}_N & \dots & \mathbf{0}_N & \mathbf{0}_N \\ \mathbf{0}_N & \mathbf{I}_N & \dots & \mathbf{0}_N & \mathbf{0}_N \\ \vdots & \vdots & \ddots & \vdots & \vdots \\ \mathbf{0}_N & \mathbf{0}_N & \dots & \mathbf{I}_N & \mathbf{0}_N \end{bmatrix}. \quad (4)$$

The innovation  $\tilde{\boldsymbol{\eta}}(t)$  is Gaussian  $\tilde{\boldsymbol{\eta}}(t) \sim \mathcal{N}(0, \mathbf{Q})$ , where the covariance matrix  $\mathbf{Q}$  is all zero, except for the first  $N$  rows and columns, which are given by  $\boldsymbol{\Lambda}^{-1}$ . For the remainder of this paper, we present the modeling and inference with respect to the time series  $\mathbf{x}(t)$ . If access to the neuronal time series  $\mathbf{s}(t)$  is required, it can easily be extracted from  $\mathbf{x}(t)$  (it simply corresponds to the first  $N$  elements of  $\mathbf{x}(t)$ ).

## 2.1. OBSERVATION MODEL

Before introducing the observation model, note that we can obtain a noisy version of the neuronal time series from the embedding process  $\mathbf{x}(t)$  as follows

$$\mathbf{z}(t) = \mathbf{B} \mathbf{x}(t) + \boldsymbol{\kappa}(t), \quad (5)$$

where  $\mathbf{B} = [\mathbf{I}_N \mathbf{0}_{N \times (P-1)N}]$  and  $\boldsymbol{\kappa}(t) \sim \mathcal{N}(0, \vartheta^{-1} \mathbf{I})$ , where  $\vartheta$  is the precision parameter. Clearly, by using very large values for  $\vartheta$ , the time series  $\mathbf{z}(t)$  approaches  $\mathbf{s}(t)$ . The introduction of this Gaussian approximation to the neuronal time series greatly improves the computational efficiency of the proposed method, as it separates the VAR model for the neuronal time series from the hemodynamic observation model. This separation leads to a reduction of the state-space dimension of the Kalman smoothing algorithm, which forms part of the inference procedure, and therefore to greatly reduced memory requirements. In addition, using the approximation allows us to perform parts of the estimation in the frequency domain, which is computationally advantageous due to the efficiency of the fast Fourier transform. The computational advantages of the proposed method will be discussed in detail in the next section.

To model the fMRI observation process, we follow the standard assumption underlying the general linear model (Friston et al., 1995), and express the fMRI observation of the  $i$ -th region as follows

$$\begin{aligned} y_i(t) &= h_i(t) * z_i(t) + \varepsilon_i(t) \\ &= \sum_{k=1}^L h_i(k) z_i(t-k+1) + \varepsilon_i(t), \end{aligned} \quad (6)$$

where  $*$  denotes the convolution operation,  $h_i(t)$  is the hemodynamic response function (HRF) of length  $L$  for the  $i$ -th region, and  $\varepsilon_i(t)$  denotes observation noise. Notice that we can arrange



the HRF  $h_i(t)$  into a  $T \times T$  convolution matrix  $\mathbf{H}_i$ , which allows us to write (Equation (6)) as

$$\mathbf{y}_i = \mathbf{H}_i \mathbf{z}_i + \boldsymbol{\varepsilon}_i, \quad (7)$$

where the  $T \times 1$  vectors  $\mathbf{y}_i$ ,  $\mathbf{z}_i$ , and  $\boldsymbol{\varepsilon}_i$  are the fMRI observation, the approximation to the neuronal signal, and the observation noise, for the  $i$ -th region, respectively.

## 2.2. VAR COEFFICIENT PRIOR MODEL

We proceed by defining priors for the VAR coefficient matrices  $\{\mathbf{A}^{(p)}\}_{p=1}^P$ . For a network consisting of a large number of regions, it can generally be assumed that the connectivity is sparse, i.e., the VAR coefficient matrices contain a small number of non-zero coefficients. In the context of inferring causal connectivity, this idea has been used in Valdés-Sosa et al. (2005), where a first order VAR model with  $\ell_1$ -norm regularization for the VAR coefficients is used to obtain a sparse solution. For higher order VAR models, it is intuitive to assume that if the VAR coefficient  $a_{ij}^{(p_1)}$  modeling the connectivity from region  $j$  to region  $i$  and lag  $p_1$  is non-zero, it is likely that also other VAR coefficients for the same connection but different lags, i.e.,  $a_{ij}^{(p_2)}$ ,  $p_2 \neq p_1$ , are also non-zero. Together with the sparsity assumption, this leads to VAR coefficient matrices with similar sparsity profiles, i.e., the coefficient matrices at different time lags have non-zero entries at mostly the same locations. In Haufe et al. (2008) this idea is formalized by using  $\ell_1 \ell_2$ -norm (group lasso) (Yuan and Lin, 2006; Meier et al., 2008) regularization for the VAR coefficients across different lags, resulting in an improved estimation performance in comparison to methods that use alternative forms of regularization, such as,  $\ell_1$ -norm or ridge regression.

We incorporate the group sparsity assumption using Gaussian priors with shared precision hyperparameters across different lags. More specifically, we use

$$p(\mathbf{A}^{(p)} | \boldsymbol{\Gamma}) = \prod_{i=1}^N \prod_{j=1}^N \mathcal{N}(a_{ij}^{(p)} | 0, \gamma_{ij}^{-1}) \quad p \in \{1, \dots, P\}, \quad (8)$$

with Jeffreys hyperpriors to the precision hyperparameters

$$p(\boldsymbol{\Gamma}) \propto \prod_{i=1}^N \prod_{j=1}^N (\gamma_{ij})^{-1}. \quad (9)$$

During estimation, most of the precision hyperparameters in  $\boldsymbol{\Gamma}$  will assume very large values, hence effectively forcing the corresponding VAR coefficients to zero. This formulation is an adaptation of sparse Bayesian learning (also known as automatic relevance determination, ARD) (Tipping, 2001) to the problem of VAR coefficient estimation and can be considered a Bayesian alternative to a deterministic  $\ell_1 \ell_2$ -norm regularization term. Formulations where shared precision hyperparameters are used to enforce group sparsity have recently been proposed for applications such as simultaneous sparse approximation (Wipf and Rao, 2007), where shared precision parameters are used to obtain solutions with similar sparsity profiles across multiple time points.

Recently, shared hyperparameters were used to model the low-rank structure of the latent matrix in matrix estimation (Babacan et al., 2012).

## 2.3. INNOVATION AND NOISE PRIOR MODELS

To complete the description of the Bayesian model, we define priors for the innovation process and the observation noise in Equations (1) and (6), respectively. We assume that the innovations are independent and identically distributed (i.i.d.) zero-mean Gaussian for each time point, i.e.,  $\boldsymbol{\eta}(t) \sim \mathcal{N}(0, \boldsymbol{\Lambda}^{-1})$  and  $\boldsymbol{\varepsilon}(t) \sim \mathcal{N}(0, \mathbf{R})$ . It has to be expected that the linear prediction model used in the proposed method cannot fully explain the relationship between the neuronal time series in different ROIs. Hence, the precision matrix  $\boldsymbol{\Lambda}$  can contain some non-zero off-diagonal elements. We model this using a Wishart prior for the precision matrix

$$p(\boldsymbol{\Lambda}) = \mathcal{W}(\boldsymbol{\Lambda} | \nu_0, \mathbf{W}_0), \quad (10)$$

where  $\nu_0$  and  $\mathbf{W}_0$  are deterministic parameters. By using a diagonal matrix for  $\mathbf{W}_0$ , we obtain a prior modeling that encourages  $\boldsymbol{\Lambda}$  to be diagonal, which is the structure usually assumed in VAR models. Another reason for choosing this prior modeling is that the Wishart distribution is the conjugate prior for the precision matrix of the Gaussian distribution, which simplifies the inference procedure.

For the observation noise, we assume that the noise in different regions is uncorrelated and use diagonal covariance matrices given by  $\mathbf{R} = \text{Diag}(\boldsymbol{\beta})^{-1}$ , where  $\boldsymbol{\beta}$  is a precision hyperparameter vector of length  $N$ . We use conjugate gamma hyperpriors for the precisions as follows

$$p(\boldsymbol{\beta}) = \prod_{i=1}^N \Gamma(\beta_i | a_\beta^0, b_\beta^0), \quad (11)$$

where the gamma distribution with shape parameter  $a$  and inverse scale parameter  $b$  is given by

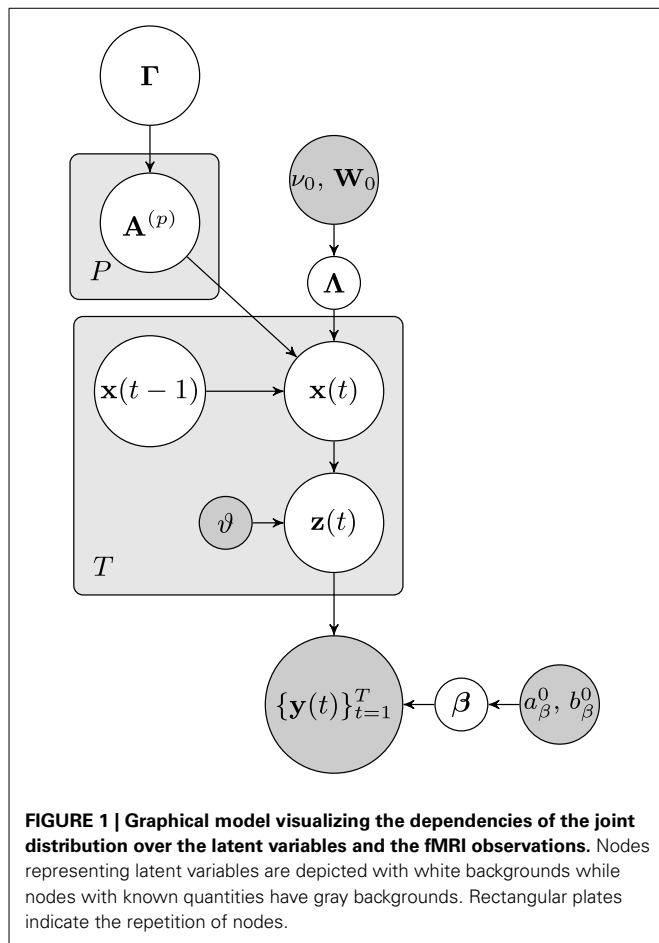
$$\Gamma(\xi | a, b) = \frac{b^a}{\Gamma(a)} \xi^{a-1} \exp(-b\xi). \quad (12)$$

We usually have some information about the fMRI observation noise and can use this knowledge to set the parameters  $a_\beta^0$  and  $b_\beta^0$ . The setting of the deterministic parameters will be discussed in more detail in the next section.

## 2.4. GLOBAL MODELING

By combining the probability distribution describing the VAR model, the fMRI observation model, and the prior model, we obtain a joint distribution over all latent variables and known quantities as

$$p(\Theta, \{\mathbf{y}(t)\}_{t=1}^T) = \left( \prod_{i=1}^N p(\mathbf{y}_i | \mathbf{z}_i, \mathbf{H}_i, \boldsymbol{\beta}_i) \right) \left( \prod_{t=1}^T p(\mathbf{z}(t) | \mathbf{x}(t), \boldsymbol{\vartheta}) \right) \times \left( \prod_{t=1}^T p(\mathbf{x}(t) | \mathbf{x}(t-1), \{\mathbf{A}^{(p)}\}_{p=1}^P, \boldsymbol{\Lambda}) \right)$$



$$\times \left( \prod_{p=1}^P p(\mathbf{A}^{(p)} | \Gamma) \right) p(\Gamma) p(\Lambda) p(\beta), \quad (13)$$

where  $\Theta$  contains all the latent variables of the model, i.e.,

$$\Theta = \{\mathbf{x}(t)_{t=1}^T, \mathbf{z}(t)_{t=1}^T, \{\mathbf{A}^{(p)}\}_{p=1}^P, \Gamma, \Lambda, \beta\}. \quad (14)$$

The dependencies of the joint distribution can be visualized as a directed acyclic graphical model, which is depicted in **Figure 1**. From the graphical model it can be seen that the node of approximate neuronal time series  $\mathbf{z}(t)$  is inserted between the nodes of the neuronal time series  $\mathbf{x}(t)$  and the observation  $\mathbf{y}(t)$ . As will be discussed in the next section, this additional node leads to important computational advantages, as it allows us to separate the hemodynamic deconvolution (estimation of  $\mathbf{z}(t)$ ) from the estimation of the neuronal time series  $\mathbf{z}(t)$  and the VAR modeling parameters.

### 3. BAYESIAN INFERENCE

We draw inference based on the posterior distribution

$$p(\Theta | \{\mathbf{y}(t)\}_{t=1}^T) = \frac{p(\Theta, \{\mathbf{y}(t)\}_{t=1}^T)}{p(\{\mathbf{y}(t)\}_{t=1}^T)}. \quad (15)$$

However, as with many probabilistic models, calculating  $p(\{\mathbf{y}(t)\}_{t=1}^T)$  and hence calculating the posterior distribution is analytically intractable. Therefore, we approximate the posterior distribution by a simpler distribution using the variational Bayesian (VB) method with the mean field approximation (Jordan et al., 1999; Attias, 2000). For the problem at hand we approximate the posterior by a distribution which factorizes over the latent variables as follows

$$q(\Theta) = q(\{\mathbf{x}(t)\}_{t=1}^T) q(\{\mathbf{z}(t)\}_{t=1}^T) q(\{\mathbf{A}^{(p)}\}_{p=1}^P) q(\Gamma, \Lambda, \beta). \quad (16)$$

Using the structure of the graphical model and the property of d-separation, it is found there are several induced factorizations when assuming the factorization given by Equation (16) (refer to Bishop, 2006 for detailed explanations). We can include the induced factorizations to further factorize the posterior as follows<sup>1</sup>

$$q(\Theta) = q(\{\mathbf{x}(t)\}_{t=1}^T) \left( \prod_{i=1}^N q(\{\mathbf{z}_i(t)\}_{t=1}^T) \right) q(\{\mathbf{A}^{(p)}\}_{p=1}^P) \\ \times \left( \prod_{i=1}^N \prod_{k=1}^N q(\gamma_{ik}) \right) q(\Lambda) \left( \prod_{i=1}^N q(\beta_i) \right). \quad (17)$$

The key ingredient of this VB method is that we only assume a specific factorization of the posterior but make no assumptions about the functional form of the distributions. Instead, we find the form of each distribution by performing a variational minimization of the Kullback–Leibler (KL) divergence between the approximation and the true posterior. The KL divergence is given by

$$C_{KL} \left( q(\Theta) \| p(\Theta | \{\mathbf{y}(t)\}_{t=1}^T) \right) = \int q(\Theta) \log \left( \frac{q(\Theta)}{p(\Theta | \{\mathbf{y}(t)\}_{t=1}^T)} \right) d\Theta \quad (18)$$

which is a non-negative measure that is only equal to zero if  $q(\Theta) = p(\Theta | \{\mathbf{y}(t)\}_{t=1}^T)$ . A standard result from VB analysis (Bishop, 2006) is that if we express [Equation (17)] as  $q(\Theta) = \prod_i q(\Phi_i)$ , i.e., we use  $q(\Phi_i)$  to denote the individual factors in [Equation (17)], the distribution for the  $i$ -th factor which minimizes [Equation (18)] is given by

$$\ln q(\Phi_i) = \left\langle \ln p(\Theta, \{\mathbf{y}(t)\}_{t=1}^T) \right\rangle_{q(\Theta \setminus \Phi_i)} + \text{const}, \quad (19)$$

where  $\langle \cdot \rangle_{q(\Theta \setminus \Phi_i)}$  denotes the expectation with respect to distributions  $q(\cdot)$  all latent variables except  $\Phi_i$ . Using this, we obtain a distribution for each factor. The VB inference algorithm sequentially updates the sufficient statistics of each distribution until

<sup>1</sup>Note that the only factorization we assume is the one in Equation (16); the induced factorizations appear in the derivation of the approximate posterior distribution and we can include them at this point to simplify the derivations.

convergence. Below we show the functional form of the variational posterior distribution for each latent variable. Due to space constraints, the derivations are not shown here and we refer to Luessi (2011) for more details.

Using Equation (19), the distribution for the neuronal time series  $q(\{\mathbf{x}(t)\}_{t=1}^T)$  is obtained from

$$\begin{aligned} \ln q(\{\mathbf{x}(t)\}_{t=1}^T) &= \left\langle \ln \prod_{t=1}^T p(\mathbf{x}(t) | \mathbf{x}(t-1), \{\mathbf{A}^{(p)}\}_{p=1}^P, \mathbf{\Lambda}) \right. \\ &\quad \left. \times p(\mathbf{z}(t) | \mathbf{x}(t), \boldsymbol{\vartheta}) \right\rangle_{q(\{\mathbf{z}(t)\}_{t=1}^T) q(\{\mathbf{A}^{(p)}\}_{p=1}^P) q(\boldsymbol{\Gamma}, \mathbf{\Lambda}, \boldsymbol{\beta})} + \text{const}, \end{aligned} \quad (20)$$

where all terms not depending on  $\{\mathbf{x}(t)\}_{t=1}^T$  have been absorbed into the additive normalization constant. Due to the conjugacy of the priors,  $q(\{\mathbf{x}(t)\}_{t=1}^T)$  is a multivariate Gaussian distribution with dimension  $TPN$ . However, this distribution has a complicated form and cannot be further factorized, which makes a direct calculation of the sufficient statistics computationally infeasible. Note that this complication is not due to the introduction of  $\mathbf{z}(t)$ ; it is also present in methods which do not employ the approximate time series  $\mathbf{z}(t)$ . Fortunately, Equation (20) has a similar form as an equation encountered in the variational Kalman smoothing algorithm (Beal and Ghahramani, 2001; Ghahramani and Beal, 2001), with the only difference that instead of using the observations we use the expectation of  $\mathbf{z}(t)$  under  $q(\{\mathbf{z}(t)\}_{t=1}^T)$ . The variational Kalman smoothing algorithm recursively estimates  $q(\mathbf{x}(t)) = \mathcal{N}(\mathbf{x}(t) | \boldsymbol{\mu}_t, \boldsymbol{\Sigma}_t)$  using a forward and a backward recursion. It is important to point out that we do not introduce an additional factorization of  $q(\{\mathbf{x}(t)\}_{t=1}^T)$  over time points, as for example done in Makni et al. (2008), which has been shown to result in an inaccurate approximation to the posterior distribution for large  $T$  (Wang and Titterton, 2004). Instead, the variational Kalman smoothing algorithm provides an efficient way for estimating  $q(\{\mathbf{x}(t)\}_{t=1}^T)$  without assuming a factorization over time points.

In our implementation we ignore the contribution from the covariances in the quadratic terms of  $\{\mathbf{A}^{(p)}\}_{p=1}^P$ , i.e., we assume  $\langle (\mathbf{A}^{(p)})^T (\mathbf{A}^{(p)}) \rangle = \langle \mathbf{A}^{(p)} \rangle^T \langle \mathbf{A}^{(p)} \rangle$ . This assumption is also made in Ryali et al. (2011) and can be expected to have only a minor influence on the performance of the proposed method. The main reason for using this approximation is that we do not need to calculate and store the covariance matrix of  $q(\{\mathbf{A}^{(p)}\}_{p=1}^P)$ , which greatly reduces the computational requirements of the method. Another effect of using this approximation is that the recursive inference algorithm becomes similar to the standard Kalman smoothing algorithm, also known as the Rauch-Tung-Striebel smoother (Rauch et al., 1965). For the forward pass, we use the initial conditions  $\boldsymbol{\mu}_0^0 = \mathbf{0}$ ,  $\boldsymbol{\Sigma}_0^0 = \mathbf{I}$  and calculate for  $t = 1, 2, \dots, T$  the following

$$\boldsymbol{\mu}_t^{t-1} = \langle \tilde{\mathbf{A}} \rangle \boldsymbol{\mu}_{t-1}^{t-1} \quad (21)$$

$$\boldsymbol{\Sigma}_t^{t-1} = \langle \tilde{\mathbf{A}} \rangle \boldsymbol{\Sigma}_{t-1}^{t-1} \langle \tilde{\mathbf{A}} \rangle^T + \langle \mathbf{Q} \rangle \quad (22)$$

$$\boldsymbol{\mu}_t^t = \boldsymbol{\mu}_t^{t-1} + \mathbf{K}_t (\langle \mathbf{z}(t) \rangle - \mathbf{B} \boldsymbol{\mu}_t^{t-1}) \quad (23)$$

$$\boldsymbol{\Sigma}_t^t = \boldsymbol{\Sigma}_t^{t-1} - \mathbf{K}_t \mathbf{B} \boldsymbol{\Sigma}_t^{t-1}, \quad (24)$$

where the Kalman gain is given by

$$\mathbf{K}_t = \boldsymbol{\Sigma}_t^{t-1} \mathbf{B}^T (\mathbf{B} \boldsymbol{\Sigma}_t^{t-1} \mathbf{B}^T + \boldsymbol{\vartheta}^{-1} \mathbf{I}_N)^{-1}. \quad (25)$$

After the forward pass, the final estimate for the last time point has been obtained, i.e., we have  $\boldsymbol{\mu}_T = \boldsymbol{\mu}_T^T$  and  $\boldsymbol{\Sigma}_T = \boldsymbol{\Sigma}_T^T$ . For the remaining time points we execute a backward pass and calculate the sufficient statistics of  $q(\mathbf{x}(t))$  for  $t = T-1, T-2, \dots, 1$  as follows

$$\boldsymbol{\mu}_t = \boldsymbol{\mu}_t^t + \mathbf{J}_t (\boldsymbol{\mu}_{t+1} - \langle \tilde{\mathbf{A}} \rangle \boldsymbol{\mu}_t^t), \quad (26)$$

$$\boldsymbol{\Sigma}_t = \boldsymbol{\Sigma}_t^t + \mathbf{J}_t (\boldsymbol{\Sigma}_{t+1}^t - \boldsymbol{\Sigma}_{t+1}^t) \mathbf{J}_t^T, \quad (27)$$

where

$$\mathbf{J}_t = \boldsymbol{\Sigma}_t^t \langle \tilde{\mathbf{A}} \rangle^T (\boldsymbol{\Sigma}_{t+1}^t)^{-1}. \quad (28)$$

As the posterior distributions of individual time points are not independent, i.e.,  $q(\{\mathbf{x}(t)\}_{t=1}^T) \neq \prod_{t=1}^T q(\mathbf{x}(t))$ , cross-time expectations contain a cross-time covariance  $\boldsymbol{\Sigma}_{t,t-1}$ , i.e.,  $\langle \mathbf{x}(t) \mathbf{x}(t-1)^T \rangle = \boldsymbol{\mu}_t \boldsymbol{\mu}_{t-1}^T + \boldsymbol{\Sigma}_{t,t-1}$ . Such cross-time covariance terms are computed as follows (see Ghahramani and Hinton, 1996)

$$\boldsymbol{\Sigma}_{t,t-1} = \boldsymbol{\Sigma}_t \mathbf{J}_{t-1}^T + \mathbf{J}_t (\boldsymbol{\Sigma}_{t+1,t} - \langle \tilde{\mathbf{A}} \rangle \boldsymbol{\Sigma}_t^t) \mathbf{J}_{t-1}^T. \quad (29)$$

The posterior distribution of the approximate time series for the  $i$ -th region  $q(\{\mathbf{z}_i(t)\}_{t=1}^T)$  is found to be a Gaussian, that is,

$$q(\{\mathbf{z}_i(t)\}_{t=1}^T) = \mathcal{N}(\mathbf{z}_i | \langle \mathbf{z}_i \rangle, \boldsymbol{\Sigma}_z^i), \quad (30)$$

with parameters

$$\langle \mathbf{z}_i \rangle = \boldsymbol{\Sigma}_z^i \left( \langle \beta_i \rangle \mathbf{H}_i^T \mathbf{y}_i + \boldsymbol{\vartheta} \langle \mathbf{x}_i \rangle \right), \quad (31)$$

$$\boldsymbol{\Sigma}_z^i = \left( \langle \beta_i \rangle \mathbf{H}_i^T \mathbf{H}_i + \boldsymbol{\vartheta} \mathbf{I}_T \right)^{-1}. \quad (32)$$

The distribution for the VAR coefficients  $\mathbf{a} = \text{vec}([\mathbf{A}^{(1)} \mathbf{A}^{(2)} \dots \mathbf{A}^{(P)}])$  is also Gaussian, the mean and covariance matrix are given by

$$\langle \mathbf{a} \rangle = \boldsymbol{\Sigma}_a \text{vec} \left( \langle \mathbf{A} \rangle \left[ \sum_{t=1}^T (\boldsymbol{\mu}_t)_{1:N} \boldsymbol{\mu}_{t-1}^T + (\boldsymbol{\Sigma}_{t,t-1})_{1:N,:} \right] \right) \quad (33)$$

$$\boldsymbol{\Sigma}_a^{-1} = \mathbf{P}_1 \otimes \langle \mathbf{A} \rangle + \text{Diag}(\mathbf{I}_P \otimes \text{vec}(\langle \boldsymbol{\Gamma} \rangle)), \quad (34)$$

where the matrix  $\mathbf{P}_1$  is given by

$$\mathbf{P}_1 = \sum_{t=1}^T \langle \mathbf{x}(t-1) \mathbf{x}(t-1)^T \rangle = \sum_{t=1}^T \boldsymbol{\mu}_{t-1} \boldsymbol{\mu}_{t-1}^T + \boldsymbol{\Sigma}_{t-1}. \quad (35)$$

Notice that the size of  $\Sigma_a^{-1}$  is  $N^2P \times N^2P$ . Hence, for large  $N$  performing a direct inversion is computationally very demanding and potentially numerically inaccurate. Moreover, storing the matrix requires large amounts of memory. Instead of directly inverting the matrix, we use a conjugate gradient (CG) algorithm to solve

$$\Sigma_a^{-1} \langle \mathbf{a} \rangle = \text{vec} \left( \langle \mathbf{A} \rangle \left[ \sum_{t=1}^T (\mu_t)_{1:N} \mu_{t-1}^T + (\Sigma_{t,t-1})_{1:N,:} \right] \right), \quad (36)$$

for  $\langle \mathbf{a} \rangle$ , which is possible since  $\Sigma_a^{-1}$  is symmetric positive definite. The CG algorithm only needs to compute matrix-vector products of the form  $\Sigma_a^{-1} \mathbf{p}$ . From the structure of  $\Sigma_a^{-1}$ , one can see that the multiplication of the diagonal matrix on the right side is simply the element-wise product of the diagonal of  $\text{Diag}(\mathbf{I}_P \otimes \text{vec}(\langle \mathbf{F} \rangle))$  and  $\mathbf{p}$ , which can be computed efficiently. Similarly,  $(\mathbf{P}_1 \otimes \langle \mathbf{A} \rangle) \mathbf{p}$  can be computed efficiently without computing the Kronecker product (Fernandes et al., 1998).

Note that computation of the gamma hyperparameters requires access to the diagonal elements of  $\Sigma_a$ . Since we do not explicitly compute  $\Sigma_a$ , we approximate the diagonal by  $\text{Diag}(\Sigma_a) \approx \text{Diag}(\text{diag}(\Sigma_a^{-1}))^{-1}$ . We performed experiments with small  $N$  where we calculated  $\Sigma_a$  directly using a matrix inversion. We found that using the CG algorithm with an approximation to the diagonal of the covariance matrix results in virtually the same estimation performance for the proposed method, while being much faster and more memory efficient.

The posterior for the noise precision  $\Lambda$  is Wishart distributed with  $q(\Lambda) = \mathcal{W}(\Lambda | \nu, \mathbf{W})$  where the parameters are given by

$$\nu = T + \nu_0, \quad (37)$$

$$\mathbf{W}^{-1} = \langle \mathbf{P}_2 \rangle + \mathbf{W}_0^{-1}. \quad (38)$$

The expectation  $\langle \mathbf{P}_2 \rangle$  is given by

$$\begin{aligned} \langle \mathbf{P}_2 \rangle = & \sum_{t=1}^T ((\mu_t)_{1:N} - \langle \bar{\mathbf{A}} \rangle \mu_{t-1}) ((\mu_t)_{1:N} - \langle \bar{\mathbf{A}} \rangle \mu_{t-1})^T \\ & - (\Sigma_{t,t-1})_{1:N,:} \langle \bar{\mathbf{A}} \rangle - \langle \bar{\mathbf{A}} \rangle^T (\Sigma_{t,t-1})_{1:N,:}^T \\ & + (\Sigma_t)_{1:N,1:N} + \langle \bar{\mathbf{A}} \rangle \Sigma_{t-1} \langle \bar{\mathbf{A}} \rangle^T, \end{aligned} \quad (39)$$

where  $\bar{\mathbf{A}} = [\mathbf{A}^{(1)} \mathbf{A}^{(2)} \dots \mathbf{A}^{(P)}]$ ,  $(\Sigma_t)_{1:N,1:N}$  is the top left  $N \times N$  block of  $\Sigma_t$ , and  $(\Sigma_{t,t-1})_{1:N,:}$  are the first  $N$  rows of  $\Sigma_{t,t-1}$ . The mean of the Wishart distribution is given by  $\langle \Lambda \rangle = \nu \mathbf{W}$ , which is the value used in the other distribution updates in the VB algorithm.

The distribution for the VAR precision hyperparameter  $q(\gamma_{ij})$  is found to be a gamma distribution with shape and inverse scale parameters

$$a_{\gamma}^{i,j} = \frac{P}{2}, \quad b_{\gamma}^{i,j} = \frac{1}{2} \sum_{p=1}^P \left( \langle a_{ij}^{(p)} \rangle^2 + \bar{a}_{ij}^{(p)} \right), \quad (40)$$

where  $\bar{a}_{ij}^{(p)}$  is the variance of  $a_{ij}^{(p)}$ , which we obtain from the approximation to the diagonal of  $\Sigma_a$ . Similarly, the posterior for the observation noise precision is a gamma distribution with the following shape parameter  $a_{\beta}^i = T/2 + a_{\beta}^0$  and inverse scale parameter

$$b_{\beta}^i = \frac{1}{2} \left[ \mathbf{y}_i^T \mathbf{y}_i - 2 \mathbf{y}_i^T \mathbf{H}_i \langle \mathbf{z}_i \rangle + \langle \mathbf{z}_i \rangle^T \mathbf{H}_i^T \mathbf{H}_i \langle \mathbf{z}_i \rangle + \text{tr}(\mathbf{H}_i^T \mathbf{H}_i \Sigma_i^z) \right] + b_{\beta}^0. \quad (41)$$

### 3.1. SELECTION OF DETERMINISTIC PARAMETERS

The proposed method has several deterministic parameters which have to be specified by the user, namely, the observation noise precision parameters  $\{a_{\beta}^0, b_{\beta}^0\}$ , the VAR model noise parameters  $\{\nu_0, \mathbf{W}_0\}$ , and the neuronal approximation precision  $\vartheta$ . Typically, an estimate of the noise variance  $\sigma^2$  present in the data is available to the user. If this case, a reasonable setting of the observation noise precision parameters is  $a_{\beta}^0 = c$ ,  $b_{\beta}^0 = c\sigma^2$ , where  $c$  is a constant related to the confidence in our initial noise estimate. For very small values of  $c$ , the observation noise precision will be estimated solely by the algorithm, while a high value forces the estimated noise precision to the value specified by the user. Unless otherwise noted, we assume throughout this work that an estimate of the noise variance is available and use  $c = 10^9$ .

On the other hand, the user typically does not have precise *a priori* knowledge of the AR innovation precision. In this case, one option is to use  $\nu_0 = 0$ ,  $\mathbf{W}_0^{-1} = \mathbf{0}$ , which is equivalent to a non-informative Jeffreys prior for the AR innovation precision matrix. However, we observed that  $\langle \Lambda \rangle$  can attain values that are too large when a non-informative prior is used. This behavior is caused by the fact that the convolution with the HRF acts as a low-pass filter and it is generally not possible to perfectly recover the high frequency content of the neuronal signal, causing an over-estimation of the AR innovation precision. We found that using  $\nu_0 = 1$  and  $\mathbf{W}_0 = 10^{-3} \mathbf{I}$ , prevents  $\langle \Lambda \rangle$  from attaining too large values and we use this setting in all experiments presented in this work. Naturally, the parameter setting depends on the scale of the fMRI observation. Throughout this work, we rescale the fMRI observation to have an RMS value of 6.0, where the root-mean-square (RMS) value is calculated as  $\text{RMS} = \sqrt{(\sum_{t=1}^T \|\mathbf{y}(t)\|_2^2) / (NT)}$ . Note that the choice of  $\text{RMS} = 6.0$  is arbitrary, i.e., different values could be used but then other deterministic parameters would have to be modified accordingly. Finally, the approximation precision parameter  $\vartheta$  plays an important role. In Equation (32) it acts similarly to a regularization parameter while having the role of the observation noise precision in the variational Kalman smoother. We heuristically found that using a value that is higher than the observation noise precision works well and we use  $\vartheta = 10/\sigma^2$  throughout this work.

### 3.2. COMPUTATIONAL ADVANTAGES OF THE PROPOSED APPROACH

To conclude this section, we highlight some important advantages in terms of computational requirements of the proposed method over previous approaches. The advantages of the proposed method are directly related to the introduction of the approximate time series  $\mathbf{z}(t)$ .



The first advantage is due to the separation of the model of the neuronal time series from the hemodynamic convolution model, which leads to a reduced state-space dimension of the Kalman smoothing algorithm. More specifically, in Smith et al. (2010); Ryali et al. (2011), the observation process is modeled as

$$\mathbf{y}(t) = \tilde{\mathbf{H}}\mathbf{x}(t) + \boldsymbol{\varepsilon}(t), \quad (42)$$

where  $\tilde{\mathbf{H}} \in \mathcal{R}^{N \times NL}$  is a matrix that contains the HRFs of all regions. This modeling requires that  $\mathbf{x}(t)$  is an embedding process over  $L$  time points, i.e., the dimension of  $\mathbf{x}(t)$  is  $D = NL$ , as opposed to  $D = NP$  in our method. The higher dimension leads to excessive memory requirements as the state-space dimension of the Kalman smoothing algorithm is increased and a total of  $2T$  covariance and cross-time covariance matrices of size  $D \times D$  need to be stored in memory. As an example, assuming double precision floating point arithmetic and  $P = 2$ ,  $L = 20$ ,  $N = 100$ ,  $T = 1000$ , the methods in Smith et al. (2010) and Ryali et al. (2011) require approximately 60 GB of memory to store the covariance matrices, whereas the proposed method only requires approximately 600 MB. The large memory consumption and the higher dimension of the required matrix inversions is the reason why previous methods become computationally infeasible for large scale problems where  $N \approx 100$  and  $T \approx 1000$ . The problem is even more severe for low TR values, since the HRF typically has a length of about 30 s and a higher sampling rate means more samples are needed to represent the HRF, thus increasing the value of  $L$ .

The second advantage due to introduction of  $\mathbf{z}(t)$  is that the approximate posterior of  $\mathbf{z}(t)$  factorizes over ROIs and we can update the posterior distribution  $q(\{z_i(t)\}_{t=1}^T)$  for each region separately using Equations (31, 32). For large numbers of time points this computation can still be expensive as the inversion of a  $T \times T$  matrix is required. However, notice that if we assume that the convolution with  $\mathbf{h}_i$  is circular, the matrix  $\mathbf{H}_i$  becomes circulant. Circulant matrices can be diagonalized by the discrete Fourier transform (see, e.g., Moon and Stirling, 2000). Hence, it is possible to perform the calculation of  $\langle \mathbf{z}_i \rangle$  in the frequency domain. In our implementation we use a fast Fourier transform (FFT) algorithm with zero-padding such that the circular convolution corresponds to a linear convolution. The resulting time complexity is  $O(T \log T)$ , compared to  $O(T^3)$  when a direct matrix inversion is used. Moreover, notice that  $\boldsymbol{\Sigma}_i^z$  is circulant as well, which allows us to reduce the computational and memory requirements by only calculating and storing the first row of  $\boldsymbol{\Sigma}_i^z$  (all other rows can be obtained by circular shifts of the first row).

#### 4. EMPIRICAL EVALUATION WITH SIMULATED DATA

In this section, we evaluate the performance of the proposed method using a number of different simulation scenarios. In all simulations, the proposed method is denoted by “VBCCA” (Variational Bayesian Causal Connectivity Analysis). For comparison purposes we include the conditional WGC analysis method implemented in the “Granger Causal Connectivity Analysis (GCCA) toolbox” (Seth, 2010), which we denote by “WGCA” (Wiener–Granger Causality Analysis). Note that we use WGCA for comparison as it is a widely used method with publicly

available implementations. More recent methods, such as the methods from Smith et al. (2010), Marinazzo et al. (2011) and Ryali et al. (2011) may offer a higher estimation performance than WGCA. However, their high computational complexity makes it difficult to apply them to large-scale problems, which is the situation where our method clearly outperforms WGCA. Nevertheless, we include a comparison with a modified version of our method, which does not use an approximation to the neuronal time series and is therefore more similar to the method from Ryali et al. (2011), and show that for small networks our method provides a comparable estimation performance.

#### 4.1. QUALITY METRICS

We use two objective metrics to evaluate the performance of the methods. The first metric serves to quantify the performance in terms of correctly detecting the presence of a connection between regions, without taking the direction of the causal influence into account. In order to do so, we calculate the area under the receiver operating characteristic (ROC) curve, which is commonly used in signal detection theory and has also previously been used to evaluate connectivity methods (Valdés-Sosa et al., 2005; Haufe et al., 2008). In the following we give a short explanation of the ROC curve and refer the reader to Fawcett (2006) for a more detailed introduction. The ROC curve is generated by applying thresholds to the estimated connectivity scores. The resulting binary masks are compared with the ground truth, resulting in a number of true positives (TP) and false positives (FP). From the TP and FP numbers, we can calculate the true positive rate (TPR) and false positive rate (FPR) as follows

$$\text{TPR} = \frac{\text{TP}}{P}, \quad \text{FPR} = \frac{\text{FP}}{N}, \quad (43)$$

where  $P$  and  $N$  are the total number of positives and negatives, respectively. For each threshold, we obtain a (FPR, TPR) point in the ROC space. By applying all possible thresholds, we can construct the ROC curve which allows us to compute the area under the curve (AUC). The AUC is the metric used here to evaluate the connection detection performance. The value of the AUC is on the interval  $[0, 1]$ , with 1.0 being perfect detection performance while 0.5 is the performance of a random detector, i.e., the AUC should always be above 0.5 and as close as possible to 1.0. To calculate the non-directional connectivity score between nodes  $i$  and  $j$  from the estimated  $N \times N$  connectivity matrix, we use the larger of the directional scores, i.e.,  $\text{con}(i, j) = \text{con}(j, i) = \max(c_{ij}, c_{ji})$ . For WGCA, the matrix  $\mathbf{C}$  is the matrix with estimated Granger causality scores, whereas for the proposed method we calculate  $\mathbf{C}$  from the estimated VAR coefficients using  $c_{ij} = \sqrt{\sum_{p=1}^P \langle a_{ij}^{(p)} \rangle}$ .

The AUC provides information on the performance in terms of detecting connections without taking directionality into account. A second metric, denoted by “d-Accuracy” (Smith et al., 2011), is used to evaluate the ability of a method to correctly identify the direction of the connection. The d-Accuracy is calculated as follows. For true connections (known from the ground truth) we compare the elements  $c_{ij}$  and  $c_{ji}$  in the connectivity matrix. We decide that the direction was estimated correctly if  $c_{ij} > c_{ji}$  and the true connection has the direction  $j \rightarrow i$ . By repeating for all

connections, we calculate the overall probability that the direction was estimated correctly, which is the d-Accuracy score. Like the AUC, the d-Accuracy lies between 0 and 1 with 1.0 indicating perfect performance and 0.5 being the performance of a random directionality detector.

## 4.2. NETWORK SIZE AND SNR

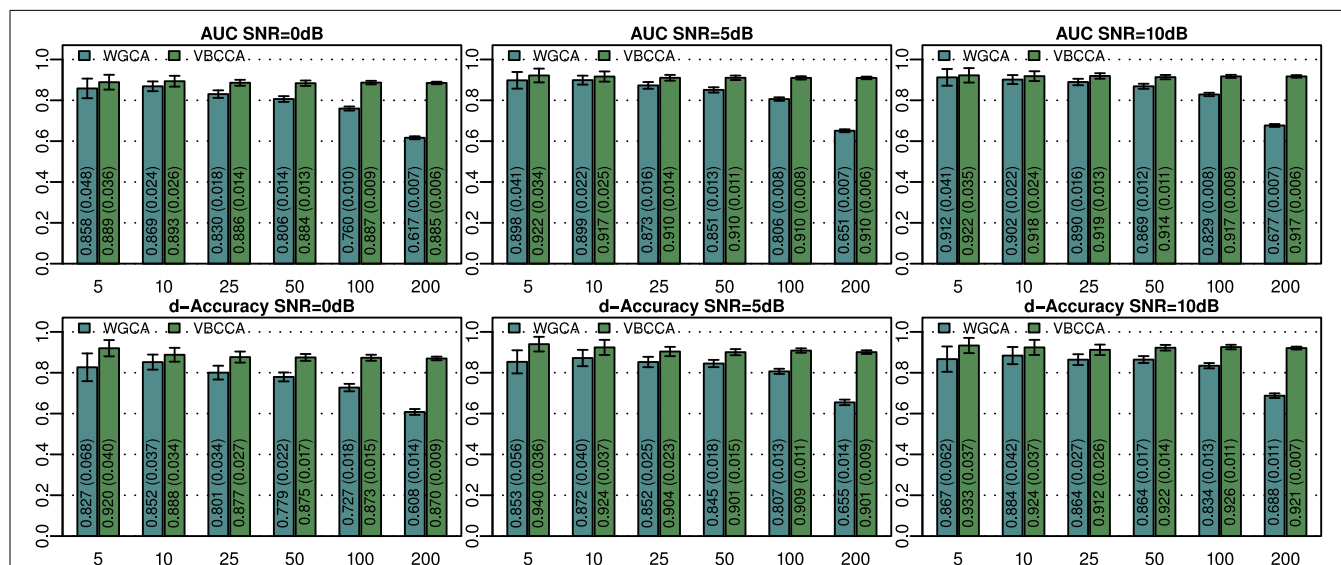
In this experiment we evaluate the performance of the proposed method for a number of networks of varying sizes and a number of different signal-to-noise ratios (SNRs). We generate neuronal time series according to Equation (1) where we simulate connectivity by randomly activating  $\lfloor N/2 \rfloor$  uni-directional connections, for which we generate the VAR coefficients according to  $a_{ij}^{(p)} \sim \mathcal{N}(0, 0.05) \forall p \in \{1, \dots, P\}$ , with  $P = 2$ . The noise term is chosen to be Gaussian with unit variance, i.e.,  $\eta(t) \sim \mathcal{N}(0, I_N)$ . Using the VAR coefficient matrices we generate a neuronal time series  $\mathbf{s}(t)$  with a total of  $T = 500$  time points. To generate the fMRI observations, we convolve the neuronal time series of each node with the canonical HRF implemented in SPM8 (<http://www.fil.ion.ucl.ac.uk/spm/>), which has a positive peak at 5 s and a smaller negative peak at 15.75 s. The HRF used has a total length of 30 s assuming a sampling rate of 1 Hz ( $L = 30$ ). Finally, to generate the noisy fMRI observation  $\mathbf{y}(t)$ , we add zero-mean, independent, identically distributed (i.i.d.) Gaussian noise with a variance  $\sigma^2$  determined by the SNR used, i.e.,  $\text{SNR}_{\text{dB}} = 10 \log_{10} \left( \left( \sum_{t=1}^T \|\mathbf{y}(t) - \bar{\mathbf{y}}(t)\|_2^2 \right) / (NT\sigma^2) \right)$ , where  $\bar{\mathbf{y}}(t)$  is the observation without additive noise.

The simulated noisy observations are used as inputs to the evaluated connectivity methods. In this experiment we use the true VAR order, i.e.,  $P = 2$ , for each evaluated method. Additionally, in the proposed method we use the same canonical HRF that is used to generate the data. Results for networks with

$N = \{5, 10, 25, 50, 100, 200\}$  nodes and SNRs of 0, 5, and 10 dB are shown in **Figure 2**. For small networks (5 and 10 nodes) both methods offer similar performance with the proposed method being slightly better. The SNR has a small influence on the performance and it can be concluded that each method performs similarly across the SNRs shown. As expected, the performance of both methods decreases with increasing network size. However, WGCA is affected drastically compared to the proposed method, which shows almost constant performance across network sizes. The proposed method clearly outperforms WGCA for large networks (more than 25 nodes). For  $N = 200$ , the AUC for WGCA is approximately 0.65, which is very poor. Therefore, for the given number of time samples, it can be concluded that WGCA is not suitable for connectivity analysis in large scale networks.

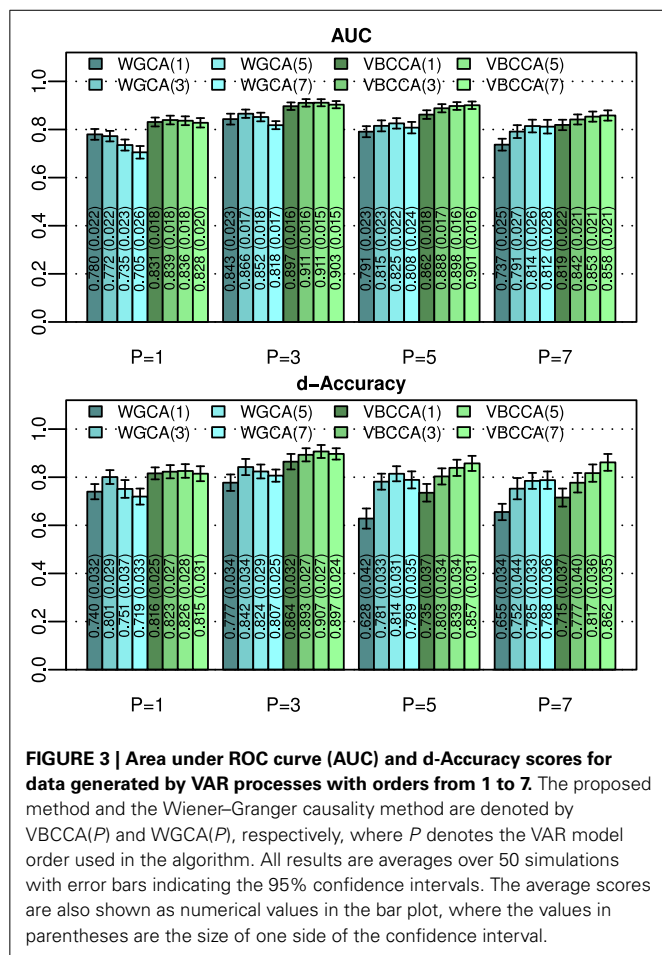
## 4.3. VAR ORDER

An important question is how the performance is affected by a mismatch in the VAR order present in the data and the VAR order assumed in the algorithm. For this evaluation we generate simulated data using the same procedure as in the first experiment for  $N = 25$  and an SNR of 0 dB, but we vary the VAR order from 1 to 7. The generated data is used as input to the evaluated methods for which we vary the VAR order used in the algorithm in the same range, i.e., from 1 to 7. Results for this simulation are shown in **Figure 3**; it can be seen that the proposed method typically outperforms the WGCA method even if there is a mismatch between the VAR order in the data and the VAR order used in the algorithm. It is also interesting to note that the proposed method typically performs well as long as the VAR order used in the algorithm is equal or higher than that present in the data. This behavior can be attributed to two factors. First, the proposed method employs a grouping of VAR coefficients across lags through shared priors, which limits the model complexity even



**FIGURE 2 |** Area under ROC curve (AUC) and d-Accuracy scores for random networks with sizes between 5 and 200 nodes and different SNRs. The proposed method is denoted by VBCCA, whereas WGCA denotes Wiener-Granger causality analysis. All results are averages over

50 simulations with error bars indicating the 95% confidence intervals. The average scores are also shown as numerical values in the bar plot, where the values in parentheses are the size of one side of the confidence interval.

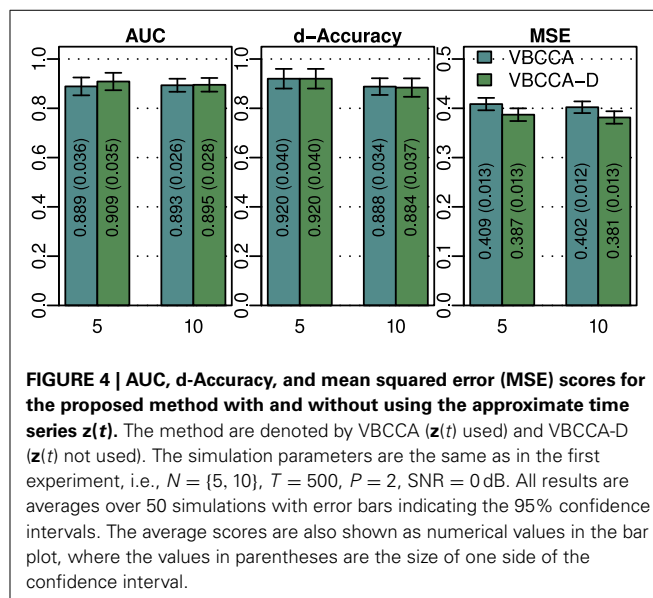


**FIGURE 3 | Area under ROC curve (AUC) and d-Accuracy scores for data generated by VAR processes with orders from 1 to 7.** The proposed method and the Wiener–Granger causality method are denoted by VBCCA( $P$ ) and WGCA( $P$ ), respectively, where  $P$  denotes the VAR model order used in the algorithm. All results are averages over 50 simulations with error bars indicating the 95% confidence intervals. The average scores are also shown as numerical values in the bar plot, where the values in parentheses are the size of one side of the confidence interval.

when the VAR order is increased. Second, we use an approximation to the posterior distribution to estimate the VAR coefficients; it is well known that methods which draw inference based on the posterior distribution are less prone to over-fitting than other methods, such as, maximum likelihood methods.

#### 4.4. EFFECT OF USING AN APPROXIMATION TO THE NEURONAL SIGNAL

As discussed in previous sections, the proposed method employs a hierarchical Bayesian model with an approximation to the neuronal time series. The approximate time series is denoted by  $\tilde{\mathbf{z}}(t)$  and is a key part of the proposed method as it enables the method to be computationally efficient through a reduction of the state space dimension used in the Kalman smoother. In addition, the time series  $\mathbf{z}(t)$  can be efficiently estimated in the frequency domain using fast Fourier transform algorithms. While the introduction of this approximation improves the computational efficiency, some reduction in the estimation performance may be caused. To quantify the influence of this approximation, we have implemented a modified version of the proposed method where  $\mathbf{z}(t)$  is not used, i.e., we increase the dimension of  $\mathbf{x}(t)$  to  $D = NL$  and model the observation process using Equation (42). This part of the modified model exactly corresponds to what is used in Smith et al. (2010) and Ryali et al. (2011). Due to the



**FIGURE 4 | AUC, d-Accuracy, and mean squared error (MSE) scores for the proposed method with and without using the approximate time series  $\mathbf{z}(t)$ .** The method are denoted by VBCCA ( $\mathbf{z}(t)$  used) and VBCCA-D ( $\mathbf{z}(t)$  not used). The simulation parameters are the same as in the first experiment, i.e.,  $N = \{5, 10\}$ ,  $T = 500$ ,  $P = 2$ ,  $\text{SNR} = 0$  dB. All results are averages over 50 simulations with error bars indicating the 95% confidence intervals. The average scores are also shown as numerical values in the bar plot, where the values in parentheses are the size of one side of the confidence interval.

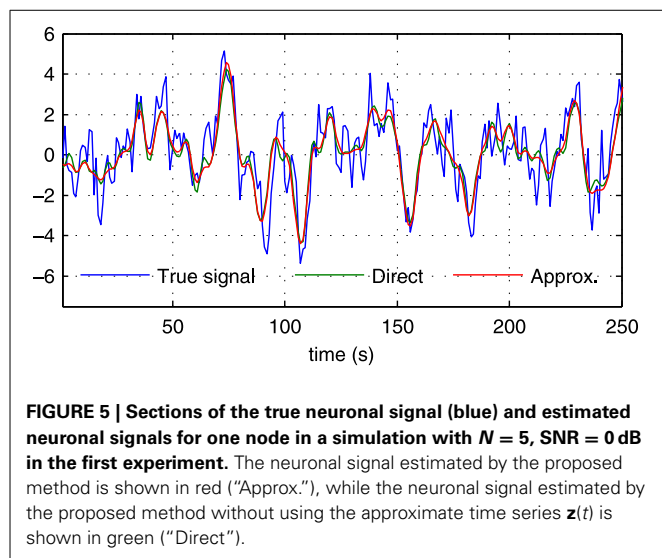
excessive memory requirements, the modified version of the proposed method, which we denote by “VBCCA-D,” can only be used for networks with small numbers of regions and HRFs consisting of a small number of time samples. We apply the method to the same data that is used in the first experiment, with  $N = \{5, 10\}$ ,  $\text{SNR} = 0$  dB. The resulting connectivity scores, as well as, the mean squared error (MSE) of the neuronal signal are shown in Figure 4. The MSE is calculated as follows

$$\text{MSE} = \left[ \sum_{t=1}^T \|\mathbf{s}(t) - \tilde{\mathbf{s}}(t)\|_2^2 \right] / \left[ \sum_{t=1}^T \|\mathbf{s}(t)\|_2^2 \right], \quad (44)$$

where  $\mathbf{s}(t)$  and  $\tilde{\mathbf{s}}(t)$  are the true and the estimated neuronal signals, respectively. It can be seen that the use of the neuronal approximation does not have a negative influence on the performance in terms of AUC while the MSE is slightly lower when the approximation is not used. The small difference in terms of MSE implies that both methods estimate the neuronal signal with similar estimation quality. This is also apparent from Figure 5, which shows the time neuronal series for one region estimated with and without the approximation.

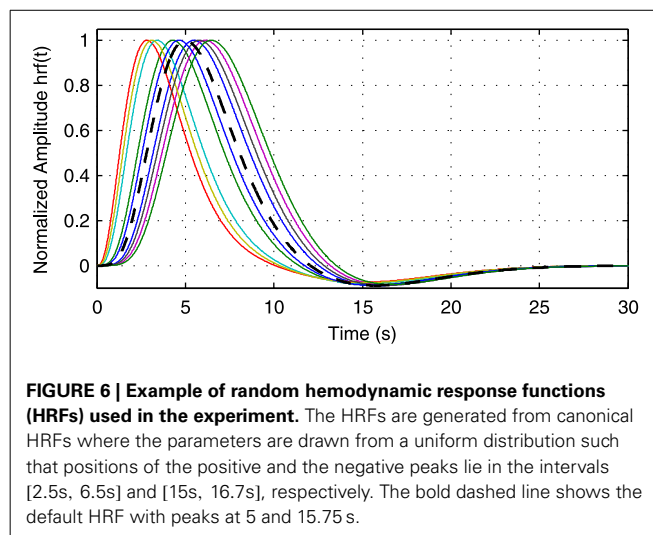
#### 4.5. DOWNSAMPLING AND HRF VARIATIONS

As processing at the neuronal level occurs at temporal scales which are orders of magnitudes faster than the sampling interval of the MRI scanner, it is important to analyze how the performance of causality based methods is affected by the low sampling rate. Another important question is the effect of HRF variability on the performance. In this experiment we analyze the influence of these effects on the estimated causality. In order to do so, we generate  $\mathbf{s}(t)$  for two regions and a single connection according to Equation (1) with zero-mean, i.i.d., Gaussian innovations, i.e.,  $\boldsymbol{\eta}(t) \sim \mathcal{N}(0, \mathbf{I})$ . The simulated sampling rate at the neuronal level is 1 kHz and we generate a total of 240 s of data. We use  $a_{1,1}^1 = a_{2,2}^1 = 0.95$  to simulate a degree of autocorrelation



within each time series. To simulate connection with a certain neuronal delay, depending of the direction of the influence we draw the value of either  $a_{1,2}^d$  or  $a_{2,1}^d$  from a uniform distribution on the interval  $[0.4, 0.9]$ . The lag parameter  $d$  is used to simulate the neuronal delay, e.g.,  $d = 10$  corresponds to a delay of 10 ms. Next, we convolve the obtained neuronal time series with an HRF for each region. In the first simulation we use the same canonical HRF with peaks at 5 and 15.75 s for both regions, whereas in the second simulation we use a randomly generated HRF for each region. To generate a random HRF, we use the HRF generation function provided in SPM8 (<http://www.fil.ion.ucl.ac.uk/spm/>). The parameter controlling the time-to-peak is drawn from a uniform distribution, such that the positions of the positive peak lies between 2.5 and 6.5 s, which is the range of peak positions reported in Handwerker et al. (2004). The parameter controlling the position of the negative peak ("undershoot") is held constant at 16 s. Due the implementation in SPM8, the negative peak of the generated HRF lies between 15 and 16.7 s, depending on the position of the positive peak. An example of HRFs used in our experiment is depicted in **Figure 6**. After each time series has been convolved with a HRF, the data is downsampled to simulate a certain TR value. Finally we add zero-mean, i.i.d., Gaussian noise such that the resulting SNR is 0 dB. To study both the influence of downsampling and the neuronal delay, we linearly vary the simulated TR between 50 ms and 2 s using a step size of 50 ms (40 points) and the delay using 40 linearly spaced values between 5 and 300 ms, resulting in a total of 1600 TR/delay combinations.

Results for the first simulation, in which the HRF is held constant, are shown in **Figure 7**. The results confirm previous findings (Seth et al., 2013) that downsampling confounds WGC. One might intuitively expect that when the neuronal delay is held constant, a lower TR will lead to a higher d-Accuracy. However, our simulations show that this is not necessarily the case; For very low delay and TR values, the WGCA method has d-Accuracy to zero, i.e., it consistently estimates a causal influence with the opposite direction of the true influence, while it approaches the

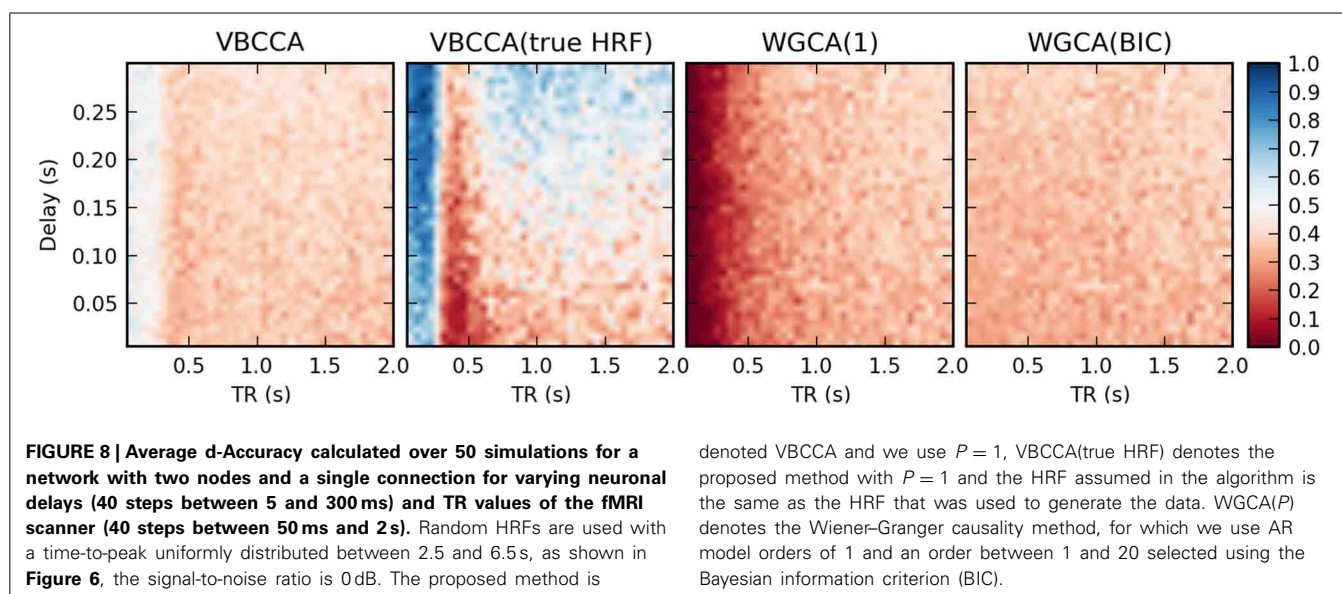
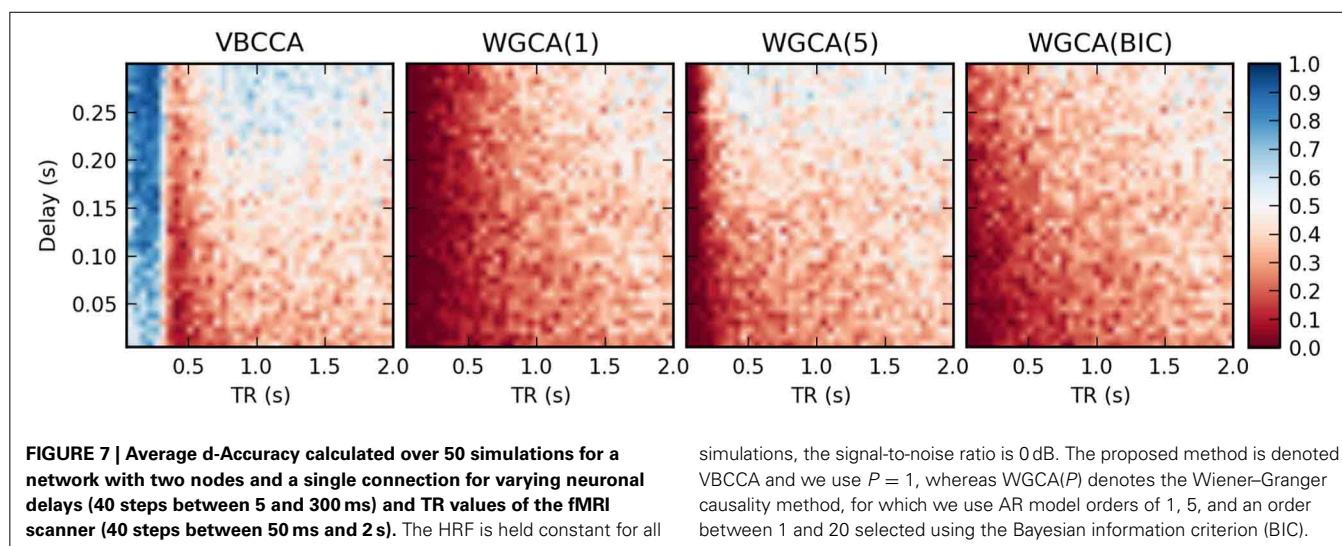


chance level (0.5) when TR is increased. The proposed method shows a similar behavior, but for TR values below 300 ms the d-Accuracy is close to 1.0. While it is difficult to assess the origin of this transition, it is likely caused by increased aliasing that occurs for larger TR values. Together with the consistent causality inversion of WGC for low TR values, it shows that causal information is still present in the data.

In the second simulation, we additionally introduce HRF variations. Results are shown in **Figure 8**. In this case, the proposed method performs poorly, even for low TR values, unless the method is provided with the true HRF for each region, in which case it can mitigate the effects of HRF variability. Somewhat surprisingly, WGCA(1) performs similarly as before when the same HRF was used for each region. However, when the BIC is used to determine the model order, the WGCA method exhibits low estimation performance for all TR and delay values. A possible explanation for this behavior is that due to the HRF convolution, the selected model order is higher than the true order and the order also depends on the HRF used (Seth et al., 2013), which results in spurious causality inversions and hence poor performance.

It is important to point out that our results should not be interpreted in the way that WGC with a fixed model order consistently estimates a causal influence with the opposite direction for low TR values; whether the inversion occurs is dependent on simulation parameters, e.g., the amount of autocorrelation in the simulated time series, the connection strength, and the signal-to-noise ratio. For example, when we repeat the first simulation with a higher signal-to-noise ratio of 20 dB, the results change drastically, as shown in **Figure 9**. The WGCA method now correctly estimates the direction of the influence except for low TR and delay values. In this case also the proposed method performs poorly for low delay values. These results show that while the proposed method performs better, especially in low-SNR situations, there is a risk of causality inversion for both methods. The superiority of the proposed method can be explained by the modeling, which explicitly takes additive noise into account. However, at the same time, both the proposed method and the WGCA method do not model the





non-linear downsampling operation and therefore can fail to correctly estimate the direction of the causal influence when the data has been downsampled.

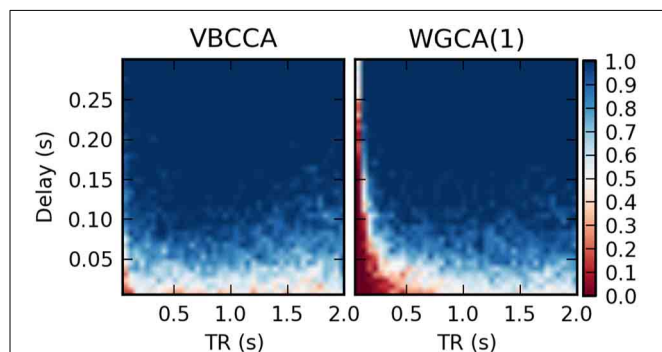
## 5. APPLICATION TO fMRI DATA

In this section, we apply the proposed method to resting-state fMRI data provided by the Human Connectome Project (HCP) (Van Essen et al., 2012). We use data from two 15 min runs of the same subject (100307), each consisting of 1200 volumes with a TR of 0.7 s. The minimally preprocessed volume data (Glasser et al., 2013) was aligned to the FreeSurfer (Fischl, 2012) “fsaverage” template and data from 148 cortical parcels from the Destrieux atlas (Destrieux et al., 2010) was extracted by averaging data across the gray matter at each vertex of the FreeSurfer surface mesh. In addition, we extracted volume data from six subcortical parcels (thalamus, caudate, putamen, pallidum, hippocampus, amygdala) for each hemisphere, resulting in a total of 160 parcels.

The extracted data was further preprocessed to reduce motion artifacts, slow drifts, and physiological artifacts. Specifically, we reduced motion artifacts and slow drifts using a linear regression for each voxel time series with three motion parameters and a cosine basis up to order 8 as nuisance regressors, where the order of the cosine basis was determined using the Bayesian Information Criterion (BIC) (Schwarz et al., 1978). To reduce physiological noise, we used a procedure similar to CompCor (Behzadi et al., 2007), i.e., we extracted data from the left and right lateral ventricles, which can be expected to not contain any signal of neuronal origin, applied the previously described detrending and motion artifact correction to it, and finally used a principal component analysis (PCA) to extract the 20 strongest temporal components. The extracted noise components were then used as nuisance regressors for each voxel time series where the number of components to use was determined using BIC. Finally, to obtain a single time series for each parcel, we computed a PCA

for the data within each parcel and retained the first principal component.

Connectivity matrices obtained by applying the proposed method and WGCA to the HCP data are shown in **Figure 10**. As a reference we also include the correlation coefficient, which is

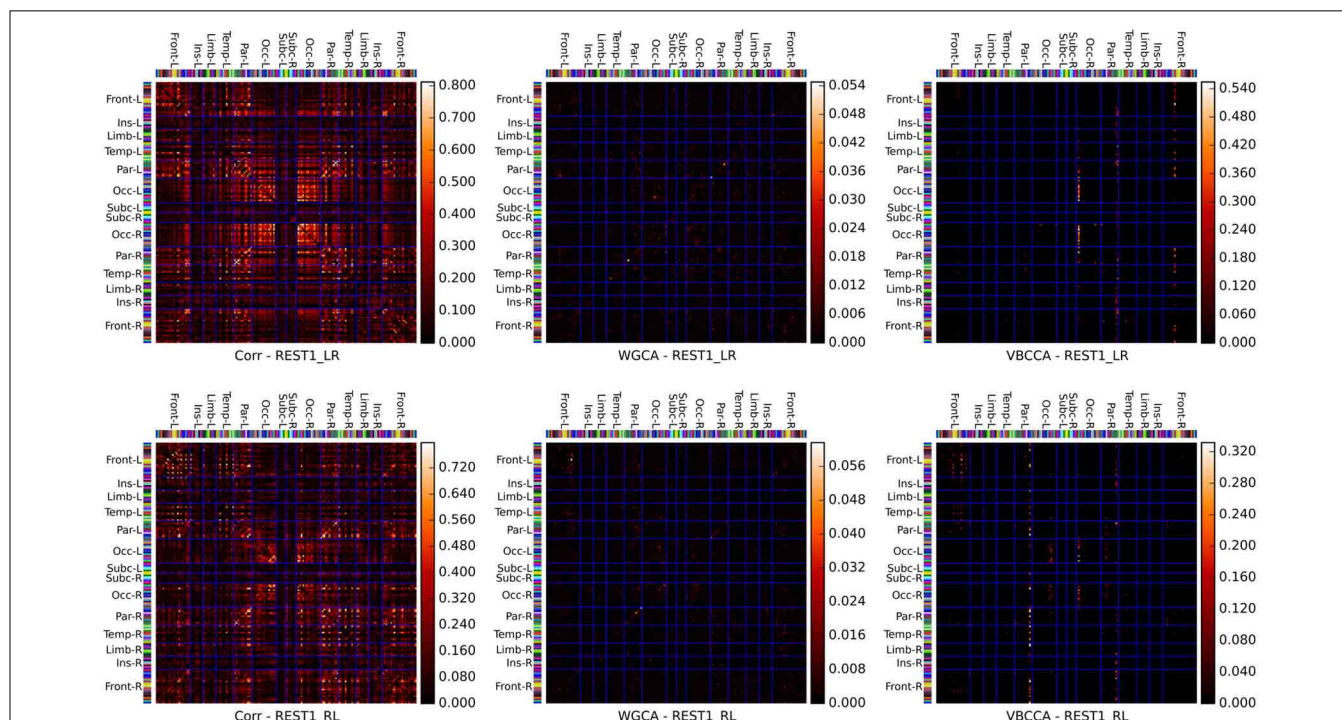


**FIGURE 9 |** Average d-Accuracy calculated over 50 simulations for a network with two nodes and a single connection for varying neuronal delays (40 steps between 5 and 300 ms) and TR values of the fMRI scanner (40 steps between 50 ms and 2 s). The HRF is held constant for all simulations, the signal-to-noise ratio is 20 dB. The proposed method is denoted VBCCA and we use  $P = 1$ , whereas WGCA(1) denotes the Wiener-Granger causality method, for which we also use AR model order of 1.

the most commonly used fMRI resting-state connectivity measure. All methods show some consistency across runs. For the proposed method and the second run, it can clearly be seen that the method finds connections between nodes that are commonly associated with resting-state networks. For example, nodes in the frontal cortices, the temporal lobes, and the parietal lobes, which are part of the default-mode network (Raichle et al., 2001). There is also strong bi-lateral connectivity between the left- and right occipital cortices, which are part of the visual resting-state network. Compared to correlation and WGCA, the VBCCA connectivity matrices are very sparse, which could indicate that there may not be enough causal information in the data to result in strong causality estimates, which would be a sensible explanation given the short propagation delays at the neuronal level and the still relatively slow sampling interval of 0.7 s. Finally, it is important to note that due to the methodological problems discussed in the previous section, it is possible that the direction of the causal influence is estimated incorrectly. The application to real fMRI data as presented here serves as a demonstration, further evaluations, e.g., using simultaneous EEG and fMRI data, are necessary to quantify the effectiveness of the proposed method when applied to real fMRI data.

## 6. CONCLUSIONS

In this paper we proposed a variational Bayesian causal connectivity method for fMRI. The method uses a VAR model for



**FIGURE 10 |** Connectivity matrices showing the absolute correlation coefficient (Corr), Wiener-Granger causality (WGCA), and causality estimated by the proposed method (VBCCA). We use the same parcel grouping and order as in Irimia et al. (2012), which groups the parcels into cortical lobes, i.e., frontal (Front), insular (Ins), limbic (Limb), temporal (Temp), parietal (Par), occipital (Occ), and subcortical (Subc). The “-L” and “-R”

suffixes indicate the left and right hemisphere, respectively. The parcel colors are the same as in the standard FreeSurfer color table. Results for the first run (REST1\_LR) and the second run (REST1\_RL) are in the top and bottom row, respectively. For WGCA and VBCCA, we use an VAR order of  $P = 1$  consistent with our simulations. For the proposed method we show  $\sqrt{c_{ij}}$  in order to better depict the estimated values within the scale of the color map.

the neuronal time series and the connectivity between regions in combination with a hemodynamic convolution model. By introducing an approximation to the neuronal time series and performing parts of the estimation in the frequency domain, our method is computationally efficient and can be applied to large scale problems with several hundred ROIs and high sampling rates.

We performed simulations with synthetic data to evaluate the performance of our method and to compare it with classical Wiener–Granger causality analysis (WGCA). There are several important findings from these simulations that need further discussion. In the first simulation, we demonstrated an important strength of our method, that is, it performs significantly better than WGCA when applied to problems with large numbers of regions. This effect is due to the use Gaussian priors for the VAR coefficients in combination with gamma priors for the precision hyperparameters. This prior has a regularizing effect by promoting sparsity for the VAR coefficients and can be seen as an adaptation of sparse Bayesian learning (Tipping, 2001) to the problem of VAR coefficient estimation. In contrast, WGCA does not use regularization for the VAR coefficients resulting in a performance degradation when the number of regions is increased. It is important to note that also the method in Ryali et al. (2011) employs Gaussian-gamma priors for the VAR coefficients. However, due to the computational complexity of the method it can only be applied to problems with small numbers of regions, where the prior is overwhelmed by the data and the sparsity promoting effect is of little benefit.

In the second set of simulations, we evaluated our method using simulated data generated by VAR processes of varying orders. Again, due to the prior for the VAR coefficients, where we group coefficients across lags together using shared precision hyperparameters, our method performed well as long as the VAR order used in the method is equal or higher than the VAR order of the data. A grouping of VAR coefficients using  $\ell_1\ell_2$ -norm regularization was first proposed in Haufe et al. (2008), in our work we propose a Bayesian formulation for this problem.

In the third simulation, we analyzed the effect of using an approximation to the neuronal time series, which is employed in our method to improve the computational efficiency, by comparing our method with a modified version of our method where the convolution with the HRF is included in the observation matrix of the linear dynamic system, as in previous methods (Smith et al., 2010; Ryali et al., 2011). The simulation results show that the approximation leads to some reduction in the quality of the estimated neuronal signal in terms of mean-squared error (MSE) but does not have a significant influence on the connectivity estimation performance. Importantly, the reduction in computational complexity resulting from the use of the approximation to the neuronal signal allows us to apply the method to large scale problems. As discussed above, the sparsity promoting priors for the VAR coefficients are of crucial importance when the method is applied to problems with large numbers of regions. The use of the approximation to the neuronal time series is therefore an important contribution of this work, as it allows us to apply the method to problem sizes where the method can benefit from the regularizing effect of the priors.

In a last set of simulations, we analyzed the effect of different downsampling ratios, simulating different TR values of the MRI scanner, the neuronal delay, and HRF variability. Perhaps not surprisingly, the proposed method is immune to HRF variability if it has access to the true HRF of each region. Clearly, in practice HRFs are subject and region dependent. However, it has been shown that HRFs are strongly correlated across subjects and regions (Handwerker et al., 2004). Hence, using data from a large number of subjects, it may be possible to construct a model describing the relationship between the HRFs in various brain regions. This “hemodynamic atlas” could then be used to approximate the HRFs in a large number of regions from a small number of estimated HRFs for each subject. We also found that the proposed method generally performs better than WGC when a significant amount of additive noise is present in the data. This finding is consistent with previous results (Seth et al., 2013) and can be explained by the model used in the proposed method which can account for additive noise. However, while the proposed method offers some benefits over WGC, we find that also the proposed method can estimate a causal influence with the opposite direction when the data has been downsampled, which is a known problem with WGC methods (David et al., 2008; Deshpande et al., 2010; Seth et al., 2013). The problem that causality estimated using a discrete-time VAR model from a sampled continuous-time VAR process can lead to opposite conclusions has been shown before (Cox, 1992). Unfortunately, this problem has received little attention in recent work on causality estimation from fMRI data, where severe downsampling is common. In Solo (2007), it is shown that while causality can be preserved under downsampling, VAR models, as used in traditional WGC analysis and the proposed method, are inadequate for estimating causality from the subsampled time series and either VAR moving average (VARMA) models or state-space (SS) models are required to correctly estimate the direction of the causal influence. This raises hopes that causality estimation from fMRI may be feasible by applying more sophisticated models to data acquired with low TR values, which may be achieved using a combination of novel acquisition sequences and MRI scanners with higher field strengths. Clearly, HRF variability will still be a problem but under certain conditions it may be possible to use a model similar to the one proposed in this work which can take into account the HRF of each region.

Finally, we applied the proposed method to real resting-state fMRI data provided by the Human Connectome Project (Van Essen et al., 2012). For this data, the proposed method finds connections between regions that are associated with known resting-state networks. However, it is important to emphasize that application to real fMRI data as presented here serves as a demonstration to show that the proposed method can be applied to real fMRI data. As the true causal relationships in real data are not known, it is not possible to determine whether the direction of causal influence is correctly estimated. As shown in our simulations, there are methodological problems which, depending on the noise level, the HRF, the TR, and the neuronal delay, can lead to causality inversions. Further experiments, e.g., using simultaneous EEG and fMRI, are necessary to quantify the effectiveness



of the proposed method to estimate the direction of the causal influence from real fMRI data.

## FUNDING

This work was partially supported by the National Institute of Child Health and Human Development (R01 HD042049). Martin Luessi was partially supported by the Swiss National Science Foundation Early Postdoc Mobility fellowship 148485. This work was supported in part by the Department of Energy under Contract DE-NA0000457, the “Ministerio de Ciencia e Innovación” under Contract TIN2010-15137, and the CEI BioTic with the Universidad de Granada Data were provided (in part) by the Human Connectome Project, WU-Minn Consortium (Principal Investigators: David Van Essen and Kamil Ugurbil; 1U54MH091657) funded by the 16 NIH Institutes and Centers that support the NIH Blueprint for Neuroscience Research; and by the McDonnell Center for Systems Neuroscience at Washington University.

## REFERENCES

- Attias, H. (2000). A variational Bayesian framework for graphical models. *Adv. Neural Inform. Process. Syst.* 12, 209–215.
- Babacan, S. D., Luessi, M., Molina, R., and Katsaggelos, A. K. (2012). Sparse bayesian methods for low-rank matrix estimation. *IEEE Trans. Signal Process.* 60, 3964–3977. doi: 10.1109/TSP.2012.2197748
- Barnett, L., and Seth, A. K. (2011). Behaviour of Granger causality under filtering: theoretical invariance and practical application. *J. Neurosci. Methods* 201, 404–419. doi: 10.1016/j.jneumeth.2011.08.010
- Beal, M. J., and Ghahramani, Z. (2001). *The Variational Kalman Smoother*. Technical Report GCNU TR 2001-003, Gatsby Computational Neuroscience Unit.
- Behzadi, Y., Restom, K., Liu, J., and Liu, T. T. (2007). A component based noise correction method (compcor) for {BOLD} and perfusion based fMRI. *Neuroimage* 37, 90–101. doi: 10.1016/j.neuroimage.2007.04.042
- Bishop, C. M. (2006). *Pattern Recognition and Machine Learning*. New York, NY: Springer.
- Bressler, S. L., and Seth, A. K. (2010). Wiener-Granger causality: a well established methodology. *Neuroimage* 58, 323–329. doi: 10.1016/j.neuroimage.2010.02.059
- Buxton, R. B., Uludag, K., Dubowitz, D. J., and Liu, T. T. (2004). Modeling the hemodynamic response to brain activation. *Neuroimage* 23, S220–S233. doi: 10.1016/j.neuroimage.2004.07.013
- Buxton, R. B., Wong, E. C., and Frank, L. R. (1998). Dynamics of blood flow and oxygenation changes during brain activation: the balloon model. *Magn. Reson. Med.* 39, 855–864. doi: 10.1002/mrm.1910390602
- Cassidy, B., Long, C., Rae, C., and Solo, V. (2012). Identifying fMRI model violations with lagrange multiplier tests. *IEEE Trans. Med. Imaging* 31, 1481–1492. doi: 10.1109/TMI.2012.2195327
- Chari, L., Vincent, T., Forbes, F., Dojat, M., and Ciuciu, P. (2013). Fast joint detection-estimation of evoked brain activity in event-related fmri using a variational approach. *IEEE Trans. Med. Imaging* 32, 821–837. doi: 10.1109/TMI.2012.2225636
- Cox, D. R. (1992). Causality: some statistical aspects. *J. R. Stat. Soc. A*, 291–301. doi: 10.2307/2982962
- David, O., Guillemain, I., and Saillet (2008). Identifying neural drivers with functional mri: an electrophysiological validation. *PLoS Biol.* 6:e315. doi: 10.1371/journal.pbio.0060315
- Deshpande, G., Sathian, K., and Hu, X. (2010). Effect of hemodynamic variability on granger causality analysis of fMRI. *Neuroimage* 52, 884–896. doi: 10.1016/j.neuroimage.2009.11.060
- Destrieux, C., Fischl, B., Dale, A., and Hagren, E. (2010). Automatic parcellation of human cortical gyri and sulci using standard anatomical nomenclature. *Neuroimage* 53, 1–15. doi: 10.1016/j.neuroimage.2010.06.010
- Fawcett, T. (2006). An introduction to roc analysis. *Pattern Recogn. Lett.* 27, 861–874. doi: 10.1016/j.patrec.2005.10.010
- Fernandes, P., Plateau, B., and Stewart, W. J. (1998). Efficient descriptor-vector multiplications in stochastic automata networks. *J. ACM (JACM)* 45, 381–414. doi: 10.1145/278298.278303
- Fischl, B. (2012). *Freesurfer*. *Neuroimage* 62, 774–781. doi: 10.1016/j.neuroimage.2012.01.021
- Frahm, J., Bruhn, H., Merboldt, K. D., and Math, D. (1992). Dynamic MR imaging of human brain oxygenation during rest and photic stimulation. *J. Magn. Reson. Imaging* 2, 501–505. doi: 10.1002/jmri.1880020505
- Friston, K. J. (1994). Functional and effective connectivity in neuroimaging: a synthesis. *Hum. Brain Mapping* 2, 56–78. doi: 10.1002/hbm.460020107
- Friston, K. J., Harrison, L., and Penny, W. (2003). Dynamic causal modelling. *Neuroimage* 19, 1273–1302. doi: 10.1016/S1053-8119(03)00202-7
- Friston, K. J., Holmes, A. P., Poline, J. B., Grasby, P. J., Williams, S. C., Frackowiak, R. S., and Turner, R. (1995). Analysis of fMRI time-series revisited. *Neuroimage* 2, 45–53. doi: 10.1006/nimg.1995.1007
- Ghahramani, Z., and Beal, M. J. (2001). Propagation algorithms for variational Bayesian learning. *Adv. Neural Inform. Process. Syst.* 13, 507–513.
- Ghahramani, Z., and Hinton, G. E. (1996). *Parameter Estimation for Linear Dynamical Systems*. University of Toronto technical report CRG-TR-96-2, 6.
- Glasser, M. F., Sotiropoulos, S. N., Wilson, J. A., Coalson, T. S., Fischl, B., Andersson, J. L., et al. (2013). The minimal preprocessing pipelines for the human connectome project. *Neuroimage* 80, 105–124. doi: 10.1016/j.neuroimage.2013.04.127
- Granger, C. (1969). Investigating causal relations by econometric models and cross-spectral methods. *Econ. J. Econ. Soc.* 37, 424–438.
- Handwerker, D. A., Ollinger, J. M., and D’Esposito, M. (2004). Variation of bold hemodynamic responses across subjects and brain regions and their effects on statistical analyses. *Neuroimage* 21, 1639–1651. doi: 10.1016/j.neuroimage.2003.11.029
- Haufe, S., Müller, K. R., Nolte, G., and Krämer, N. (2008). “Sparse causal discovery in multivariate time series,” in *NIPS Workshop on Causality*, (Whistler).
- Irimia, A., Chambers, M. C., Torgerson, C. M., Filippou, M., Hovda, D. A., Alger, J. R., et al. (2012). Patient-tailored connectomics visualization for the assessment of white matter atrophy in traumatic brain injury. *Front. Neurol.* 3:10. doi: 10.3389/fneur.2012.00010
- Jordan, M. I., Ghahramani, Z., Jaakkola, T. S., and Saul, L. K. (1999). An introduction to variational methods for graphical models. *Mach. Learn.* 37, 183–233. doi: 10.1023/A:1007665907178
- Luessi, M. (2011). *Bayesian Approaches to Inverse Problems in Functional Neuroimaging*. Ph.D. thesis, Northwestern University.
- Makni, S., Beckmann, C., Smith, S., and Woolrich, M. (2008). Bayesian deconvolution fMRI data using bilinear dynamical systems. *Neuroimage* 42, 1381–1396. doi: 10.1016/j.neuroimage.2008.05.052
- Marinazzo, D., Liao, W., Chen, H., and Stramaglia, S. (2011). Nonlinear connectivity by granger causality. *Neuroimage* 58, 330–338. doi: 10.1016/j.neuroimage.2010.01.099
- Meier, L., Geer, S., and Bühlmann, P. (2008). The group lasso for logistic regression. *J. R. Stat. Soc. Ser. B* 70, 53–71. doi: 10.1111/j.1467-9868.2007.00627.x
- Moon, T. K., and Stirling, W. C. (2000). *Mathematical Methods and Algorithms for Signal Processing*, Vol. 204. New York, NY: Prentice Hall.
- Ogawa, S., Lee, T., Kay, A., and Tank, D. (1990). Brain magnetic resonance imaging with contrast dependent on blood oxygenation. *Proc. Natl. Acad. Sci. U.S.A.* 87, 9868. doi: 10.1073/pnas.87.24.9868
- Penny, W., Ghahramani, Z., and Friston, K. (2005). Bilinear dynamical systems. *Philos. Trans. R. Soc. B Biol. Sci.* 360, 983. doi: 10.1098/rstb.2005.1642
- Raichle, M. E., MacLeod, A. M., Snyder, A. Z., Powers, W. J., Gusnard, D. A., and Shulman, G. L. (2001). A default mode of brain function. *Proc. Natl. Acad. Sci. U.S.A.* 98, 676–682. doi: 10.1073/pnas.98.2.676
- Rauch, H. E., Tung, F., and Striebel, C. (1965). Maximum likelihood estimates of linear dynamic systems. *AIAA J.* 3, 1445–1450. doi: 10.2514/3.3166
- Roebroeck, A., Seth, A. K., and Valdes-Sosa, P. (2011). Causal time series analysis of functional magnetic resonance imaging data. *J. Mach. Learn. Res. (Workshop and Conference Proceedings. Causality in Time Series)* 12, 65–94.
- Ryali, S., Supekar, K., Chen, T., and Menon, V. (2011). Multivariate dynamical systems models for estimating causal interactions in fMRI. *Neuroimage* 54, 807–823. doi: 10.1016/j.neuroimage.2010.09.052
- Schwarz, G. et al. (1978). Estimating the dimension of a model. *Annal. Stat.* 6, 461–464.



- Seth, A. K. (2010). A matlab toolbox for granger causal connectivity analysis. *J. Neurosci. Methods* 186, 262–273. doi: 10.1016/j.neuroimage.2010.08.063
- Seth, A. K., Chorley, P., and Barnett, L. C. (2013). Granger causality analysis of fMRI BOLD signals is invariant to hemodynamic convolution but not downsampling. *Neuroimage* 65, 540–555. doi: 10.1016/j.neuroimage.2012.09.049
- Smith, J. F., Pillai, A., Chen, K., and Horwitz, B. (2010). Identification and validation of effective connectivity networks in functional magnetic resonance imaging using switching linear dynamic systems. *Neuroimage* 52, 1027–1040. doi: 10.1016/j.neuroimage.2009.11.081
- Smith, S. M., Miller, K. L., Salimi-Khorshidi, G., Webster, M., Beckmann, C. F., Nichols, T. E., et al. (2011). Network modelling methods for fMRI. *Neuroimage* 54, 875–891. doi: 10.1016/j.neuroimage.2010.08.063
- Solo, V. (2007). “On causality I: sampling and noise,” 2007 46th IEEE Conference on Decision and Control (New Orleans, LA), 3634–3639. doi: 10.1109/CDC.2007.4434049
- Tipping, M. E. (2001). Sparse bayesian learning and the relevance vector machine. *J. Mach. Learn. Res.* 1, 211–244.
- Valdés-Sosa, P. A., Sánchez-Bornot, J. M., Lage-Castellanos, A., Vega-Hernández, M., Bosch-Bayard, J., Melie-García, L., et al. (2005). Estimating brain functional connectivity with sparse multivariate autoregression. *Philos. Trans. R. Soc. B Biol. Sci.* 360, 969. doi: 10.1098/rstb.2005.1654
- Van Essen, D. C., Ugurbil, K., Auerbach, E., Barch, D., Behrens, T. E. J., Bucholz, R., et al. (2012). The human connectome project: a data acquisition perspective. *Neuroimage* 62, 2222–2231. doi: 10.1016/j.neuroimage.2012.02.018
- Wang, B., and Titterton, D. (2004). Lack of consistency of mean field and variational bayes approximations for state space models. *Neural Process. Lett.* 20, 151–170. doi: 10.1007/s11063-004-2024-6
- Weigend, A. S., and Gershenfeld, N. A. (1994). *Time Series Prediction: Forecasting the Future and Understanding the Past*. Reading, MA: Addison-Wesley.
- Wiener, N. (1956). *The theory of prediction*. Modern mathematics for engineers. New York, NY: McGraw-Hill.
- Wipf, D. P., and Rao, B. D. (2007). An Empirical Bayesian Strategy for Solving the Simultaneous Sparse Approximation Problem. *IEEE Trans. Signal Process.* 55, 3704–3716. doi: 10.1109/TSP.2007.894265
- Yuan, M., and Lin, Y. (2006). Model selection and estimation in regression with grouped variables. *J. R. Stat. Soc. Ser. B (Statistical Methodology)* 68, 49–67. doi: 10.1111/j.1467-9868.2005.00532.x

**Conflict of Interest Statement:** The authors declare that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.

Received: 05 January 2014; accepted: 01 April 2014; published online: 05 May 2014.

Citation: Luessi M, Babacan SD, Molina R, Booth JR and Katsaggelos AK (2014) Variational Bayesian causal connectivity analysis for fMRI. *Front. Neuroinform.* 8:45. doi: 10.3389/fninf.2014.00045

This article was submitted to the journal *Frontiers in Neuroinformatics*.

Copyright © 2014 Luessi, Babacan, Molina, Booth and Katsaggelos. This is an open-access article distributed under the terms of the Creative Commons Attribution License (CC BY). The use, distribution or reproduction in other forums is permitted, provided the original author(s) or licensor are credited and that the original publication in this journal is cited, in accordance with accepted academic practice. No use, distribution or reproduction is permitted which does not comply with these terms.



# Canonical information flow decomposition among neural structure subsets

Daniel Y. Takahashi<sup>1</sup>, Luiz A. Baccalá<sup>2\*</sup> and Koichi Sameshima<sup>3</sup>

<sup>1</sup> Psychology Department, Neuroscience Institute, Princeton University, Princeton, NJ, USA

<sup>2</sup> Telecommunications and Control Department, Escola Politécnica, University of São Paulo, São Paulo, Brazil

<sup>3</sup> Department of Radiology and Oncology, Faculdade de Medicina, University of São Paulo, São Paulo, Brazil

## Edited by:

Daniele Marinazzo, University of Gent, Belgium

## Reviewed by:

Katerina Hlavackova-Schindler, University of Life Sciences, Austria  
Angeliki Papana, University of Macedonia, Greece

## \*Correspondence:

Luiz A. Baccalá,  
Telecommunications and Control  
Department, Escola Politécnica,  
University of São Paulo, Av. Prof.  
Luciano Gualberto, trav. 3, #158,  
São Paulo, SP 05508-900, Brazil  
e-mail: baccala@lcs.poli.usp.br

Partial directed coherence (PDC) and directed coherence (DC) which describe complementary aspects of the directed information flow between pairs of univariate components that belong to a vector of simultaneously observed time series have recently been generalized as bPDC/bDC, respectively, to portray the relationship between subsets of component vectors (Takahashi, 2009; Faes and Nollo, 2013). This generalization is specially important for neuroscience applications as one often wishes to address the link between the set of time series from an observed ROI (region of interest) with respect to series from some other physiologically relevant ROI. bPDC/bDC are limited, however, in that several time series within a given subset may be irrelevant or may even interact opposingly with respect to one another leading to interpretation difficulties. To address this, we propose an alternative measure, termed cPDC/cDC, employing canonical decomposition to reveal the main frequency domain modes of interaction between the vector subsets. We also show bPDC/bDC and cPDC/cDC are related and possess mutual information rate interpretations. Numerical examples and a real data set illustrate the concepts. The present contribution provides what is seemingly the first canonical decomposition of information flow in the frequency domain.

**Keywords:** directed connectivity measures, canonical decomposition, frequency domain, information flow, generalized coherence

## 1. INTRODUCTION

Human behavior is primarily thought as a property that emerges from the interaction of several brain areas, body parts, and the environment. Understanding how these elements dynamically interact is one of major themes of systems neuroscience. Several multivariate time series methods—old and new—have been introduced to describe the interdependence between brain areas using signal modalities like EEG, BOLD signals, MEG and LFP—and are collectively called connectivity measures. *Partial directed coherence* (PDC) (Baccalá and Sameshima, 2001) and *directed coherence/directed transfer function* (DC/DTF) (Kamiński and Blinowska, 1991) are two examples of such connectivity measures. Both describe complementary aspects (see Baccalá and Sameshima, 2014 for an in depth discussion) of how information flows between pairs of univariate time series components that belong to a multivariate vector of simultaneously observed time series (Takahashi et al., 2010). Recently, PDC and DC have been generalized (as bPDC/bDC, respectively) to describe how subsets (blocks) of components within a time series vector interrelate (Takahashi, 2009; Faes and Nollo, 2013). This is specially important for neuroscience applications as one often wants to investigate the interaction between sets of time series that are circumscribed to an observed region of interest (ROI) with respect to another physiologically relevant ROI (Nedungadi et al., 2011). The potential relevance of this type of question alone justifies looking for their deeper meaning in terms of information theoretical quantities.

Despite their practical importance, bPDC/bDC suffer from the limitation that several time series within a given subset may be irrelevant or interact in opposition to one another thereby posing interpretation difficulties. Also, in several situations, a researcher may be interested in just the few “best” descriptions of interaction between two sets of time series but not in the total amount of information flowing between them. For a more concrete example, assume that two brain areas interact and that bPDC is large. In this situation, it does not straightforwardly follow that all brain region components are interacting in the same way, or even whether some such components may be ignored. One way to address this limitation is to decompose bPDC/bDC into different components weighed according to relevance.

The aim of this article is twofold: (a) to provide a proper information theoretic interpretation for bPDC/bDC and (b) to introduce a canonical decomposition of information flows, henceforth termed, respectively, canonical PDC/DC (cPDC/cDC). These new decompositions allow us to closely mimic classical canonical correlation analysis so that different dynamically relevant interaction modes between brain areas can be exposed. Due to PDC interpretability in terms of Granger causality (Baccalá and Sameshima, 2014), a consequence of the present formulation is that cPDC represents a long sought frequency domain counterpart to time domain canonical decompositions of Granger causality (Sato et al., 2010; Ashrafulla et al., 2013).

The article is organized as follows. We first introduce the background and notation necessary for the rest of the article

(section 2). In the results section (section 3), we first show that both bPDC and bDC between two subsets of processes are block coherences between suitably defined underlying processes. Then, we demonstrate that such coherences are nothing but monotonic transformations of the mutual information rate between the respective processes (Gelfand and Yaglom, 1959; Takahashi et al., 2010; Nedungadi et al., 2011) leading immediately to their interpretability as mutual information rates. Next, we introduce cPDC and cDC and prove that they are the non-zero eigenvalues of the matrices whose determinants underlie the respective bPDC and bDC definitions (section 4). Using simulated examples and publicly available data we illustrate the usefulness of cPDC/cDC (section 5) followed by a brief discussion (section 6). Proof details are left to the Appendix.

## 2. BACKGROUND

Let  $X_1, \dots, X_K$  be  $K$  distinct multivariate time series vectors with dimension  $M_1, \dots, M_K$ . Using  $T$  to indicate matrix transposition, let  $X(t) = [X_1(t)^T, \dots, X_K(t)^T]^T$  for each time  $t \in \mathbb{Z}$  be a second order stationary time series with spectral density matrix  $S(\omega)$  at each frequency  $\omega \in [-\pi, \pi)$ . To justify our formal computation, we assume that  $S(\omega)$  is uniformly bounded from below and above and invertible at all frequencies (Hannan, 1970). This is called the *boundedness* condition which guarantees that the following autoregressive (AR) representation of  $X$  holds in the mean square sense

$$X(t) = \sum_{l=1}^{+\infty} A(l)X(t-l) + \epsilon(t), \quad (1)$$

where  $\epsilon(t) = [\epsilon_1(t)^T \dots \epsilon_K(t)^T]^T$  stands for a zero mean innovation process, i.e.,  $\mathbb{E}[\epsilon(t)\epsilon(t)^T] = \Sigma$  and  $\mathbb{E}[\epsilon(t)\epsilon(l)^T] = 0$  for  $l \neq t$ . For  $l \geq 1$ ,  $A(l)$  are  $(M_1 + \dots + M_K)^2$ -dimensional matrices. Let  $A_{pq}(l)$  for  $p, q \in \{1, \dots, K\}$  and  $l \geq 1$  be  $M_p \times M_q$ -dimensional matrices so that  $A(l)$  has the following structure

$$A(l) = \begin{bmatrix} A_{11}(l) & \dots & A_{1M}(l) \\ \vdots & \ddots & \vdots \\ A_{M1}(l) & \dots & A_{MM}(l) \end{bmatrix}$$

We define  $\bar{A}(\omega) = I - \sum_{l \geq 1} A(l)e^{-\sqrt{-1}\omega l}$ .

Under the boundedness condition, the following moving average (MA) mean square sense representation for the process  $X$  also holds

$$X(t) = \sum_{l=0}^{+\infty} H(l)\epsilon(t-l), \quad (2)$$

where  $H(l)$  for  $l \geq 0$  are  $(M_1 + \dots + M_K)^2$ -dimensional matrices. Let  $\bar{H}(\omega) = \sum_{l \geq 0} H(l)e^{-\sqrt{-1}\omega l}$ . We have that  $\bar{A}^*(\omega) = \bar{H}^{-1}(\omega)$  for all  $\omega \in [-\pi, \pi)$ . The superscript  $*$  indicates the matrix complex conjugate.

Let  $P(\omega) = S^{-1}(\omega)$ . bPDC from the multivariate process  $X_j$  to the process  $X_i$  at frequency  $\omega$ , denoted  $\pi_{ij}^{(b)}(\omega)$ , is defined

(Takahashi, 2009; Faes and Nollo, 2013) by

$$\pi_{ij}^{(b)}(\omega) = 1 - \det(P_{jj}(\omega) - \bar{A}_{jj}^*(\omega)\Sigma_{ii}^{-1}\bar{A}_{ij}(\omega)) \det(P_{jj}(\omega))^{-1}, \quad (3)$$

where  $\det$  indicates the determinant and the subscript indices relate to the natural block structure associated with the matrices.

Let  $\Theta = \Sigma^{-1}$ . bDC from the multivariate process  $X_j$  to the process  $X_i$  at frequency  $\omega$ , denoted  $\gamma_{ij}^{(b)}(\omega)$ , is defined (Takahashi, 2009; Faes and Nollo, 2013) by

$$\gamma_{ij}^{(b)}(\omega) = 1 - \det(S_{ii}(\omega) - \bar{H}_{ij}(\omega)\Theta_{jj}^{-1}\bar{H}_{ij}^*(\omega)) \det(S_{ii}(\omega))^{-1}. \quad (4)$$

Note that the present bDC definition differs slightly from the one in Faes and Nollo (2013). We removed the unnecessary condition of strict causality, i.e., diagonality of  $\Sigma$ , simply by substituting  $\Sigma_{jj}^{-1}$  by  $\Theta_{jj}^{-1}$  in their definition of bDC as it is more suited for formulating information theoretic results as shown ahead.

Consider a second-order stationary multivariate process  $W(t) = [Y(t)^T Z(t)^T]^T$ . The block coherence between  $Y$  and  $Z$  at frequency  $\omega$  is defined as (Nedungadi et al., 2011)

$$C_{YZ}^{(b)}(\omega) = 1 - \det(S_{WW}(\omega)) \det(S_{YY}(\omega))^{-1} \det(S_{ZZ}(\omega))^{-1}. \quad (5)$$

Observe that we used the process name in the subscript of the power spectrum  $S$  to indicate the corresponding spectral density matrices. In the rest of the article, we will use interchangeably the process name or the corresponding indices in the subscript whenever there is no ambiguity.

Another important definition is that of mutual information rate (MIR) between two multivariate strictly stationary processes  $Y$  and  $Z$  is

$$\text{MIR}_{YZ} = \lim_{t \rightarrow +\infty} \frac{1}{t} \mathbb{E} \left[ \log \frac{d\mathbb{P}(Y(1), \dots, Y(t), Z(1), \dots, Z(t))}{d\mathbb{P}(Y(1), \dots, Y(t))d\mathbb{P}(Z(1), \dots, Z(t))} \right]. \quad (6)$$

The classical relationship between block coherence (Equation 5) and mutual information rate (Equation 6) follows from

**Theorem.** (Gelfand and Yaglom, 1959; Pinsker, 1964) *If  $Y$  and  $Z$  are jointly stationary Gaussian processes satisfying the boundedness condition, we have that the MIR between  $Y$  and  $Z$  is given by*

$$\text{MIR}_{YZ} = -\frac{1}{4\pi} \int_{-\pi}^{\pi} \log(1 - C_{YZ}^{(b)}(\omega)) d\omega. \quad (7)$$

Now, following Takahashi et al. (2010), we define, for  $i \in \{1, \dots, K\}$ , the partialized process  $\eta_i$  by

$$\eta_i(t) = X_i(t) - \mathbb{E}[X_i(t) | \{X_j(l), j \neq i, l \in \mathbb{Z}\}], \quad (8)$$

where  $\mathbb{E}[\odot | \odot]$  henceforth denotes the best linear conditional predictor. Likewise the partialized innovation process  $\zeta_i$  for  $i \in \{1, \dots, K\}$  is

$$\zeta_i(t) = \epsilon_i(t) - \mathbb{E}[\epsilon_i(t) | \{\epsilon_j(t), j \neq i\}]. \quad (9)$$

Observe that both partialized process and partialized innovation process were defined in Takahashi et al. (2010) but for the special case of scalar  $\eta_i$  and  $\xi_i$ .

### 3. RELATION BETWEEN bPDC/bDC AND MUTUAL INFORMATION RATE

Our first result establishes the relationship between bPDC and block coherence and is analogous to Theorem 1 in Takahashi et al. (2010).

**Theorem 1.** *Let  $X$  satisfy the boundedness condition. For all  $i, j \in \{1, \dots, K\}$  and all frequencies  $\omega \in [-\pi, \pi]$  we have that*

$$\pi_{ij}^{(b)}(\omega) = C_{\epsilon_i \eta_j}^{(b)}(\omega). \quad (10)$$

A straightforward corollary is

**Corollary 1.** *Let  $X$  be a stationary Gaussian process and satisfy the boundedness condition. For all  $i, j \in \{1, \dots, K\}$  we have that*

$$\text{MIR}_{\epsilon_i \eta_j} = -\frac{1}{4\pi} \int_{-\pi}^{\pi} \log(1 - \pi_{ij}^{(b)}(\omega)) d\omega. \quad (11)$$

Similar results also hold for bDC.

**Theorem 2.** *Let  $X$  satisfy the boundedness condition. For all  $i, j \in \{1, \dots, K\}$  and all frequencies  $\omega \in [-\pi, \pi]$  we have that*

$$\gamma_{ij}^{(b)}(\omega) = C_{X_i \xi_j}^{(b)}(\omega). \quad (12)$$

and

**Corollary 2.** *Let  $X$  be a stationary Gaussian process and satisfy the boundedness condition. For all  $i, j \in \{1, \dots, K\}$ , we have that*

$$\text{MIR}_{X_i \xi_j} = -\frac{1}{4\pi} \int_{-\pi}^{\pi} \log(1 - \gamma_{ij}^{(b)}(\omega)) d\omega. \quad (13)$$

### 4. CANONICAL PDC AND DC

Canonical correlation is a classical method developed initially by Hotelling (1936) to address the relationship between random vectors. Brillinger (1981) generalized the method for time series and gave an excellent account of the relationship between canonical correlation analysis and different ideas in multivariate statistics. Our formulation of canonical coherence is equivalent to the definition introduced by Brillinger (1981).

Let  $Y$  and  $Z$  be respectively  $M_Y$ - and  $M_Z$ -dimensional jointly second order stationary processes. To better understand the relationship between  $Y$  and  $Z$ , we can ask the following question: Which components of  $Y$  and  $Z$  are most representative of the interaction between the processes? One way to formalize this is to consider filtering matrices  $B_Y(l)$  ( $1 \times M_Y$ ) and  $B_Z(l)$  ( $1 \times M_Z$ ), for all  $l \in \mathbb{Z}$  and define the scalar processes  $b_Y$  and  $b_Z$  by

$$b_Y(t) = \sum_{l \in \mathbb{Z}} B_Y(l) Y(t-l) \quad (14)$$

and

$$b_Z(t) = \sum_{l \in \mathbb{Z}} B_Z(l) Z(t-l), \quad (15)$$

so that  $C_{b_Y b_Z}(\omega)$  is maximized for all  $\omega \in [-\pi, \pi]$ . If furthermore  $Y$  and  $Z$  are jointly stationary Gaussian processes, then this is equivalent to maximizing  $\text{MIR}_{b_Y b_Z}$ .

Following the above idea, we define the first canonical coherence between  $Y$  and  $Z$  at frequency  $\omega$  by

$$C_{YZ}^{(c_1)}(\omega) = \sup_{B_Y, B_Z} C_{b_Y b_Z}(\omega). \quad (16)$$

Assume that the supremum (Equation 16) is achieved for  $\bar{b}_Y$  and  $\bar{b}_Z$ , which we call *first canonical time series*. Consider the residual processes  $Y^1(t) = Y(t) - \mathbb{E}[Y(t) | \{\bar{b}_Y(l), l \in \mathbb{Z}\}]$  and  $Z^1(t) = Z(t) - \mathbb{E}[Z(t) | \{\bar{b}_Z(l), l \in \mathbb{Z}\}]$ . Observe that  $Y^1$  and  $Z^1$  are uncorrelated to the processes  $\bar{b}_Y$  and  $\bar{b}_Z$ , respectively. The second canonical coherence  $C_{YZ}^{(c_2)}(\omega)$  is defined recursively on the residues by  $C_{YZ}^{(c_2)}(\omega) = C_{Y^1 Z^1}^{(c_1)}(\omega)$ .

Analogously, for  $2 \leq m \leq \min\{M_Y, M_Z\}$ , considering the residual processes

$$Y^m(t) = Y^{m-1}(t) - \mathbb{E}[Y^{m-1}(t) | \{\bar{b}_{Y^k}(l), l \in \mathbb{Z}, k \in \{1, \dots, m-1\}\}]$$

and

$$Z^m(t) = Z^{m-1}(t) - \mathbb{E}[Z^{m-1}(t) | \{\bar{b}_{Z^k}(l), l \in \mathbb{Z}, k \in \{1, \dots, m-1\}\}],$$

one may define the  $m$ -th canonical coherence as

$$C_{YZ}^{(c_m)}(\omega) = C_{Y^{m-1} Z^{m-1}}^{(c_1)}(\omega). \quad (17)$$

In this way, it is possible to construct a hierarchy of coherences where each element captures the dependence structure that is not explained by the other elements.

Finally, we introduce cPDC and cDC. For  $m \leq \min\{M_i, M_j\}$ , the  $m$ -th canonical PDC from  $j$  to  $i$  at frequency  $\omega$  denoted  $\pi_{ij}^{(c_m)}(\omega)$  is defined by

$$\pi_{ij}^{(c_m)}(\omega) = C_{\epsilon_i \eta_j}^{(c_m)}(\omega). \quad (18)$$

Similarly, the  $m$ -th canonical DC from  $j$  to  $i$  at frequency  $\omega$  denoted  $\gamma_{ij}^{(c_m)}(\omega)$  is defined by

$$\gamma_{ij}^{(c_m)}(\omega) = C_{X_i \xi_j}^{(c_m)}(\omega). \quad (19)$$

At first sight, it is unclear whether the canonical PDC and DC exist at all or even if they are uniquely defined. More importantly, nor is it obvious that it is possible to compute them. Despite these initial uncertainties, we show next that canonical coherences are consistently defined as the non-null eigenvalues of some specific matrices.



Let  $\lambda^m(Q)$  denote its  $m$ -th eigenvalue from matrix  $Q$  ordered from its largest to its smallest value. The following theorem furnishes a practical way to calculate cPDC and cDC.

**Theorem 3.** Under the boundedness condition for  $X$  the following identities hold:

$$\pi_{ij}^{(c_m)}(\omega) = \lambda^m \left( \tilde{A}_{ij}^*(\omega) \Sigma_{ii}^{-1} \tilde{A}_{ij}(\omega) P_{jj}^{-1}(\omega) \right) \quad (20)$$

and

$$\gamma_{ij}^{(c_m)}(\omega) = \lambda^m \left( S_{ii}^{-1}(\omega) \tilde{H}_{ij}(\omega) \Theta_{jj}^{-1} \tilde{H}_{ij}^*(\omega) \right). \quad (21)$$

Furthermore it is possible to relate bPDC/bDC and cPDC/cDC via

**Theorem 4.** Under the same conditions of Theorem 3 the following identities hold:

$$\pi_{ij}^{(b)}(\omega) = 1 - \prod_{m=1}^{\min\{M_i, M_j\}} \left( 1 - \pi_{ij}^{(c_m)}(\omega) \right) \quad (22)$$

and

$$\gamma_{ij}^{(b)}(\omega) = 1 - \prod_{m=1}^{\min\{M_i, M_j\}} \left( 1 - \gamma_{ij}^{(c_m)}(\omega) \right). \quad (23)$$

A simple consequence of Equations (22), (23) is that for stationary Gaussian processes satisfying the boundedness condition, we now have a decomposition of the mutual information rates

$$\text{MIR}_{\epsilon_i \eta_j} = \sum_{m=1}^{\min\{M_i, M_j\}} -\frac{1}{4\pi} \int_{-\pi}^{\pi} \log \left( 1 - \pi_{ij}^{(c_m)}(\omega) \right) d\omega \quad (24)$$

and

$$\text{MIR}_{X_i \zeta_j} = \sum_{m=1}^{\min\{M_i, M_j\}} -\frac{1}{4\pi} \int_{-\pi}^{\pi} \log \left( 1 - \gamma_{ij}^{(c_m)}(\omega) \right) d\omega. \quad (25)$$

Note how the quantities being summed in Equations (24), (25) are formally themselves contributions to the mutual information written in terms of their canonical coherence contributions.

## 5. ILLUSTRATIONS

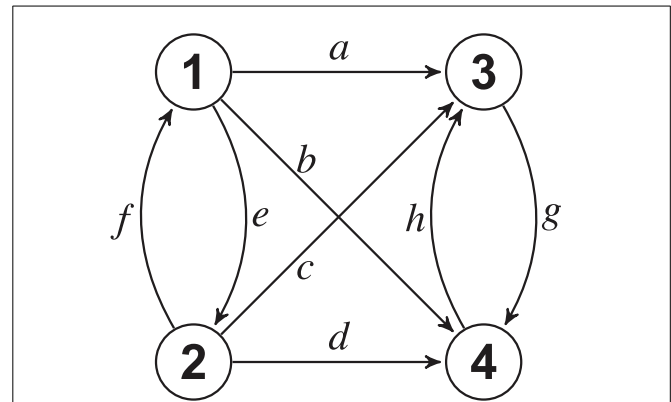
### 5.1. SIMULATED MODELS

*Example 1.* To provide insight into cPDC behavior, we begin with a very simple example that can be fully and explicitly solved.

Let a vector of observed time series  $[Y_1, Y_2, Y_3, Y_4]$  be a real valued autoregressive process of order  $p = 1$  and  $\Sigma = I$ . The autoregressive coefficients of the model are described by

$$A(1) = \begin{pmatrix} .5 & f & 0 & 0 \\ e & .5 & 0 & 0 \\ a & b & .5 & h \\ c & d & g & .5 \end{pmatrix}, \quad (26)$$

as in **Figure 1**.



**FIGURE 1 | Connectivity diagram for Example 1.** The number of canonical components depends on the value of  $ad - bc$ .

By adopting time series blocks as  $X_1 = [Y_1 \ Y_2]$  and  $X_2 = [Y_3 \ Y_4]$ , when  $e = f = g = h = 0$ , direct computation shows that the canonical PDC from block  $X_2$  to  $X_1$  is zero, i.e.,  $\pi_{12}^{(c_1)}(\omega) = \pi_{12}^{(c_2)}(\omega) = 0$  for all  $\omega$  (reflecting the nullity of the  $2 \times 2$   $A(l)$  right side upper block), whereas the coupling in the opposite direction contributes two distinct components:

$$\pi_{21}^{(c_1)}(\omega) = \frac{a^2 + b^2 + c^2 + d^2 + \sqrt{(a^2 + b^2 + c^2 + d^2)^2 - 4(ad - bc)^2}}{2.5 - 2 \cos(\omega)} \quad (27)$$

and

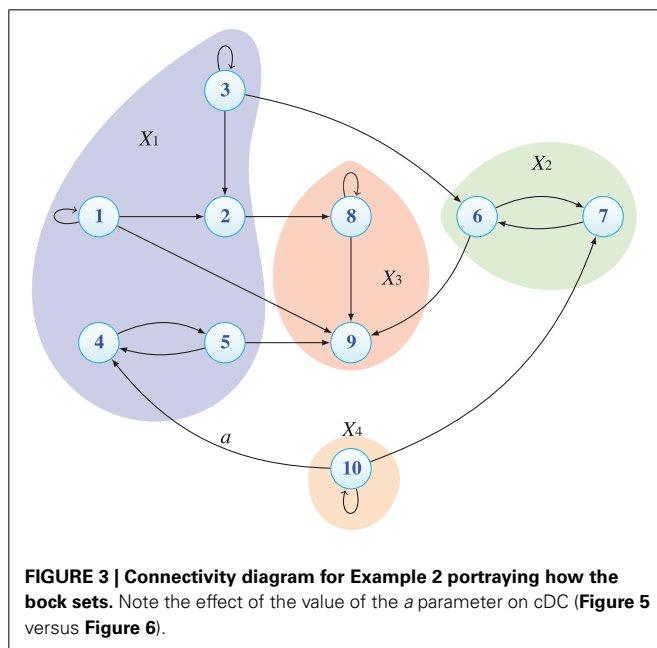
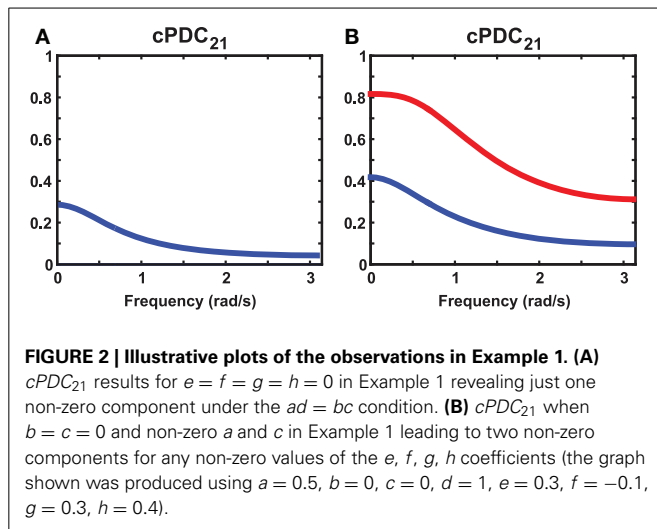
$$\pi_{21}^{(c_2)}(\omega) = \frac{a^2 + b^2 + c^2 + d^2 - \sqrt{(a^2 + b^2 + c^2 + d^2)^2 - 4(ad - bc)^2}}{2.5 - 2 \cos(\omega)}. \quad (28)$$

For  $ad = bc$ —i.e., if the lower left  $2 \times 2$  block determinant of  $A(l)$  is zero as well, the total number of non-zero cPDC components reduces to just 1.

Even if  $e, f, g, h$  are non-zero, i.e., regardless of intrablock dynamics,  $a = b = 0$  suffices to produce the single non-zero  $\pi_{21}^{(c_1)}(\omega)$  component (shown in **Figure 2A**) since block  $X_1$  interacts with block  $X_2$  exclusively through  $Y_4$ , i.e.,  $\pi_{21}^{(c_2)}(\omega) \equiv 0$ . In this case, since only  $Y_4$  is directly impacted by the interaction, only one combined source of variance exists even though two links exist between the blocks. Likewise if  $b = d = 0$ , even though two links leave  $X_1$ , there is only one dynamical component that counts.

This contrasts with the situation when  $b = c = 0$  where two non-zero  $\pi_{21}^{(c_2)}(\omega)$  coexist (**Figure 2B**) regardless of the values of  $e, f, g, h$  which, nonetheless, contribute to the relative size of the components.

*Example 2.* In the next example, a 10-variate time series  $(Y_1, \dots, Y_{10})$  follows the connectivity diagram represented in **Figure 3**. The multivariate time series is divided into four blocks ( $X_1, X_2, X_3$ , and  $X_4$ ), where  $X_4$  only sends information and  $X_3$ , which is an integrative block, only receives information. Block  $X_1$  has two functionally distinct internal parts, and only one is reached by outside influence. The scenario is fairly complicated and we next illustrate cPDC/cDC usefulness for understanding the underlying dynamic interaction between blocks.



To help interpret the results, we begin by describing the non-zero model coefficients and their dynamical effects. Observe that the model subscript indices in this example indicate the corresponding scalar process and not the block number.

#### 1. Block $X_1 = [Y_1 \ Y_2 \ Y_3 \ Y_4 \ Y_5]$

$$A_{1,1}(1) = 1.98 \cos(\pi/50), A_{1,1}(2) = -(.99)^2, \quad (\text{low frequency oscillator in } Y_1) \quad (29)$$

$$A_{2,3}(1) = 1, \quad (30)$$

$$A_{3,3}(1) = 1.98 \cos(\pi/2), A_{3,3}(2) = -(.99)^2, \quad (\text{oscillator at midband } (\pi/2) \text{ in } Y_3) \quad (31)$$

$$A_{5,4}(1) = .99, A_{4,5}(1) = -.99, \quad (\text{oscillator at midband in } [Y_4 \ Y_5]) \quad (32)$$

$$A_{8,2}(1) = 1, A_{8,2}(3) = 1, \quad (\text{midband notch}) \quad (33)$$

$$A_{6,3}(1) = 1, A_{6,3}(3) = 1, \quad (\text{midband notch}) \quad (34)$$

$$A_{9,1}(1) = 1, A_{9,5}(1) = 1. \quad (35)$$

#### 2. Block $X_2 = [Y_6 \ Y_7]$

$$A_{6,7}(1) = .99, A_{7,6}(1) = -.99, \quad (\text{oscillator identical to the } [Y_4 \ Y_5]) \quad (36)$$

$$A_{9,6}(1) = 1. \quad (37)$$

#### 3. Block $X_3 = [Y_8 \ Y_9]$

$$A_{8,8}(1) = -1, A_{9,8}(1) = .5. \quad (38)$$

#### 4. Block $X_4 = [Y_{10}]$

$$A_{10,10}(1) = 1.98 \cos(2\pi/3), A_{10,10}(2) = -(.99)^2, \quad (\text{high frequency oscillator in } Y_{10}) \quad (39)$$

$$A_{4,10}(1) = a, \quad (40)$$

$$A_{7,10}(1) = 1. \quad (41)$$

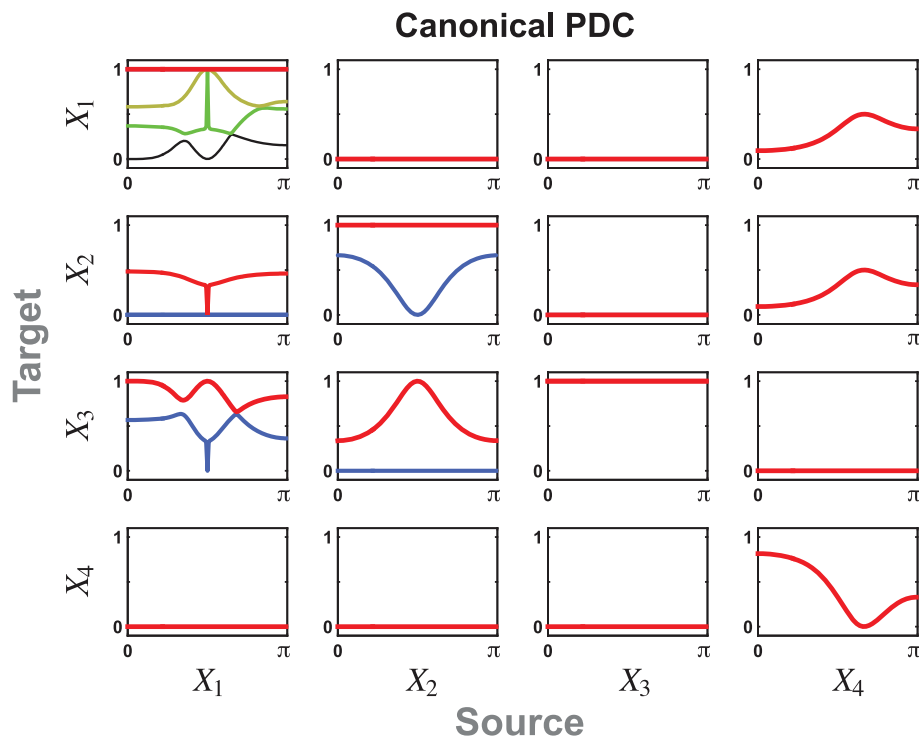
The resulting cPDC components can be appreciated in **Figure 4** for  $|a| = 1$ . Among their interesting features is the existence of the notch filtered link from  $X_1$  to  $X_2$  and to  $X_3$  at midband. The effects of the low frequency dynamics due to  $Y_1$  and the midband resonance due to  $[Y_4$  and  $Y_5]$  manifests itself as the strongest component from  $X_1$  to  $X_3$ . Likewise the single link effect from  $X_2$  to  $X_3$  is readily apparent as the higher frequency resonances from  $X_4$  toward both  $X_1$  and  $X_2$ . Both  $X_3$  components are identically equal to 1 since nothing leaves the block.

The corresponding cDCs are portrayed in **Figure 5** for  $a = -1$  with no signal reachability from  $X_4$  to  $X_3$ . This contrasts markedly with **Figure 6** for  $a = 1$  where  $X_4$ 's indirect effects on  $X_3$  are not balanced out.

The effects of the notch connections are readily apparent in both cases. For example, the power associated with the notch frequencies are the local components to  $X_2$  and  $X_3$  and cannot be attributed to outside influence. For block  $X_1$  only one of the five components is different from 1 reflecting the contribution coming from  $X_4$ .

## 5.2. EMPIRICAL DATA

This example is based on EEG data borrowed from Sameshima et al. (2014) (Ex. 7.7), which describes a left mesial temporal ictal episode monitored using an extended 10–20 system. The midline electrodes were excluded and left (L) and right (R) side electrodes were grouped as to whether they were frontal (F), central (C), parietal (P), temporal (T) or occipital (O) leading to the canonical PDCs portrayed in **Figure 7** where the most important connecting blocks share a dominant low pass frequency canonical component of fairly identical shape pointing to the existence



**FIGURE 4 | The cPDC for Example 2 reflects the existence of the notch connecting filters from  $X_1$  to  $X_2$  and to  $X_3$ .** The intrinsic dynamics of the oscillators from a subregion of  $X_1$  into  $X_3$  is apparent in the resonances of the largest cPDC component. The resonance within block  $X_2$  manifest itself in the single non-zero component into

$X_3$  while the effect of  $X_4$  reaches symmetrically into  $X_1$  and  $X_2$  via its single dynamic component. In this and following two figures, each subfigure may contain up to five cPDC/cDC components, given by  $\min(M_i, M_j)$  as in Equations (22)/(23), represented in red, blue, yellow, green, or black lines in decreasing order of magnitude.

of a shared dominant connectivity dynamics behind the observation, see **Figure 8A**. Their connectivity is further summarized in **Figure 8B**.

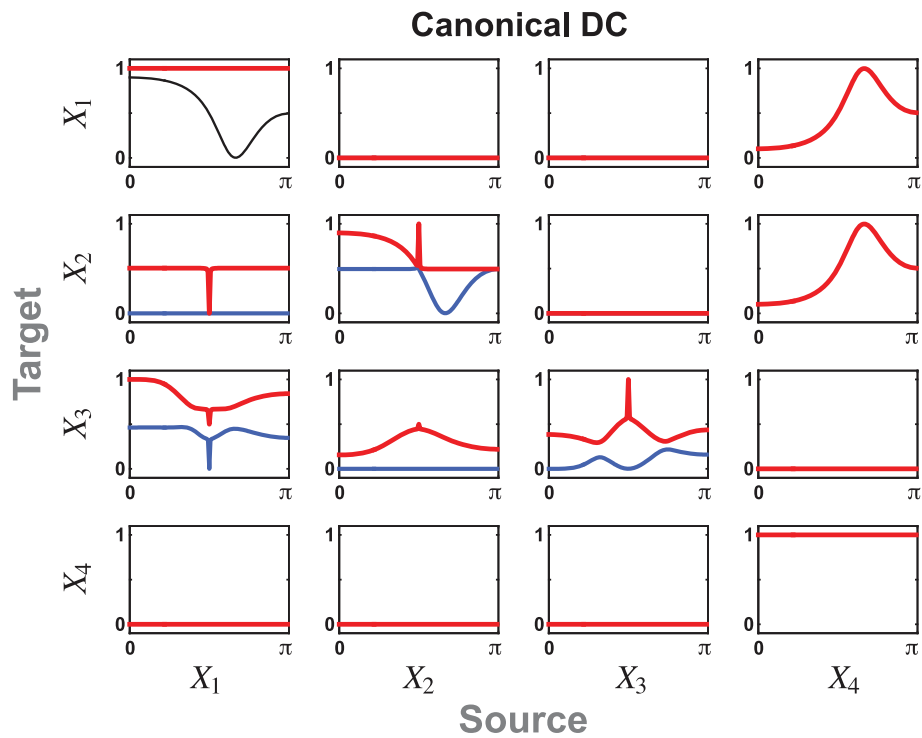
## 6. DISCUSSION

We showed that bPDC/bDC introduced in Takahashi (2009) and Faes and Nollo (2013) are block coherences between properly chosen vector time series. When the time series are Gaussian, this implies that bPDC/bDC represent mutual information rates between well defined underlying vector time series. This fully generalizes the results presented in Takahashi et al. (2010). To enhance the understanding of the possibly complex interaction between multiple time series and overcome some bPDC/bDC limitations, we showed that the latter can be decomposed in canonical terms that we call cPDC/cDC. These decompositions represent the various different modes of interaction whereby sets of time series interact. We introduced an explicit way to compute these new quantities and proved some of their properties. The usefulness of cPDC/cDC was illustrated by three examples.

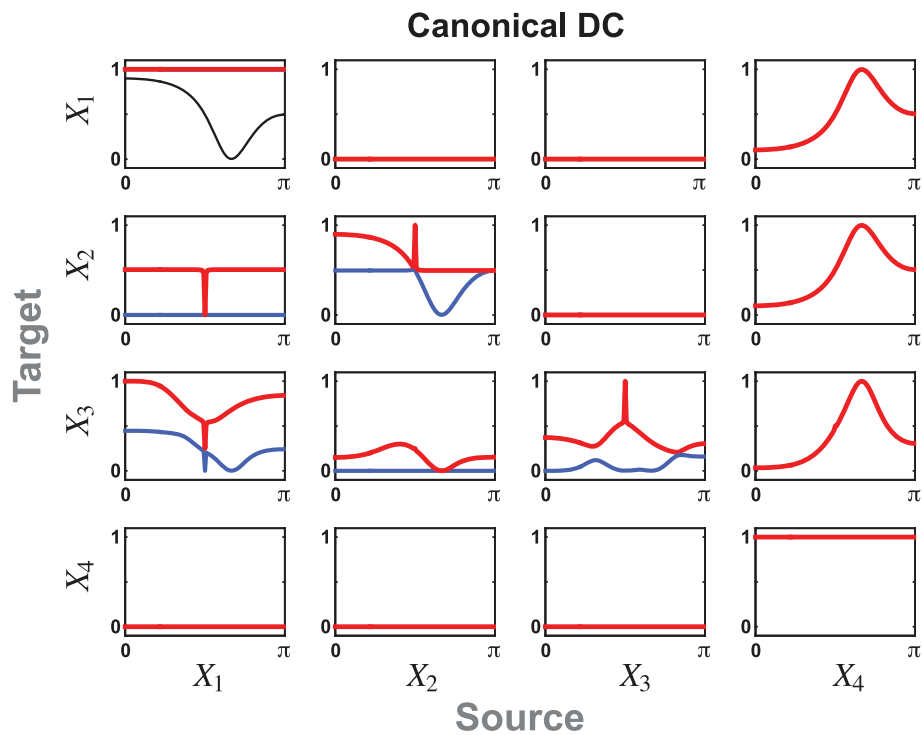
### 6.1. bPDC AND bDC AS BLOCK COHERENCES

Takahashi et al. (2010) showed that PDC from the  $j$ -th scalar time series to the  $i$ -th scalar time series is the coherence between the  $i$ -th innovation process and the  $j$ -th partialized process with a similar result for DC. It is natural to ask whether an analogous

result holds for bPDC and bDC. We showed that this is indeed the case where bPDC/bDC represent block coherences relating subsets of adequately defined innovations/partialization processes (Takahashi, 2009; Nedungadi et al., 2011). At first sight these identities may seem surprising as both bPDC and bDC are fully multivariate and directional measures of dependence, whereas block coherences are at once block-pairwise and symmetric measures of dependence. Yet careful reading of Theorems 1 and 2 highlights that bPDC/bDC from  $j$  to  $i$  and bPDC/bDC from  $i$  to  $j$  are, in general, block coherences between distinct pairs of vector processes which explains their asymmetric nature and lends them their directed connectivity character. Also, we note that for both bPDC and bDC, the coherences involve innovation process subsets which explains their fully multivariate characteristic as measures. Another interesting observation is that since the innovation processes are uncorrelated to the past of the partialized processes by construction, in the case of bPDC only innovations in the past of the partialized process contribute to the coherence which explains why bPDC is a directed measure of dependence. An analogous observation holds for bDC. In the Gaussian case, the bPDC/bDC representation as a block coherence allows relating them to the mutual information rate between suitably chosen time series. Formally this justifies the idea that these quantities are *de facto* measures of information flow. For an interesting comparison between bPDC/bDC and Geweke's measure of linear feedback see Faes

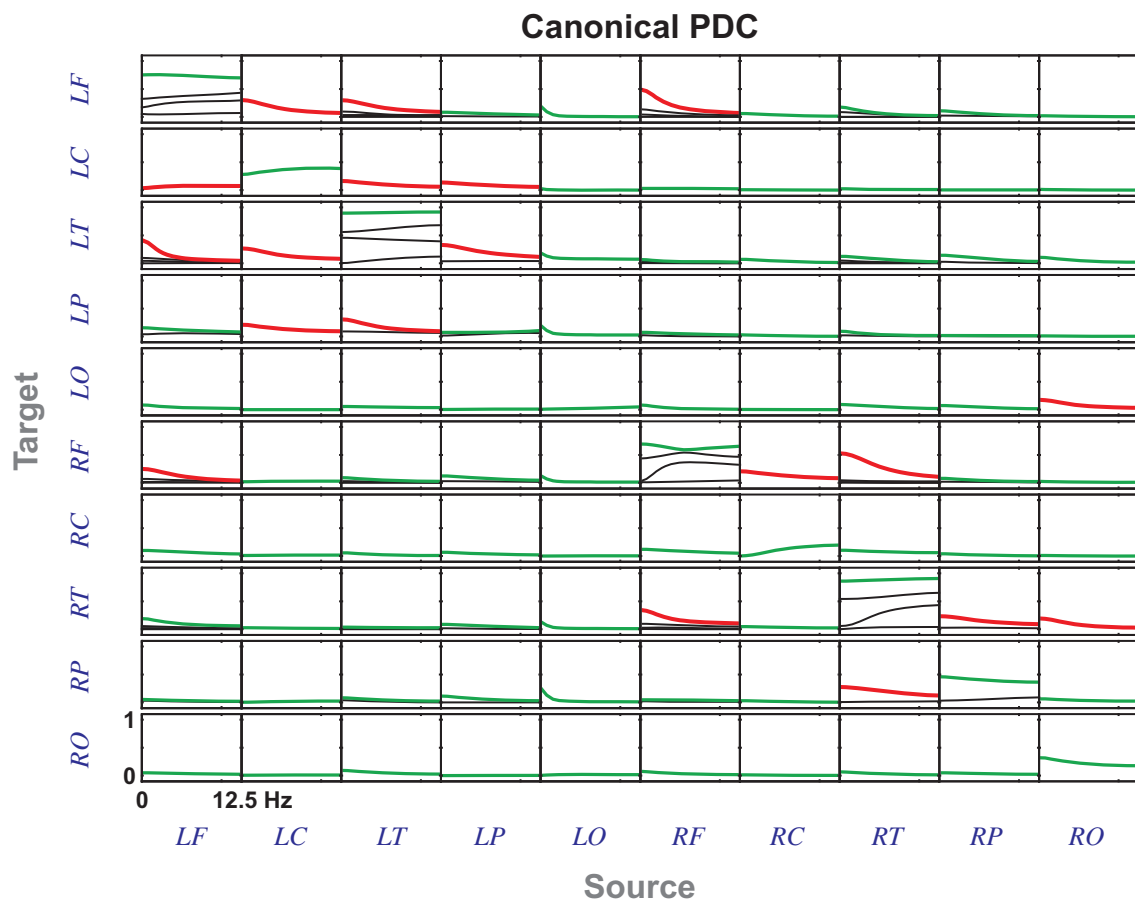


**FIGURE 5 | cDC for Example 2 for  $a = -1$  leading to a cancellation of the effect of  $X_4$  on  $X_3$  as the signal travels indirectly through two exactly identical structures but with opposite phases before reaching  $X_3$ . The notch filtering action is also apparent from the cDCs from  $X_1$  to  $X_2$  and  $X_3$ .**



**FIGURE 6 | cDC for Example 2 with  $a = 1$  which differs from Figure 5 in the effect from  $X_4$  to  $X_3$  which no longer cancels out.**





**FIGURE 7 | cPDC from the Empirical Data example (section 5.2) from the left mesial ictal episode where the largest components are represented either in red or green. cPDC values in red were arbitrarily considered significant and were pictorially summarized in Figure 8B.**

and Nollo (2013). As a small note for the reader, we observe that our definition of bPDC/bDC is slightly more general than the one proposed by Faes and Nollo (2013) because the covariance matrix of the innovations does not need to be diagonal as they assumed.

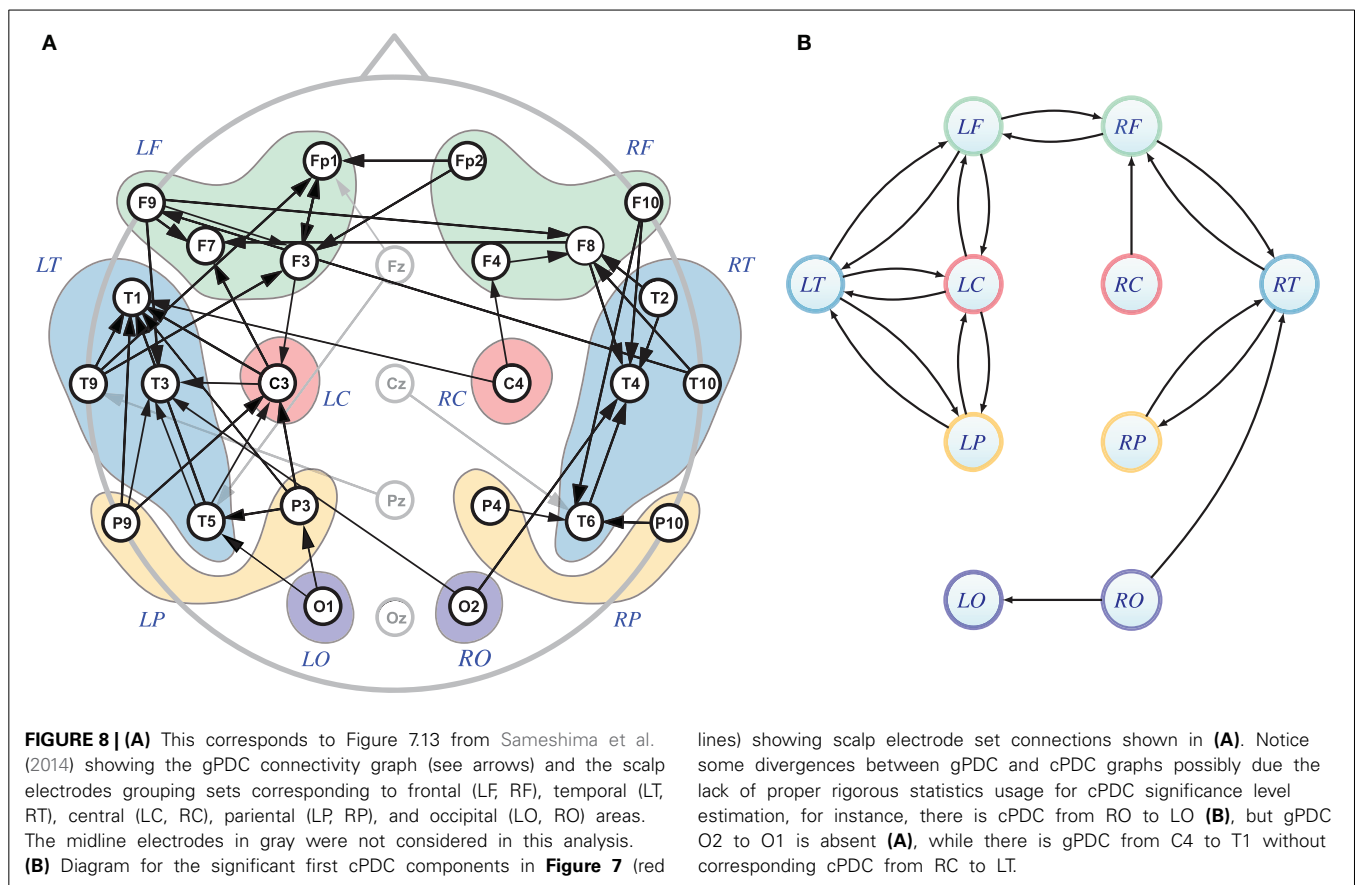
## 6.2. CANONICAL DECOMPOSITION OF DIRECTIONAL MEASURES

Given a pair of random vectors, it is natural to ask how to measure/represent dependence between them. In statistics, there are two main methods, both inspired by the basic Pearson correlation, to address this. The first one generalizes Pearson correlation directly using the determinants of the covariance matrix between and within each set of random variables. For time series, the equivalent measure in the frequency domain is the block coherence and the directed versions are bPDC and bDC. A second generalization rests on the idea of canonical correlation introduced by Hotelling (1936). There are several generalizations of canonical correlation for time series tailored specifically to infer Granger causality in the time domain (Sato et al., 2010; Wu et al., 2011), but, to the best of our knowledge, cPDC and cDC are the first proposals of canonical measures of directed dependence in the frequency domain.

One advantage of cPDC/cDC over bPDC/bDC is that canonical decomposition allows inferring the various different existing modes of interaction between sets of time series in close analogy to what is done for classical canonical correlation and principal component analyses. One should expect this to be useful when several signals are redundant, generated by similar mechanisms, or when there are several time series that do not significantly contribute to the interaction between sets of time series, e.g., when there are many brain areas that are not interacting with each other during some specific behavior. Besides, as we show in Theorem 4, we can recover the bPDC/bDC from the cPDC/cDC.

## 6.3. INTERPRETING cPDC/cDC

The main practical interest of cPDC/cDC is to allow the simplification of connectivity interpretations whilst giving new insights into the dynamical interaction between neural structures. We illustrated the achievable simplification using an EEG data set from an epileptic patient. We also showed how cPDC is related to the number of “modes” of interaction between sets of time series through the simple numerical Example 1 and via the slightly more complex Example 2. We expect that cPDC/cDC together



with bPDC/bDC become useful tools for handling high dimensional data sets that are increasingly being recorded by several researchers.

We propose that a reasonable way to understand the usefulness of cPDC/cDC is to make an analogy with classical principal component and canonical correlation analyses. Therefore, similar heuristics could be applied in practical situations, for example, to decide the number of different components to include in the interpretation. The canonical time series  $\bar{b}_Y$  and  $\bar{b}_Z$  from section 4 (see also Brillinger, 1981) are analogous to the canonical variables from the classical canonical correlation analysis and can play a similar role for result interpretation.

Finally we remark that the computational procedures used for the present paper will be made available the PDC homepage at <http://www.lcs.poli.usp.br/~baccala/pdc/canon> together with the data used in section 5.2.

## ACKNOWLEDGMENTS

CNPq Grants 307163/2013-0 to Luiz A. Baccalá and 309381/2012-6 to Koichi Sameshima are also gratefully acknowledged and to NAPNA—Núcleo de Neurociência Aplicada from the University of São Paulo. Part of this work took place during FAPESP Grant 2005/56464-9 (CInAPCe). Daniel Y. Takahashi was partially supported by Pew Latin American

Fellowship and Ciência sem Fronteiras Fellowship-CNPq grant (246778/2012-1).

## REFERENCES

- Ashrafulla, S., Haldar, J. P., Joshi, A. A., and Leahy, R. M. (2013). Canonical Granger causality between regions of interest. *Neuroimage* 83, 189–199. doi: 10.1016/j.neuroimage.2013.06.056
- Baccalá, L. A., and Sameshima, K. (2001). Partial directed coherence: a new concept in neural structure determination. *Biol. Cybern.* 84, 463–474. doi: 10.1007/PL00007990
- Baccalá, L. A., and Sameshima, K. (2014). “Multivariate time series brain connectivity: a sum up,” in *Methods in Brain Connectivity Inference Through Multivariate Time Series Analysis*, eds K. Sameshima and L. A. Baccalá (Boca Raton: CRC Press), 245–251. doi: 10.1201/b16550-18
- Brillinger, D. R. (1981). *Time Series: Data Analysis and Theory. Classics in applied mathematics*. Vol. 36. San Francisco, CA: SIAM, Society for Industrial and Applied Mathematics; Holden-Day.
- Faes, L., and Nollo, G. (2013). Measuring frequency domain Granger causality for multiple blocks of interacting time series. *Biol. Cybern.* 107, 217–232. doi: 10.1007/s00422-013-0547-5
- Gelfand, I. M., and Yaglom, A. M. (1959). Calculation of amount of information about a random function contained in another such function. *Am. Math. Soc. Transl. Ser. 2*, 3–52.
- Hannan, E. J. (1970). *Multiple Time Series (Wiley Series in Probability and Mathematical Statistics)*. New York, NY: Wiley. doi: 10.1002/9780470316429
- Hotelling, H. (1936). Relations between two sets of variates. *Biometrika* 28, 321–377. doi: 10.2307/2333955
- Kamiński, M., and Blinowska, K. J. (1991). A new method of the description of the information flow in brain structures. *Biol. Cybern.* 65, 203–210. doi: 10.1007/BF00198091

- Lütkepohl, H. (1996). *Handbook of Matrices*. Chichester: John Wiley.
- Nedungadi, A. G., Ding, M., and Rangarajan, G. (2011). Block coherence: a method for measuring the interdependence between two blocks of neurobiological time series. *Biol. Cybern.* 104, 197–207. doi: 10.1007/s00422-011-0429-7
- Pinsker, M. S. (1964). *Information and Information Stability of Random Variables and Processes*. San Francisco, CA: Holden-Day.
- Sameshima, K., Takahashi, D. Y., and Baccalá, L. A. (2014). “Asymptotic PDC properties,” in *Methods in Brain Connectivity Inference through Multivariate Time Series Analysis*, eds K. Sameshima and L. A. Baccalá (Boca Raton: CRC Press), 113–131. doi: 10.1201/b16550-9
- Sato, J. R., Fujita, A., Cardoso, E. F., Thomaz, C. E., Brammer, M. J., and Amaro, E. Jr. (2010). Analyzing the connectivity between regions of interest: an approach based on cluster Granger causality for fMRI data analysis. *Neuroimage* 52, 1444–1455. doi: 10.1016/j.neuroimage.2010.05.022
- Takahashi, D. Y. (2009). *Medidas de Fluxo de Informação com Aplicação em Neurociência*. Ph.D. thesis, University of São Paulo. Available online at: <http://www.teses.usp.br/teses/disponiveis/95/95131/tde-07062011-115256/en.php>
- Takahashi, D. Y., Baccalá, L. A., and Sameshima, K. (2010). Information theoretic interpretation of frequency domain connectivity measures. *Biol. Cybern.* 103, 463–469. doi: 10.1007/s00422-010-0410-x
- Wu, G. R., Chen, F., Kang, D., Zhang, X., Marinazzo, D., and Chen, H. (2011). Multiscale causal connectivity analysis by canonical correlation: theory and application to epileptic brain. *IEEE Trans. Biomed. Eng.* 58, 3088–3096. doi: 10.1109/TBME.2011.2162669

**Conflict of Interest Statement:** The authors declare that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.

Received: 17 January 2014; accepted: 23 April 2014; published online: 30 May 2014.

Citation: Takahashi DY, Baccalá LA and Sameshima K (2014) Canonical information flow decomposition among neural structure subsets. *Front. Neuroinform.* 8:49. doi: 10.3389/fninf.2014.00049

This article was submitted to the journal *Frontiers in Neuroinformatics*.

Copyright © 2014 Takahashi, Baccalá and Sameshima. This is an open-access article distributed under the terms of the Creative Commons Attribution License (CC BY). The use, distribution or reproduction in other forums is permitted, provided the original author(s) or licensor are credited and that the original publication in this journal is cited, in accordance with accepted academic practice. No use, distribution or reproduction is permitted which does not comply with these terms.

## A. APPENDIX

### A.1 PROOF OF THEOREMS 1 AND 2 AND COROLLARIES 1 AND 2

The proofs in this section follow the pattern of those in Takahashi et al. (2010). The chief difference lies in the care needed regarding the order of the products between the defining matrices. Here we exhibit the main proof ingredients for reader convenience, with further details available in Takahashi (2009) and Takahashi et al. (2010).

*Proof of Theorem 1 and Corollary 1.* Let  $W = [Y^T \ Z^T]^T$  be a second order stationary process satisfying the boundedness condition, using the following well known identity for determinants (Lütkepohl, 1996)

$$\det(S_W(\omega)) = \det(S_{ZZ}(\omega) - S_{ZY}(\omega)S_{YY}^{-1}(\omega)S_{YZ}(\omega)) \det(S_{YY}(\omega)) \quad (\text{A1})$$

leads to

$$\begin{aligned} C_{YZ}^{(b)}(\omega) &= 1 - \det(S_{ZZ}(\omega) \\ &\quad - S_{ZY}(\omega)S_{YY}^{-1}(\omega)S_{YZ}(\omega)) \det(S_{ZZ}^{-1}(\omega)) \quad (\text{A2}) \\ &= 1 - \det(S_{YY}(\omega) \\ &\quad - S_{YZ}(\omega)S_{ZZ}^{-1}(\omega)S_{ZY}(\omega)) \det(S_{YY}^{-1}(\omega)), \quad (\text{A3}) \end{aligned}$$

under Equation (5).

Rewrite bPDC as

$$\begin{aligned} \pi_{jj}^{(b)}(\omega) &= 1 - \det(P_{jj}^{-1}(\omega) \\ &\quad - P_{jj}^{-1}(\omega)\bar{A}_{ij}^*(\omega)\Sigma_{ii}^{-1}\bar{A}_{ij}(\omega)P_{jj}^{-1}(\omega)) \det(P_{jj}(\omega)), \quad (\text{A4}) \end{aligned}$$

so that using the following identities proved in Takahashi et al. (2010)

$$P_{jj}(\omega) = S_{\eta_j\eta_j}^{-1}(\omega), \quad (\text{A5})$$

$$S_{\epsilon_i\eta_j}(\omega) = \bar{A}_{ij}(\omega)S_{\eta_j\eta_j}(\omega), \quad (\text{A6})$$

and for all  $\omega \in [-\pi, \pi)$

$$S_{\epsilon_i\epsilon_i}(\omega) = \Sigma_{ii}, \quad (\text{A7})$$

back substituted into Equation (A4) leads to

$$\begin{aligned} \pi_{jj}^{(b)}(\omega) &= 1 - \det(S_{\eta_j\eta_j}(\omega) \\ &\quad - S_{\eta_j\epsilon_i}(\omega)\Sigma_{ii}^{-1}S_{\epsilon_i\eta_j}(\omega)) \det(S_{\eta_j\eta_j}^{-1}(\omega)), \quad (\text{A8}) \end{aligned}$$

so that using Equation (A2) shows that the right-hand side of Equation (A8) actually is  $C_{\epsilon_i\eta_j}^{(b)}(\omega)$  as we set out to prove. Corollary 1 is immediate from Theorem 1 and Equation (7).  $\square$

*Proof of Theorem 2 and Corollary 2.* Theorem 2 is obtained by rewriting  $C_{X_i\zeta_j}^{(b)}(\omega)$  using Equation (A3) noting that

$$S_{X_i\zeta_j}(\omega) = \bar{H}_{ij}(\omega)S_{\zeta_j\zeta_j}(\omega) \quad (\text{A9})$$

and for all  $\omega \in [-\pi, \pi)$

$$S_{\zeta_j\zeta_j}(\omega) = \Theta_{jj}^{-1}. \quad (\text{A10})$$

Corollary 2 follows from Theorem 2 and Equation (7).  $\square$

### A.2 PROOF OF THEOREMS 3 AND 4

Brillinger (1981, chapter 10) introduced the idea of canonical coherence for time series. We restate his result under our notation as the following theorem.

**Theorem 5** (Brillinger, Theorem 10.3.2). *Let  $X$  and  $Y$  be  $m_1$  and  $m_2$ -dimensional time-series jointly satisfying the boundedness condition. For  $m \leq \min\{m_1, m_2\}$ , the following identity holds:*

$$C_{XY}^{(cm)}(\omega) = \lambda^m(S_{YY}^{-1}(\omega)S_{YX}(\omega)S_{XX}^{-1}(\omega)S_{XY}(\omega)) \quad (\text{A11})$$

$$= \lambda^m(S_{XX}^{-1}(\omega)S_{XY}(\omega)S_{YY}^{-1}(\omega)S_{YX}(\omega)). \quad (\text{A12})$$

*Proof of Theorem 3.* From Equations (18), (A11), we have

$$C_{\epsilon_i\eta_j}^{(cm)}(\omega) = \lambda^m(S_{\eta_j\eta_j}^{-1}(\omega)S_{\eta_j\epsilon_i}(\omega)S_{\epsilon_i\epsilon_i}^{-1}(\omega)S_{\epsilon_i\eta_j}(\omega)). \quad (\text{A13})$$

Now, from Equations (A5), (A6), and (A7) it follows that

$$S_{\eta_j\eta_j}^{-1}(\omega)S_{\eta_j\epsilon_i}(\omega)S_{\epsilon_i\epsilon_i}^{-1}(\omega)S_{\epsilon_i\eta_j}(\omega) = \bar{A}_{ij}^*(\omega)\Sigma_{ii}^{-1}\bar{A}_{ij}(\omega)P_{jj}^{-1}(\omega), \quad (\text{A14})$$

which proves Equation (20).

To prove Equation (21), we use Equations (19), (A12) to obtain

$$C_{X_i\zeta_j}^{(cm)}(\omega) = \lambda^m(S_{X_iX_i}^{-1}(\omega)S_{X_i\zeta_j}(\omega)S_{\zeta_j\zeta_j}^{-1}(\omega)S_{\zeta_jX_i}(\omega)). \quad (\text{A15})$$

Finally, from Equations (A9), (A10), we have

$$S_{X_iX_i}^{-1}(\omega)S_{X_i\zeta_j}(\omega)S_{\zeta_j\zeta_j}^{-1}(\omega)S_{\zeta_jX_i}(\omega) = S_{ii}^{-1}(\omega)\bar{H}_{ij}(\omega)\Theta_{jj}^{-1}\bar{H}_{ij}^*(\omega), \quad (\text{A16})$$

which concludes the proof.  $\square$

*Proof of Theorem 4.* Rewrite bPDC as

$$1 - \pi_{ij}^{(b)}(\omega) = \det(I - \bar{A}_{ij}^*(\omega)\Sigma_{ii}^{-1}\bar{A}_{ij}(\omega)P_{jj}^{-1}(\omega)). \quad (\text{A17})$$

Now, Equation (22) is a straightforward consequence of the relationship between eigenvalues and the determinant of a matrix. A similar argument proves Equation (23).  $\square$





# Algorithms of causal inference for the analysis of effective connectivity among brain regions

Daniel Chicharro<sup>1\*</sup> and Stefano Panzeri<sup>1,2</sup>

<sup>1</sup> Neural Computation Laboratory, Center for Neuroscience and Cognitive Systems@UniTn, Istituto Italiano di Tecnologia, Rovereto, Italy

<sup>2</sup> Institute of Neuroscience and Psychology, University of Glasgow, Glasgow, UK

## Edited by:

Miguel Angel Muñoz, Universidad de Granada, Spain

## Reviewed by:

Thomas Natschläger, Software Competence Center Hagenberg GmbH, Austria

Gautam Agarwal, University of California Berkeley, USA

## \*Correspondence:

Daniel Chicharro, Neural Computation Laboratory, Center for Neuroscience and Cognitive Systems@UniTn, Istituto Italiano di Tecnologia, Corso Bettini 31, Rovereto 38068, Italy  
e-mail: daniel.chicharro@iit.it

In recent years, powerful general algorithms of causal inference have been developed. In particular, in the framework of Pearl's causality, algorithms of inductive causation (IC and IC\*) provide a procedure to determine which causal connections among nodes in a network can be inferred from empirical observations even in the presence of latent variables, indicating the limits of what can be learned without active manipulation of the system. These algorithms can in principle become important complements to established techniques such as Granger causality and Dynamic Causal Modeling (DCM) to analyze causal influences (effective connectivity) among brain regions. However, their application to dynamic processes has not been yet examined. Here we study how to apply these algorithms to time-varying signals such as electrophysiological or neuroimaging signals. We propose a new algorithm which combines the basic principles of the previous algorithms with Granger causality to obtain a representation of the causal relations suited to dynamic processes. Furthermore, we use graphical criteria to predict dynamic statistical dependencies between the signals from the causal structure. We show how some problems for causal inference from neural signals (e.g., measurement noise, hemodynamic responses, and time aggregation) can be understood in a general graphical approach. Focusing on the effect of spatial aggregation, we show that when causal inference is performed at a coarser scale than the one at which the neural sources interact, results strongly depend on the degree of integration of the neural sources aggregated in the signals, and thus characterize more the intra-areal properties than the interactions among regions. We finally discuss how the explicit consideration of latent processes contributes to understand Granger causality and DCM as well as to distinguish functional and effective connectivity.

**Keywords:** causal inference, brain effective connectivity, Pearl causality, Granger causality, Dynamic Causal Models, graphical models, latent processes, spatial aggregation

## INTRODUCTION

The need to understand how the interactions and coordination among brain regions contribute to brain functions has led to an ever increasing attention to the investigation of brain connectivity (Bullmore and Sporns, 2009; Friston, 2011). In addition to anatomical connectivity, two other types of connectivity that regard how the dynamic activity of different brain regions is interrelated have been proposed. *Functional connectivity* refers to the statistical dependence between the activity of the regions, while *effective connectivity* refers, in a broad sense, to the causal influence one neural system exerts over another (Friston, 2011).

Attempts to go beyond the study of dynamic correlations to investigate the causal interactions among brain regions have made use of different approaches to study causality developed outside neuroscience (Granger, 1963, 1980). Granger causality was proposed in econometrics to infer causality from time-series and has been widely applied in neuroscience as a model-free approach to study causal interactions among brain regions (see Bressler and Seth, 2011, for an overview). It has been applied to

different types of neural data, from intracranial electrophysiological recordings (e.g., Bernasconi and König, 1999; Besserve et al., 2010), Magnetoencephalography recordings (e.g., Vicente et al., 2011), to functional magnetic resonance imaging (fMRI) measures (e.g., Roebroeck et al., 2005; Mäki-Marttunen et al., 2013; Wu et al., 2013). New approaches have been also developed within neuroscience, such as Dynamic Causal Modeling (DCM) (Friston et al., 2003) which explicitly models the biophysical interactions between different neural populations as well as the nature of the recorded neural signals (Friston et al., 2013).

Separately, in the field of artificial intelligence, another approach to causal analysis has been developed by Pearl and coworkers. Pearl's approach combines causal models and causal graphs (Spirtes et al., 2000; Pearl, 2009). The fundamental difference with the approaches currently used to study the brain's effective connectivity (Granger causality and DCM) is that the understanding of causation in Pearl's framework ultimately relies on the notion of an external intervention that actively perturbs the system. This notion of intervention provides a rigorous

definition of the concept of causal influence but at the same time illustrates the limitations of causal analysis from observational studies.

The analysis of the causal influence one neural system exerts over another (i.e., effective connectivity) requires considering causation at different levels (Chicharro and Ledberg, 2012a), in particular distinguishing between causal inference and quantification or modeling of causal effects (Pearl, 2009). At the most basic level, *causal inference* deals with assessing which causal connections exist and which do not exist, independently of their magnitude or the mechanisms that generate them. At a higher level, the quantification of the magnitude implies selecting a measure of the strength of the causal effect, and the characterization of the mechanisms implies implementing a plausible model of how the dynamics of the system are generated. Recently, it has been pointed out that the existence of causal connections should be distinguished from the existence of causal effects, and in particular that only in some cases it is meaningful to understand the interactions between subsystems in terms of the causal effect one exerts over another (Chicharro and Ledberg, 2012a). Furthermore, the possibility and the limitations to quantify causal influences with Granger causality has been examined (Lizier and Prokopenko, 2010; Chicharro and Ledberg, 2012b; Chicharro, 2014b).

In this work we focus on the basic level of causal analysis constituted by causal inference. In particular, we investigate how some general algorithms of causal inference (IC and IC\* algorithms) developed in the Pearl's framework (Verma and Pearl, 1990; Pearl, 2009) can be applied to infer causality between dynamic processes and thus used for the analysis of effective connectivity. This algorithmic approach relies on the evaluation of the statistical dependencies present in the data, similarly to the non-parametric formulation of Granger causality. Its particularity is that it explicitly considers the impact of latent (unobserved) processes as well as the existence of different causal structures which are equivalent in terms of the statistical dependencies they produce. Accordingly, it provides a principled procedure to evaluate the discrimination power of the data with respect to the possible causal structures underlying the generation of these data.

Although these causal algorithms do not assume any constraint on the nature of the variables to which they are applied, their application to dynamic processes has yet to be investigated. The main goal of this paper is to study the extension of Pearl's causal approach to dynamic processes and to evaluate conceptually how it can contribute to the analysis of effective neural connectivity. To guide the reader, we provide below an overview of the structure of this article.

## OVERVIEW OF THE STRUCTURE OF THE ARTICLE

We start by reviewing the approach to causal inference of Pearl (2009) and Granger (1963, 1980) and we then focus on the analysis of temporal dynamics. In the first part of our Results we investigate the application to dynamic processes of the algorithms of causal inference proposed by Pearl. We then recast their basic principles combining them with Granger causality into a new algorithm which, as the IC\* algorithm, explicitly deals with latent processes but furthermore provides a more suited

output representation of the causal relations among the dynamic processes.

In the second part of our Results, we shift the focus from the inference of an unknown causal structure to studying how statistical dependencies can be predicted from the causal structure. In particular, for a known (or hypothesized) causal structure underlying the generation of the recorded signals, we use graphical criteria to identify the statistical dependencies between the signals. We specifically consider causal structures compatible with the state-space models which have recently been recognized as an integrative framework in which refinements of Granger causality and DCM converge (Valdes-Sosa et al., 2011). This leads us to reformulate in a general unifying graphical approach different effects relevant for the analysis of effective connectivity, such as those of measurement noise (Nalatore et al., 2007), of hemodynamic responses (e.g., Seth et al., 2013), and of time aggregation (e.g., Smirnov, 2013). We especially focus on the effect of spatial aggregation caused by the superposition in the recorded signals of the massed activity of the underlying sources of neural activity interacting at a finer scale.

Finally, in Discussion we discuss the necessity to understand how causal interactions propagate from the microscopic to the macroscopic scale. We indicate that, although the algorithms here discussed constitute a non-parametric approach to causal inference, our results are also relevant for modeling approaches such as DCM and help to better understand how difficult it is in practice to distinguish functional and effective connectivity.

## REVIEW OF RELEVANT CONCEPTS OF CAUSAL MODELS

In this section, we lay the basis for the novel results by reviewing the approach to causal inference of Pearl (2009) and Granger (1963, 1980).

### MODELS OF CAUSALITY

We begin reviewing the models of causality described by Pearl (2009) and relating them to DCM (Friston et al., 2003). For simplicity, we restrict ourselves to the standard Pearl models which are the basis of the IC and IC\* algorithm, without reviewing extensions of these models such as settable systems (White and Chalak, 2009), which are suitable for a broader set of systems involving, e.g., optimization and learning problems.

A *Causal Model*  $M$  is composed by a set of  $n$  stochastic variables  $V_k$ , with  $k \in \{1, \dots, n\}$  which are endogenous to the model, and a set of  $n'$  stochastic variables  $U_{k'}$ , with  $k' \in \{1, \dots, n'\}$ , which are exogenous to the model. Endogenous variables are those explicitly observed and modeled. For example, when studying the brain's effective connectivity, these variables may be the neural activity of a set of  $n$  different regions. The exogenous variables correspond to sources of variability not explicitly considered in the model, which can for example correspond to sources of neuromodulation, uncontrolled variables related to changes in the cognitive state (Masquelier, 2013), or activity of brain areas not recorded. Accordingly, for each variable  $V_k$  the model contains a function  $f_k$  such that

$$V_k = f_k(pa(V_k), U_k, \theta_k) \quad (1)$$

That is, the value of  $V_k$  is assigned by a function  $f_k$  determined by a set  $\theta_k$  of constant parameters and taking as arguments a subset of the endogenous variables which is called the *parents* of  $V_k$  ( $pa(V_k)$ ), as well as a subset of the exogenous variables  $U_k$ . In general, in Pearl's formulation the exogenous variables are considered as noise terms which do not introduce dependencies between the endogenous variables, so that a single variable  $U_k$  can be related to each  $V_k$ . Causality from  $V_j$  to  $V_{j'}$  is well-defined inside the model:  $V_j$  is directly causal to  $V_{j'}$  if it appears as an argument of the function  $f_{j'}$ , that is, if  $V_j$  is a parent of  $V_{j'}$  ( $V_j \in pa(V_{j'})$ ). However, whether the inside-model causal relation correctly captures some real physical causality depends on the goodness of the model. To complete the model the probability distribution  $p(\{U\})$  of the exogenous variables is required, so that the joint distribution of the endogenous variables  $p(\{V\})$  is generated using the functions. Accordingly,  $p(\{V\})$  can be decomposed in a Markov factorization that reflects the constraints in terms of conditional independence that result from the functional model:

$$p(V_1, \dots, V_n) = \prod_{k=1}^n p(V_k | pa(V_k)). \quad (2)$$

Each causal model  $M$  has an associated graphical representation called *causal structure*  $G(M)$ . A causal structure is a directed acyclic graph (DAG) in which each endogenous variable  $V_k$  corresponds to a node and an arrow pointing to  $V_k$  from each of its parents is added. A *path* between nodes  $V_j$  and  $V_{j'}$  is a sequence of arrows linking  $V_j$  and  $V_{j'}$ . It is not required to follow the direction of the arrows, and a path that respects their direction is called a *directed path*. A causal structure reflects the parental structure in the functional model, and thus indicates some constraints to the set  $\Theta = \{\theta_1, \dots, \theta_n\}$  of constant parameters used to construct the functions. The factorization of Equation (2) is reflected in  $V_k$  being conditionally independent from any other of its ancestors once conditioned on  $pa(V_k)$ , where the ancestors of  $V_k$ —i.e.,  $an(V_k)$ —are defined in the graph as those nodes that can be attained by following backwards any directed path that arrives to  $V_k$ .

In the formulation of Pearl no constraints concern the nature of the variables in the causal model. However, in the presentation of Pearl's framework (Pearl, 2009) dynamic variables are seldom used. This fact, together with the fact that the causal graphs associated with the causal models are acyclic, has sometimes lead to erroneously think that the Pearl's formulation is not compatible with processes that involve feedback connections, since they lead to cyclic structures in the graph (see Valdes-Sosa et al., 2011, for discussion). However, cycles only appear when not considering the dynamic nature of the causal model underlying the graphical representation. For dynamic variables, the functional model consists of a set of differential equations, DCM state equations being a well-known example (Valdes-Sosa et al., 2011). In particular, in a discretized form, the state equations are expressed as

$$V_{k,i+1} = f_k(pa(V_{k,i+1}), U_{k,i}; \theta_k); \quad (3)$$

where  $V_{k,i+1}$  is the variable associated with the time sampling  $i+1$  of process  $k$ . In general, the parents of  $V_{k,i+1}$  include  $V_{k,i}$  and can

comprise several sampling times from other processes, depending on the delay in the interactions. Depending on the type of DCM models used, deterministic or stochastic, the variables  $\{U\}$  can comprise exogenous drivers or noise processes. It is thus clear that the models of causality described by Pearl are general and comprise models of the form used in DCM.

## STATISTICAL INDEPENDENCIES DETERMINED BY CAUSAL INTERACTIONS

As mentioned above, a causal structure is a graph that represents the structure of the parents in a causal model. Pearl (1986) provided a graphical criterion for DAGs called *d-separation*—where *d* stands for directional—to check the independencies present in any model compatible with a causal structure. Its definition relies on the notion of *collider* on a path, a node on a path for which, when going along the path, two arrows point toward the node ( $\rightarrow V \leftarrow$ ). The criterion of d-separation states:

### D-separation

Two nodes  $V_j, V_{j'}$  are d-separated by a set of nodes  $C$  if and only if for every path between  $V_j, V_{j'}$  one of the following conditions is fulfilled:

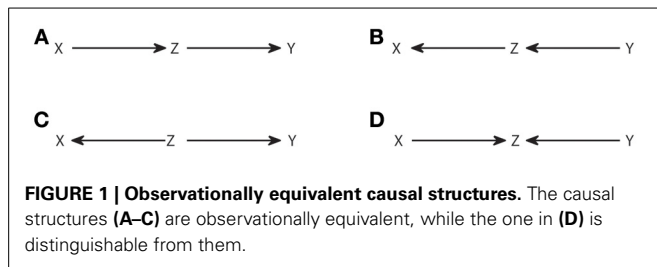
- (1) The path contains a non-collider  $V_k$  ( $\rightarrow V_k \rightarrow$  or  $\leftarrow V_k \leftarrow$ ) which belongs to  $C$ .
- (2) The path contains a collider  $V_k$  ( $\rightarrow V_k \leftarrow$ ) which does not belong to  $C$  and  $V_k$  is not an ancestor of any node in  $C$ .

For a causal model compatible with a causal structure the d-separation of  $V_j$  and  $V_{j'}$  by  $C$  is a sufficient condition for  $V_j$  and  $V_{j'}$  being conditional independent given  $C$ , that is

$$V_j \perp_G V_{j'} | C \Rightarrow V_j \perp_M V_{j'} | C \quad (4)$$

where  $\perp_G$  indicates d-separation in the causal structure  $G$  and  $\perp_M$  independence in the joint probability distribution of the variables generated by the causal model  $M$ . This sufficient condition can be converted into an if and only if condition if further assuming *stability* (Pearl, 2009)—or equivalently *faithfulness* (Spirtes et al., 2000)—, which states that conditional independence between the variables does not result from a particular tuning of the parameters  $\Theta$ , which would disappear if those were infinitesimally modified.

Considering the correspondence between d-separation and conditional independence, an important question is the degree to which the underlying causal structure can be inferred from the set of conditional independencies present in an observed joint distribution. The answer is that there are classes of causal structures which are observationally equivalent, that is, they produce exactly the same set of conditional independencies observable from the joint distribution. Consider, for example, the four causal structures of **Figure 1**. Each causal structure is characterized by a list of all the conditional independencies compatible with it. Applying d-separation it can be checked that for **Figures 1A–C** we have that  $X$  and  $Y$  are d-separated by  $Z$  ( $X \perp Y | Z$ ), while in **Figure 1D**  $X$  and  $Y$  are d-separated by the empty set ( $X \perp Y$ ). Therefore, we can discriminate **Figures 1A–C** from **Figure 1D**, but not among



**Figures 1A–C.** Statistical dependencies, the only type of available information when recording the variables, only retain limited information about how the variables have been generated.

Verma and Pearl (1990) provided the conditions for two DAGs to be *observationally equivalent*. Two DAGs are observationally equivalent if and only if they have the same skeleton and the same v-structures, where the skeleton refers to the links without considering the direction of the arrows, and a v-structure refers to three nodes such that two arrows point head to head to the central node, while the other two nodes are *non-adjacent*, i.e., not directly linked (as in **Figure 1D**). It is clear from this criterion that the structures in **Figures 1A–C** are equivalent and the one in **Figure 1D** is not.

## CAUSAL INFERENCE

### Causal inference without latent variables, the IC algorithm

Given the existence of observationally equivalent classes of DAGs, it is clear that there is an intrinsic fundamental limitation to the inference of a causal structure from recorded data. This is so even assuming that there are no latent variables. Here we review the IC algorithm (Verma and Pearl, 1990; Pearl, 2009), which provides a way to identify with which equivalence class a joint distribution is compatible, given the conditional independencies it contains. The input to the algorithm is the joint distribution  $p(\{V\})$  on the set  $\{V\}$  of variables, and the output is a graphical pattern that reflects all and no more conditional independencies than the ones in  $p(\{V\})$ . These independencies can be read from the pattern applying d-separation. The algorithm is as following:

### IC ALGORITHM (INDUCTIVE CAUSATION)

- (1) For each pair of variables  $a$  and  $b$  in  $\{V\}$  search for a set  $S_{ab}$  such that conditional independence between  $a$  and  $b$  given  $S_{ab}$  ( $a \perp b | S_{ab}$ ) holds in  $p(\{V\})$ . Construct an undirected graph linking the nodes  $a$  and  $b$  if and only if  $S_{ab}$  is not found.
- (2) For each pair of non-adjacent nodes  $a$  and  $b$  with a common adjacent node  $c$  check if  $c$  belongs to  $S_{ab}$ . If it does, then continue. If it does not, then add arrowheads pointing at  $c$  to the edges (i.e.,  $a \rightarrow c \leftarrow b$ ).
- (3) In the partially oriented graph that results, orient as many edges as possible subject to two conditions: (i) Any alternative orientation would yield a new v-structure. (ii) Any alternative orientation would yield a directed cycle.

The algorithm is a straightforward application of the definition of observational equivalence. Step 1 recovers the skeleton of the graph, linking those nodes that are dependent in any context.

Step 2 identifies the v-structures and Step 3 prevents creating new ones or cycles. A more procedural formulation of Step 3 was proposed in Verma and Pearl (1992). As an example, in **Figure 2** we show the output from the IC algorithm that would result from joint distributions compatible with causal structures of **Figure 1**. Note that throughout this work, unless otherwise stated, conditional independencies are not evaluated by estimating the probability distributions, but graphically identified using Equation (4). The causal structures of **Figures 2A,C** result in the same pattern (**Figures 2B,D**, respectively), which differ from the one that results from **Figure 2E** (**Figure 2F**).

The output pattern is not in general a DAG because not all links are arrows. It is a partial DAG which constitutes a graphical representation of the conditional independencies. D-separation is applicable, but now it has to be considered that non-colliders comprise edges without arrows, while the definition of collider remains the same. Note that, to build any causal structure that is an element of the class represented by a pattern, one has to continue adding arrows to the pattern subject to not creating v-structures or cycles. For example, the pattern of **Figure 2B** can be completed to lead to any causal structure of **Figures 1A–C**, but one cannot add head to head arrows, because this would give a non-compatible causal structure which corresponds to the pattern of **Figure 2F**.

### CAUSAL INFERENCE WITH LATENT VARIABLES: THE IC\* ALGORITHM

So far we have addressed the case in which the joint distribution  $p(\{V\})$  includes all the variables of the model. Now we consider that only a subset  $\{V_O\}$  is observed. We have seen that while a causal structure corresponds to a unique pattern which represents the equivalence class, a pattern can represent many causal structures. The size of the equivalence class generally increases with the number of nodes. This means that when latent variables are not excluded, if no constraints are imposed to the structure of the latent variables, the size of the class grows infinitely. For example, if the latent variables are interlinked, the unobserved part of the causal structure may contain many conditional independencies that we cannot test. To handle this, Verma (1993) introduced the notion of a *projection* and proved that any causal structure with a subset  $\{V_O\}$  of observable nodes has a dependency-equivalent projection, that is, another causal structure compatible with the same set of conditional independencies involving the observed variables, but for which all unobserved nodes are not linked between them and are parents of exactly two observable nodes. Accordingly, the objective of causal inference with the IC\* algorithm is to identify with which dependency-equivalent class of projections a joint distribution  $p(\{V_O\})$  is compatible. In the next section we will discuss how relevant it is for the application to dynamic processes the restriction of inference to projections instead of more general causal structures.

The input to the IC\* algorithm (Verma, 1993; Pearl, 2009) is  $p(\{V_O\})$ . The output is an *embedded pattern*, a hybrid acyclic graph that represents all and no more conditional independencies than the ones contained in  $p(\{V_O\})$ . While the patterns that result from the IC algorithm are partial DAGs which only contain arrows that indicate a causal connection, or undirected edges to be completed, the embedded patterns obtained with the IC\*

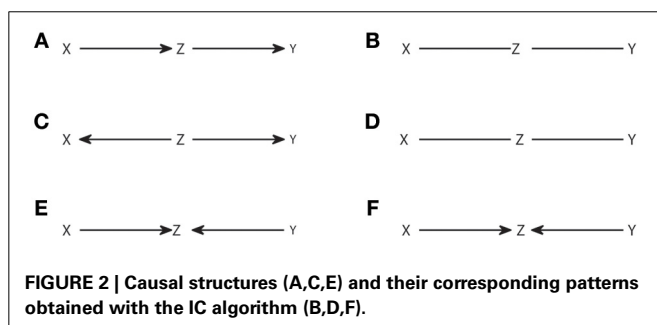


algorithm are hybrid acyclic graphs because they can contain more types of links: genuine causal connections are indicated by solid arrows ( $a \rightarrow b$ ). These are the only causal connections that can be inferred with certainty from the independencies observed. *Potential causes* are indicated by dashed arrows ( $a \dashrightarrow b$ ), and refer to a possible causal connection ( $a \rightarrow b$ ), or to a possible latent common driver ( $a \leftarrow \alpha \rightarrow b$ ), where greek letters are used for latent nodes. Furthermore, bidirectional arrows indicate certainty about the existence of a common driver. Undirected edges indicate a link yet to be completed. Therefore, there is a hierarchy of inclusion of the links, going from completely undefined, to completely defined identification of the source of the dependence: Undirected edges subsume potential causes, which subsume genuine causes and common drivers.

Analogously to the patterns of the IC algorithm, the embedded patterns are just a graphical representation of the dependency class. Their main property is that using d-separation one can read from the embedded pattern all and no more than the conditional independencies compatible with the class. In the case of the embedded patterns, d-separation has to be applied extending the definition of collider to any head to head arrows of any of the type present in the hybrid acyclic graphs.

#### IC\* ALGORITHM (INDUCTIVE CAUSATION WITH LATENT VARIABLES)

- (1) For each pair of variables  $a$  and  $b$  in  $\{V_O\}$  search for a set  $S_{ab}$  such that conditional independence between  $a$  and  $b$  given  $S_{ab}$  ( $a \perp b \mid S_{ab}$ ) holds in  $p(\{V_O\})$ . Construct an undirected graph linking the nodes  $a$  and  $b$  if and only if  $S_{ab}$  is not found.
- (2) For each pair of non-adjacent nodes  $a$  and  $b$  with a common adjacent node  $c$  check if  $c$  belongs to  $S_{ab}$ .  
If it does, then continue.  
If it does not, then substitute the undirected edges by dashed arrows pointing at  $c$ .
- (3) Recursively apply the following rules:
  - 3R<sub>1</sub>: if  $a$  and  $b$  are non-adjacent, they have a common adjacent node  $c$ , if the link between  $a$  and  $c$  has an arrowhead into  $c$  and the link between  $b$  and  $c$  has no arrowhead into  $c$ , then substitute the link between  $c$  and  $b$  (either an undirected edge or a dashed arrow) by a solid arrow from  $c$  to  $b$ , indicating a genuine causal connection ( $c \rightarrow b$ ).
  - 3R<sub>2</sub>: if there is a directed path from  $a$  to  $b$  and another path between them with a link that renders this path compatible with a directed path in the opposite direction, substitute the

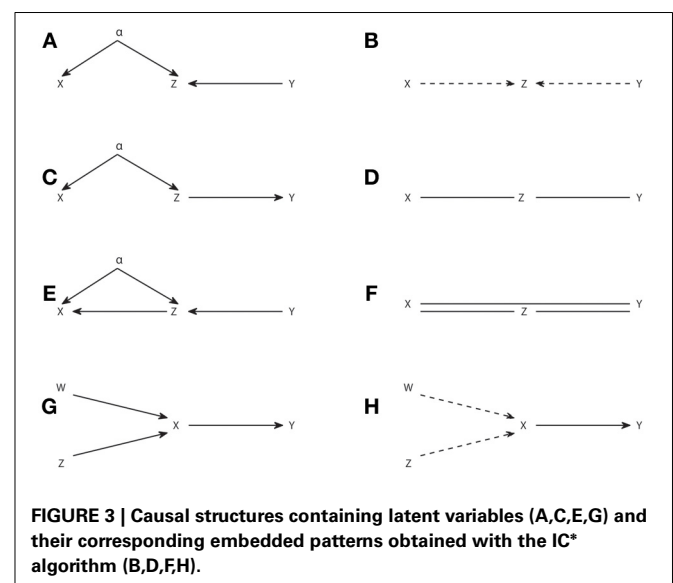


type of link by the one immediately below in the hierarchy that excludes the existence of a cycle.

Steps 1 and 2 of the algorithm are analogous to the steps of the IC algorithm, except that now in Step 2 dashed arrows are introduced indicating potential causes. The application of step 3 is analogous to the completion in Step 3 of the IC algorithm, but adapted to consider all the types of links that are now possible. In 3R<sub>1</sub> a causal connection ( $c \rightarrow b$ ) is identified because either a causal connection on the opposite direction or a common driver would create a new v-structure. In 3R<sub>2</sub> cycles are avoided.

As an example of the application of the IC\* algorithm in **Figure 3** we show several causal structures and their corresponding embedded patterns. The causal structure of **Figure 3A** results in an embedded pattern with two potential causes pointing to Z (**Figure 3B**), while the one of **Figure 3C** results in an embedded pattern with undirected edges (**Figure 3D**). The embedded pattern of **Figure 3B** can be seen as a generalization, when latent variables are considered, of the pattern of **Figure 2F**. Similarly, the pattern of **Figure 3D** is a generalization of **Figures 2B,D**. In the case of these embedded patterns a particular causal structure from the dependency class can be obtained by selecting one of the connections compatible with each type of link, e.g., a direct arrow or to add a node that is a common driver for the case of dashed arrows indicating a potential cause. Furthermore, like for the completion of patterns obtained from the IC algorithm, no new v-structures or cycles can be created, e.g., in **Figure 3D** the undirected edges cannot be both substituted by head to head arrows.

However, in general for the embedded patterns, not all the elements of the dependency class can be retrieved by completing the links, even if one restricts itself to projections. For example, consider the causal structure of **Figure 3E** and its corresponding embedded pattern in **Figure 3F**. In this case the embedded pattern does not share the skeleton with the causal structure, since a link  $X-Y$  is present indicating that  $X$  and  $Y$  are adjacent. This makes the mapping of the embedded pattern to the underlying



causal structure less intuitive and further highlights that the patterns and embedded patterns are just graphical representations of a given observational and dependency class, respectively.

As a last example in **Figures 3G,H** we show a causal structure and its corresponding embedded pattern where a genuine causal structure is inferred by applying the rule  $3R_1$ . A genuine cause from  $X$  to  $Y$  ( $X \rightarrow Y$ ) is the only possibility since a genuine cause from  $Y$  to  $X$  ( $X \leftarrow Y$ ), as well as a common driver ( $X \leftarrow \alpha \rightarrow Y$ ) would both create a new v-structure centered at  $X$ . Therefore, rule  $3R_1$  reflects that even if allowing for the existence of latent variables, it is sometimes possible to infer a genuine causation just from observations, without having to manipulate the system. As described in rule  $3R_1$ , inferring genuine causation from a variable  $X$  to a variable  $Y$  always involves a third variable and requires checking at least two conditional independencies. See the Supplementary Material for details of a sufficient condition of genuine causation (Verma, 1993; Pearl, 2009) and how it is formulated in terms of Granger causality when examining dynamic processes.

### THE CRITERION OF GRANGER CAUSALITY FOR CAUSAL INFERENCE

So far we have reviewed the approach of Pearl based on models of causality and graphical causal structures. The algorithms of causal inference proposed in this framework are generic and not conceived for a specific type of variables. Conversely, Granger (1963, 1980) proposed a criterion to infer causality specifically between dynamic processes. The criterion to infer causality from process  $X$  to process  $Y$  is based on the extra knowledge obtained about the future of  $Y$  given the past of  $X$ , in a given context  $Z$ . In its linear implementation, this criterion results in a comparison of prediction errors, however, as already pointed out by Granger (1980), a strong formulation of the criterion is expressed as a condition of independence

$$p(Y_{i+1}|\{V\}^i) = p(Y_{i+1}|\{V\}^i \setminus X^i), \quad (5)$$

where the superindex  $i$  refers to the whole past of a process up to and including sample  $i$ ,  $\{V\}$  refers to the whole system  $\{X, Y, Z\}$ , and  $\{V\}^i \setminus X^i$  refers to the past of the whole system excluding the past of  $X$ . That is,  $X$  is Granger non-causal to  $Y$  given  $Z$  if the equality above holds. Granger (1980) indicated that Granger causality is context dependent, i.e., adding or removing other processes from the context  $Z$  affects the test for causality. In particular, genuine causality could only be checked if  $Z$  was including all the processes that have a causal link to  $X$  and  $Y$ , otherwise a hidden common driver or an intermediate process may be responsible for the dependence. Latent variables commonly result in the existence of instantaneous correlations, which are for example reflected in a non-zero cross-correlation of the innovations when multiple regression is used to analyze linear Granger causality. In its strong formulation (Granger, 1980) the existence of instantaneous dependence is tested with the criterion of conditional independence

$$p(X_{i+1}, Y_{i+1}|\{V\}^i) = p(X_{i+1}|\{V\}^i)p(Y_{i+1}|\{V\}^i), \quad (6)$$

called by Granger *instantaneous causality* between  $X$  and  $Y$ . Both criteria of Granger causality and instantaneous causality can be

generally tested using the conditional Kullback-Leibler divergence (Cover and Thomas, 2006)

$$KL(p(Y|X); q(Y|X)) = \sum_{x,y} p(x, y) \log \frac{p(y|x)}{q(y|x)}. \quad (7)$$

The KL-divergence is non-negative and only zero if the distributions  $p$  and  $q$  are equal. Accordingly, plugging into Equation (7) the probability distributions of the criterion of Granger causality of Equation (5) we get (Marko, 1973).

$$\begin{aligned} T_{X \rightarrow Y|Z} &= I(Y_{i+1}, X^i|Y^i, Z^i) \\ &= KL(p(Y_{i+1}|Y^i, Z^i, X^i); p(Y_{i+1}|Y^i, Z^i)), \end{aligned} \quad (8)$$

which is a conditional mutual information often referred to as transfer entropy (Schreiber, 2000). Analogously, a general information-theoretic measure of instantaneous causality is obtained plugging the probabilities of Equation (6) into Equation (7) (e.g., Rissanen and Wax, 1987; Chicharro and Ledberg, 2012b):

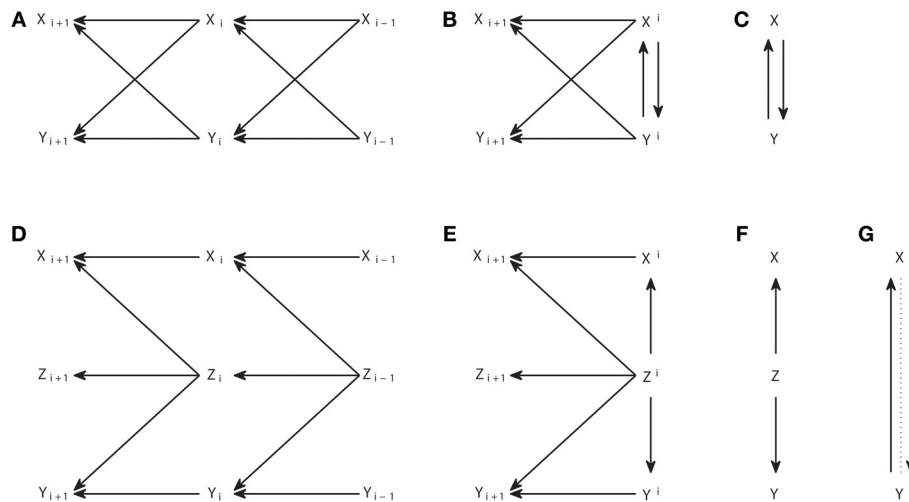
$$\begin{aligned} T_{X \cdot Y|Z} &= I(X_{i+1}; Y_{i+1}|X^i, Y^i, Z^i) \\ &= KL(p(Y_{i+1}|X_{i+1}, X^i, Y^i, Z^i); p(Y_{i+1}|X^i, Y^i, Z^i)). \end{aligned} \quad (9)$$

Note that here we use Granger causality to refer to the criterion of conditional independence of Equation (5), and not to the particular measure resulting from its linear implementation (Bressler and Seth, 2011). In that sense, we include in the Granger causality methodology not only the transfer entropy but also other measures developed for example to study causality in the spectral domain (Chicharro, 2011, 2014a).

### GRAPHICAL REPRESENTATIONS OF CAUSAL INTERACTIONS

Causal representations are also commonly used when applying Granger causality analysis. However, we should distinguish other types of causal graphs from the *causal structures*. The connections in a causal structure are such that they reflect in a unique way the arguments of the functions in the causal model which provides a mechanistic explanation of the generation of the variables. This means that, for processes, when the functional model consists of differential equations that in their discretized form are like in Equation (3), the causal structure comprises the variables corresponding to all sampling times, explicitly reflecting the temporal nature of the processes. **Figures 4A,D** show two examples of interacting processes, the first with two bidirectionally connected processes and the second with two processes driven by a common driver.

The corresponding causal structures constitute a *microscopic* representation of the processes and their interactions, since they contain the detailed temporal information of the exact lags at which the causal interactions occur. However, when many processes are considered together, like in a brain connectivity network, this representation becomes unmanageable. Chicharro and Ledberg (2012b) showed that an intermediate *mesoscopic* representation is naturally compatible with Granger causal analysis, since it contains the same groups of variables used in Equations



**FIGURE 4 | Graphical representations of interacting processes at different scales. (A–C)** Represent the same bivariate process at a micro, meso, and macroscopic scale. **(D–F)** Represent another process also at these

different scales, and **(G)** represents the Granger causal and instantaneous causality relations when only  $X$  and  $Y$  are included in the Granger causality analysis.

(5, 6). These graphs are analogous to the *augmentation graphs* used in Dahlhaus and Eichler (2003). At the mesoscopic scale the detailed information of the lags of the interactions is lost and thus also is lost the mapping to the parental structure in the causal model, so that an arrow cannot be associated with a particular causal mechanism. Accordingly, the mesoscopic graphs are not in general DAGs, as illustrated by Figure 4B.

*Macroscopic* graphs offer an even more schematized representation (Figures 4C,F) where each process corresponds to a single node. Moreover, the meaning of the arrows changes depending on the use given to the graph. If one is representing some known dynamics, for example when studying some simulated system, then the macroscopic graph can be just a summary of the microscopic one. On the other hand, for experimental data, the graph can be a summary of the Granger causality analysis and then the arrows represent the connections for which the measure of Granger causality, e.g., the transfer entropy, gives a non-zero value. Analogously, Granger instantaneous causality relations estimated as significant can be represented in the graphs with some undirected link. For example, Figure 4F summarizes the Granger causal relations of the system  $\{X, Y, Z\}$  when all variables are observed, and Figure 4G is a summary of the Granger causal relations (including instantaneous), when the analysis is restricted to the system  $\{X, Y\}$ , taking  $Z$  as a latent process. In Figure 4G the instantaneous causality is indicated by an undirected dotted edge. *Mixed graphs* of this kind have been studied to represent Granger causality analysis, e.g., Eichler (2005, 2007). Furthermore, graph analysis with macroscopic graphs is also common to study structural or functional connectivity (Bullmore and Sporns, 2009).

Apart from the correspondence to a causal model, which is specific of causal structures, it is important to determine for the other graphical representations if it is possible to still apply d-separation or an analogous criterion to read conditional independencies present in the associated probability

distributions. Without such a criterion the graphs are only a basic sketch to gain some intuition about the interactions. For mesoscopic graphs, a criterion to derive Granger causal relations from the graph was proposed by Dahlhaus and Eichler (2003) using moralization (Lauritzen, 1996). Similarly, a criterion of separation was proposed in Eichler (2005) for the mixed graphs representing Granger causality and instantaneous Granger causality. However, in both cases these criteria provide only a sufficient condition to identify independencies, even if stability is assumed, in contrast to d-separation for causal structures or patterns, which under stability provides an if and only if condition.

## EXTENSION OF PEARL'S CAUSAL MODELS TO DYNAMIC SYSTEMS AND RELEVANCE TO STUDYING THE BRAIN'S EFFECTIVE CONNECTIVITY

Above we have reviewed two different approaches to causal inference. The approach by Pearl is based on causal models and explicitly considers the limitations of causal inference, introducing the notion of observational equivalence and explicitly addressing the consequences of potential latent variables in the algorithm IC\*. Conversely, Granger causality more operationally provides a criterion of causality between processes specific for a context, and does not explicitly handle latent influences. Moreover, the Pearl's approach is not restricted with respect to the nature of the variables and should thus be applicable also to processes. Since this approach is more powerful in how it treats latent variables and in how it indicates the limits of what can be learned, in the following we investigate how the IC and IC\* algorithms can be applied to dynamic processes and how they are related to Granger causality.

## CAUSAL INFERENCE WITHOUT LATENT VARIABLES FOR DYNAMIC PROCESSES

We here reconsider the IC algorithm for the especial case of dynamic processes. Of course one can apply the IC algorithm directly, since there are no assumptions about the nature of the

variables. However, the causal structures associated with dynamic processes (e.g., the microscopic graphs in **Figures 4A,D**) have a particular structure which can be used to simplify the algorithm. In particular, the temporal nature of causality assures that all the arrows should point from a variable at time  $i$  to another at time  $i + d$ , with  $d > 0$ . This means that the arrows can only have one possible direction. Therefore, once Step 1 has been applied to identify the skeleton of the pattern, all the edges can be assigned a head directly, without necessity to apply Steps 2 and 3. Furthermore, even Step 1 can be simplified, since the temporal precedence give us information of which variables should be used to search for an appropriate set  $S_{ab}$  that renders  $a$  and  $b$  conditionally independent. In particular, for  $V_{j,i}$  and  $V_{j',i+d}$ , indicating the variable of process  $j$  at the time instant  $i$  and the variable of process  $j'$  at time  $i + d$ , respectively, the existence of  $V_{j,i} \rightarrow V_{j',i+d}$  can be inferred testing if it does not hold

$$p(V_{j',i+d}|\{V\}^{i+d-1}) = p(V_{j',i+d}|\{V\}^{i+d-1} \setminus V_{j,i}), \quad (10)$$

where  $\{V\}^{i+d-1} \setminus V_{j,i}$  means the whole past of the system at time  $i + d$  excluding  $V_{j,i}$ . This is because conditioning on the rest of the past blocks any path that can link the two nodes except a direct arrow. Therefore,  $S_{ab} = \{V\}^{i+d-1} \setminus V_{j,i}$  is always a valid set to check if  $V_{j,i}$  and  $V_{j',i+d}$  are conditionally independent, even if considerations about the estimation of the probability distributions lead to seek for smaller sets (e.g., Faes et al., 2011; Marinazzo et al., 2012).

Note that the combination of the assumption of no latent variables with the use of temporal precedence to add the direction of the arrows straightforwardly after Step 1 of the IC algorithm leads to patterns that are always complete DAGs. This straightforward completion indicates that there is a unique relation between the pattern and the underlying causal structure, that is, there are no two different causal structures sharing the same pattern. For example, from the three causal structures that are observationally equivalent in **Figures 1A–C**, if only one direction of the arrows is allowed (from right to left for consistency with **Figure 4**) then only the causal structure of **Figure 1B** is possible.

There is a clear similarity between the criterion of Equation (10) to infer the existence of a single link in the causal structure and the criterion of Granger causality in Equation (5). In particular, Equation (10) is converted into Equation (5) by two substitutions: (i) taking  $d = 1$  and (ii) taking the whole past  $V_j^{i+d-1}$  instead of a single node  $V_{j,i}$ . Both substitutions reflect that Granger causality analysis does not care about the exact lag of the causal interactions. It allows representing the interactions in a mesoscopic or macroscopic graph, but is not enough to recover the detailed causal structure. By taking  $d = 1$  and taking the whole past one is including any possible node that can have a causal influence from process  $j$  to process  $j'$ . The Granger causality criterion combines in a single criterion the pile of criteria of Equation (10) for different  $d$ . Accordingly, in the absence of latent variables, Granger causality can be considered as a particular application of the IC algorithm, simplified accordingly to the objectives of characterizing the causal relations between the processes. Note that this equivalence relies on the stochastic nature of the endogenous variables in Pearl's model (Equation 1).

Furthermore, it is consistent with the relation between Granger causality and notions of structural causality as discussed in White and Lu (2010).

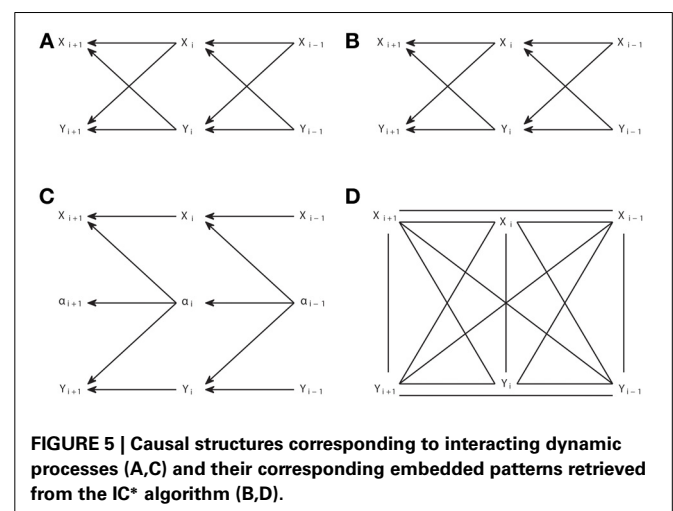
### CAUSAL INFERENCE WITH LATENT VARIABLES FOR DYNAMIC PROCESSES

We have shown above that in the absence of latent processes adding temporal precedence as a constraint tremendously simplifies the IC algorithm and creates a unique mapping between causal structures and patterns. Adding temporal precedence makes causal inference much easier because time provides us with extra information and, in the absence of latent variables, no complications are added when dealing with dynamic processes.

We now show that this simplification does not hold anymore when one considers the existence of latent processes. We start with two examples in **Figure 5** that illustrate how powerful or limited can be the application of the IC\* algorithm to dynamic processes. Note that the IC\* algorithm is applied taking the causal structures in **Figures 5A,C** as an interval of stationary processes, so that the same structure holds before and after the nodes displayed.

In **Figure 5A** we display a causal structure of two interacting processes without any latent process, and in **Figure 5B** the corresponding embedded pattern. We can see that, even allowing for the existence of latent processes, the IC\* algorithm can result in a DAG which completely retrieves the underlying causal structure. In this case the output of the IC algorithm and of the IC\* algorithm are the same pattern, but the output of the IC\* algorithm is actually a much stronger result, since it states that a bidirectional genuine causation must exist between the processes even if one considers that some other latent processes exist.

Conversely, consider the causal structure of **Figure 5C** in which  $X$  and  $Y$  are driven by a hidden process. The resulting embedded pattern is a completely filled undirected graph, in which all nodes are connected to all nodes since there are no conditional independencies. Further using the extra information provided by temporal precedence—by substituting all horizontal undirected links by dashed arrows pointing to the left and vertical links by bidirectional arrows—does not allow us to better retrieve



**FIGURE 5 | Causal structures corresponding to interacting dynamic processes (A,C) and their corresponding embedded patterns retrieved from the IC\* algorithm (B,D).**



the underlying causal structure since, unlike the patterns resulting from the IC algorithm, the embedded patterns resulting from the IC\* algorithm do not have to share the skeleton with the causal structures belonging to their dependency equivalence class.

The IC\* algorithm is not suited to study dynamic processes for two main reasons. First, the embedded pattern chosen as a representation of the dependency class is strongly determined by the selection of projections as the representative subset of the class. The projections exclude connections between the latent variables or latent variables connected to more than two observed variables. By contrast, a latent process generally consists *per se* in a complex structure of latent variables. In particular, commonly causal interactions exist between the latent nodes, since most latent processes will have a causal dependence on their own past, and each node does not have a causal influence on only two observable nodes.

Second, the IC\* algorithm is designed to infer the causal structure associated with the causal model. This means that, for dynamic processes, for which generally an acyclic directed graph is only obtained when explicitly considering the dynamics, the IC\* algorithm necessarily infers the microscopic representation of the causal interactions. In contrast to the case of the IC algorithm in which there are no latent variables, it is not possible to establish an immediate correspondence with Granger causality analogous to the relation between Equation (5) and Equation (10). The fact that the IC\* algorithm necessarily has to infer the microscopic causal structure is not desirable for dynamic processes. This is because of several reasons related to the necessity to handle a much higher number of variables (nodes). In first instance, it requires the estimation of many more conditional independencies in Step 1 of the algorithm, which is a challenge for practical implementations (see Supplementary Material for discussion of the implementation of the algorithms). In second instance, the microscopic embedded pattern, as for example the one in **Figure 5D**, can be too detailed without actually adding any information about the underlying causal structure but, on the contrary, rendering the reading of its basic structure less direct.

Here we propose a new algorithm to obtain a representation of the dependency class when studying dynamic processes. The new algorithm recasts the basic principles of the IC\* algorithm but has the advantage that it avoids the assumptions related to the projections, and allows to study causal interactions between the processes at a macroscopic level, without necessarily examining the lag structure of the causal interactions. With respect to usual applications of Granger causality, the new algorithm has the advantage that it explicitly considers the existence of potential latent processes. It is important to note that the new algorithm is not supposed to outperform the IC\* algorithm in the inference of the causal interactions. They differ only in the number of conditional independencies that have to be tested, much lower for the new algorithm since only the macroscopic causal structure is examined, and in the form of the embedded pattern chosen to represent the dependency equivalent class. In simpler terms, for dynamic processes, the new algorithm offers a more appropriate representation of the class of networks compatible with the estimated conditional independencies. Both algorithms rely on the same framework to infer causality from

conditional independencies, and theoretically their performance is only bounded by the existence of observationally equivalent causal structures. None of the two algorithms addresses the practical estimation of the conditional independencies, and thus any evaluation of their practical performance is specific to the particular choice of how to test conditional independence (see Supplementary Material for discussion of the implementation).

In comparison to the assumptions related to projections, the new algorithm assumes that any latent process is such that its present state depends in a direct causal way on its own past, that is, that its autocorrelation is not only indirectly produced by the influence of other processes. In practice, this means that we are excluding cases like an uncorrelated white noise that is a common driver of two observable processes. The reason for this assumption is that, excluding these processes without auto-causal interactions, we have (Chicharro and Ledberg, 2012b) that there is a clear difference between the effect of hidden common drivers and the effect of hidden processes that produce indirect causal connections (i.e.,  $X \rightarrow \alpha \rightarrow Y$ ). In particular, if we have a system composed by two observable processes  $X$  and  $Y$  such that a hidden process  $\alpha$  mediates the causal influence from  $X$  to  $Y$ , we have that

$$X \rightarrow \alpha \rightarrow Y \Rightarrow T_{X \rightarrow Y} > 0 \wedge T_{X \cdot Y} = 0, \quad (11)$$

where  $\wedge$  indicates conjunction. Conversely, if the system  $\alpha$  is a common driver we have that

$$X \leftarrow \alpha \rightarrow Y \Rightarrow T_{X \rightarrow Y} > 0 \wedge T_{X \cdot Y} > 0, \quad (12)$$

We see that common drivers and mediators have a different effect regarding the induction of instantaneous causality. This difference generalizes to multivariate systems with any number of observed or latent processes (see Supplementary Material). Common drivers are responsible for instantaneous causality. In fact, if there is no set of observable processes such that when conditioning on it the instantaneous causality is canceled, then some latent common drivers must exist since *per se* causality cannot be instantaneous unless we think about entanglement of quantum states. Accordingly,

$$\forall S T_{X \cdot Y|S} > 0 \Leftrightarrow \text{common driver latent processes cause instantaneous causality}, \quad (13)$$

where one or more common driver latent processes may be involved. Properties in Equations (11–13) are used in the new algorithm. The input is the joint distribution that includes the variables corresponding to sampling time  $i + 1$  and to the past of the observable processes  $V_O$ , i.e.,  $p(\{V_{O,i+1}\}, \{V_O^i\})$ . The output is a macroscopic graph which reflects all and no more Granger causality and instantaneous causality relationships than the ones present in  $p(\{V_{O,i+1}\}, \{V_O^i\})$ . The algorithm proceeds as follows:

#### ICG\* ALGORITHM (INDUCTIVE CAUSATION WITH LATENT VARIABLES USING GRANGER CAUSALITY)

- (1) For each pair of processes  $a$  and  $b$  in  $\{V_O\}$  search for a set  $S_{ab}$  of processes such that  $T_{a \cdot b|S_{ab}} = 0$  holds in  $p(\{V_O\})$ , i.e.,

there is no instantaneous causality between  $a$  and  $b$  given  $S_{ab}$ . Construct a macroscopic graph with each process represented by one node and linking the nodes  $a$  and  $b$  with a bidirectional arrow  $a \leftrightarrow b$  if and only if  $S_{ab}$  is not found.

- (2) For each pair  $a$  and  $b$  not linked by a bidirectional arrow search for a set  $S_{ab}$  of processes such that  $T_{a \rightarrow b|S_{ab}} = 0$  holds in  $p(\{V_O\})$ , i.e., there is no Granger causality from  $a$  to  $b$  given  $S_{ab}$ . Link the nodes  $a$  and  $b$  with a unidirectional arrow  $a \rightarrow b$  if and only if  $S_{ab}$  is not found.
- (3) For each pair  $a$  and  $b$  not linked by a bidirectional arrow search for a set  $S_{ab}$  of processes such that  $T_{b \rightarrow a|S_{ab}} = 0$  holds in  $p(\{V_O\})$ , i.e., there is no Granger causality from  $b$  to  $a$  given  $S_{ab}$ . Link the nodes  $a$  and  $b$  with a unidirectional arrow  $a \leftarrow b$  if and only if  $S_{ab}$  is not found.

The zero values of the Granger measures indicate the existence of some conditional independencies. Step 1 identifies the existence of latent common drivers whenever Granger instantaneous causality exists and marks it with a bidirectional arrow. Steps 2 and 3 identify Granger causality in each direction when there is no Granger instantaneous causality. In fact Granger causality will also be present for the bidirectionally linked nodes, but there is no need to check it separately, given Equation (12). Steps 1–3 are analogous to Step 1 of the IC\* algorithm since conditioning sets of different size have to be screened, but now the conditional independencies examined are not between single variables but between processes and this is why Granger causality measures are used.

The algorithm differs in two principle ways from how Granger causality is commonly used. First, Granger causality is not applied once for each pair of nodes, but one has to search for a context that allows assessing if a conditional independence exists. This is different from applying bidirectional Granger causality to all combinations of nodes, and also from applying to all combinations of nodes conditional Granger causality conditioning on the whole rest of the system. The reason is that, as discussed in Hsiao (1982) and Ramb et al. (2013), when latent processes exist, further adding new processes to the conditioning can convert a zero Granger causality into positive.

Second, an explicit consideration of the possible existence of latent processes is incorporated, to our knowledge for the first time, when applying Granger causality. A bidirectional arrow indicates that the dependencies between the processes can only be explained by latent common drivers. We should note that this does not discard that in addition to common drivers there are directed causal links between the processes, in the same way that unidirectional arrows do not discard that the causal influence is not direct but through a mediator latent processes. This is because the output of the algorithm is again a representation of a class of causal structures and thus these limitations are common to the IC\* algorithm which also implicitly allows the existence of multiple hidden paths between two nodes or of latent mediators. Of course, when studying brain connectivity it can be relevant to establish for example if two regions are directly causally connected, but this cannot be done without recording from the potential intermediate regions, or using some heuristic knowledge of the anatomical connectivity.

The output of the ICG\* algorithm most often is more intuitive about the causal influences between the processes than the embedded pattern resulting from the IC\* algorithm and does not need to consider the microscopic structure. For example, while for the causal structure of Figure 5C we found that the IC\* algorithm provides as output the embedded pattern of Figure 5D (which has a lot of edges that are not in the underlying causal structure so that a direct mapping is not possible), we found that the ICG\* algorithm simply provides as output  $X \leftrightarrow Y$  thereby revealing synthetically, directly, and correctly the existence of at least one latent common driver.

However, to be meaningful as a representation of the conditional independencies associated with the Granger causality relationships, we need to complement the algorithm with a criterion of separation analogous to the one available for the patterns and embedded patterns obtained from the IC and IC\* algorithms, respectively. In particular, d-separation can be again used, now considering a collider on a path to be any node with two head to head arrows on the path, where the heads can belong to the two types of arrows, i.e., unidirectional or bidirectional. Accordingly, the subsequent sufficient conditions can be applied to read the Granger causal relations from the graph:

#### Graphical sufficient condition for Granger non-causality

$X$  is d-separated from  $Y$  by  $S$  on each path between  $X$  and  $Y$  with an arrow pointing to  $Y \Rightarrow T_{X \rightarrow Y|S} = 0$ .

#### Graphical sufficient condition for instantaneous non-causality

$X$  is d-separated from  $Y$  by  $S$  on each path between  $X$  and  $Y$  with an arrow pointing to  $X$  and an arrow pointing to  $Y \Rightarrow T_{X \cdot Y|S} = 0$ .

Proofs for these conditions are provided in the Supplementary Material. As in general for d-separation, these conditions become if and only if conditions if further assuming *stability*. The conditions here introduced for the graphs resulting from the ICG\* algorithm are very similar to the ones proposed by Eichler (2005) for mixed graphs. Also for mixed graphs Eichler (2009) proposed an algorithm of identification of Granger causality relationships. The critical difference with respect to this previous approach is that here instantaneous causality is considered explicitly as the result of existing latent variables, according to Equations (11–13), while in the mixed graphs there is no explanation of how it arises from the underlying dynamics.

#### ANALYSIS OF THE EFFECT OF LATENT VARIABLES

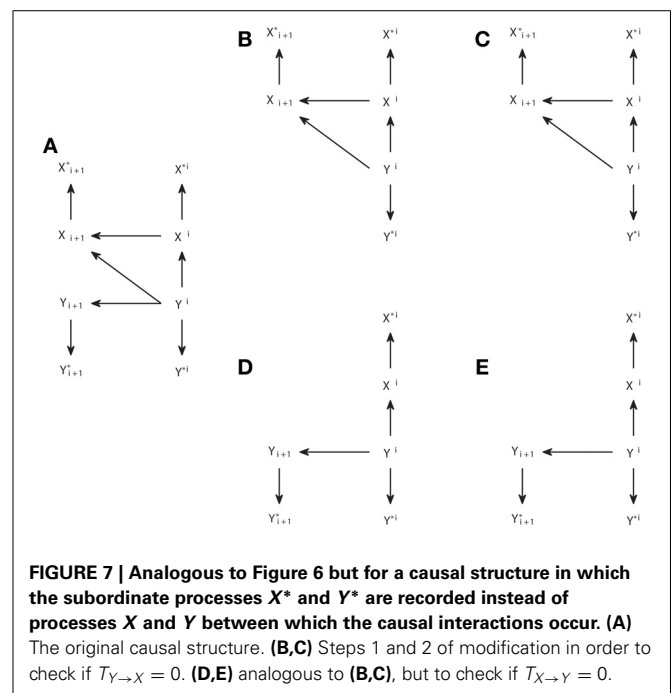
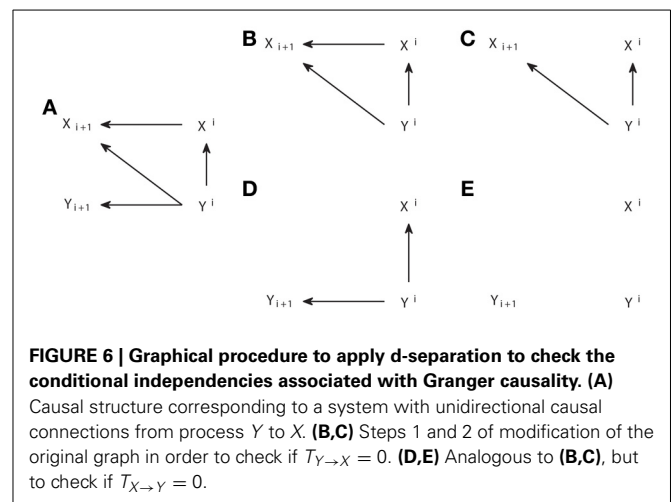
The results above concern the application of general algorithms of causal inference to dynamic processes, and how these algorithms are related to the Granger causality analysis. The perspective was focused on how to learn the properties of an unknown causal structure from the conditional independencies contained in a probability distribution obtained from recorded data. In this section we address the opposite perspective, i.e., we assume that we know a causal structure and we focus on examining what we learn by reading the conditional independencies that are present in any distribution compatible with the structure. We will see that a simple analysis applying d-separation can explain in a simple way

many of the scenarios in which Granger causality analysis can lead to inconsistent results about the causal connections. We here term the positive values of Granger causality that do not correspond to arrows in the causal structure as *inconsistent positives*. These are to be distinguished from *false positives* as commonly understood in hypothesis testing, since the inconsistent positives do not result from errors related to estimation, but, as we show below, they result from the selection of subordinate signals as the ones used to carry out the causal inference analysis.

The definition of d-separation does not provide a procedure to check if all paths between the two variables which conditional independence is under consideration have been examined. However, a procedure based on graphical manipulation exists that allows checking all the paths simultaneously (Pearl, 1988; Kramers, 1998). We here illustrate this procedure to see how it supports the validity of Granger causality for causal inference when there are no latent processes and then apply it to gain more intuition about different scenarios in which inconsistent positive values are obtained. The procedure works as follows: to check if  $X$  is d-separated from  $Y$  by a set  $S$ , first create a subgraph of the complete structure including only the nodes and arrows that are attained moving backward from  $X$ ,  $Y$  or the nodes in  $S$  (i.e., only the ancestors  $\text{an}(X, Y, S)$  appear in the subgraph); second, delete all the arrows coming out of the nodes belonging to  $S$ ; finally, check if there is still any path connecting  $X$  and  $Y$  and if such a path does not exist,  $X$  and  $Y$  are separated by  $S$ .

In **Figure 6** we display the modifications of the graph performed to examine the conditional independencies associated with the criterion of Granger causality. In **Figure 6A** we show the mesoscopic graph of a system with unidirectional causal interactions from  $Y$  to  $X$ . In **Figures 6B,C** we show the two subsequent modifications of the graph required to check if  $T_{Y \rightarrow X} = 0$ , while in **Figures 6D,E** we show the ones required to check if  $T_{X \rightarrow Y} = 0$ . In **Figure 6B** the subgraph is selected moving backward from  $\{X_{i+1}, X^i, Y^i\}$ , the nodes involved in the corresponding criterion in Equation (5). In **Figure 6C** the arrow leaving the conditioning variable  $X^i$  is removed. The analogous procedure is followed in **Figures 6D,E**. It can be seen that in **Figure 6C**  $Y^i$  and  $X_{i+1}$  are still linked, indicating that  $T_{Y \rightarrow X} > 0$ , while there is no link between  $X^i$  and  $Y_{i+1}$  in **Figure 6E**, indicating that  $T_{X \rightarrow Y} = 0$ .

Therefore, d-separation allows us to read the Granger causal relations from the structure of **Figure 6A**. One may ask why we should care about d-separation providing us with information which is already apparent from the original causal structure in **Figure 6A** that we assume to know. The answer is that, when one constructs a causal structure to reproduce the setup in which the observable data are recorded, the Granger causal relations between those are generally not so obvious from the causal structure. To illustrate that, we consider below a quite general case in which the Granger causality analysis is not applied to the actual processes between which the causal interactions occur, but to some time series derived from them. In **Figure 7A** we display the same system with a unidirectional causal interaction from  $Y$  to  $X$ , but now adding the extra processes  $X^*$  and  $Y^*$ , which are obtained by some processing of  $X$  and  $Y$ , respectively. If only the processes  $X^*$  and  $Y^*$  are observable, and the Granger causality analysis is applied to them, this case comprises scenarios such as



the existence of measurement noise, or the case of fMRI in which the observed BOLD responses only indirectly reflect the hidden neuronal states (Friston et al., 2003; Seth et al., 2013).

We can see in **Figure 7C** that  $T_{Y^* \rightarrow X^*} > 0$ , as if the analysis was done on the original underlying processes  $X$  and  $Y$ , for which  $T_{Y \rightarrow X} > 0$ . However, in the opposite direction we see in **Figure 7E** that an inconsistent positive value appears, since also  $T_{X^* \rightarrow Y^*} > 0$ , while  $T_{X \rightarrow Y} = 0$ . We can see that this happens because  $Y^i$  acts as a common driver of  $Y_{i+1}^*$  and  $X_{i+1}^*$ , through the paths  $Y^i \rightarrow Y_{i+1} \rightarrow Y_{i+1}^*$  and  $Y^i \rightarrow X^i \rightarrow X_{i+1}^*$ , respectively. This case, in which the existence of a causal interaction in one direction leads to an inconsistent positive in the opposite direction when there is an imperfect observation of the driven system (here  $Y$ ), has been recently discussed in Smirnov (2013). Smirnov (2013) has exemplified that the effect of measurement noise or

time aggregation—due to low sampling—can be understood in this way. However, the illustration in Smirnov (2013) is based on the construction of particular examples and requires complicated calculations to obtain analytically the Granger causality values. With our approach, general conclusions are obtained more easily by applying d-separation to a causal structure that correctly captures how the data analyzed are obtained. Nonetheless, the use of graphical criteria and exemplary simulations is complementary, since one advantage of the examples in Smirnov (2013) is that it is shown that the non-negative values of the Granger causality measure in the opposite direction can have a magnitude comparable or even bigger than those in the correct direction.

In **Table 1** we summarize some paradigmatic common scenarios in which a latent process acts as a common driver leading to inconsistent positives in Granger causality analysis. In all these cases Granger causality can easily be assessed in a general way from the corresponding causal structure that includes the latent process. First, when non-stationarities exist, time can act as a common driver since the time instant provides information about the actual common dynamics. This is the case for example of integrated processes, for which an adapted formulation of Granger causality has been proposed (Lütkepohl, 2005). Also event-related setups may produce a common driver, since the changes in the ongoing state from trial to trial can simultaneously affect the two processes (e.g., Wang et al., 2008).

The other cases listed in **Table 1** are analogous to the one illustrated in **Figure 7**. Discretizing continuous signals can induce inconsistent positives (e.g., Kaiser and Schreiber, 2002) and also measurement noise (e.g., Nalatore et al., 2007). In both cases Granger causality is calculated from subordinate signals, obtained after binning or after noise contamination, which constitute a voluntary or unavoidable processing of the underlying interacting processes. Similarly, the hemodynamic responses  $h(X)$  and  $h(Y)$  only provide with a subordinate processed signal from the neural states (e.g., Roebroeck et al., 2005; Deshpande et al., 2010).

**Table 1 | Cases in which a hidden common driver leads to inconsistent positive Granger causality from the observed process derived from process  $X$  to the observed process derived from process  $Y$  when there are unidirectional causal connections from  $Y$  to  $X$  (or processes  $Y_k$  to  $X_k$ ).**

	Observed variables	Common driver
1 Non-stationarity	$X_i$ and $Y_i$	Time
2 Event-related setup	$X_i$ and $Y_i$	Trial ongoing state
3 Discretizing	$\text{Bin}(X)_i$ and $\text{Bin}(Y)_i$	Underlying process $Y$
4 Measurement noise	$X_i^* = X_i + \varepsilon_{X,i}$ and $Y_i^* = Y_i + \varepsilon_{Y,i}$	Underlying process $Y$
5 fMRI analysis	$h(X)_i$ and $h(Y)_i$	Underlying process $Y$
6 Time aggregation	$X_{T_i}$ and $Y_{T_i}$	Unsampled time instants of $Y$
7 Spatial aggregation	$X_i^* = \sum_k X_{k,i}$ and $Y_i^* = \sum_k Y_{k,i}$	Underlying processes $Y_k$

In the case of time aggregation, the variables corresponding to unsampled time instants are the ones acting as common drivers (Granger, 1963). The continuous temporal nature of the processes has been indicated as a strong reason to advocate for the use of DCM instead of autoregressive modeling (see Valdes-Sosa et al., 2011 for discussion). Finally, aggregation also takes place in the spatial domain. To our knowledge, the consequences of spatial aggregation for the interpretation of the causal interactions have been studied less extensively so far than those posed by time aggregation, and thus we focus on spatial aggregation in the section below.

## THE CASE OF SPATIAL AGGREGATION

We next investigate what happens when it is not possible to measure directly the activity of the neural sources among which the causal interactions occur because only spatially aggregated signals that aggregate many different neural sources are recorded. For example, a single fMRI voxel reflects the activity of thousands of neurons (Logothetis, 2008), or the local Field Potential amplitude measured at a cortical location captures contributions from several sources spread over several hundreds of microns (Einevoll et al., 2013). The effect of spatial aggregation on stimulus coding and information representations has been studied theoretically (Scannell and Young, 1999; Nevado et al., 2004), but its effect on causal measures of the kind considered here still needs to be understood in detail.

Possible distortions introduced by spatial aggregation depend on the nature of the processes and the scale at which the analysis is done. In particular, neuronal causal interactions occur at a much more detailed scale (e.g., at the level of synapses) than the scale corresponding to the signals commonly analyzed. It is not clear, and to our knowledge it has not been addressed, how causal relations at a detailed scale are preserved or not when zooming out to a more macroscopic representation of the system. As we will discuss in more depth in the Discussion, the fact that a macroscopic model provides a good representation of macroscopic variables derived from the dynamics does not assure that it also provides a good understanding of the causal interactions.

In general, the effect of spatial aggregation on causal inference can be understood examining a causal structure analogous to the one of **Figure 7**, but where instead of a single pair of underlying processes  $X$  and  $Y$  there are two sets  $X_k$ ,  $k = 1, \dots, N$ , and  $Y_{k'}$ ,  $k' = 1, \dots, N'$  between which the causal interactions occur. The signals observed are just an average or a sum of the processes,  $X^* = \sum_{k=1}^N X_k$  and  $Y^* = \sum_{k=1}^{N'} Y_k$ . For example, in the case of the brain, the processes can correspond to the firing activity of individual neurons, and the recorded signals to some measure of the global activity of a region, like the global rates  $r_X$  and  $r_Y$ . Even if for each pair  $X_k$ ,  $Y_{k'}$  a unidirectional causal connection exists, the Granger causality between  $r_X$  and  $r_Y$  will be positive in both directions, as can be understood from **Figure 7**.

We will now examine some examples of spatial aggregation. As we mentioned in the Introduction, here we specifically focus on causal inference, i.e., determining which causal interactions exist. We do not address the issue of further quantifying the magnitude of causal effects, since this is generally more difficult (Chicharro and Ledberg, 2012b; Chicharro, 2014b) or even



in some cases not meaningful (Chicharro and Ledberg, 2012a). In the case of spatial aggregation, the fact that Granger causality calculated from the recorded signals has always positive values in both directions is predicted by the graphical analysis based on d-separation. However, in practice the conditional independencies have to be tested from data instead of derived using Equation (4). When tested with Granger causality measures, the magnitude of the measure is relevant, even if not considered as a quantification of the strength of the causal effect, because it can determine the significance of a non-negative value. The relation between magnitude and significance depends on the estimation procedure and on the particular procedure used to assess the significance levels (e.g., Roebroeck et al., 2005; Besserve et al., 2010). It is not on the focus of this work to address a specific implementation of the algorithms of causal inference, which requires specifying these procedures (see Supplementary Material for discussion). Nonetheless, we now provide some numerical examples following the work of Smirnov (2013) to illustrate the impact of spatial aggregation on the magnitude of the Granger causality measures and we show that the inconsistent positives can have comparable or even higher magnitude than the consistent positives, and thus are expected to impair the causal inference performance.

In **Figure 8A** we show the macroscopic graph representing the spatial aggregation of two processes in two areas, respectively. The processes are paired, so that a unidirectional interaction from  $X_k$  to  $Y_k$  exists, but the signals recorded on each area are a weighted sum of the processes, that is, we have  $X = m_x X_1 + (1 - m_x) X_2$ , and analogously for  $Y$  with  $m_y$ . This setup reproduces some basic properties of neural recordings, in which different sources contribute with different intensity to the signal recorded in a position. To be able to calculate analytically the Granger causality measures we take, as a functional model compatible with the causal structure that corresponds to **Figure 8A**, a multivariate linear Gaussian autoregressive process. Considering the whole dynamic process  $W = \{X_1, X_2, Y_1, Y_2\}$ , the autoregressive process is expressed as

$$\begin{pmatrix} X_{1i+1} \\ X_{2i+1} \\ Y_{1i+1} \\ Y_{2i+1} \end{pmatrix} = \begin{pmatrix} c_{11} & c_{12} & 0 & 0 \\ c_{21} & c_{22} & 0 & 0 \\ 0.8 & 0 & 0.8 & 0 \\ 0 & 0.8 & 0 & 0.8 \end{pmatrix} \begin{pmatrix} X_{1i} \\ X_{2i} \\ Y_{1i} \\ Y_{2i} \end{pmatrix} + \begin{pmatrix} \varepsilon_{x1i} \\ \varepsilon_{x2i} \\ \varepsilon_{y1i} \\ \varepsilon_{y2i} \end{pmatrix}, \quad (14)$$

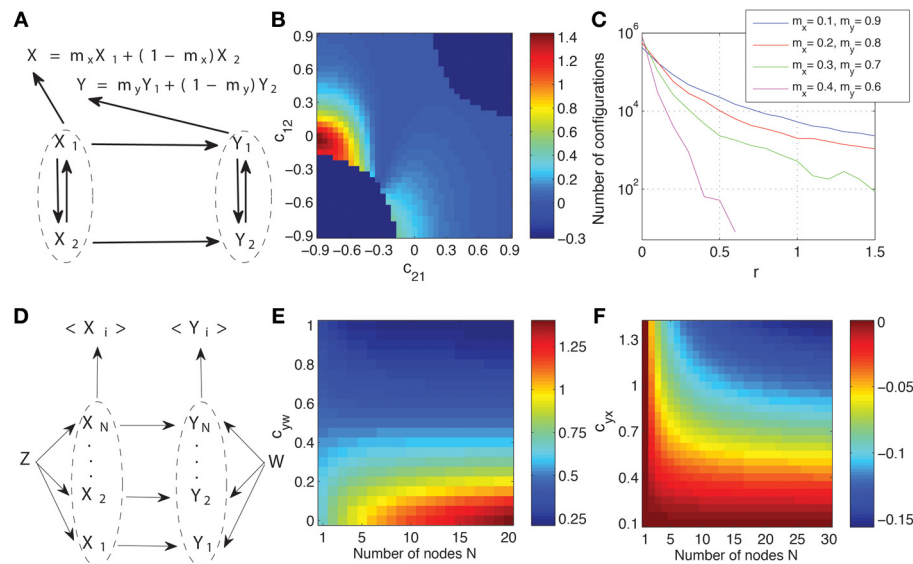
where  $C$  is the matrix that determines the connectivity. For example, the coefficient  $c_{12}$  indicates the coupling from  $X_2$  to  $X_1$ . Matrix  $C$  is compatible with the graph of **Figure 8A**: we fix  $c_{13} = c_{14} = c_{23} = c_{24} = c_{32} = c_{41} = 0$  so that inter-areal connections are unidirectional from  $X_k$  to  $Y_k$ . Furthermore, to reduce the dimensions of the parameter space to be explored, we also fix  $c_{34} = c_{43} = 0$ , so that  $Y_1$  and  $Y_2$  are not directly connected, and  $c_{31} = c_{42} = c_{33} = c_{44} = 0.8$ . The autoregressive process is of order one because the future values at time  $i + 1$  only depend on time at  $i$ . We assume that there are no latent influences and thus the different components of the noise term  $\varepsilon$  are uncorrelated, i.e., the innovations have a diagonal covariance matrix. We fix the variance of all innovations to 1. Accordingly, the parameter space that we explore involves the coefficients  $c_{11}$ ,  $c_{22}$ ,  $c_{12}$ , and  $c_{21}$ . We exclude those configurations which are non-stationary.

The observed signals are obtained from the dynamics as a weighted average. The Granger causality measures can then be calculated analytically from the second order moments (see Chicharro and Ledberg, 2012b and Smirnov, 2013 for details). In all cases 20 time lags of the past are used, which is enough for convergence. If the Granger causality measures were calculated for each pair of underlying processes separately, we would get always  $T_{X_k \rightarrow Y_k} > 0$  and  $T_{Y_k \rightarrow X_k} = 0$ . However, for the observed signals  $X$  and  $Y$ , inconsistent positives are expected. To evaluate the magnitude of these inconsistent positives we calculate their relative magnitude.

$$r = T_{Y \rightarrow X} / T_{X \rightarrow Y}. \quad (15)$$

In **Figure 8B** we show the values of  $r$  in the space of  $c_{12}$ ,  $c_{21}$ , fixing  $c_{11} = 0.8$  and  $c_{22} = 0.2$ . Furthermore, we fix  $m_x = 0.3$  and  $m_y = 0.7$ . This means that  $X_2$  has a preeminent contribution to  $X$  while  $Y_1$  has a preeminent contribution to  $Y$ . We indicate the excluded regions where non-stationary processes are obtained with  $r = -0.3$ . In the rest of the space  $r$  is always positive, but can be low ( $\sim 10^{-5}$ ). However, for some regions  $r$  is on the order of 1, and even bigger than 1. In particular, this occurs around  $c_{12} = 0$ , where  $T_{X \rightarrow Y}$  is small, but also around  $c_{21} = 0$ , where  $T_{X \rightarrow Y}$  is high. Here we only intend to illustrate that non-negligible high values of  $r$  are often obtained, and we will not discuss in detail why some particular configurations enhance the magnitude of the inconsistent positives (a detailed analysis of the dependencies can be found in Chicharro and Ledberg, 2012b and Smirnov, 2013). In **Figure 8C** we show the number of configurations in the complete space of the parameters  $c_{11}$ ,  $c_{22}$ ,  $c_{12}$ , and  $c_{21}$  in which a given  $r$ -value is obtained. We show the results for four combinations of weights. We see that the presence of values  $r > 0.1$  is robust in this space, and thus it is not only for extreme cases that the inconsistent positives would be judged as having a non-negligible relative magnitude. In particular, for this example,  $r$  increases when the weights at the two areas differ, consistently with the intuition that the underlying interactions can be characterized worse when processes from different pairs are preeminently recorded in each area. Note that none of the algorithms of causal inference, including in particular the ICG\*, can avoid obtaining such inconsistent positives. In fact, for the examples of **Figure 8**, in which the only two analyzed signals are those that are spatially aggregated, the ICG\* algorithm is reduced to the calculation of  $T_{X \rightarrow Y}$ ,  $T_{Y \rightarrow X}$ , and  $T_{X,Y}$  for these two signals. This illustrates that no algorithm of causal inference can overcome the limitation of not having access to the sources between which the causal interactions actually occur.

In the example above we focused on evaluating the relative magnitude of inconsistent positives of Granger causality. However, spatial aggregation also affects the magnitude of Granger causality in the direction in which a true underlying causal connection exists. We also examine these effects since, although as we mentioned above it may not be safe to use this magnitude as a measure of the strength of the causal effect, it has been widely used with this purpose or more generally as a measure of directional connectivity (see Bressler and Seth, 2011 for a review). To appreciate this, we examine a system sketched in the macroscopic graph of **Figure 8D**. Here we consider two areas  $X$  and  $Y$  each comprising  $N$  processes. For simplification, instead of considering causal connections internal to each area, the degree of



**FIGURE 8 | Effects of spatial aggregation on Granger causality. (A)** Causal graph representing two areas composed each of two processes and from which signals are recorded as a weighted sum. See the text for details of how a system compatible with the graph is generated as a multivariate linear Gaussian autoregressive process. **(B)** Dependence of the relative magnitude  $r$  of the inconsistent positives of Granger causality (Equation 15) on the space formed by coupling coefficients between  $X_1$  and  $X_2$ . **(C)** Number of configurations with a given value  $r$  for all stationary configurations in the space of the parameters  $c_{11}$ ,  $c_{22}$ ,  $c_{12}$ , and  $c_{21}$  and for different weights

combinations. **(D)** Another example of causal graph where spatial aggregation is present in the recording of the signals from the two areas. The system is again generated as a multivariate autoregressive process with identical connections from  $Z$  to each  $X_k$ , identical from  $W$  to each  $Y_k$ , and identical from each  $X_k$  to each  $Y_k$  (see the main text for details). **(E)** The Granger causality measure  $T_{<X>-><Y>}$  as a function of the coefficient  $c_{yw}$  and the number of processes  $N$ . **(F)** The relative changes  $\Delta T'$  (Equation 16) of the Granger causality measure as a function of the coefficient  $c_{yx}$  and the number of processes  $N$ .

integration within each area is determined by a common driver to all the processes of one area,  $Z$  for  $X_k$  and  $W$  for  $Y_k$ . The coupling between the areas is unidirectional for the pairs  $X_k \rightarrow Y_k$ , and only the average of all the processes is recorded from each area,  $\langle X \rangle$  and  $\langle Y \rangle$ . We now focus on examining how  $T_{<X>-><Y>}$  depends on the number of processes  $N$ . Again, the processes are generated with a multivariate autoregressive process for which the entries of the coefficient matrix  $C$  are compatible with the connections of **Figure 8D**:

$$\begin{pmatrix} X_{1i+1} \\ \vdots \\ X_{Ni+1} \\ Z_{i+1} \\ Y_{1i+1} \\ \vdots \\ Y_{Ni+1} \\ W_{i+1} \end{pmatrix} = \begin{pmatrix} c_{xx} & \cdots & 0 & c_{xz} & 0 & \vdots & 0 & 0 \\ \vdots & \ddots & \vdots & \vdots & \vdots & \ddots & \vdots & \vdots \\ 0 & \cdots & c_{xx} & c_{xz} & 0 & \cdots & 0 & 0 \\ 0 & \cdots & 0 & c_{zz} & 0 & \cdots & 0 & 0 \\ c_{yx} & \cdots & 0 & 0 & c_{yy} & \cdots & 0 & c_{yw} \\ \vdots & \ddots & \vdots & \vdots & \vdots & \ddots & \vdots & \vdots \\ 0 & \cdots & c_{yx} & 0 & 0 & \cdots & c_{yy} & c_{yw} \\ 0 & \cdots & 0 & 0 & 0 & \cdots & 0 & c_{ww} \end{pmatrix} \begin{pmatrix} X_{1i} \\ \vdots \\ X_{Ni} \\ Z_i \\ Y_{1i} \\ \vdots \\ Y_{Ni} \\ W_i \end{pmatrix} + \begin{pmatrix} \varepsilon_{x1i} \\ \vdots \\ \varepsilon_{xNi} \\ \varepsilon_{zi} \\ \varepsilon_{y1i} \\ \vdots \\ \varepsilon_{yNi} \\ \varepsilon_{wi} \end{pmatrix} \quad (16)$$

Furthermore, the innovations covariance matrix is again an identity matrix. In **Figure 8E** we fix all the non-zero coefficients to 0.8 except  $c_{xz}$  and  $c_{yw}$ , which determine the degree of integration in area  $X$  due to the common driver  $Z$ , and of area  $Y$  due to common driver  $W$ , respectively. We then display  $T_{<X>-><Y>}$  as a function of  $c_{yw}$  and  $N$  fixing  $c_{xz} = 0.5$ , in the middle of the interval  $[0, 1]$  examined for  $c_{yw}$ . We see that  $T_{<X>-><Y>}$  either increases or decreases with  $N$  depending on which coupling

is stronger,  $c_{xz}$  or  $c_{yw}$ . This means that,  $T_{<X>-><Y>}$ , which is commonly interpreted as a measure of the strength of the connectivity between the areas, is highly sensitive to properties internal to each of the region when evaluated at a macroscopic scale at which spatial aggregation is present. Changes in the level of intra-areal integration could be interpreted as changes in the inter-areal interactions, but in fact  $T_{X_k \rightarrow Y_k}$  is constant for all the configurations shown in **Figure 8E**.

In **Figure 8F** we examine how vary, depending on the number of processes  $N$ , the changes of  $T_{<X>-><Y>}$  as a function of the actual coupling coefficient between the areas at the lower scale ( $c_{yx}$ ). We again fix all the non-zero coefficients to 0.8 except  $c_{xz} = 1.4$ ,  $c_{xx} = 0.2$ , and  $c_{yx} \in [0.1, 1.4]$ . Since  $c_{xz} > c_{yw}$  the Granger causality increases with  $N$ . We examine if this increase is different depending on  $c_{yx}$ . For that purpose, for each value of  $N$  we take as a reference the Granger causality calculated for the lowest coupling  $c_{yx} = 0.1$ . We then calculate  $T'_{<X>-><Y>}(c_{yx}, N) = T_{<X>-><Y>}(c_{yx}, N) / T_{<X>-><Y>}(0.1, N)$ , that is, the proportion of the Granger causality for each  $c_{yx}$  with respect to the one for  $c_{yx} = 0.1$ . We then consider the relative changes of  $T'_{<X>-><Y>}(c_{yx}, N)$  depending on  $N$ :

$$\Delta T'(c_{yx}, N) = \frac{T'_{<X>-><Y>}(c_{yx}, N) - T'_{<X>-><Y>}(c_{yx}, 1)}{T'_{<X>-><Y>}(c_{yx}, 1)} \quad (17)$$

We see in **Figure 8F** that the changes of Granger causality with  $c_{yx}$  are different for different  $N$ . This means that if we want

to compare different connections with different strength (determined by  $c_{yx}$ ), the results will be affected by the degree of spatial aggregation. However, as illustrated in **Figure 8F** the influence of changes in the actual coupling strength  $c_{yx}$  is low compared to the influence of the intra-areal integration, as shown in **Figure 8E**. These results were robust for other configurations of the setup represented in **Figure 8D**.

Altogether, we have shown that spatial aggregation can produce inconsistent positives of a high relative magnitude, and renders the measures of connectivity particularly sensitive to intra-areal properties, because these properties determine the resulting signals after spatial aggregation.

## DISCUSSION

We started by reviewing previous work about causal inference, comprising Granger causality (Granger, 1980) and causal models (Pearl, 2009). In particular, we described how causal models are associated with graphical causal structures, we indicated that Dynamic Causal Models (DCM) (Friston et al., 2003) are subsumed in the causal models described by Pearl, and that Pearl's approach does not exclude feedback connections because feedback interactions can be represented in acyclic graphs once the temporal dynamics are explicitly considered. Furthermore, we reviewed the criterion of d-separation to graphically read conditional independencies, and the algorithms proposed by Pearl and collaborators (Pearl, 2009) for causal inference without (IC algorithm) and with (IC\* algorithm) the existence of latent variables being considered. These algorithms have as output a graphical pattern that represents the class of all observationally equivalent causal structures compatible with the conditional independencies present in the data.

We then investigated the application of these algorithms to infer causal interactions between dynamic processes. We showed that Granger causality is subsumed by the IC algorithm. From our analysis it is also clear that other recent proposals to decompose Granger causality in different contributions or to identify the delay of the interactions (Runge et al., 2012; Wibral et al., 2013) are also subsumed by the IC algorithm. Moreover, we illustrated that the IC\* algorithm provides an output representation not suited for the analysis of dynamic processes, since it assumes the lack of structure of the latent variables. Accordingly, we proposed an alternative new algorithm based on the same principles of the IC\* algorithm but specifically designed to study dynamic processes. We did not conceive the new algorithm intending to outperform the IC\* algorithm, whose performance is theoretically optimal given the bounds imposed by the existence of observationally equivalent classes. Rather the new algorithm intends to provide a more appropriate and concise representation of the causal structures for dynamic processes. Furthermore, the algorithm integrates Pearl's algorithmic approach with the use of Granger causality. To our knowledge, this new algorithm is the first to use Granger causality explicitly considering the existence of latent processes. This improvement can be very helpful to assess how informative are the observed Granger causality relations to identify the actual causal structure of the dynamics.

Furthermore, we showed that an adequate graphical model of the setup in which some data are recorded is enough to predict,

without any numerical calculation, the existent Granger causality relationships using d-separation. We used this graphical analysis to explain, in a unified way, scenarios in which inconsistent positives of Granger causality have been reported. These comprise non-stationary correlated trends (Lütkepohl, 2005), related ongoing state variability (Wang et al., 2008), discretization (Kaiser and Schreiber, 2002), measurement noise (Nalatore et al., 2007), hemodynamic responses (Deshpande et al., 2010), time aggregation (Granger, 1963; Valdes-Sosa et al., 2011), and spatial aggregation. Regarding the effect of hemodynamic responses, our results may seem contradictory to the recent study of Seth et al. (2013) which shows that Granger causality is invariant when the hemodynamic response is an invertible filter. We note that the graphical analysis with d-separation is suited for stochastic variables, such as the ones in the causal models described in section "Models of Causality." The invariance of Granger causality is lost if noise variability is incorporated to the hemodynamic response.

We specifically focused on the effect of spatial aggregation of the underlying neural sources between which the causal interactions occur. The effects of spatial aggregation concern virtually all measures of causation calculated from neuroimaging data, and to those obtained with intracranial massed signals such as LFP. Yet, to our knowledge, this problem still remains to be fully understood. We showed that spatial aggregation can induce inconsistent positive Granger causality values of a magnitude comparable to the consistent ones. More generally, it renders Granger causality particularly sensitive to the degree of integration of the processes spatially aggregated. This means that in the presence of spatial aggregation Granger causality, independently of being used for causal inference or as a measure of functional connectivity (Valdes-Sosa et al., 2011; Friston et al., 2013), may reflect more the intra-areal properties of the system than inter-areal interactions.

In this work we followed the framework of Pearl based on causal models and associated graphical causal structures, in which a non-parametric approach to causal inference is proposed that is based on evaluating conditional independencies. In neuroscience applications, and in particular in fMRI analysis, there has been a recent controversy comparing Granger causality and DCM (Valdes-Sosa et al., 2011; Friston et al., 2013). We pointed out that both approaches are theoretically subsumed by Pearl's framework. In fact, much more relevant than this comparison is the distinction between non-parametric causal inference and model-based causal inference. Granger causality can be calculated in a model-based way, with autoregressive or more refined models (Lütkepohl, 2005), or it can be estimated in a non-parametric way using transfer entropy (e.g., Besserve et al., 2010). The motivation of using a generative model of the observed signals from underlying processes, which is at the core of DCM, is the same of proposing Kalman filters to improve the estimation of Granger causality (Winterhalder et al., 2005; Nalatore et al., 2007).

All the considerations regarding the limitations of causal inference due to observational equivalence and latent variables also hold for model-based approaches like DCM. In DCM the identification of the model causal structure is partially done a priori, by the selection of the priors of the parameters in the model, and partially carried out together with the parameters estimation. Therefore, the model selected (and thus the corresponding causal

structure) is not chosen only based on capturing the conditional independencies observed in the data, but also on optimizing some criterion of fitting to the actual data. Given the sophistication of the procedure of model inference, it is not straightforward to evaluate how the selected DCM model reflects the observed conditional independencies (and this may vary across different types of DCM models). Furthermore, the framework of network discovery within DCM (Friston et al., 2011) is very powerful evaluating the posterior probability—evidence—for different models, but still does not incorporate an evaluation of the influence of latent variables, like they do the algorithms of causal inference.

Modeling goes beyond causal inference. A good model gives us information not only about the causal structure, but also about the actual mechanisms that generate the dynamics. But a model can be good in terms of statistical prediction without being an appropriate causal model. That is, the effect of latent processes can be captured indirectly so that the parameters reflect not only the interactions between the observed processes but also the hidden ones. Therefore, even if by definition inside-model causality is well-defined in any DCM model, obtaining a good causal model is much harder than a good statistical model, and cannot be evaluated without interventions on the system. This means that, in the same sense that the Granger causality measures are measures of functional connectivity which, in some cases, can be used to infer causal relations, DCM models are functional connectivity models which, to the extent to which they increasingly reproduce the biophysical mechanisms generating the data, converge to causal models.

The issue of spatial aggregation we addressed here is particularly relevant for causal models, and not only to infer the causal structure. This is because it regards the nature of each node in the graph and requires understanding how causal mechanisms that certainly operate at a finer scale can be captured and are meaningful for macroscopic variables. That is, to which degree can we talk about a *causal* model between variables representing the activity of large brain areas? This is a crucial question for the mechanistical—and not only statistical—interpretation of DCM models, which, despite their increasing level of biological complexity, necessarily stay at a quite macroscopic level of description.

## ACKNOWLEDGMENTS

We acknowledge the financial support of the SI-CODE project of the Future and Emerging Technologies (FET) program within the Seventh Framework Programme for Research of the European Commission, under FET–Open Grant number: FP7-284553, and of the European Community's Seventh Framework Programme FP7/2007–2013 under Grant agreement number PITN-GA-2011–290011. The research was also funded by the Autonomous Province of Trento, Call “Grandi Progetti 2012,” project “Characterizing and improving brain mechanisms of attention—ATTEND.” We are grateful to Anders Ledberg for his valuable comments, and Ariadna Soy for carefully reading a draft of this manuscript.

## SUPPLEMENTARY MATERIAL

The Supplementary Material for this article can be found online at: <http://www.frontiersin.org/journal/10.3389/fninf.2014.00064/abstract>

## REFERENCES

- Bernasconi, C., and König, P. (1999). On the directionality of cortical interactions studied by structural analysis of electrophysiological recordings. *Biol. Cybern.* 81, 199–210. doi: 10.1007/s004220050556
- Besserve, M., Scholkopf, B., Logothetis, N. K., and Panzeri, S. (2010). Causal relationships between frequency bands of extracellular signals in visual cortex revealed by an information theoretic analysis. *J. Comput. Neurosci.* 29, 547–566. doi: 10.1007/s10827-010-0236-5
- Bressler, S. L., and Seth, A. K. (2011). Wiener-Granger causality: a well-established methodology. *Neuroimage* 58, 323–329. doi: 10.1016/j.neuroimage.2010.02.059
- Bullmore, E., and Sporns, O. (2009). Complex brain networks: graph theoretical analysis of structural and functional systems. *Nat. Rev. Neurosci.* 10, 187–198. doi: 10.1038/nrn2575
- Chicharro, D. (2011). On the spectral formulation of Granger causality. *Biol. Cybern.* 105, 331–347. doi: 10.1007/s00422-011-0469-z
- Chicharro, D. (2014a). “Parametric and non-parametric criteria for causal inference from time-series,” in *Directed Information Measures in Neuroscience*, Understanding Complex Systems, eds M. Wibral, R. Vicente, and J. T. Lizier (Berlin; Heidelberg: Springer-Verlag), 195–219. doi: 10.1007/978-3-642-54474-3\_8
- Chicharro, D. (2014b). A causal perspective on the analysis of signal and noise correlations and their role in population coding. *Neural Comput.* 26, 999–1054. doi: 10.1162/NECO\_a\_00588
- Chicharro, D., and Ledberg, A. (2012a). When two become one: the limits of causality analysis of brain dynamics. *PLoS ONE* 7:e32466. doi: 10.1371/journal.pone.0032466
- Chicharro, D., and Ledberg, A. (2012b). Framework to study dynamic dependencies in networks of interacting processes. *Phys. Rev. E Stat. Nonlin. Soft Matter Phys.* 86:041901. doi: 10.1103/PhysRevE.86.041901
- Cover, T. M., and Thomas, J. A. (2006). *Elements of Information Theory*. New York, NY: Wiley.
- Dahlhaus, R., and Eichler, M. (2003). “Causality and graphical models for time series,” in *Highly Structured Stochastic Systems*, eds P. Green, N. Hjort, and S. Richardson (Oxford: University Press), 115–137.
- Deshpande, G., Sathian, K., and Hu, X. P. (2010). Effect of hemodynamic variability on Granger causality analysis of fMRI. *Neuroimage* 52, 884–896. doi: 10.1016/j.neuroimage.2009.11.060
- Eichler, M. (2005). A graphical approach for evaluating effective connectivity in neural systems. *Philos. Trans. R. Soc. B Biol. Sci.* 360, 953–967. doi: 10.1098/rstb.2005.1641
- Eichler, M. (2007). Granger causality and path diagrams for multivariate time series. *J. Econom.* 137, 334–353. doi: 10.1016/j.jeconom.2005.06.032
- Eichler, M. (2009). “Causal inference from multivariate time series: what can be learned from Granger causality,” in *Logic, Methodology and Philosophy of Science. Proceedings of the 13th International Congress*, eds C. Glymour, W. Wang, and D. Westerstaal (London: College Publications).
- Einevoll, G. T., Kayser, C., Logothetis, N., and Panzeri, S. (2013). Modeling and analysis of local field potentials for studying the function of cortical circuits. *Nat. Rev. Neurosci.* 14, 770–785. doi: 10.1038/nrn3599
- Faes, L., Nollo, G., and Porta, A. (2011). Information-based detection of nonlinear Granger causality in multivariate processes via a nonuniform embedding technique. *Phys. Rev. E Stat. Nonlin. Soft Matter Phys.* 83:051112. doi: 10.1103/PhysRevE.83.051112
- Friston, K. J. (2011). Functional and effective connectivity: a review. *Brain Connect.* 1, 13–36. doi: 10.1089/brain.2011.0008
- Friston, K. J., Harrison, L., and Penny, W. (2003). Dynamic causal modelling. *Neuroimage* 19, 1273–1302. doi: 10.1016/S1053-8119(03)00202-7
- Friston, K. J., Li, B., Daunizeau, J., and Stephan, K. E. (2011). Network discovery with DCM. *Neuroimage* 56, 1202–1221. doi: 10.1016/j.neuroimage.2010.12.039
- Friston, K. J., Moran, R., and Seth, A. K. (2013). Analysing connectivity with Granger causality and dynamic causal modelling. *Curr. Opin. Neurobiol.* 23, 172–178. doi: 10.1016/j.conb.2012.11.010
- Granger, C. W. J. (1963). Economic processes involving feedback. *Inf. Control* 6, 28–48. doi: 10.1016/S0019-9958(63)90092-5
- Granger, C. W. J. (1980). Testing for causality - a personal viewpoint. *J. Econ. Dyn. Control* 2, 329–352. doi: 10.1016/0165-1889(80)90069-X
- Hsiao, C. (1982). Autoregressive modeling and causal ordering of economic variables. *J. Econ. Dyn. Control* 4, 243–259. doi: 10.1016/0165-1889(82)90015-X



- Kaiser, A., and Schreiber, T. (2002). Information transfer in continuous processes. *Physica D* 166, 43–62. doi: 10.1016/S0167-2789(02)00049-0
- Kramers, G. (1998). *Directed Information for Channels with Feedback*. Ph.D thesis, Swiss Federal Institute of Technology, Zurich.
- Lauritzen, S. L. (1996). *Graphical Models*. Oxford: Clarendon Press.
- Lizier, J. T., and Prokopenko, M. (2010). Differentiating information transfer and causal effect. *Eur. Phys. J. B* 73, 605–615. doi: 10.1140/epjb/e2010-00034-5
- Logothetis, N. K. (2008). What we can do and what we cannot do with fMRI. *Nature* 453, 869–878. doi: 10.1038/nature06976
- Lütkepohl, H. (2005). *New Introduction to Multiple Time Series Analysis*. Berlin: Springer-Verlag.
- Mäki-Marttunen, V., Diez, I., Cortes, J. M., Chialvo, D. R., and Villarreal, M. (2013). Disruption of transfer entropy and inter-hemispheric brain functional connectivity in patients with disorder of consciousness. *Front. Neuroinform.* 7:24. doi: 10.3389/fninf.2013.00024
- Marinazzo, D., Pellicoro, M., and Stramaglia, S. (2012). Causal information approach to partial conditioning in multivariate data sets. *Comput. Math. Methods Med.* 2012:303601. doi: 10.1155/2012/303601
- Marko, H. (1973). The bidirectional communication theory—A generalization of information theory. *IEEE Trans. Commun.* 21, 1345–1351. doi: 10.1109/TCOM.1973.1091610
- Masquelier, T. (2013). Neural variability, or lack thereof. *Front. Comput. Neurosci.* 7:7. doi: 10.3389/fncom.2013.00007
- Nalatore, H., Ding, M. Z., and Rangarajan, G. (2007). Mitigating the effects of measurement noise on Granger causality. *Phys. Rev. E Stat. Nonlin. Soft Matter Phys.* 75:031123. doi: 10.1103/PHYSREVE.75.031123
- Nevado, A., Young, M. P., and Panzeri, S. (2004). Functional imaging and neural information coding. *Neuroimage* 21, 1083–1095. doi: 10.1016/j.neuroimage.2003.10.043
- Pearl, J. (1986). Fusion, propagation, and structuring in belief networks. *Artif. Intell.* 29, 241–288. doi: 10.1016/0004-3702(86)90072-X
- Pearl, J. (1988). *Probabilistic Reasoning in Intelligent Systems*. San Mateo, CA: Morgan Kaufman.
- Pearl, J. (2009). *Causality: Models, Reasoning, Inference*. New York, NY: Cambridge University Press.
- Ramb, R., Eichler, M., Ing, A., Thiel, M., Weiller, C., Grebogi, C., et al. (2013). The impact of latent confounders in directed network analysis in neuroscience. *Philos. Trans. R. Soc. A* 371:20110612. doi: 10.1098/rsta.2011.0612
- Rissanen, J., and Wax, M. (1987). Measures of Mutual and Causal Dependence between two Time-Series. *IEEE Trans. Inform. Theory* 33, 598–601. doi: 10.1109/TIT.1987.1057325
- Roebroeck, A., Formisano, E., and Goebel, R. (2005). Mapping directed influence over the brain using Granger causality and fMRI. *Neuroimage* 25, 230–242. doi: 10.1016/j.neuroimage.2004.11.017
- Runge, J., Heitzig, J., Petoukhov, V., and Kurths, J. (2012). Escaping the curse of dimensionality in estimating multivariate transfer entropy. *Phys. Rev. Lett.* 108:258701. doi: 10.1103/PhysRevLett.108.258701
- Scannell, J. W., and Young, M. P. (1999). Neuronal population activity and functional imaging. *Philos. Trans. R. Soc. B Biol. Sci.* 266, 875–881. doi: 10.1098/rspb.1999.0718
- Schreiber, T. (2000). Measuring information transfer. *Phys. Rev. Lett.* 85, 461–464. doi: 10.1103/PhysRevLett.85.461
- Seth, A. K., Chorley, P., and Barnett, L. C. (2013). Granger causality analysis of fMRI BOLD signals is invariant to hemodynamic convolution but not downsampling. *Neuroimage* 65, 540–555. doi: 10.1016/j.neuroimage.2012.09.049
- Smirnov, D. A. (2013). Spurious causalities with transfer entropy. *Phys. Rev. E Stat. Nonlin. Soft Matter Phys.* 87:042917. doi: 10.1103/PhysRevE.87.042917
- Spirtes, P., Glymour, C. N., and Scheines, R. (2000). *Causation, Prediction, and Search*. Cambridge, MA: MIT Press.
- Valdes-Sosa, P. A., Roebroeck, A., Daunizeau, J., and Friston, K. (2011). Effective connectivity: influence, causality and biophysical modeling. *Neuroimage* 58, 339–361. doi: 10.1016/j.neuroimage.2011.03.058
- Verma, T. (1993). *Graphical Aspects of Causal Models, Technical Report R-191*. Los Angeles, CA: Computer Science Department, UCLA.
- Verma, T., and Pearl, J. (1990). “Equivalence and synthesis of causal models,” in *Proceedings of the Sixth Conference on Uncertainty in Artificial Intelligence* (Cambridge, MA), 220–227.
- Verma, T., and Pearl, J. (1992). “An algorithm for deciding if a set of observed independencies has a causal explanation,” in *Proceedings of the Eighth Conference on Uncertainty in Artificial Intelligence* (Stanford, CA), 323–330.
- Vicente, R., Wibral, M., Lindner, M., and Pipa, G. (2011). Transfer entropy—a model-free measure of effective connectivity for the neurosciences. *J. Comput. Neurosci.* 30, 45–67. doi: 10.1007/s10827-010-0262-3
- Wang, X., Chen, Y. H., and Ding, M. Z. (2008). Estimating Granger causality after stimulus onset: a cautionary note. *Neuroimage* 41, 767–776. doi: 10.1016/j.neuroimage.2008.03.025
- Wibral, M., Pampu, N., Priesemann, V., Siebenhühner, F., Seiwert, H., Lindner, M., et al. (2013). Measuring information-transfer delays. *PLoS ONE* 8:e55809. doi: 10.1371/journal.pone.0055809
- White, H., and Chalak, K. (2009). Settable systems: an extension of Pearl’s causal model with optimization, equilibrium, and learning. *J. Mach. Learn. Res.* 10, 1759–1799. doi: 10.1145/1577069.1755844
- White, H., and Lu, X. (2010). Granger causality and dynamic structural systems. *J. Financ. Econom.* 8, 193–243. doi: 10.1093/jfinc/nbq006
- Winterhalder, M., Schelter, B., Hesse, W., Schwab, K., Leistriz, L., Klan, D., et al. (2005). Comparison of linear signal processing techniques to infer directed interactions in multivariate neural systems. *Signal Process.* 85, 2137–2160. doi: 10.1016/j.sigpro.2005.07.011
- Wu, G., Liao, W., Chen, H., Stramaglia, S., and Marinazzo, D. (2013). Recovering directed networks in neuroimaging datasets using partially conditioned Granger causality. *Brain Connect.* 3, 294–301. doi: 10.1089/brain.2013.0142

**Conflict of Interest Statement:** The authors declare that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.

Received: 06 January 2014; accepted: 12 June 2014; published online: 02 July 2014.

Citation: Chicharro D and Panzeri S (2014) Algorithms of causal inference for the analysis of effective connectivity among brain regions. *Front. Neuroinform.* 8:64. doi: 10.3389/fninf.2014.00064

This article was submitted to the journal *Frontiers in Neuroinformatics*.

Copyright © 2014 Chicharro and Panzeri. This is an open-access article distributed under the terms of the Creative Commons Attribution License (CC BY). The use, distribution or reproduction in other forums is permitted, provided the original author(s) or licensor are credited and that the original publication in this journal is cited, in accordance with accepted academic practice. No use, distribution or reproduction is permitted which does not comply with these terms.



# Sample entropy reveals high discriminative power between young and elderly adults in short fMRI data sets

Moses O. Sokunbi<sup>1,2\*</sup>

<sup>1</sup> MRC Centre for Neuropsychiatric Genetics and Genomics, Institute of Psychological Medicine and Clinical Neurosciences, Cardiff School of Medicine, Cardiff University, Cardiff, UK

<sup>2</sup> Imaging Science, Cardiff University Brain Research Imaging Centre, Cardiff University, Cardiff, UK

## Edited by:

Daniele Marinazzo, University of Gent, Belgium

## Reviewed by:

Daniele Marinazzo, University of Gent, Belgium

Jennifer Yentes, University of Nebraska at Omaha, USA

## \*Correspondence:

Moses O. Sokunbi, Imaging Science, Cardiff University Brain Research Imaging Centre, Cardiff University, 70 Park Place, Cardiff CF10 3AT, UK  
e-mail: sokunbimo@cardiff.ac.uk

Some studies have placed Sample entropy on the same data length constraint of  $10^m$ – $20^m$  ( $m$ : pattern length) as approximate entropy, even though Sample entropy is largely independent of data length and displays relative consistency over a broader range of possible parameters ( $r$ , tolerance value;  $m$ , pattern length;  $N$ , data length) under circumstances where approximate entropy does not. This is particularly erroneous for some fMRI experiments where the working data length is less than 100 volumes (when  $m = 2$ ). We therefore investigated whether Sample entropy is able to effectively discriminate fMRI data with data length,  $N$  less than  $10^m$  (where  $m = 2$ ) and  $r = 0.30$ , from a small group of 10 younger and 10 elderly adults, and the whole cohort of 43 younger and 43 elderly adults, that are significantly ( $p < 0.001$ ) different in age. Ageing has been defined as a loss of entropy; where signal complexity decreases with age. For the small group analysis, the results of the whole brain analyses show that Sample entropy portrayed a good discriminatory ability for data lengths,  $85 \leq N \leq 128$ , with an accuracy of 85% at  $N = 85$  and 80% at  $N = 128$ , at  $q < 0.05$ . The regional analyses show that Sample entropy discriminated more brain regions at  $N = 128$  than  $N = 85$  and some regions common to both data lengths. As data length,  $N$  increased from 85 to 128, the noise level decreased. This was reflected in the accuracy of the whole brain analyses and the number of brain regions discriminated in the regional analyses. The whole brain analyses suggest that Sample entropy is relatively independent of data length, while the regional analyses show that fMRI data with length of 85 volumes is consistent with our hypothesis of a loss of entropy with ageing. In the whole cohort analysis, Sample entropy discriminated regionally between the younger and elderly adults only at  $N = 128$ . The whole cohort analysis at  $N = 85$  was indicative of the ageing process but this indication was not significant ( $p > 0.05$ ).

**Keywords:** ageing, blood oxygen level dependent (BOLD), data length, functional magnetic resonance imaging (fMRI), noise level, sample entropy

## INTRODUCTION

Recently, the application of entropy measures to investigate signal complexity and irregularity in human data has become quite popular (Yentes et al., 2013). Entropy values reflect the number of times the patterns in a signal are repeated and thus measure the randomness and predictability of stochastic process and in more general terms, increase with greater randomness (Sokunbi et al., 2013). The computation of entropy in biological data processing became a possible solution to the shortcomings posed by some metrics of nonlinear time series analysis techniques such as correlation dimension (Pritchard et al., 1994) and Lyapunov exponent (Wolf et al., 1985), which require a large data set (Eckmann and Ruelle, 1992) and assume that the time series is stationary (Grassberger and Procaccia, 1983), a feature normally not true for biological data. Approximate entropy (ApEn) (Pincus, 1991) and sample entropy (SampEn) (Richman and Moorman, 2000) are a few of the different types of entropy measures that have

evolved from the concept of entropy. Regularity and complexity statistics such as ApEn and SampEn are measures without the shortcomings that correlation dimension and Lyapunov exponent possess (Richman and Moorman, 2000). ApEn and SampEn can effectively discriminate both stochastic processes and noisy deterministic data sets in instances where measures such as spectral and autocorrelation analyses exhibit minimal distinctions (Pincus, 2001). They are also nearly unaffected by low level noise, are robust to occasional, very large or small artifacts and give meaningful information with a reasonable number of data points, and are finite for both stochastic and deterministic processes (Zhang and Roy, 2001).

The ApEn algorithm counts each sequence as matching itself to avoid the occurrence of  $\ln(0)$  in the calculations, which led to the discussion of the bias of ApEn (Pincus, 1995). This bias causes ApEn to be heavily dependent on data length and uniformly lower than expected for short data lengths. Also, ApEn lacks relative

consistency. To reduce this bias, SampEn was introduced as an improvement of ApEn where self-matches are excluded, i.e., vectors are not compared to themselves (Richman and Moorman, 2000). SampEn is the negative natural logarithm of the conditional probability that two sequences remain similar at the next point, where self-matches are not included in calculating the probability (Richman and Moorman, 2000). Hence, a lower value of SampEn also indicates more self-similarity in the time series. The algorithm of SampEn is simpler than the ApEn algorithm, requiring less time for computation. SampEn is largely independent of data length and displays relative consistency over a broader range of possible parameters ( $r$ , tolerance value;  $m$ , pattern length;  $N$ , data length) under circumstances where ApEn does not (Richman and Moorman, 2000).

SampEn has been used to characterize human data from a number of imaging modalities. To mention a few, it has been used to analyze the electroencephalogram (EEG) background activity in Alzheimer's disease patients (Abasolo et al., 2006). It has further been used to analyse the spontaneous magnetoencephalography (MEG) signals in patients with ADHD (Gomez et al., 2011) and to probe the complexity of resting state fMRI activity in adult patients with ADHD (Sokunbi et al., 2013). More recently, it has been used to examine the whole brain entropy patterns of a large cohort of normal subjects using fMRI (Wang et al., 2014). In all three brain imaging modalities, fMRI had the shortest data length. Since there are no laid down guidelines for choosing parameters to compute SampEn for all modalities of biomedical signals, some investigators have made suggestions for selecting parameters to use. Abasolo et al. (2006) suggested that to estimate SampEn of EEG accurately, a data length of  $10^m$ – $20^m$  is required. Here, they used parameters  $m = 1$ ,  $r = 0.25$ , and  $N = 1280$  data length. In a recent study, Yentes et al. (2013) examined the robustness of ApEn and SampEn algorithms by exploring the effect of changing parameter values on short data sets using both theoretical and experimental data (musculoskeletal data with a data length of 200). In conclusion, they suggested to use a data length larger than 200, an  $m$  of 2, and to examine several  $r$ -values before selecting parameters. However, they also noted that SampEn was less sensitive to changes in data length and demonstrated fewer problems with relative consistency. Also, in another recent study of fMRI multiscale sample entropy analysis, SampEn was placed at the same data length threshold of  $10^m$ – $20^m$  with ApEn (Yang et al., 2013), even though it is largely independent of data length and displays relative consistency under circumstances where ApEn does not (Richman and Moorman, 2000).

The developers of SampEn (Richman and Moorman, 2000) tested the consistency of SampEn for very short data sets using theoretical data (independent, identically distributed (i.i.d) Gaussian numbers) and found that SampEn statistics deviated from predictions for very short data sets. They calculated the biased results of SampEn ( $2, 0.2, N$ ) for the range of  $4 \leq N \leq 102$ . For Gaussian random numbers with  $m = 2$  and  $r = 0.2$ , they found that the deviation was less than 3% for data lengths greater than 100 points but as high as 35% for data length of 15 points. They found that the bias of SampEn for very small data sets is largely due to “non-independence of templates” (Richman and

Moorman, 2000) and that this bias appears to be present only for very small data lengths. They did not suggest or recommend a data length constraint for estimating SampEn.

fMRI is a potent research tool and has found more applications in research than clinical use. In contrast to EEG and MEG, fMRI possesses poor temporal resolution (in order of seconds) but excellent spatial specificity. As a result, most fMRI experiments are usually short, in the range of 100–200 data lengths. Prior data analysis, standard fMRI data processing requires that the first 3 or 4 volumes (data lengths) of fMRI data are discarded to enable signal conditioning. For fMRI data acquisitions of 100 data length, this results in a data length of 97 or 96. Our experience of characterizing fMRI data with SampEn shows that it is possible to obtain reliable results while using robust and optimal parameters such as  $m = 2$ ,  $r = 0.46$  (a high  $r$ -value) and a data length less than 100 (97 data points) (Sokunbi et al., 2013). We further tested the ability of SampEn to effectively discriminate fMRI data with data length,  $N$  less than  $10^m$  (where  $m = 2$ ) using a resting state fMRI data set from a small group of 10 healthy right-handed younger and 10 right-handed elderly adults that are significantly ( $p < 0.001$ ) different in age, extracted from the International Consortium for Brain Mapping (ICBM) resting state dataset. We also investigated the discriminatory ability of SampEn on the whole ICBM resting state cohort of 43 younger and 43 elderly adults that are significantly ( $p < 0.001$ ) different in age. We used  $m = 2$  which is superior to  $m = 1$  since it allows more detailed reconstruction of the joint probabilistic dynamics of the time series (Pincus and Goldberger, 1994).

With normal ageing, there are declines in mental domains such as processing speed, reasoning, memory and executive functions, some of which is underpinned by a decline in a general cognitive factor (Deary et al., 2009). The bases for this decline are not fully understood. There has been progress in normal cognitive ageing from genetics, general health, biological processes, neurobiological changes, diet, lifestyle and many other areas of biomedical and psychosocial sciences. For example, the complexity of longitudinal physiological measurements such as EEG has been shown to vary with age and disease (Gaal et al., 2010). Complexity can be described as the difficulties associated with predicting a signal and this can be estimated by measuring the signal's entropy (Lu et al., 2008). Some studies have suggested that the characterization and analysis of the brain's output in terms of its complexity may reveal a better understanding of an individual's health and robustness (Goldberger et al., 2002), adaptive capacity in terms of brain ageing (Sokunbi et al., 2011) and diseases (Sokunbi et al., 2013, 2014), and *in-vivo* effect of drugs (Ferenets et al., 2007). Healthy systems portray chaotic and complex behaviors whereas pathological states exhibit predictable behaviors (Pool, 1989). Estimating the complexity of the blood oxygen level dependent (BOLD) fMRI signals can help to probe different aspects of complex signals brought about by ageing and disease, revealing subtle patterns which may provide fundamental insights that can lead to clinical and biomedical applications.

Investigators have argued that the pathway of change in the behavior and physiology of an organism with age and disease can either result in a decrease or an increase in the complexity of the system's output (Vaillancourt and Newell, 2002; Sokunbi

et al., 2014). Vaillancourt and Newell (2002) postulate that the directional change in output complexity of a physiological or behavioral system with ageing or disease depends on the system having an underlying fixed point or an oscillatory attractor determining output. An attractor is the state to which a system returns to after perturbation (Vaillancourt and Newell, 2002). In the fixed-point attractor system, complexity decreases with age and disease (Sokunbi et al., 2013) while in the oscillatory attractor system complexity increases with age and disease (Sokunbi et al., 2014). Ageing has been defined as a loss of entropy (Lipsitz, 2004) and specific brain regions have been implicated in the ageing process (Craik and Salthouse, 2000). Also, functional entropy has been shown to increase with age (Yao et al., 2013). In the present analysis, we expect SampEn to decrease with age according to Lipsitz's (2004) entropy definition of ageing and Vaillancourt and Newell's (2002) fixed-point attractor postulate. Most importantly, we expect SampEn results at  $N$  less than 100 to be indicative of this ageing process since it is largely independent of data length and displays relative consistency (Richman and Moorman, 2000).

## MATERIALS AND METHODS

### SUBJECTS

A small group of 10 healthy right-handed younger adults [5 male, mean age ( $22.40 \pm 3.44$ )] and 10 healthy right-handed elderly adults [5 male, mean age ( $69.60 \pm 9.25$ )] with significant ( $p < 0.001$ ) age difference were extracted from the ICBM resting state dataset made publicly available in the 1000 Functional Connectomes project. The subjects used for the small group analysis are listed in the supplementary data, Table S1. The whole ICBM resting state cohort of 43 younger adults [21 male, mean age ( $29.05 \pm 8.66$ )] and 43 elderly adults [20 male, mean age ( $59.33 \pm 10.27$ )] with significant ( $p < 0.001$ ) age difference was also investigated. The study was approved by the local research ethics committee and subjects had no history of neurological or psychiatric disorders. Written informed consent was obtained from the subjects. Information regarding this dataset is available at [https://www.nitrc.org/projects/fcon\\_1000/](https://www.nitrc.org/projects/fcon_1000/).

### BRAIN IMAGING

Functional MR images were acquired with a  $T_2^*$  weighted gradient echo echo-planar imaging sequence (EPI) using a standard head coil on a 3T scanner. A total of 23 axial slices were obtained for each of 133 volumes using a TR of 2 s and matrix  $64 \times 64$ . A total of 128 volumes of fMRI data remained after discarding the first five volumes to allow for signal conditioning. Subjects were asked to lie in the scanner with their eyes closed.

### IMAGE PRE-PROCESSING

fMRI data pre-processing were performed using SPM8 software (The Wellcome Department of Imaging Neuroscience, UCL, London, UK). The images were realigned to correct for head movement distortion. Temporal high pass filtering was performed (128 s) to reduce low frequency noise and spatial smoothing was performed to reduce white noise using the full-width at half maximum (FWHM) of the Gaussian smoothing kernel [8 8 8]. Each voxel time series was standardized to a mean of zero and standard

deviation of unity. This allows a signal value of  $r$  (tolerance) to be used for all voxels independent of amplitude and variance.

### COMPUTATION OF SampEn

The SampEn of a time series of length  $N$  ( $x_1, x_2, \dots, x_N$ ) can be computed from the given sets of equations (Sokunbi et al., 2013):

$$\text{SampEn}(m, r, N) = -\ln \left[ \frac{U^{m+1}(r)}{U^m(r)} \right]$$

$$U^m(r) = [N - m\tau]^{-1} \sum_{i=1}^{N-m\tau} C_i^m(r) \quad (1)$$

Where

$$C_i^m(r) = \frac{B_i}{N - (m + 1)\tau}$$

$$B_i = \text{number of } j \text{ where } d[X_i, X_j] \leq r \quad (2)$$

$$X_i = (x_i, x_{i+\tau}, \dots, x_{i+(m-1)\tau}) \quad (3)$$

$$X_j = (x_j, x_{j+\tau}, \dots, x_{j+(m-1)\tau}) \quad (4)$$

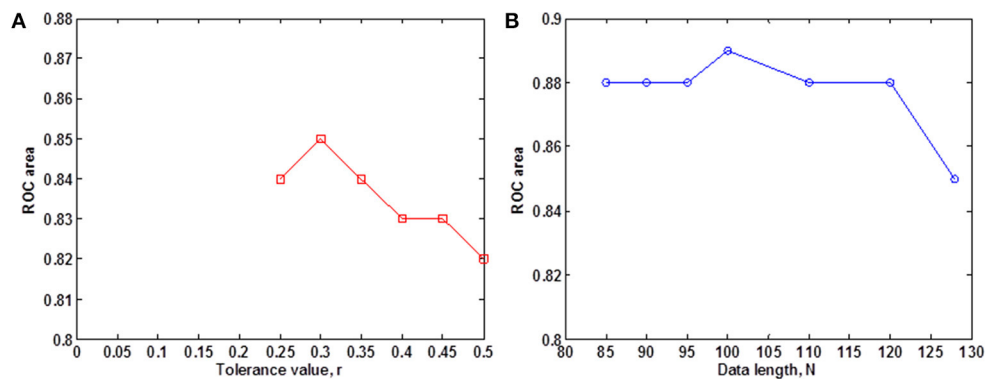
$$1 \leq j \leq N - m\tau, j \neq i$$

$N$  specifies the data length,  $m$  is the pattern length,  $r$  is the tolerance value, and  $\tau$  is the time delay as shown in Equation (1). In Equation (2), the two patterns  $i$  and  $j$  of  $m$  measurements of the time series are similar if the difference,  $d[X_i, X_j]$ , between any pair of corresponding measurements of  $X_i$  and  $X_j$  is less than, or equal to,  $r$ . In Equations (3 and 4),  $X_i$  and  $X_j$  are pattern vectors (length  $m$ ) whose components are time-delayed versions of the elements in the original time series with time delay,  $\tau$ .

We evaluated the ability of SampEn to discriminate the younger from the elderly adults, using the value of the receiver operating characteristic (ROC) area (Zweig and Campbell, 1993). ROC areas are used as a guide to classify the precision of a diagnostic test. Areas with values between 0.90 and 1 indicate that the precision of the diagnostic test is excellent, when the values are between 0.80 and 0.89, it means the test is good. It is fair if the area values are between 0.70 and 0.79, poor when the area is between 0.60 and 0.69 and bad for values ranging from 0.50 to 0.59. Using the small group of 10 younger and 10 elderly adults, we determined the optimal  $r$ -value where this discrimination occurs by computing the ROC area for a range of  $r$ -values. The ROC area was computed from the mean whole brain SampEn values of each subject in the small group using a robust value of  $m = 2$  (Pincus and Goldberger, 1994), data length  $N = 128$  and by varying the  $r$ -value from 0.05 to 0.5 at intervals of 0.05. **Figure 1A** shows that this optimal  $r$ -value occurred at  $r = 0.30$ .

Whole brain SampEn was computed for each subject in the small group using  $m = 2$ , the optimal  $r$ -value of 0.30 (**Figure 1A**), multiplied by the SD of the fMRI time series,  $\tau = 1$  and fMRI data lengths of 128, 120, 110, 100, 90, 95, and 85. Only data lengths





**FIGURE 1 | Small group analysis.** (A) ROC area for determining the optimal  $r$ -value for fMRI data of 128 volumes, for  $0.05 \leq r \leq 0.5$  at intervals of 0.05. The optimal  $r$ -value was obtained at  $r = 0.30$ ;

(B) ROC area of SampEn ( $m = 2$ ,  $r = 0.30$ ,  $85 \leq N \leq 128$ ) for fMRI data lengths  $N$ . SampEn shows good discriminating ability and relative consistency for all the data lengths.

where all 20 subjects returned SampEn values were included in the study. Data lengths less than 85 could not be included in the study because some of the subjects did not return SampEn values as a result of a lack of templates to compare. Whole brain SampEn maps were generated on a voxel by voxel basis using the same approach as Sokunbi et al. (2011) on a MATLAB and C platform. A threshold of 0.1 times the maximum signal was used to exclude voxels being calculated outside the brain. The mean whole brain SampEn value for each subject was computed. Also, the ROC area for discriminating between both groups was computed from the mean whole brain SampEn value of each subject in both groups for all the data lengths. SampEn showed good discriminating ability for  $85 \leq N \leq 128$  as shown in Figure 1B.

Similarly, whole brain SampEn maps were generated for the cohort of 43 younger and 43 elderly adults using  $m = 2$ , the optimal  $r$ -value of 0.30 (Figure 1A), multiplied by the SD of the fMRI time series,  $\tau = 1$  and fMRI data lengths of 128 and 85. The ROC area for discriminating between the cohort of 43 younger and 43 elderly adults was computed from the mean whole brain SampEn value of each subject in both groups for data lengths  $N = 128$  and  $N = 85$ .

## STATISTICAL ANALYSIS

The ROC analyses were performed on the mean whole brain SampEn values using the International Business Machines Corporation (IBM) Statistical Package for Social Sciences (SPSS 20.0; New York, USA) software. Independent  $t$ -tests for the different data lengths,  $N$ , were performed between the mean whole brain SampEn values of both groups using SPSS software. Also, correlations using the Pearson correlation analyses between the mean whole brain SampEn and age for the whole population were performed in SPSS, for the different data lengths,  $N$ . False discovery rate (FDR) for multiple comparisons correction ( $q < 0.05$ ) in R-Statistics (<http://www.r-project.org/>) was used to correct the  $p$ -values of the independent  $t$ -tests and  $p$ -values of the Pearson's correlation analyses. The Pearson's correlation coefficients ( $r$ -values) were interpreted using Dancey and Reidy's categorisation (Dancey and Reidy, 2004). Here,  $r$ -value of  $\pm 1$  is interpreted as a perfect correlation,  $r$ -values

between  $\pm 0.7$  to  $\pm 0.9$  are interpreted as strong correlations,  $r$ -values in the range  $\pm 0.4$  to  $\pm 0.6$  are categorized as moderate correlations,  $r$ -values between  $\pm 0.1$  to  $\pm 0.3$  are weak correlations and an  $r$ -value of 0 is zero correlation, implying there is no correlation.

The SampEn map of each subject was normalized to a standard echo planar imaging (EPI) template, and a regional (spatial) analysis was performed using the two-sample  $t$ -test in SPM8, comparing the SampEn maps of the younger and elderly adults at a family-wise error (FWE) corrected cluster level significance of  $p < 0.05$  and threshold  $p = 0.005$ . This was only done for data lengths  $N = 85$  and  $N = 128$ . Correlations between the SampEn maps and age for the whole population were tested using multiple regression approach in SPM8.

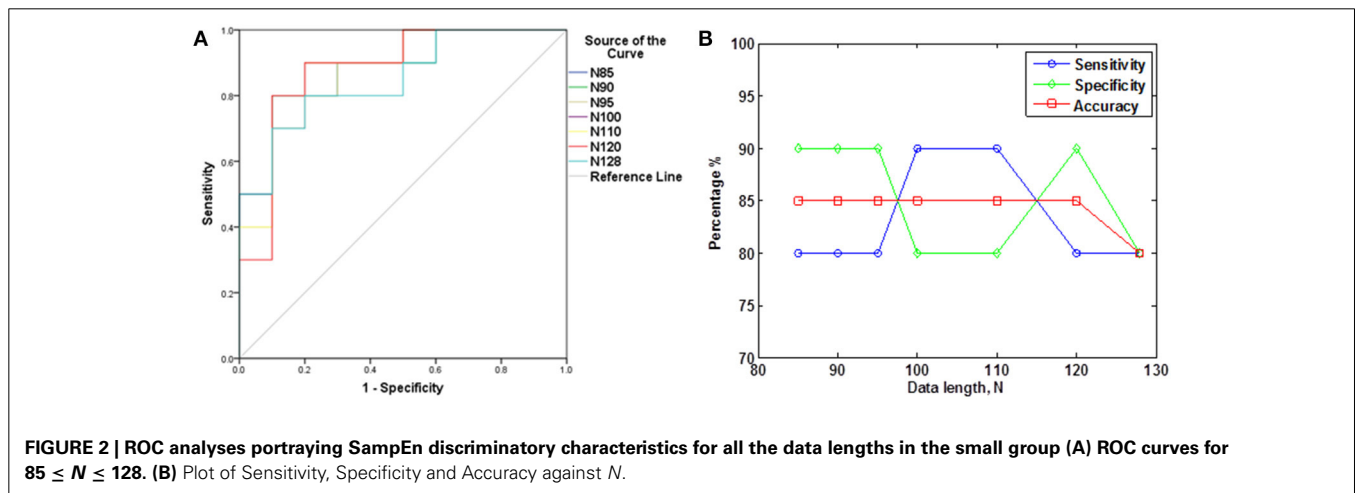
## RESULTS

### SMALL GROUP OF 10 YOUNGER AND 10 ELDERLY ADULTS

The subjects' characteristics and SampEn measures for the small group of 10 younger and 10 elderly adults are shown in Table 1. The ROC results of the mean whole brain SampEn for data lengths  $85 \leq N \leq 128$  were in the range 0.850–0.890. This implies that the ability of SampEn to effectively discriminate the younger from the elderly adults across all the data lengths is good and that this ability is not dependent on data length. The ROC curves and characteristics for  $85 \leq N \leq 128$  are shown in Figure 2A. The sensitivity and specificity obtained from the ROC analysis ranged between 80 and 90% for all the data lengths, while the accuracy was 85% for all data lengths except for  $N = 128$  where the accuracy dropped to 80% (see Figure 2B and Table 2). For data lengths  $85 \leq N \leq 128$ , the mean whole brain SampEn values of the younger adults were significantly ( $p < 0.05$ ) higher than the mean whole brain SampEn values of the elderly adults. After corrections for multiple comparisons using the FDR, the mean whole brain differences for all the data lengths remained significantly ( $q < 0.05$ ) higher. The mean whole brain differences between the younger and elderly adults for all the data lengths are shown in Figure 3. Moderate negative correlations ( $r$ -values between  $-0.581$  and  $-0.626$ ) were obtained at  $p < 0.01$  between the mean whole brain SampEn values and the age of the

**Table 1 | Subjects' characteristics and SampEn measures for the small group of 10 younger and 10 elderly adults.**

	Younger adults	Elderly adults	Significance ( <i>p</i> -values)	Significance FDR corrected ( <i>q</i> -values)
Age (years)	22.40 ± 3.44	69.60 ± 9.25	$p < 0.001$	
Sex (M/F)	5/5	5/5		
SampEn at $N = 85$	1.7413 ± 0.0298	1.6888 ± 0.0400	$p = 0.004$	$q = 0.007$
SampEn at $N = 90$	1.7354 ± 0.0280	1.6779 ± 0.04631	$p = 0.003$	$q = 0.007$
SampEn at $N = 95$	1.7309 ± 0.0260	1.6729 ± 0.0472	$p = 0.003$	$q = 0.007$
SampEn at $N = 100$	1.7258 ± 0.0268	1.6687 ± 0.0458	$p = 0.003$	$q = 0.007$
SampEn at $N = 110$	1.7164 ± 0.0278	1.6595 ± 0.0506	$p = 0.006$	$q = 0.007$
SampEn at $N = 120$	1.7082 ± 0.0288	1.6489 ± 0.0529	$p = 0.006$	$q = 0.007$
SampEn at $N = 128$	1.6980 ± 0.0359	1.6407 ± 0.0517	$p = 0.010$	$q = 0.010$

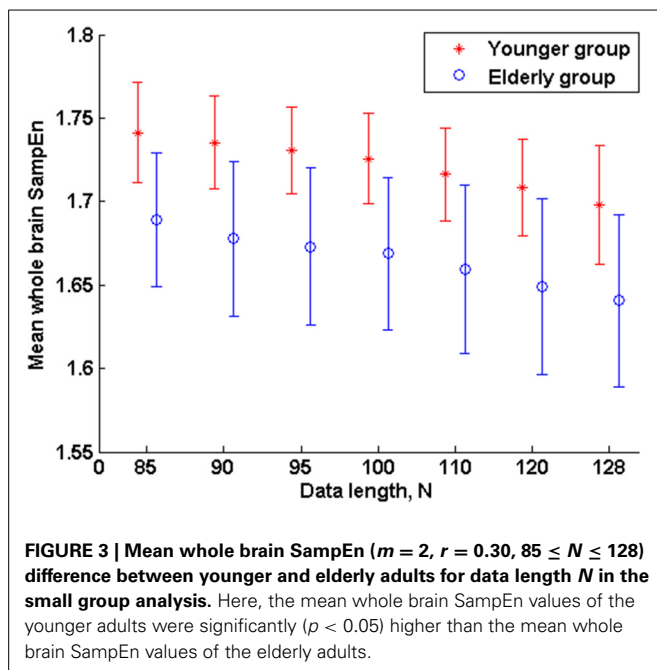
**FIGURE 2 | ROC analyses portraying SampEn discriminatory characteristics for all the data lengths in the small group (A) ROC curves for  $85 \leq N \leq 128$ . (B) Plot of Sensitivity, Specificity and Accuracy against  $N$ .****Table 2 | ROC characteristics for the small group of 10 younger and 10 elderly adults.**

Data length, $N$	Threshold	Sensitivity (%)	Specificity (%)	Accuracy (%)	Area under the ROC curve
85	1.7332	80	90	85	0.880
90	1.7244	80	90	85	0.880
95	1.7183	80	90	85	0.880
100	1.7026	90	80	85	0.890
110	1.6902	90	80	85	0.880
120	1.6888	80	90	85	0.880
128	1.6710	80	80	80	0.850

population, for all the data lengths ( $85 \leq N \leq 128$ ). Also, after corrections for multiple comparisons using FDR, the moderate negative correlations between the mean whole brain SampEn values and age remained significant ( $q < 0.05$ ). This implies that for all the data lengths SampEn decreased with age. **Table 3** shows the Pearson's correlation coefficients,  $r$ , the  $p$ -values and the  $q$ -values (FDR) for data lengths  $85 \leq N \leq 128$ . **Figures 4A–G** shows the regression curve estimation between SampEn and age for the population. A graph was plotted to further investigate how the Pearson's correlation coefficients,  $r$  (correlation of SampEn

and age) varied with the different data lengths  $85 \leq N \leq 128$ . The graph shown in **Figure 5** shows that the Pearson's correlation coefficients,  $r$  remained relatively constant with the different data lengths. This implies that the correlation between SampEn and age was relatively consistent with the changes in data length.

To investigate regional differences and similarities in data lengths, the whole brain SampEn maps for the minimum and maximum data lengths ( $85 \leq N \leq 128$ ) were tested regionally with a family-wise error (FWE) corrected cluster level significance of  $p < 0.05$  using the two-sample  $t$ -test in SPM8. The results consistent with that of the mean whole brain analysis show that the younger adults exhibited significantly ( $p < 0.05$ ) higher SampEn values than the elderly adults at a threshold of  $p = 0.005$  with corresponding discriminated brain regions. For data length  $N = 85$ , only the frontal lobe of the brain was discriminated while for  $N = 128$ , the frontal lobe and parietal lobe were discriminated. These discriminated brain regions are listed in **Table 4**. **Figure 6** shows the rendered images of the two-sample  $t$ -tests between the younger and elderly adults, for data lengths,  $N = 85$  and  $N = 128$ . Also, correlations between the whole brain SampEn maps and age, of the whole population, for data lengths,  $N = 85$  and  $N = 128$  were performed using multiple regression analysis in SPM8. Again, SampEn portrayed a significant ( $p < 0.05$ ) negative correlation with age, for both data lengths as shown by the rendered images in **Figure 7**. For  $N = 85$ , the frontal, limbic and



parietal lobes were discriminated while for  $N = 128$  the frontal lobe, limbic lobe, parietal lobe and sub-lobar brain regions were discriminated. See **Table 5** for a list of the discriminated brain regions.

#### COHORT OF 43 YOUNGER AND 43 ELDERLY ADULTS

The subjects' characteristics and SampEn measures for the whole ICBM resting state cohort of 43 younger and 43 elderly adults are shown in **Table 6**. The ROC results of the mean whole brain SampEn for data lengths  $N = 85$  and  $N = 128$  were 0.600 and 0.603 respectively. This implies that the ability of SampEn to effectively discriminate the younger from the elderly adults of both data lengths is poor. For data length  $N = 85$ , the sensitivity was 65.10%, the specificity was 53.50% and accuracy was 59.30% at a threshold of 1.7298. While for data length  $N = 128$ , the sensitivity was 58.10%, the specificity was 58.10% and accuracy was 58.10% at a threshold of 1.6986. For both data lengths, the mean whole brain SampEn values of the younger and elderly adults were not significantly ( $p > 0.05$ ) different but the younger adults had higher mean whole brain SampEn values than the elderly adults. Weak negative correlations,  $r$ -values of  $-0.078$  and  $-0.099$  were obtained at  $p > 0.05$  between the mean whole brain SampEn values and the age of the population, for data lengths  $N = 85$  and  $N = 128$  respectively.

For data length,  $N = 128$ , the result of the regional analysis show that the younger adults exhibited higher SampEn values than the elderly adults at a threshold of  $p = 0.005$  with a family-wise error (FWE) corrected cluster level significance of  $p < 0.05$  at the parietal and frontal lobes. These discriminated brain regions are listed in **Table 7**. For data length,  $N = 85$ , the younger adults also exhibited higher SampEn values than the elderly adults at the left parietal lobe ( $-24, -48, 54$ , Sub-Gyrus, White Matter;  $-22, -52, 44$ , Precuneus, White Matter;  $-32, -40$ ,

**Table 3 | Correlation of SampEn with age for the small group of 10 younger and 10 elderly adults.**

	Pearson's correlation ( $r$ -values)	Significance ( $p$ -values)	Significance FDR corrected ( $q$ -values)
SampEn at $N = 85$	$-0.602$	$p = 0.005$	$q = 0.006$
SampEn at $N = 90$	$-0.624$	$p = 0.003$	$q = 0.006$
SampEn at $N = 95$	$-0.626$	$p = 0.003$	$q = 0.006$
SampEn at $N = 100$	$-0.624$	$p = 0.003$	$q = 0.006$
SampEn at $N = 110$	$-0.599$	$p = 0.005$	$q = 0.006$
SampEn at $N = 120$	$-0.608$	$p = 0.004$	$q = 0.006$
SampEn at $N = 128$	$-0.581$	$p = 0.007$	$q = 0.007$

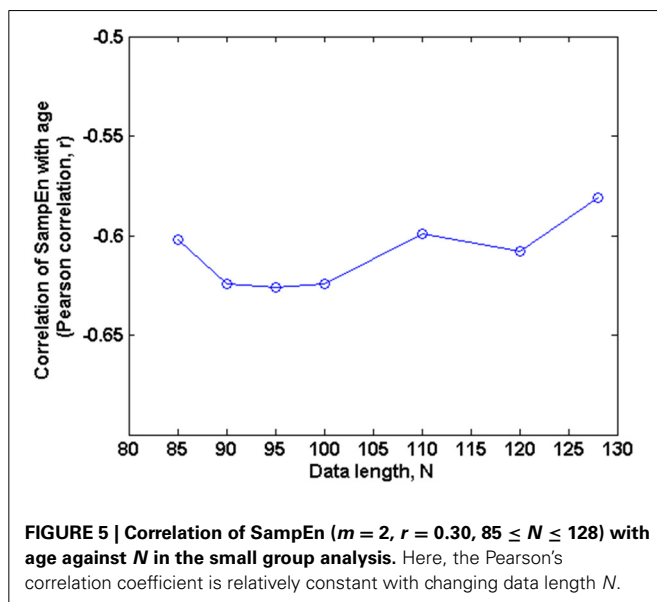
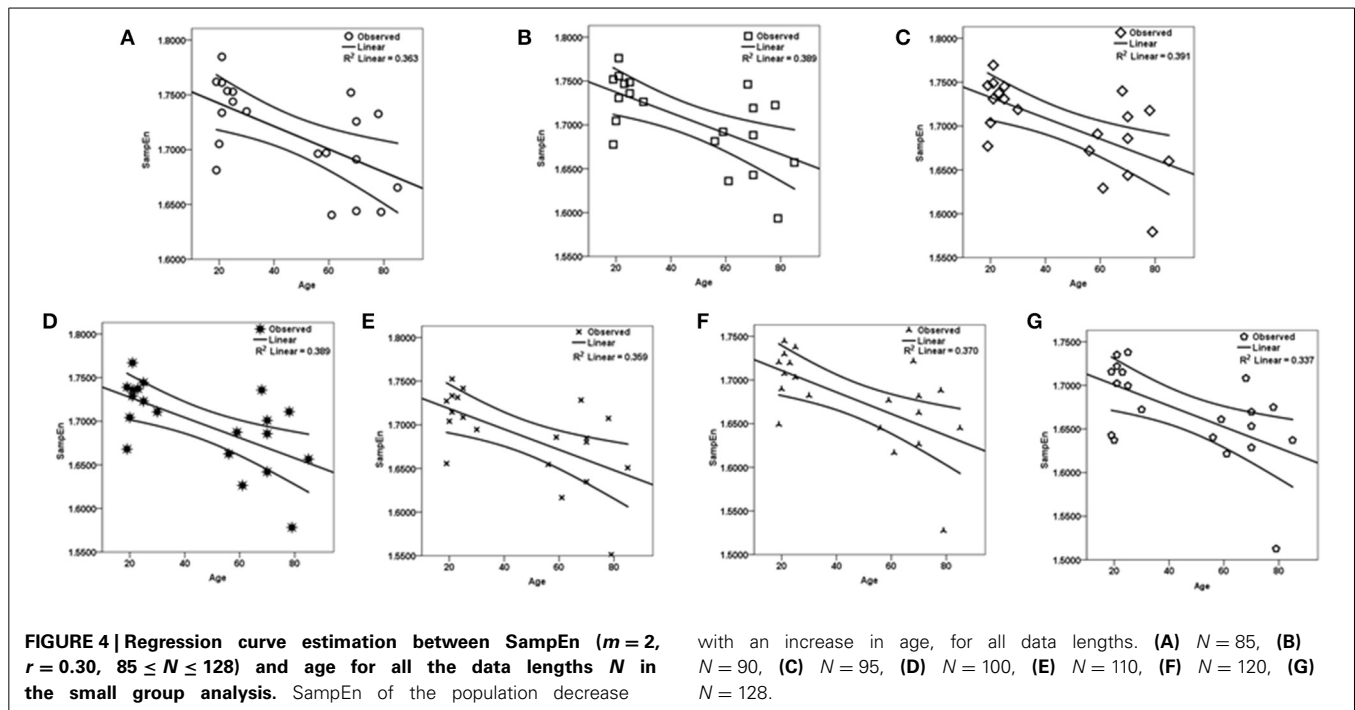
52, Postcentral Gyrus, White Matter) with a threshold of  $p = 0.005$  and at an uncorrected  $p$ -value of 0.005. When the analysis at  $N = 85$  was corrected for multiple comparisons, the discriminated brain region was not significant ( $p > 0.05$ ). There were no significant ( $p > 0.05$ ) correlations between the whole brain SampEn maps and age, of the whole population, for both data lengths ( $N = 85$  and  $N = 128$ ).

**Figure 8** shows the rendered images of the two-sample  $t$ -tests between the younger and elderly adults, for the small group (10 younger and 10 elderly adults) and the whole cohort (43 younger and 43 elderly adults) at data length  $N = 128$ . The images show that both analyses had overlapping discriminated brain regions between the frontal and parietal lobes.

#### DISCUSSION

The aim of this study was to test the ability of SampEn to effectively discriminate between two different age groups of resting state fMRI data with data length,  $N$  less than  $10^m$  (where  $m = 2$ ). For the small group analysis, the results of the whole brain analyses shows that the ROC areas for  $N = 85, 90$ , and  $95$  were the same (0.880), the ROC area for  $N = 100$  was 0.890, the areas for  $N = 110$  and  $120$  were 0.880, and for  $N = 128$  was 0.850. The disproportionality of these ROC areas to the respective data lengths is in line with the notion that SampEn is largely independent of data length. Furthermore, the same level of accuracy (85%) exhibited by all the data lengths with the exception of  $N = 128$  having accuracy of 80%, indicates that SampEn displays some relative consistency. Also, the mean whole brain SampEn of the younger adults was significantly ( $p < 0.05$ ) higher than the elderly adults across data lengths,  $85 \leq N \leq 128$ . There were also moderate negative correlations ( $r$ -values between  $-0.581$  and  $-0.626$ ) (see **Table 3**) between the mean whole brain SampEn values and age for  $85 \leq N \leq 128$  at  $q < 0.05$ . Wang et al. (2014) showed that data length has only a minor effect on SampEn, which ensured including all the resting state fMRI data at the 1000 Functional Connectomes project repository, even with different time points for their brain entropy (BEN) mapping.

In the regional analyses of the small group, the younger adults exhibited significantly higher SampEn than the elderly adults, only at the frontal lobe for  $N = 85$ , and at the frontal and parietal lobes for  $N = 128$ . For  $N = 85$ , there was a significant negative correlation between SampEn and age at the frontal, limbic



and parietal lobes while for  $N = 128$ , this negative correlation occurred at the frontal lobe, limbic lobe, parietal lobe and sub-lobar region. These associations indicate that there is reduction in entropy with increase in age. This reduction in entropy is common to both analyses (at  $N = 85$  and  $N = 128$ ), independent of the different data lengths and overlaps at the frontal, limbic and parietal lobes of the brain. The frontal lobe has been implicated in age-related processes resulting in a decline in memory functions (Craik and Salthouse, 2000). In a diffusion tensor imaging (DTI) study of a healthy population of 25–70 years, the limbic system which is responsible for emotion processing and

memory function has been shown to undergo degradation with ageing (Gunbey et al., 2014). The sub-lobar brain region has been implicated in white matter structures associated with cognitive ageing (Staff et al., 2006). Also, decreased fractional anisotropy (FA) measurements in the frontal and parietal lobes has been associated with poorer cognitive performance in a study investigating the relationship between FA and selected measures of cognition across a broad age group (20–73 years of healthy subjects) to explore a possible structural basis for cognitive changes with age (Grieve et al., 2007). Our findings of decrease in entropy with age are consistent with Lipsitz's (2004) entropy definition of ageing (loss of entropy) and Vaillancourt and Newell's (2002) fixed-point attractor postulate where complexity decreases with age and disease.

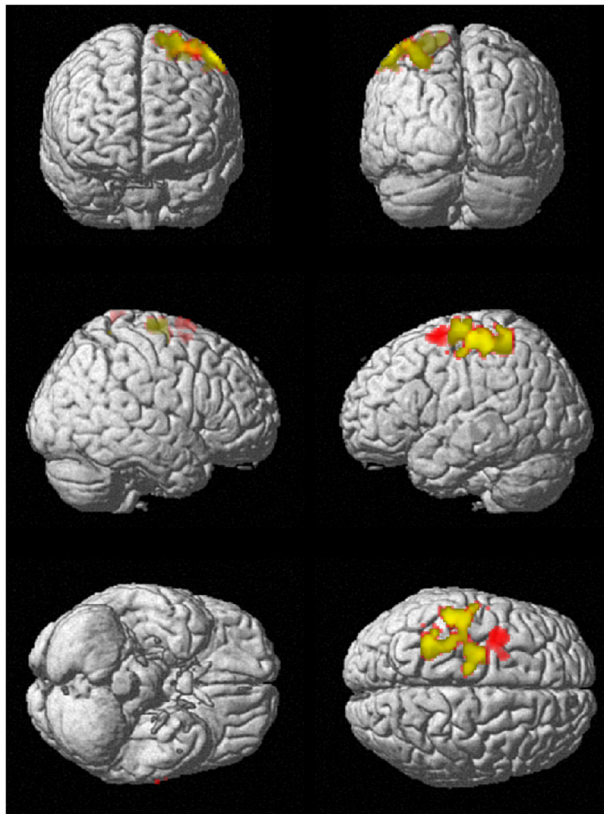
Comparing the whole cohort (43 younger and 43 elderly adults) to the small group (10 younger and 10 elderly adults) analysis at data lengths  $N = 85$  and  $128$ , the small group analysis discriminated between the younger and elderly adults, and showed that the fMRI brain complexity decreases with age at both data lengths. The whole cohort analysis only discriminated between the younger and elderly adults at  $N = 128$ . The whole cohort analysis at  $N = 85$  was indicative of the ageing process but this indication was not significant ( $p > 0.05$ ). The inability of SampEn to portray the same discriminatory effect for both the small group and whole cohort analyses may be due to two factors. Firstly, it may be due to the variance in the heterogeneous distribution of the subjects' ages in both datasets. For the small group, the mean age of the younger and elderly adults is ( $22.40 \pm 3.44$ ) and ( $69.60 \pm 9.25$ ) respectively, while in the whole cohort the mean age of the younger and elderly adults is ( $29.05 \pm 8.66$ ) and ( $59.33 \pm 10.27$ ) respectively. Clearly, there is disparity in the mean and SD of the younger and elderly adults between the



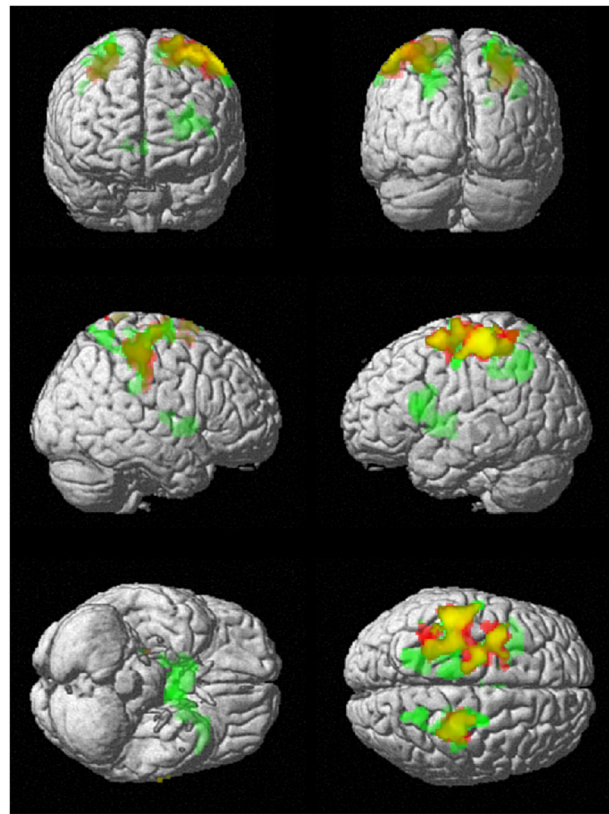
**Table 4 | SampEn differences for the small group of 10 younger and 10 elderly adults.**

Data length, <i>N</i>	Cluster number and extent	Brain region	Talairach coordinate ( <i>X, Y, Z</i> )	Brain label	Tissue type	Cluster <i>p</i> -value (FWE corrected)	Voxel <i>t</i> -value
85	Cluster 1 Extent = 2181	Frontal lobe	−34, 2, 66	Left middle frontal gyrus	Gray matter	$p < 0.001$	5.47
128	Cluster 1 Extent = 889	Frontal lobe	−22, −14, 66	Left middle frontal gyrus	Gray Matter	$p = 0.007$	4.02
		Parietal lobe	−28, −44, 56	Left inferior parietal lobule	White matter	$p = 0.007$	4.26
		Parietal lobe	−46, −22, 60	Left post-central gyrus	Gray matter	$p = 0.007$	3.90

Location coordinates are those of the peak significance in each region (threshold  $p = 0.005$ , FWE corrected cluster  $p < 0.05$ ).



**FIGURE 6 | SampEn ( $m = 2$ ,  $r = 0.30$ ,  $N$ ) differences between younger and elderly adults for the small group analysis.  $N = 85$  is red and  $N = 128$  is green. Overlap is yellow. SampEn values of the younger adults were significantly ( $p < 0.05$ ) higher than SampEn values of the elderly adults with the corresponding brain regions as shown.**



**FIGURE 7 | Correlation of SampEn ( $m = 2$ ,  $r = 0.30$ ,  $N$ ) with age for the small group analysis.  $N = 85$  is red and  $N = 128$  is green. Overlap is yellow. SampEn for the population decrease as age increase with corresponding brain regions as depicted.**

small group and whole cohort. The second factor may be due to the limited discriminatory ability of SampEn. This study was conducted with SampEn on a single scale, a multiscale SampEn analysis is superior to a single scale analysis and portrays a superior discriminatory ability (Costa et al., 2002; Yang et al., 2013). Another approach which may show superior discriminatory ability to SampEn is single scale Fuzzy approximate entropy (fApEn) (Xie et al., 2010), which has not been investigated in comparison to SampEn and in fMRI datasets.

An increase in functional entropy with age (Yao et al., 2013) was found in a recent study, where Shannon entropy; a measure of information, choice and uncertainty (in bits) (Shannon, 1948) was used as a bivariate measure to characterize the correlation coefficient (considered as a random variable) of a distinct pair of brain regions. The resulting entropy measure in bits was called functional entropy. The functional entropy measured the dispersion (or spread) of functional connectivity that exists within the brain. At the population level, they found that the functional

**Table 5 | SampEn correlation with age for  $N = 85$  and  $N = 128$ , for the small group of 10 younger and 10 elderly adults.**

Data length, $N$	Cluster number and extent	Brain region	Talairach coordinate ( $X, Y, Z$ )	Brain label	Tissue type	Cluster $p$ -value (FWE corrected)	Voxel $t$ -value
85	Cluster 1 Extent = 768	Frontal lobe	36, -22, 48	Right post-central gyrus	White matter	$p = 0.015$	6.96
		Frontal lobe	30, -22, 38	Right sub-gyral	White matter	$p = 0.015$	5.30
		Limbic lobe	20, -24, 40	Right cingulate gyrus	White matter	$p = 0.015$	5.14
	Cluster 2 Extent = 3320	Frontal lobe	-34, 2, 66	Left middle frontal gyrus	Gray matter	$p < 0.001$	5.69
		Parietal lobe	-46, -22, 60	Left post-central gyrus	Gray matter	$p < 0.001$	5.42
128	Cluster 1 Extent = 1247	Frontal lobe	-30, 16, 16	Left sub-gyral	White matter	$p = 0.004$	8.40
		Limbic lobe	2, 2, -4	Right anterior cingulate	Gray matter	$p = 0.004$	5.28
		Sub-lobar	-6, -2, 4	Left extra-nuclear	White matter	$p = 0.004$	5.26
	Cluster 2 Extent = 3406	Parietal lobe	-26, -42, 56	Left sub-gyral	White matter	$p < 0.001$	5.90
		Parietal lobe	-20, -54, 40	Left pre-cuneus	White matter	$p < 0.001$	5.32
		Parietal lobe	-50, -28, 58	Left post-central gyrus	Gray matter	$p < 0.001$	4.85
	Cluster 3 Extent = 1246	Parietal lobe	32, -34, 54	Right post-central gyrus	Gray matter	$p = 0.004$	5.09
		Frontal lobe	20, -18, 64	Right middle frontal gyrus	White matter	$p = 0.004$	4.65
		Parietal lobe	28, -28, 48	Right sub-gyral	White matter	$p = 0.004$	4.48

Location coordinates are those of the peak significance in each region (threshold  $p = 0.005$ , FWE corrected cluster  $p < 0.05$ ).

**Table 6 | Subjects' characteristics and SampEn measures for the whole ICBM resting state cohort of 43 younger and 43 elderly adults.**

	Younger adults	Elderly adults	Significance ( $p$ -values)
Age (years)	29.05 $\pm$ 8.66	59.33 $\pm$ 10.27	$p < 0.001$
Sex (M/F)	21/22	20/23	
SampEn at $N = 85$	1.7387 $\pm$ 0.0526	1.7172 $\pm$ 0.0597	$p = 0.080$
SampEn at $N = 128$	1.6979 $\pm$ 0.0545	1.6735 $\pm$ 0.0655	$p = 0.065$

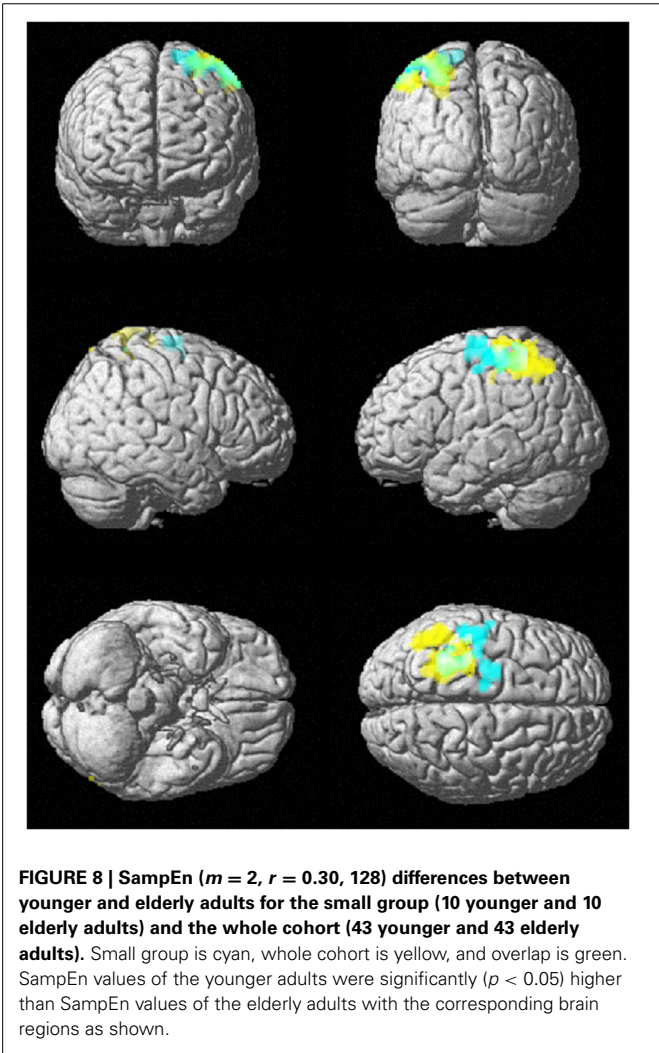
entropy of the human brain increases with age where a higher level of randomness reflected the way different brain-regions functionally interacted with one another. At the regional level, they found some regions where the functional entropy increases, decreases and where it remains almost constant. They noted a decrease in functional entropy with age in the left and right insulars. Furthermore, a computational model based on DTI was used to investigate the origins of the relationship between functional entropy and age. The model implicated a brain entropy that decreases when the excitatory connection strength and neuron number in each brain region are simultaneously reduced. In the present study, our analysis entailed a univariate characterization of a voxel with SampEn. Here, SampEn is used as an estimate of complexity and returns a dimensionless numerical value. Our results showed that sample entropy decrease with age. SampEn and ApEn are not the same as Shannon entropy, they are used to indicate system complexity because both of them were defined as approximates to the Kolmogorov complexity (Wang et al., 2014).

In the small group analysis, the reduction in the accuracy of SampEn to effectively discriminate the younger from the elderly adults (in the mean whole brain analyses) from 85% for data lengths  $85 \leq N \leq 120$  to 80% for data length  $N = 128$  may be attributed to the “averaging effect” which is basically the simplest form of a digital filter and is a means of reducing the effect of random noise (Smith, 1999). Averaging the BOLD fMRI response of a voxel over a number of data lengths can help to improve the BOLD signal to noise ratio. The amount of noise reduction that this “averaging effect” can produce is equal to the square-root of the data length in the average (Smith, 1999). For example, data lengths of  $N = 85, 90, 95, 100, 110, 120$ , and 128 of BOLD fMRI signal would reduce the noise by a factor of 9.22, 9.49, 9.75, 10.00, 10.49, 10.96, and 11.31 respectively. As a result of this, the level of noise in data length  $N = 128$  is less compared to data length  $N = 85$  and vice versa. The level of noise in data length  $N = 85$  is higher than  $N = 128$ . Since noise is the signal with the most complex dynamics and highest measured entropy (Lu et al., 2008), it is expected that the entropy of the younger and elderly adults for data length  $N = 85$  would be higher than the corresponding groups in data length  $N = 128$  and was therefore reflected in the measured accuracies. This is evident in the mean whole brain SampEn measurements for  $85 \leq N \leq 128$  in **Table 1**. Here it can be clearly seen that the measured SampEn values decreases as the data length increases from  $N = 85$  to  $N = 128$ , implying that the level of noise decrease from  $N = 85$  to  $N = 128$ . Another obvious evidence suggesting the influence of noise in the accuracy was demonstrated in the regional analyses where noise played an opposite effect. Here, Sample entropy discriminated more brain regions at  $N = 128$  than  $N = 85$ . The difference in the discriminated brain regions can be attributed to the influence of a higher noise level in  $N = 85$  than  $N = 128$ .

Table 7 | SampEn differences for the whole ICBM resting state cohort of 43 younger and 43 elderly adult.

Data length, <i>N</i>	Cluster number and extent	Brain region	Talairach coordinate ( <i>X, Y, Z</i> )	Brain label	Tissue type	Cluster <i>p</i> -value (FWE corrected)	Voxel <i>t</i> -value
128	Cluster 1 Extent = 2251	Parietal lobe	−24, −46, 56	Left sub-gyral	Gray matter	<i>p</i> < 0.001	4.41
		Parietal lobe	−24, −56, 52	Left precuneus	White matter	<i>p</i> < 0.001	3.58
		Parietal lobe	−46, −22, 60	Left inferior parietal lobule	Gray matter	<i>p</i> < 0.001	3.11
		Frontal lobe	−26, −30, 66	Left precentral gyrus	Gray matter	<i>p</i> < 0.001	3.00
		Frontal lobe	−28, −24, 46	Left sub-gyral	White matter	<i>p</i> < 0.001	2.95

Location coordinates of the significant regions (threshold *p* = 0.005, FWE corrected cluster *p* < 0.05).



**FIGURE 8 |** SampEn (*m* = 2, *r* = 0.30, 128) differences between younger and elderly adults for the small group (10 younger and 10 elderly adults) and the whole cohort (43 younger and 43 elderly adults). Small group is cyan, whole cohort is yellow, and overlap is green. SampEn values of the younger adults were significantly (*p* < 0.05) higher than SampEn values of the elderly adults with the corresponding brain regions as shown.

Sample entropy (an optimized approximate entropy) is nearly unaffected by low level noise, is robust to occasional very large or small artifacts, gives meaningful information with a reasonable number of data lengths, and is finite for both stochastic and deterministic processes (Zhang and Roy, 2001).

In the computation of Sample entropy from an fMRI signal, a high noise level is a potential confounder and may prevent

Sample entropy from discriminating effectively between system complexities. The noise present in fMRI data consists of system noise (white noise), arising from both thermal noise and hardware imperfections, and 1/*f* low-frequency noise, physiological fluctuations from respiratory and cardiac activities. The noise level can be reduced as we have done by applying high pass filtering to reduce the low frequency components of the noise and spatial smoothing to reduce the system noise. With the level of noise reduced, an optimized and robust computation of Sample entropy can be implemented with an appropriate tolerance value, *r*. To avoid a significant contribution from noise in the calculation of the entropy, one must choose *r* larger than most of the noise (Pincus, 1991). A higher *r*-value shows better robustness to reduced noise in distinguishing the nonlinear system dynamics (Xie et al., 2010) of the experimental and control groups. When a small *r*-value is used, the algorithm identifies two sections being compared as dissimilar when the difference may be brought about by noise. Using a larger *r* avoids the misclassification. Using a large *r*, however, may result in some signal detail being lost. The selection of the appropriate *r* is essentially a compromise between these two phenomena: i.e., an *r* large enough that allows the Sample entropy algorithm to distinguish the system signal from noise, but small enough to allow the algorithm to assess the detail present in the signal (Chen et al., 2009). We have used a higher *r*-value to obtain an optimized and robust computation of Sample entropy in the presence of minimal noise. The *r*-value (*r* = 0.30) we used showed better robustness to reduced noise in distinguishing the nonlinear system dynamics of both younger and elderly adults (Figure 1A).

Some studies have suggested that the bias of SampEn from short data lengths may be compensated for by using a small pattern length (*m* = 1) and a relatively large similarity factor (tolerance value), *r*, to accommodate the short and noisy BOLD data (Yang et al., 2013). The choice of *m* = 2 is superior to *m* = 1 because it allows more detailed reconstruction of the joint probabilistic dynamics of the time series (Pincus and Goldberger, 1994). It has also been shown that using *m* = 2 is more consistent than *m* = 1 over a wider range of tolerance values, *r* (Sokunbi et al., 2013). Using *m* = 2 implies that the SampEn of fMRI data with data length less than 100 can be computed with robust and optimized parameter contrary to the suggestion of others (Abasolo et al., 2006; Yang et al., 2013), avoiding erroneous data length constraint. Also, *m* = 2 has been



used for data length  $N = 50$  of i.i.d uniform random numbers (Chen et al., 2009).

Richman and Moorman (2000) concluded that the SampEn ( $m, r, N$ ) statistics are not completely unbiased under all conditions. They found that the bias of SampEn was less than 3% for data lengths greater than 100 but as high as 35% for data length of 15 points and that the bias of SampEn for very small data sets is largely due to non-independence of templates. They suggested that one method of removing this bias would be to partition the time series but noted that this unbiased approach has the potentially severe limitation of reducing the number of possible template matches and enlarging the confidence intervals about the SampEn estimate. They also argue that because this bias appears to be present only for very small  $N$ , the disjoint template approach does not appear necessary in usual practice. One notable limitation of the present study is that we would expect the bias of our fMRI SampEn (2, 0.30,  $85 \leq N \leq 128$ ) analyses to be in the proximity of the bias of less than 3% for data lengths greater than 100. Another limitation of SampEn is that SampEn values for data lengths less than 85 could not be obtained because of a lack of templates to compare.

## CONCLUSION

The small group fMRI SampEn analyses provided additional evidence that it is possible to obtain good discriminating feature from fMRI data with data lengths less than 100, indicating that SampEn is largely independent on changes in data length and displays some relative consistency. While it is better to acquire data with longer data lengths for best analysis results, low noise level and minimum bias, it is not always possible to do this with fMRI data because of the nature of some fMRI experiments and its temporal limitation. SampEn is a possible analysis tool amongst time series analysis techniques because it is less sensitive to changes in data length and relatively consistent. SampEn is well suited for short data sets like fMRI data, though a compromise has to be made with the increase in noise level as data length decreases. The heterogeneous distribution of the subjects ages in the whole cohort ages compared to the small group ages may have limited the single scale discriminatory ability of SampEn in the whole cohort analyses. A multiscale SampEn analysis may portray a superior discriminatory ability. In the present study, using  $m = 2$  ensures that SampEn is computed for fMRI data (having data length less than 100) with robust and optimized parameter thereby avoiding the erroneous data length constraint of  $10^m - 20^m$ . Finally, before characterizing data sets, especially short data sets with SampEn, we would recommend using optimal parameters; an  $m$  of 2 or as appropriate and to determine the  $r$ -value (by examining several  $r$ -values) where SampEn displays its best discriminating ability.

## ACKNOWLEDGMENT

We acknowledge the 1000 Functional Connectomes project ([https://www.nitrc.org/projects/fcon\\_1000/](https://www.nitrc.org/projects/fcon_1000/)) for the use of the publicly available International Consortium for Brain Mapping (ICBM) resting state dataset.

## SUPPLEMENTARY MATERIAL

The Supplementary Material for this article can be found online at: <http://www.frontiersin.org/journal/10.3389/fninf.2014.00069/abstract>

## REFERENCES

- Abasolo, D., Hornero, R., Espino, P., Álvarez, D., and Poza, J. (2006). Entropy analysis of the EEG background activity in Alzheimer's disease patients. *Physiol. Meas.* 27, 241–253. doi: 10.1088/0967-3334/27/3/003
- Chen, W., Zhuang, J., Yu, W., and Wang, Z. (2009). Measuring complexity using FuzzyEn, ApEn, and SampEn. *Med. Eng. Phys.* 31, 61–68. doi: 10.1016/j.medengphy.2008.04.005
- Costa, M., Goldberger, A. L., and Peng, C. K. (2002). Multiscale entropy analysis of complex physiologic time series. *Phys. Rev. Lett.* 89:068102. doi: 10.1103/PhysRevLett.89.068102
- Craik, F. I. M., and Salthouse, T. A. (eds.). (2000). *The Handbook of Aging and Cognition*. 2nd Edn. New Jersey: Lawrence Erlbaum Associates, Inc.
- Dancey, C., and Reidy, J. (2004). *Statistics Without Maths for Psychology: Using SPSS for Windows*. London: Prentice Hall.
- Deary, I. J., Corley, J., Gow, A. J., Harris, S. E., Houlihan, L. M., Marioni, R. E., et al. (2009). Age-associated cognitive decline. *Br. Med. Bull.* 92, 135–152. doi: 10.1093/bmb/ldp033
- Eckmann, J. P., and Ruelle, D. (1992). Fundamental limitations for estimating dimensions and Lyapunov exponents in dynamical system. *Physica D* 56, 185–187. doi: 10.1016/0167-2789(92)90023-G
- Ferenets, R., Vanluchene, A., Lipping, T., Heyse, B., and Struys, M. M. (2007). Behavior of entropy/complexity measures of the electroencephalogram during propofol-induced sedation: dose-dependent effects of remifentanyl. *Anesthesiology* 106, 696–706. doi: 10.1097/01.anes.0000264790.07231.2d
- Gaal, Z. A., Boha, R., Stam, C. J., and Molnár, M. (2010). Age-dependent features of EEG- reactivity-Spectral, complexity, and network characteristics. *Neurosci. Lett.* 479, 79–84. doi: 10.1016/j.neulet.2010.05.037
- Goldberger, A. L., Peng, C., and Lipsitz, L. A. (2002). What is physiologic complexity and how does it change with aging and disease? *Neurobiol. Aging* 23, 23–26. doi: 10.1016/S0197-4580(01)00266-4
- Gomez, C., Poza, J., Garcia, M., Fernandez, A., and Hornero, R. (2011). “Regularity analysis of spontaneous MEG activity in Attention-Deficit/Hyperactivity Disorder,” in *33rd Annual International Conference of the IEEE EMBS* (Boston, MA), 1765–1768.
- Grassberger, P., and Procaccia, I. (1983). Characterization of strange attractors. *Phys. Rev. Lett.* 50, 346–349. doi: 10.1103/PhysRevLett.50.346
- Grieve, S. M., Williams, L. M., Paul, R. H., Clark, C. R., and Gordon, E. (2007). Cognitive aging, executive function, and fractional anisotropy: a diffusion tensor MR imaging study. *Am. J. Neuroradiol.* 28, 226–235.
- Gunbey, H. P., Ercan, K., Findikoglu, A. S., Bulut, H. T., Karaoglanoglu, M., and Arslan, H. (2014). The limbic degradation of aging brain: a quantitative analysis with diffusion tensor imaging. *Sci. World J.* 2014:196513. doi: 10.1155/2014/196513
- Lipsitz, L. A. (2004). Physiological complexity, aging, and the path to frailty. *Sci. Aging knowledge Environ.* 2004:pe16. doi: 10.1126/sageke.2004.16.pe16
- Lu, S., Chen, X., Kanter, J. K., Solomon, I. C., and Chon, K. H. (2008). Automatic selection of the threshold value  $r$  for approximate entropy. *IEEE Trans. Biomed. Eng.* 55, 1966–1972. doi: 10.1109/TBME.2008.919870
- Pincus, S. (1995). Approximate entropy (ApEn) as a complexity measure. *Chaos* 5, 110–117. doi: 10.1063/1.166092
- Pincus, S. M. (1991). Approximate entropy as a measure of system complexity. *Proc. Natl. Acad. Sci. U.S.A.* 88, 2297–2301. doi: 10.1073/pnas.88.6.2297
- Pincus, S. M. (2001). Assessing serial irregularity and its implications for health. *Ann. N.Y. Acad. Sci.* 954, 245–267. doi: 10.1111/j.1749-6632.2001.tb02755.x
- Pincus, S. M., and Goldberger, A. L. (1994). Physiological time-series analysis: what does regularity quantify? *Am. J. Physiol.* 266, H1643–H1656.
- Pool, R. (1989). Is it healthy to be chaotic? *Science* 243, 604–607. doi: 10.1126/science.2916117
- Pritchard, W. S., Duke, D. W., Coburn, K. L., Moore, N. C., Tucker, K. A., Jann, M. W., et al. (1994). EEG-based neural-net predictive classification of Alzheimer's



- disease versus control subjects is augmented by non-linear EEG measures. *Electroencephalogr. Clin. Neurophysiol.* 91, 118–130. doi: 10.1016/0013-4694(94)90033-7
- Richman, J. S., and Moorman, J. R. (2000). Physiological time-series analysis using approximate and sample entropy. *Am. J. Physiol.* 278, H2039–H2049.
- Shannon, C. E. (1948). A mathematical theory of communication. *Bell Syst. Tech. J.* 27, 379–656. doi: 10.1002/j.1538-7305.1948.tb01338.x
- Smith, S. W. (1999). *The Scientist and Engineer's Guide to Digital Signal Processing*. 2nd Edn. San Diego, CA: California Technical Publishing.
- Sokunbi, M. O., Fung, W., Sawlani, V., Choppin, S., Linden, D. E. J., and Thome, J. (2013). Resting state fMRI entropy probes complexity of brain activity in adults with ADHD. *Psychiatry Res.* 214, 341–348. doi: 10.1016/j.psychres.2013.10.001
- Sokunbi, M. O., Gradin, V. B., Waiter, G. D., Cameron, G. G., Ahearn, T. S., Murray, A. D., et al. (2014). Nonlinear complexity analysis of brain fMRI signals in schizophrenia. *PLoS ONE* 9:e95146. doi:10.1371/journal.pone.0095146
- Sokunbi, M. O., Staff, R. T., Waiter, G. D., Ahearn, T. S., Fox, H. C., Deary, I. J., et al. (2011). Inter-individual differences in fMRI entropy measurements in old age. *IEEE Trans. Biomed. Eng.* 58, 3206–3214. doi: 10.1109/TBME.2011.2164793
- Staff, R. T., Murray, A. D., Deary, I. J., and Whalley, L. J. (2006). Generality and specificity in cognitive aging: a volumetric brain analysis. *Neuroimage* 30, 1433–1440. doi: 10.1016/j.neuroimage.2005.11.004
- Vaillancourt, D. E., and Newell, K. M. (2002). Changing complexity in human behavior and physiology through aging and disease. *Neurobiol. Aging* 23, 1–11. doi: 10.1016/S0197-4580(01)00247-0
- Wang, Z., Li, Y., Childress, A. R., and Detre, J. A. (2014). Brain entropy mapping using fMRI. *PLoS ONE* 9:e89948. doi:10.1371/journal.pone.0089948
- Wolf, A., Swift, J. B., Swinney, H. L., and Vastano, J. A. (1985). Determining Lyapunov exponents from a time series. *Physica D* 16, 285–317. doi: 10.1016/0167-2789(85)90011-9
- Xie, H. B., Guo, J. Y., and Zheng, Y. P. (2010). Fuzzy approximate entropy analysis of chaotic and natural complex systems: detecting muscle fatigue using electromyography signals. *Ann. Biomed. Eng.* 38, 1483–1496. doi: 10.1007/s10439-010-9933-5
- Yao, Y., Lu, W. L., Xu, B., Li, C. B., Lin, C. P., Waxman, D., et al. (2013). The increase of the functional entropy of the human brain with age. *Nat. Sci. Rep.* 3, 2853. doi: 10.1038/srep02853
- Yang, A. C., Huang, C. C., Yeh, H. L., Liu, M. E., Hong, C. J., Tu, P. C., et al. (2013). Complexity of spontaneous BOLD activity in default mode network is correlated with cognitive function in normal male elderly: a multiscale entropy analysis. *Neurobiol. Aging* 34, 428–438. doi: 10.1016/j.neurobiolaging.2012.05.004
- Yentes, J. M., Hunt, N., Schmid, K. K., Kaipust, J. P., McGrath, D., and Stergiou, N. (2013). The appropriate use of approximate entropy and sample entropy with short data sets. *Ann. Biomed. Eng.* 41, 349–365. doi: 10.1007/s10439-012-0668-3
- Zhang, X., and Roy, R. J. (2001). Derived fuzzy knowledge model for estimating the depth of Anesthesia. *IEEE Trans. Biomed. Eng.* 48, 312–323. doi: 10.1109/10.914794
- Zweig, M. H., and Campbell, G. (1993). Receiver-operating characteristic (ROC) plots: a fundamental evaluation tool in clinical medicine. *Clin. Chem.* 39, 561–577.

**Conflict of Interest Statement:** The author declares that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.

Received: 12 March 2014; accepted: 03 July 2014; published online: 23 July 2014.

Citation: Sokunbi MO (2014) Sample entropy reveals high discriminative power between young and elderly adults in short fMRI data sets. *Front. Neuroinform.* 8:69. doi: 10.3389/fninf.2014.00069

This article was submitted to the journal *Frontiers in Neuroinformatics*.

Copyright © 2014 Sokunbi. This is an open-access article distributed under the terms of the Creative Commons Attribution License (CC BY). The use, distribution or reproduction in other forums is permitted, provided the original author(s) or licensor are credited and that the original publication in this journal is cited, in accordance with accepted academic practice. No use, distribution or reproduction is permitted which does not comply with these terms.



# Multi-scale integration and predictability in resting state brain activity

Artemy Kolchinsky<sup>1,2</sup>, Martijn P. van den Heuvel<sup>3</sup>, Alessandra Griffo<sup>4,5</sup>, Patric Hagmann<sup>4,5</sup>, Luis M. Rocha<sup>1,2</sup>, Olaf Sporns<sup>6</sup> and Joaquín Goñi<sup>6\*</sup>

<sup>1</sup> Department of Informatics, School of Informatics and Computing, Indiana University, Bloomington, IN, USA

<sup>2</sup> Instituto Gulbenkian de Ciência, Oeiras, Portugal

<sup>3</sup> Department of Psychiatry, Rudolf Magnus Institute of Neuroscience, University Medical Center Utrecht, Utrecht, Netherlands

<sup>4</sup> Signal Processing Laboratory 5, Ecole Polytechnique Fédérale de Lausanne, Lausanne, Switzerland

<sup>5</sup> Department of Radiology, Lausanne University Hospital (CHUV) and University of Lausanne (UNIL), Lausanne, Switzerland

<sup>6</sup> Department of Psychological and Brain Sciences, Indiana University, Bloomington, IN, USA

## Edited by:

Daniele Marinazzo, University of Ghent, Belgium

## Reviewed by:

Petra Ritter, Charité – Universitätsmedizin Berlin, Germany  
Jaroslav Hlinka, Academy of Sciences of the Czech Republic, Czech Republic

## \*Correspondence:

Joaquín Goñi, Department of Psychological and Brain Sciences, Indiana University, 1101 E 10th St., Bloomington, IN 47405, USA  
e-mail: jgonicor@indiana.edu

The human brain displays heterogeneous organization in both structure and function. Here we develop a method to characterize brain regions and networks in terms of information-theoretic measures. We look at how these measures scale when larger spatial regions as well as larger connectome sub-networks are considered. This framework is applied to human brain fMRI recordings of resting-state activity and DSI-inferred structural connectivity. We find that strong functional coupling across large spatial distances distinguishes functional hubs from unimodal low-level areas, and that this long-range functional coupling correlates with structural long-range efficiency on the connectome. We also find a set of connectome regions that are both internally integrated and coupled to the rest of the brain, and which resemble previously reported resting-state networks. Finally, we argue that information-theoretic measures are useful for characterizing the functional organization of the brain at multiple scales.

**Keywords:** human connectome, resting-state, integrative regions, information theory, multivariate mutual information, complexity measures

## INTRODUCTION

The human brain is characterized by complex functional and structural organization at different scales. Both structural and functional aspects of large-scale brain organization can be studied using magnetic resonance imaging (MRI) technology. On the one hand, functional activity can be estimated from the blood-oxygen-level dependent (BOLD) signal recorded by functional MRI (fMRI) of gray matter. The pattern of correlations between the BOLD activities of pairs of regions determines the *functional connectivity*. On the other hand, the *structural connectivity*, or network of anatomical connections between brain regions also called the *human connectome* (Sporns et al., 2005; Sporns, 2013), can be inferred from the orientation of constrained diffusion throughout the brain as measured by diffusion spectrum imaging (DSI).

Recent research has sought to characterize the different functional and structural properties of different brain regions, both in “bottom-up” terms, by assigning distinct roles to localized regions, as well as in “top-down” terms, by decomposing the entire brain into interpretable networks or subsystems. For example, many functional studies have investigated *functional hubs*, or regions that maintain strong correlations with many other regions (Achard et al., 2006; van den Heuvel and Sporns, 2013). Other studies have decomposed resting-state time series into maximally independent components, where regions within the same components display correlated patterns of activation (Beckmann et al., 2005; Damoiseaux et al., 2006; Fox and Raichle, 2007; Smith et al.,

2009; Yeo et al., 2011; Moussa et al., 2012). Similarly, structural studies of the connectome have found differences among regions in features such as degree, strength, betweenness and k-coreness (Hagmann et al., 2008; van den Heuvel and Sporns, 2013). They have also identified important structural subsystems, including communities (Hagmann et al., 2008; Betzel et al., 2014) and a densely-interconnected “rich club” backbone that ties together distant hubs (van den Heuvel et al., 2012). These findings are generally in accordance with a view of the brain as being organized along hierarchical lines, with segregated low-level processing of unimodal information taking place in the primary visual, auditory, sensory and motor cortices, higher-level representation and association of modal information taking place in the secondary cortices, and multisensory areas integrating information between distinct modalities across large-scale networks (Felleman and Essen, 1991; Yeo et al., 2011).

In this work, we propose a method to characterize the information-theoretic properties of local brain regions as well as networks of regions, here referred to as *subsystems*. Our method employs functional time series in conjunction with spatial and structural connectivity data. It uses both structure and function data in a complementary manner, as opposed to studies that assess structural or functional domains separately, or that use one domain to predict the other, such as recent work in predicting functional from structural connectivity (Honey et al., 2010; Abdelnour et al., 2014; Goñi et al., 2014).

Specifically, we looked at the amount of *functional coupling* that holds between brain regions of interest (ROI), as quantified by *predictability* or the number bits of mutual information provided about the activity of one set of regions given knowledge of the activity of another set of regions. We also looked at the amount of *integration*, or internal functional coupling within subsystems as quantified by a multivariate generalization of mutual information.

In addition, we measured the *scaling* of predictability and integration by quantifying the growth of these measures as increasingly large sets of regions are considered, an approach motivated by previous work on multi-scale integration in complex multivariate systems (Grassberger, 1986; Tononi et al., 1994; Bialek et al., 2001). In this work, subsystems were defined with respect to the structural and physical organization of the brain. In particular, three different metrics were used to define subsystems (which may overlap): *Euclidean subsystems* are maximally compact in terms of physical distance; *Connectome subsystems* are maximally compact according to shortest path distances on the connectome; and *Randomized subsystems* are maximally compact according to shortest path distances on a randomly rewired version of the Connectome.

Our methodology combines data from resting-state functional MRI [fMRI] as well as from structural deterministic fiber tractography based on DSI. We first explored the scaling of measures of predictability of individual ROIs using subsystems of different sizes chosen using Euclidean, Connectome and Randomized metrics, where larger scales correspond to subsystems containing more ROIs. Then ROIs were characterized in terms of their functional coupling to the rest of their corresponding hemisphere, as well as in terms of the Euclidean spatial range at which they maintained long-distance functional coupling. An analysis of the correlation between functional coupling and a structural measure of shortest-path efficiency between ROIs and distant neighbors was performed across different scales. Finally, we looked at scaling of multivariate measures of integration within subsystems and of functional coupling of subsystems with the rest of the hemisphere. We identified a set of Connectome-based networks whose subsystems showed a combination of high internal integration and high coupling with the rest of the hemisphere.

The rest of the paper is organized as follows. The MRI data is described in the section MRI Data and section Distance Metrics describes the three different structural metrics considered in this study, corresponding to physical proximity (Euclidean), anatomical connectivity (Connectome), and a randomized control of the Connectome (Randomized). In section Information-Theoretic Measures and Efficiency, we describe our information-theoretic measures of the predictability of ROIs and subsystems at multiple spatial scales defined by the three metrics. In section Results, we report average measures of information-theoretic scaling, variation of these measures across the cortical surface, the relationship between long-range functional and structural shortest-paths, and identify Connectome subsystems that are both internally integrated and coupled to the rest of their hemispheres. In section Discussion, we discuss the use of information theory for studying the functional organization of the brain, interpret our results in the context of the integrative functions of the cortex, and

overview some methodological considerations. We finish by suggesting possible avenues for future development of our approach.

## MATERIALS AND METHODS

### MRI DATA

Forty healthy subjects underwent an MRI session on a 3T Siemens Trio scanner with a 32-channel head-coil. Magnetization Prepared Rapid Gradient Echo (MPRAGE) sequence was 1 mm in-plane resolution and 1.2 mm slice thickness, with a FOV of  $256 \times 240$  mm, and included 160 slices. Diffusion Spectrum Imaging (DSI) sequence included 128 diffusion weighted volumes + 1 reference  $b_0$  volume, with maximum  $b$ -value  $b = 8000$  s/mm<sup>2</sup>,  $2.2 \times 2.2 \times 3.0$  mm voxel size,  $212 \times 212$  mm FOV, and 34 slices. Functional MRI Echo Planar (EPI) sequence was 3.3 mm in-plane resolution with 3.3 mm slice thickness and 0.3 mm slice gap,  $212 \times 193$  mm FOV, 32 slices, and TR 1920 ms. DSI, resting-state fMRI and MPRAGE data were processed using the Connectome Mapping Toolkit (Daducci et al., 2012). All the processing steps were performed in the individual subject space with no spatial normalization.

Segmentation of gray and white matter was based on MPRAGE volumes. The parcellation used for all the analyses in this work divides the GM cortex into 448 ROIs (Cammoun et al., 2012); one ROI was eliminated due to signal acquisition errors, resulting in a final analysis on 447 ROIs (see Figure S1). Subcortical regions were not considered in this study. For reporting purposes, ROIs within each hemisphere were grouped into 34 larger, physically-compact *anatomical areas* corresponding to a GM anatomical atlas (Desikan et al., 2006). Figure S2 in the Supplementary Material shows the assignment of ROIs to anatomical areas.

During the resting-state fMRI acquisition, subjects were lying in the scanner with eyes open, resting but awake and cognitively alert, for a period of approximately 9 min. Functional data preprocessing included motion correction, regression of white matter, cerebrospinal fluid and movement signals, linear detrending, motion scrubbing and low-pass filtering (Fox et al., 2009; Power et al., 2012), producing a 280-sample time series for each ROI of each subject. The first four samples were removed to allow for signal stabilization, resulting in a final time series length of 276 samples per ROI per subject. Some subjects were found to have spikes in across-ROI variance of fMRI signal; maximum across-ROI variance over all time points was computed for all subjects and three subjects with outlier maximum variance were removed (outliers chosen according to Tukey's rule threshold of  $upper\text{-}quartile + 1.5 \times inter\text{-}quartile\text{-}range$ ). This resulted in a final dataset containing 37 subjects (16 female,  $25.3 \pm 5.0$  years old). The data used for the findings reported here were not processed with global signal regression. However, when global signal regression was applied, none of the reported results changed qualitatively (data not shown).

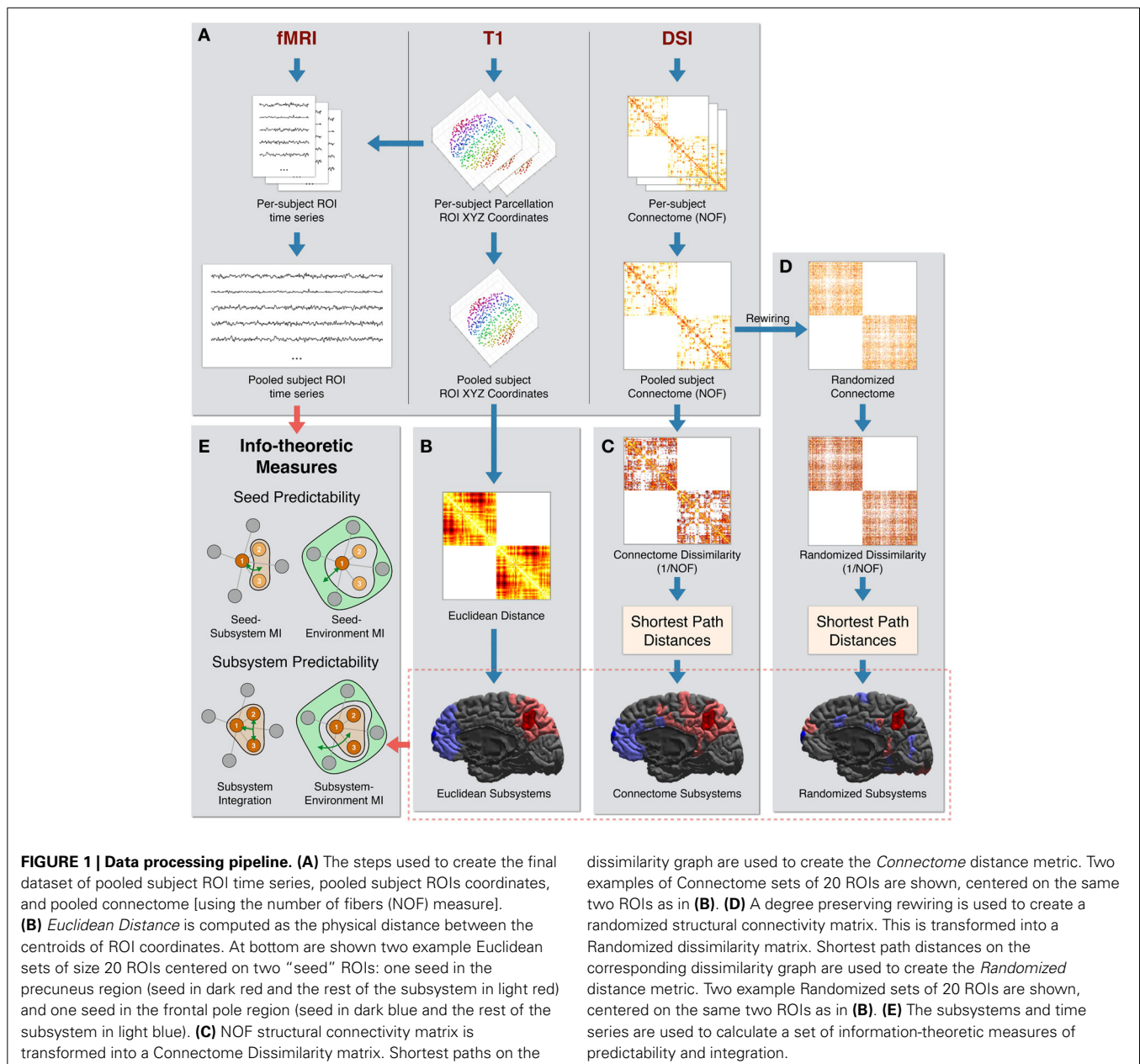
Whole brain streamline tractography was performed on reconstructed DSI data (Wedeen et al., 2008), resulting in a structural connectivity matrix where each entry reflects the number of fibers (Hagmann et al., 2008), denoted by NOF in this paper. This dataset was also assessed in two other studies (Betzel et al., 2014; Goñi et al., 2014). In this work, we did not consider inter-hemispheric connections, which pose difficulties for DWI-based

deterministic fiber tractography (Gong et al., 2009) and which may be systematically underrepresented in connectomes constructed using such methods.

Finally, subjects' fMRI time series and DSI connectomes were combined into a single "pooled" subject. Though no inter-subject spatial normalization was performed, subject-wise functional time series and structural connectivity can be pooled together because they were evaluated on the basis of the same anatomical atlas registered to each individual subject space. The time series for each ROI of each subject was mean-centered, rescaled to standard deviation 1, and concatenated across subjects to yield a single time series of  $276 \times 37 = 10,212$  samples. For the structural connectivity matrices, entries in the pooled matrices were taken to be means of the corresponding

connectivity values across the individual subject connectivity matrices. Though the data for the pooled subject does not correspond to any real subject, it is more robust and generates more stable and reliable statistics, important for computing the kinds of information-theoretic measures considered in this work (see Section Methodological Considerations). For these reasons, this kind of subject-pooling is frequently performed in computational neuroscience (van den Heuvel and Sporns, 2011; Deco et al., 2013; Haimovici et al., 2013; Betzel et al., 2014; Goñi et al., 2014).

The processing steps used to create the final dataset of pooled subject ROI time series, pooled subject ROIs coordinates, and pooled subject structural connectivity are diagrammed in Figure 1A.





## DISTANCE METRICS

As will be described in the next section, we computed measures of predictability in terms of a given ROI and its nearest neighbor ROIs. Nearest neighbors rankings were defined according to three different distance metrics: *Euclidean*, *Connectome*, and *Randomized*.

The *Euclidean* metric was defined as the Euclidean distance between the centroid coordinates of pairs of ROIs. The Euclidean neighbors of a given ROI were thus the most physically proximate ROIs. This is diagrammed in **Figure 1B**.

The *Connectome* metric was defined using anatomical connectivity inferred from DSI data. For the weights of structural connections linking ROIs, we used the NOF between ROIs as identified by the tractography algorithm (Hagmann et al., 2008). Since higher values of this measure indicate greater connectivity, we computed Connectome dissimilarity between anatomically connected ROIs as the inverse of the connectivity values between them (i.e.,  $1/\text{NOF}$ ). Connectome neighbors of a given ROI were other ROIs most proximate in terms of shortest-path distances on the Connectome dissimilarity graph. These processing steps are diagrammed in **Figure 1C**.

Finally, the *Randomized* metric was defined by first performing a degree-preserving rewiring (Maslov and Sneppen, 2002) of the Connectome graph. This rewiring method creates a randomized symmetric graph that preserves the density of the network (i.e., the number of direct connections), the degree (number of connections per ROI), and the overall distribution of NOF values. As performed in the Connectome metric, dissimilarity was computed as the inverse of the (rewired) NOF values. Analogously to the other metrics, Randomized neighbors of a given ROI were the most proximate ROIs in terms of shortest-path distances on the Randomized dissimilarity graph. These processing steps are diagrammed in **Figure 1D**.

As mentioned in the last section, due to possible confounding errors in inferring inter-hemispheric structural connectivity, each hemisphere was analyzed separately and only neighbors from the same hemisphere were considered for a given ROI.

We illustrate some examples of the subsystems defined according to these metrics in the bottom sections of **Figures 1B–D**. For each of the three metrics, two sets of 20 ROIs (a seed ROI and its 19 nearest neighbors) centered on two right-hemisphere ROIs are colored: one in the precuneus region (seed in dark red and rest of the subsystem in light red) and one in the frontal pole region (seed in dark blue and the rest of the subsystem in light blue).

As expected, sets of Euclidean neighbors (bottom of **Figure 1B**) are physically contiguous and compact. Connectome neighbors (bottom of **Figure 1C**) also tend to cluster spatially but are more distributed, with connections that span large physical distances present. In addition, according to the Connectome metric, the precuneus is close to the entire medial portion of the hemisphere (and far from more lateral regions) while the frontal pole is closer to superior and medial frontal as well as inferior temporal regions. Finally, the ROIs comprising Randomized neighbors (bottom of **Figure 1D**) are scattered throughout the hemisphere.

The Euclidean distance matrix as well as the Connectome and Randomized connectivity matrices are shown in Figure S3

in Supplementary Material. That figure also shows the neighbor ranks of all ROIs for all given seeds. Notably, while distances between ROIs are symmetric, ranks are not necessarily so (if one ROI is the  $k$ th neighbor of another ROI, the second is not necessarily the  $k$ th neighbor of the first).

## INFORMATION-THEORETIC MEASURES AND EFFICIENCY

Our information-theoretic measures of predictability were computed in terms of mutual information (MI) between different sets of ROIs. Mutual information is defined as

$$I(X; Y) := H(X) + H(Y) - H(X, Y)$$

where  $H(\cdot)$  stands for the entropy function. In addition, we used a multivariate generalization of MI known as *total correlation* (Ay et al., 2006), defined as the sum of the marginal entropies for a set of random variables minus their joint entropy:

$$TC(X_1, \dots, X_k) := \sum_{i=1}^k H(X_i) - H(X_1, \dots, X_k)$$

Total correlation,  $TC(\cdot)$ , quantifies the degree of multivariate correlation present in a subsystem and can be interpreted as the bits of compression gained by encoding the joint outcome of a set of random variables as opposed to encoding each variable's outcome independently. It is large when individual variables have high individual variance but are jointly correlated (for example, if all variables are copies of each other).

In this work, we considered the entropy of fMRI-recorded BOLD time-series of different brain regions. Because this data is continuous, we computed our information-theoretic measures using differential entropy (Cover and Thomas, 2012). For a random variable  $X$  with probability density function  $p(x)$ , differential entropy is defined as:

$$h(X) = - \int p(x) \log p(x) dx$$

To estimate differential entropies, we used a multivariate Gaussian assumption and employed the uniformly minimum-variance unbiased estimator of multivariate Gaussian entropy (Ahmed and Gokhale, 1989). If  $X$  is a  $k$ -by- $n$  matrix representing  $n$  samples from a  $k$ -dimensional multivariate Gaussian (for example, corresponding to samples of the activity of a group of  $k$  ROIs), this method estimates the entropy in bits of the underlying distribution as:

$$\frac{1}{\ln 2} \left( \frac{k}{2} \ln e\pi + \frac{1}{2} \ln |XX'| - \frac{1}{2} \sum_{i=1}^k \psi \left( \frac{n+1-i}{2} \right) \right)$$

where  $\psi$  is the digamma function. By itself, differential entropy is not guaranteed to be positive nor invariant to one-to-one coordinate transforms such as rescalings. However, mutual information and total correlation values computed using differential entropies are always positive and invariant to coordinate changes (Cover and Thomas, 2012).

We measured the information shared between sets of ROIs defined as follows. Any given ROI  $i$  can be considered as the target

of prediction, in which case it's called the *seed*. The ROI which is the  $k$ th ranked neighbor of  $i$  is indicated by  $n_i(k)$  (as described in the previous section, these can be chosen according to one of three different metrics: physical Euclidean distance, Connectome distance, or Randomized connectome distance). The seed together with its  $k - 1$  most proximate neighbor ROIs comprise the *subsystem of size  $k$*  centered on  $i$ , indicated by  $S_i(k)$ . For a given seed, all of the ROIs in the same hemisphere except those that are in its subsystem of size  $k$  (in other words, those that are further than its  $k$ th neighbor, according to a given metric) belong to the *environment*, indicated by  $E_i(k)$ .

Given these definitions of seed, neighbor, subsystem and environment, we defined the following five measures of ROI predictability:

**Pairwise MI**,  $I(i;j)$ , is the MI between the activity of any two individual ROIs  $i$  and  $j$ . One particular type of Pairwise MI we consider in detail is the **Seed-Neighbor MI**,  $I(i; n_i(k))$ , which is the MI between the activity of a seed ROI and the seed's  $k$ -ranked neighbor.

**Seed-Subsystem MI**,  $I(i; S_i(k) \setminus i)$ , is the MI between the activity of the seed ROI  $i$  and the joint activity of the rest of its size- $k$  subsystem. This measures how well the ROIs in a seed's size- $k$  subsystem collectively predict the seed. This measure is illustrated in schematic form in **Figure 1E**.

**Total MI**,  $I(i; V \setminus i)$ , where  $V$  represents the set of all the ROIs in the hemisphere, is the total amount of prediction possible about the seed using all other ROIs in the hemisphere. It is equivalent to the Seed-Subsystem MI when the subsystem corresponds to the entire hemisphere.

**Seed-Environment MI**,  $I(i; E_i(k))$ , is the MI between the activity of the seed and the joint activity of the ROIs in the environment. This measure quantifies how well ROIs in the environment predict the activity of the seed ROI. This measure is illustrated in schematic form in **Figure 1E**.

**Euclidean Coupling Range** is the neighbor number at which Seed-Environment MI drops below a specific threshold. This quantifies the smallest spatial scale at which a seed becomes effectively functionally decoupled from the environment.

In addition, we defined two multivariate measures for measuring the integration and predictability of entire subsystems. As before, we chose  $k$ -sized subsystems that are centered on a given seed ROI, and we again defined the environment as the set of ROIs in a hemisphere that are not members of a given subsystem. We considered two multivariate measures:

**Subsystem Integration**,  $TC(S_i(k))$ , is the total correlation of the activity of the set of ROIs in a size- $k$  subsystem centered on ROI  $i$ . This measure is high when ROI activity is individually varied but collectively correlated, and is illustrated in schematic form in **Figure 1E**.

**Subsystem-Environment MI**,  $I(S_i(k); E_i(k))$ , is the MI between the joint activity of the set of ROIs in a size- $k$  subsystem and the joint activity of the set of ROIs in the environment. This measure is high when there is strong functional coupling between subsystem and environment, and low when there is high functional segregation between the subsystem and the environment. This measure is illustrated in schematic form in **Figure 1E**.

When reporting these two subsystem predictability measures for subsystems of different sizes, we normalized them by subsystem size. This resulted in measures of **Subsystem Integration per ROI** and **Subsystem-Environment per ROI**.

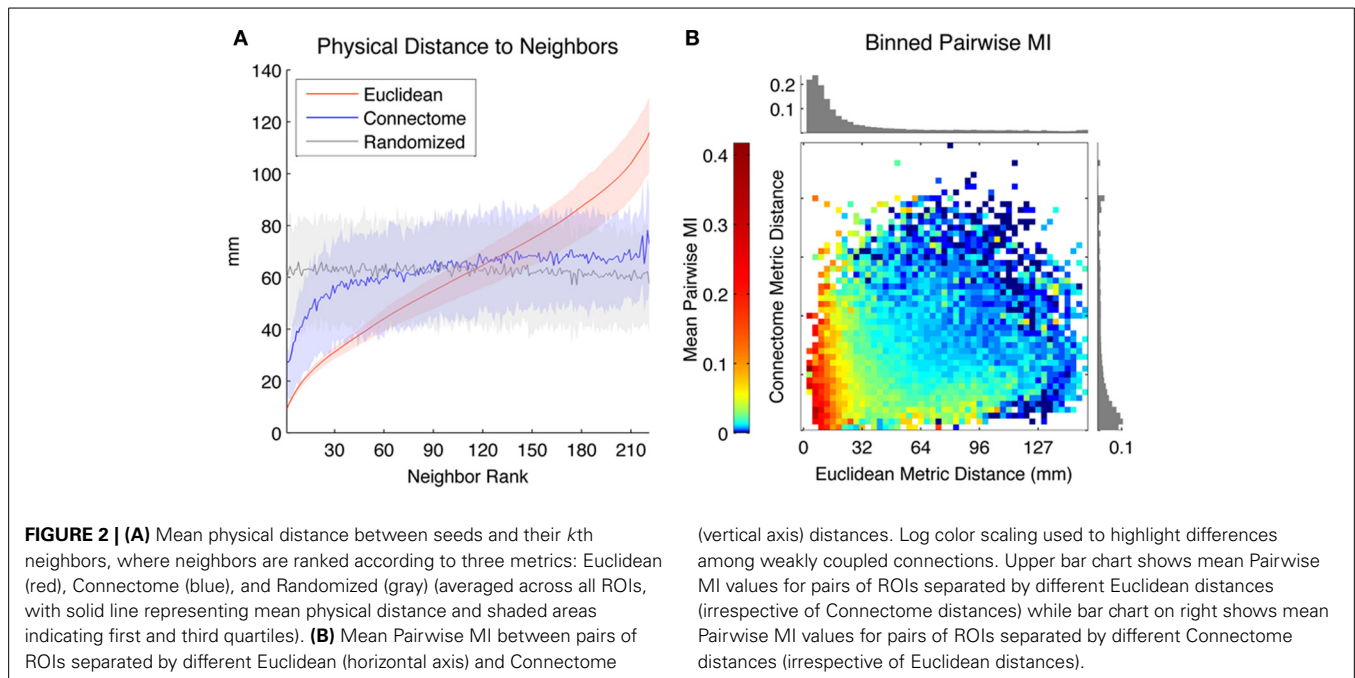
Finally, we also computed correlations between ROI predictability measures and one measure reflecting long-range efficiency. **Global efficiency** (Latora and Marchiori, 2001) is the average of the inverse of all shortest-path distances between pairs of vertices. We define **long-range efficiency** for an ROI within a subsystem as the mean inverse shortest-path between the ROI and its Euclidean environment ROIs (i.e., the ROIs outside of its Euclidean subsystem). The long-range efficiency between the seed and the ROIs in the seed's Euclidean environment was computed using shortest paths defined by the three aforementioned metrics: Euclidean, Connectome, and Randomized.

## RESULTS

As discussed in previous sections, for each ROI taken as the seed we obtained a list of neighbor ROIs ranked from most proximate to most distant according to three distance metrics (Euclidean, Connectome and Randomized). Figure S3 shows the ranks of neighbors for each seed and metric that were used to compute scaling properties of the information-theoretic measures. We looked at predictability of seed activity given the activity of neighbors, subsystems and environments of different sizes.

We first characterized the distance in physical space between seeds and neighbors ranked according to different metrics (Euclidean, Connectome and Randomized). In **Figure 2A**, the Y-axis depicts the Euclidean distance (mm) between seed ROIs and the  $k$ th-neighbor (X-axis) chosen according to the three metrics, averaged across all seed ROIs in both hemispheres (shaded areas reflect 1st and 3rd quartiles). The physical distance to nearby Connectome neighbors tends to be small, though highly variable across seeds and not as small as to Euclidean neighbors, which are by definition maximally proximate in physical space. Randomized neighbors display no spatial regularity, with average distance to neighbor of any rank corresponding to the expected Euclidean distance separating randomly chosen pairs of ROIs ( $\sim 65$  mm).

We used Pairwise MI to measure functional connectivity between pairs of ROIs as a function of their separation according to both Euclidean and Connectome distance. Euclidean and Connectome distances were divided into 50 equal-width bins and mean Pairwise MI between intra-hemispheric pairs of ROIs corresponding to each Connectome and Euclidean bin was computed. **Figure 2B** shows a heat map of mean Pairwise MI values within each bin (log color scaling used to better highlight differences among weakly coupled connections). The bar chart at the top of the heat map shows mean MI values for pairs of ROIs separated by different Euclidean distances (irrespective of Connectome distances), while the bar chart at the right of the heat map shows mean MI values for pairs of ROIs separated by different Connectome distances (irrespective of Euclidean distances). Overall, mean Pairwise MI tends to decrease monotonically with increasing Euclidean distance as well as with increasing Connectome distance. Pairs of ROIs that are distant according to both metrics tend to be weakly coupled (mean MI below 0.01 bits). The most strongly coupled pairs of ROIs (mean MI



above 0.2 bits) are those separated by small Euclidean distances, irrespectively of Connectome distance. However, ROIs that are distant in Euclidean space but proximate on the Connectome also tend to have higher coupling (mean MI  $\sim 0.03$  bits) than those that are distant in both metrics.

We next report the scaling of ROI-based predictability measures defined in section Information-Theoretic Measures and Efficiency, namely Seed-Neighbor MI, Seed-Subsystem MI and Seed-Environment MI.

**Figure 3A** shows Seed-Neighbor MI between seeds and their neighbors chosen according to the three distance metrics, averaged over all ROIs in both hemispheres as seeds. ROIs that are closer in Euclidean and Connectome space have a higher MI, with closely ranked Euclidean neighbors (up to neighbor  $\sim 8$ ) showing a higher coupling than Connectome neighbors (this reproduces the effect seen in **Figure 2B**, where proximate Euclidean and Connectome pairs tend to have higher Pairwise MI). As expected, Pairwise MI with Randomized neighbors displays no systematic regularity with neighbor rank. Mean Seed-Neighbor MI for Euclidean neighbors becomes most similar to the mean Seed-Neighbor MI for Randomized neighbors at approximately the 50th neighbor (for Euclidean neighbors, this corresponds to a distance of approximately 40 mm). This is the Euclidean scale at which functional correlations between pairs of physically proximate ROIs decay to baseline levels.

**Figure 3B** shows the scaling of Seed-Subsystem MI with increasing subsystems averaged over all ROIs in both hemispheres as seeds. The illustration in the top left corner shows in schematic form how this measure is computed (dark brown is the seed, light brown is subsystem, and green arrow is MI). Seed-Subsystem MI grows monotonically with increasingly large subsystems as more subsystem ROIs become available to predict the activity of the seed. On average, seeds have the strongest coupling to

Euclidean subsystems, closely followed by Connectome subsystems. However, across the full range of subsystem sizes, there is great overlap in the distribution of Seed-Subsystem MI values for subsystems defined according to these two metrics. In contrast, Randomized subsystems display much less Seed-Subsystem MI over the entire range of subsystem sizes. The three measures converge once subsystems begin to overlap and grow toward including the entire hemisphere.

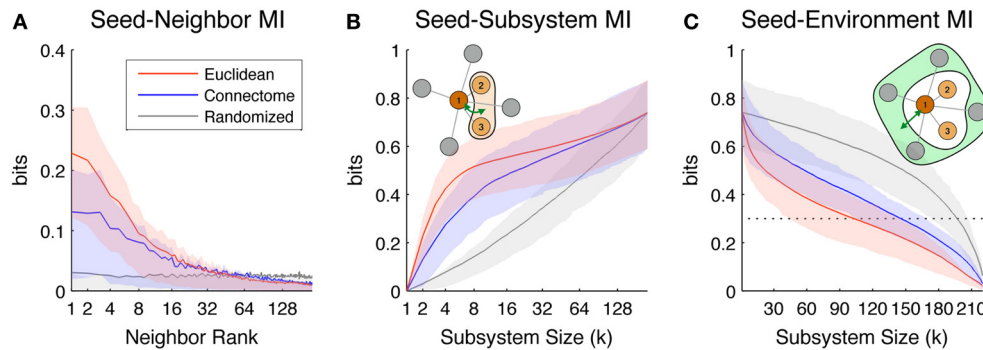
**Figure 3C** shows scaling of Seed-Environment MI, the multivariate coupling between the seed and the environment. The illustration in the top right corner shows in schematic form how this measure is computed (dark brown is the seed, light green is environment, and green arrow is MI). Note that since the environment is defined as the set of hemispheric ROIs outside of the subsystem, environment size decreases with increasing subsystem size. For this reason, Seed-Environment MI always decreases monotonically with increasing subsystem size, as less and less environmental ROIs are available for predicting the seed. On average, Euclidean environments tend to have less predictability about seeds than Connectome environments, indicating that sets of ROIs that are distant in space tend to be less functionally coupled to seeds than sets of ROIs distant on the Connectome. However, there is again a large overlap between Seed-Environment MI values over the range of environment sizes. Randomized environments tend to have the highest values of Seed-Environment MI. This is due to the fact that Randomized environments include more spatially- and structurally-proximate ROIs to the seeds (which tend to be highly functionally coupled; **Figure 2A**) than Euclidean and Connectome environments that by definition do not include ROIs that are, respectively, proximate in space or on the Connectome.

Next, we looked at how predictability of different ROIs varies across the cortical surface using two measures defined in section

Information-Theoretic Measures and Efficiency: Total MI and Euclidean Coupling Range.

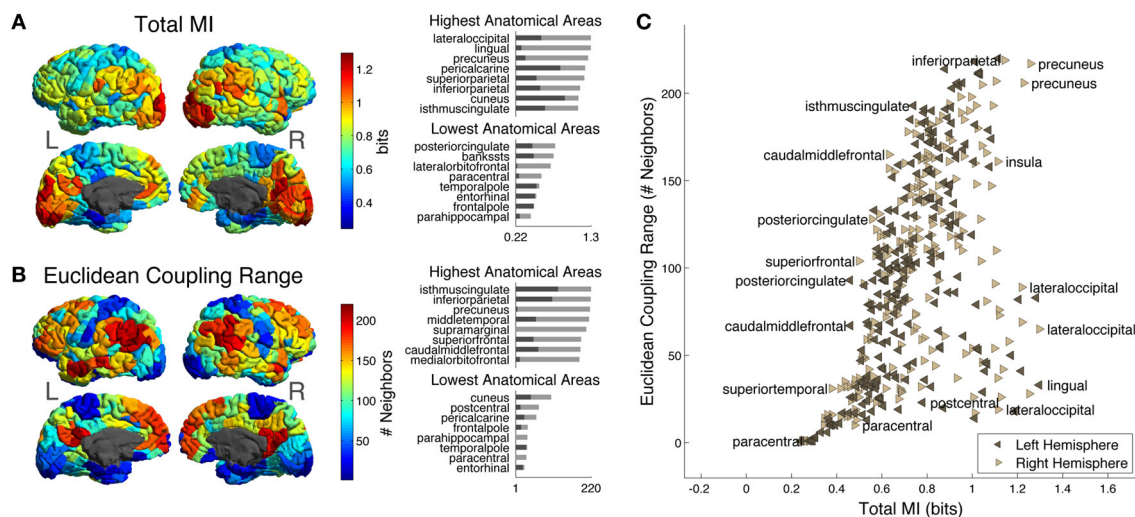
**Figure 4A** shows Total MI, or total amount of predictability available about the activity of each ROI given the rest of ROIs in the hemisphere (note that this measure does not depend on choice of distance metric). In addition, we show the distribution of Total MI in different anatomical areas. As described in

section MRI Data, ROIs in each hemisphere are grouped into 34 larger-scale “anatomical areas” that correspond to the FreeSurfer parcellation (Desikan et al., 2006). The bar chart on the upper right of the cortical Total MI plot shows the top 8 anatomical areas arranged according to maximum Total MI of ROIs within each area (maximum Total MI of ROIs in each area indicated by light gray bars; the minimum indicated by dark gray bars). The



**FIGURE 3 | Scaling of the information-theoretic measures of seed predictability.** Colored lines indicate mean values across all seed ROIs in both hemispheres, while shaded areas indicate values within 1st and 3rd quartile. Colors indicate values for neighbors/subsystems/environments chosen according to Euclidean (red), Connectome (blue), and Randomized (gray) distance metrics. **(A)** Average Seed-Neighbor MI between seeds and their corresponding  $k$ th rank neighbors chosen according to the three distance metrics. **(B)** Seed-Subsystem MI between seeds and subsystems

built according to the three distance metrics. The illustration in the top left corner diagrams how this measure is computed for a given seed and subsystem of size 3. **(C)** Seed-Environment MI between seeds and environments built according to the three distance metrics. The illustration in the top right corner diagrams how this measure is computed for a given seed and subsystem size 3 (environment size 4). The horizontal dotted line indicates 0.3 bits of Seed-Environment MI, a threshold used later in our definition of Euclidean Coupling Range.



**FIGURE 4 | (A)** Cortical distribution of Total MI, the total amount of predictability available about each ROI from the ROIs in the rest of the hemisphere. On the upper right are the top 8 anatomical areas arranged according to maximum Total MI of ROIs with each area (maximum Total MI of intra-area ROIs indicated by light gray bars; the minimum indicated by dark gray bars) while on the lower right are the bottom 8 anatomical areas arranged according to maximum Total MI of ROIs within each area (maximum Total MI of intra-area ROIs indicated by light gray bars; the minimum indicated by dark gray bars). **(B)** Cortical distribution of Euclidean Coupling Range, the neighbor number at which Euclidean

Seed-Environment MI drops below a threshold of 0.3 bits (see text for details). On the upper right are the top 8 anatomical areas arranged according to maximum Euclidean Coupling Range of ROIs with each area [light and dark gray bars indicating maximum and minimum values as in **(A)**] while on the lower right are the bottom 8 anatomical areas arranged according to maximum Euclidean Coupling Range of ROIs with each area [light and dark gray bars indicating maximum and minimum values as in **(A)**]. **(C)** Scatter plot of Total MI vs. Euclidean Coupling Range for left and right hemisphere ROIs. A few ROIs are labeled with the names of their corresponding anatomical areas.



areas with the 8 largest maximum Total MIs are lateral occipital, lingual, precuneus, pericalcarine, superior parietal, inferior parietal, cuneus, and isthmus cingulate. The bar chart on the lower right of the cortical Total MI plot shows the lowest 8 anatomical areas arranged according to maximum Total MI of ROIs within each area (maximum Total MI of ROIs within each area indicated by light gray bars; the minimum indicated by dark gray bars). The areas with the 8 lowest maximum Total MIs are posterior cingulate, banks of the superior temporal sulcus (bankssts), lateral orbitofrontal, paracentral, temporal pole, entorhinal, frontal pole, and parahippocampal.

We next investigated how Euclidean Coupling Range is distributed across the cortical surface, as well as its correlation with a structural measure.

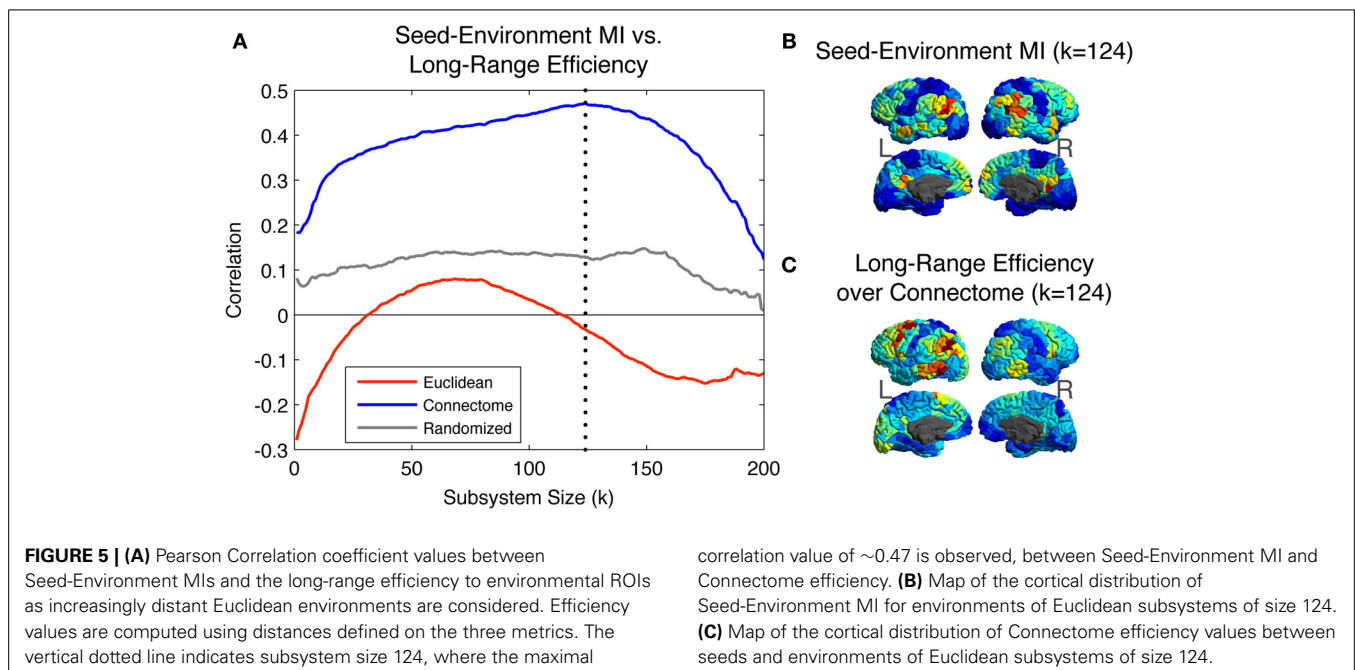
**Figure 4B** shows the Euclidean Coupling Range of each ROI on the cortical surface. Euclidean Coupling Range, defined as the Euclidean neighbor number at which Seed-Environment MI drops below a given threshold, quantifies the maximal spatial scale at which a given amount of functional coupling with the environment is maintained. We used a threshold amount of 0.3 bits, which is the average Seed-Environment MI when half-hemisphere-sized subsystems/environments ( $\sim 110$  ROIs) are considered (see horizontal dotted line in **Figure 3C**). On the upper right are shown the top 8 anatomical areas arranged according to maximum Euclidean Coupling Range of ROIs within each area (light and dark gray bars indicating maximum and minimum values as in **Figure 4A**). The areas containing the 8 highest maximum Euclidean Coupling Range are isthmus cingulate, inferior parietal, precuneus, middle temporal, supramarginal, superior frontal, caudal middle frontal and medial orbitofrontal. On the lower right are shown the bottom 8 anatomical areas arranged according to maximum Euclidean Coupling Range of ROIs within each area (light and dark gray bars indicating maximum and

minimum values as in **Figure 4A**). The areas containing the 8 lowest maximum Euclidean Coupling Range are cuneus, postcentral, pericalcarine, frontal pole, parahippocampal, temporal pole, paracentral and entorhinal.

In **Figure 4C**, we contrast these two measures using a scatter plot of Total MI (X-axis) vs. Euclidean Coupling Range (Y-axis) values for all left- and right-hemisphere ROIs. Several ROIs are labeled with the names of corresponding anatomical areas in order to indicate which areas tend to have high and low values of these two measures.

We then looked at the relationship between Seed-Environment MI, a measure of functional coupling, and long-range efficiency, a measure of structural connectivity, in order to assess whether structural features may be driving long-range functional coupling. Long-range efficiency (see Section Information-Theoretic Measures and Efficiency) is defined as the mean inverse shortest-path lengths between each seed ROI and the set of ROIs in its Euclidean environment (that is, its long-Euclidean-range neighbors). Seed ROIs with greater efficiency values are more proximate, according to some metric, to their long-range Euclidean neighbors than those with lower efficiency ones. To compare the accessibility of long-Euclidean-range neighbors over Connectome space vs. Euclidean and Randomized space, we computed different efficiency values corresponding to shortest-path lengths to those neighbors on the three different distance metrics.

**Figure 5A** shows the Pearson correlation values between the Seed-Environment MI and the three long-range efficiency measures as increasingly long Euclidean distances are considered (with increasing subsystem size on the X-axis, environments become increasingly small and distant). Correlations are computed separately across all seed ROIs within each hemisphere and then averaged between hemispheres. Correlations are highest between Seed-Environment MI and long-range efficiency values



over the Connectome metric. They reach a peak correlation value of  $\sim 0.47$  at  $k = 124$  (vertical dotted line), corresponding to environments composed of ROIs located further than  $\sim 65$  mm from the seed. Such a strong correlation was not observed for efficiency values computed using either of the other two metrics at any scale.

In **Figure 5B**, we plot for different seed ROIs the Seed-Environment MI of Euclidean environments corresponding to subsystems of size 124 (when the Connectome structural vs. functional correlation is maximal; vertical dotted line in **Figure 5A**). In **Figure 5C**, we plot the cortical distribution of the corresponding Connectome efficiency values between seed ROIs and the Euclidean environments. It can be seen that these two measures display a highly similar spatial distribution, indicating that at this scale ROIs with the highest functional coupling to long-Euclidean-range ROIs also tend to be the most efficiently connected to them over the Connectome.

So far we have looked at the predictability of individual ROIs considered as seeds. We now look at two (normalized) multivariate measures of the predictability of joint activity of entire subsystems: Subsystem Integration per ROI and Subsystem-Environment MI per ROI.

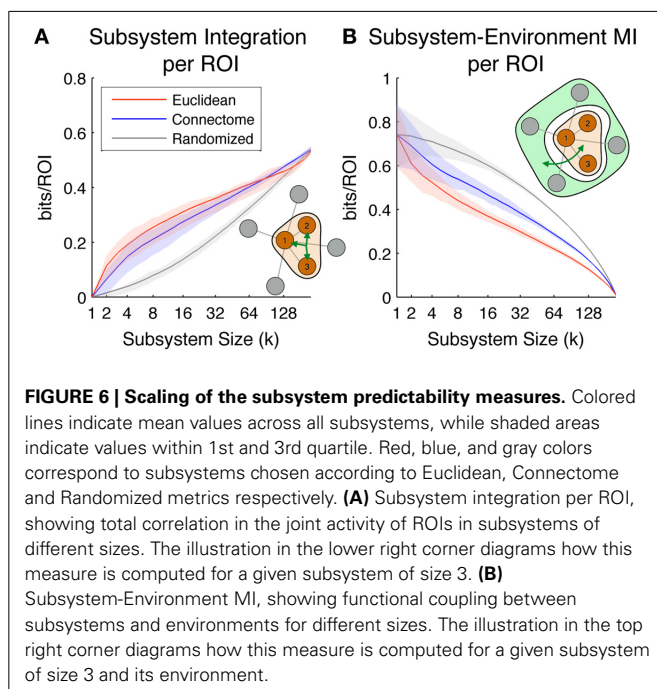
**Figure 6A** shows the Subsystem Integration per ROI, which quantifies the amount of total correlation of subsystem activity (divided by subsystem size for normalization purposes). The diagram in the lower right of the figure shows in schematic form how this measure is computed (brown is the subsystem, and the three-pointed green arrow is total correlation). On average, the most integrated subsystems up to size  $\sim 90$  ROIs are those defined according to the Euclidean metric (size-90 Euclidean subsystems have a radius of  $\sim 55$  mm), while subsystems defined according to the Connectome are on average the most integrated for larger subsystem sizes. As expected, subsystems selected according to the Randomized

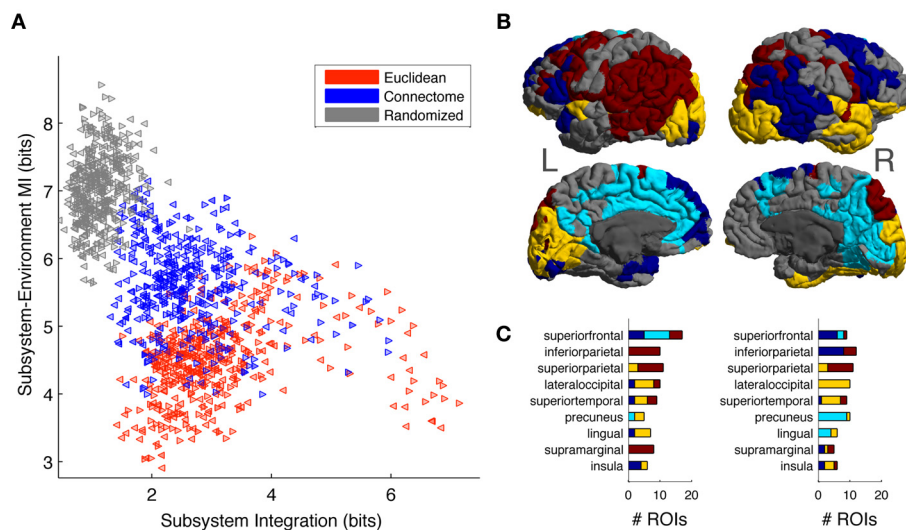
metric, which are neither spatially co-located nor densely structurally interconnected, display a much lower level of multivariate integration.

**Figure 6B** shows Subsystem-Environment MI per ROI, a measure of the mutual information between subsystems and their environments (divided by subsystem size for normalization purposes). The diagram in the upper right of the figure shows in schematic form how this measure is computed (brown is subsystem, light green is environment, and the green arrow is MI). On average this measure is lowest for Euclidean subsystems, indicating that these are more functionally segregated from the rest of the hemisphere than subsystems defined according to the other metrics. Interestingly, Connectome subsystems are nearly as segregated as Euclidean ones at small scales (up to subsystem size  $\sim 10$ ), but at larger scales they are more functionally coupled to the rest of the hemisphere. Randomized subsystems have the highest Subsystem-Environment MI for all the scales, since they are composed of groups of ROIs scattered through the brain and their boundaries are spanned by many pairs of ROIs separated by short Euclidean and Connectome distances (which tend to have high functional connectivity).

Overall, **Figure 6** shows that Connectome subsystems exhibit both high Subsystem-Environment MI and high Subsystem Integration. We explored this finding in more depth in the following figure. First, we selected all Connectome subsystems of size of 11, corresponding to a volume of approximately 5% of each hemisphere (as seen in **Figure 6A**, at this size Connectome subsystems are on average nearly as integrated as Euclidean subsystems but, as **Figure 6B** shows, contain much information about their environments). **Figure 7A** shows the scatter plot of Subsystem Integration (X-Axis) vs. Subsystem-Environment MI (Y-Axis) for size-11 subsystems defined according to Euclidean, Connectome and Randomized metrics. Randomized Subsystems (gray) tend to cluster in regions of the scatter plot characterized by high Subsystem-Environment MI (lack of segregation from environment) and low Subsystem Integration (lack of internal integration). Euclidean Subsystems (red) tend to occupy regions of the scatter plot characterized by low Subsystem-Environment MI (high segregation from environment) and high Subsystem Integration (high internal integration). Connectome Subsystems (blue), however, occupy intermediate regions of the scatter plot, demonstrating significant amounts of both Subsystem-Environment MI (thus not being functionally segregated from the rest of the hemisphere) while also having significant Subsystem Integration (thus also having internal integration).

We investigated which specific Connectome subsystems maximize both Subsystem Integration and Subsystem-Environment MI. First, Connectome subsystems that were in the upper 50 percentile of both measures in each hemisphere were selected. Next, because these subsystems overlapped (contained some of the same ROIs; see **Figure S4A**), we clustered them into a smaller number of minimally-overlapping “subsystem communities.” To do so, for each hemisphere we computed a subsystem-by-subsystem Overlap Matrix whose entries measured the proportion of ROIs shared between each pair of subsystems





**FIGURE 7 | (A)** Scatter plot of Subsystem Integration vs. Subsystem-Environment MI for subsystems of size 11, with red, blue and gray colors correspond to subsystems chosen according to Euclidean, Connectome and Randomized metrics respectively. Left-hemisphere subsystems are indicated with left-pointing triangles and right-hemisphere subsystems are indicated with right-pointing triangles. **(B)** Connectome subsystems in the upper 50 percentile of both Subsystem Integration and Subsystem-Environment MI were chosen and allowed us to

identify four minimally overlapping “subsystem communities” in the left and right hemispheres. ROIs are colored according to community membership (color arbitrary); gray ROIs are those that did not belong to any high-Subsystem-Integration, high-Subsystem-Environment MI subsystem. **(C)** The distribution of subsystem communities across anatomical areas. Bar chart shows the number of ROIs from each community that are contained in different anatomical areas for the top 9 represented anatomical areas. Bar chart colors correspond to the colors used on the cortical map.

(see Figure S4B, S4C for the left and right hemisphere Overlap Matrices). A community-detection algorithm (Blondel et al., 2008) was run on this matrix to provide a partition of the subsystems into communities.

The community-detection algorithm identified four communities in the left hemisphere and another four in the right hemisphere. **Figure 7B** shows cortical surface of the left and right hemisphere, with each ROI colored according to its membership in a subsystem community (colors arbitrary but selected so that communities that have similar spatial distributions in both hemispheres have the same color). ROIs that belong to more than one selected subsystem were assigned to their most frequent community. Gray colored ROIs are those that were not part of any subsystem that was in the top 50 percentiles according to the two MI measurements.

Finally, we looked at how the subsystem communities obtained were distributed across anatomical areas. Anatomical areas were ranked in terms of their participation in the subsystems maximizing Subsystem Integration and Subsystem-Environment MI. **Figure 7C** lists the top 9 anatomical areas: superior frontal, inferior parietal, superior parietal, lateral occipital, superior temporal, precuneus, lingual, supramarginal, and insula. The stacked bar charts indicate, for both hemispheres, the number of ROIs from each subsystem community that are contained in each anatomical area, with bar chart colors corresponding to the colors used on the cortical map. We discuss the distribution of these subsystem communities across anatomical areas in more detail in the section Subsystem Predictability and Integration vs. Segregation Trade-Off.

## DISCUSSION

In this work, we characterized brain regions and networks in terms of their information-theoretic measures by using both functional and structural information in a complementary manner. The measures presented here quantify the amount of functional coupling between sets of ROIs as well as integration within sets of ROIs. Sets of ROIs form subsystems which are selected according to three different possible distance metrics: Euclidean (reflecting the physical spatial embedding of brain regions), Connectome (reflecting the anatomical structural connectivity of the brain), and Randomized (a comparison condition based on a rewired version of the Connectome graph; see Section Distance Metrics). We also investigated the scaling of these measures, in the sense of their growth as larger subsystems are considered.

In section Information-Theoretic Measures for Studying the Organization of the Brain, we discuss the use of information-theoretic measures for characterizing the brain and the need for such measures to account for the brain’s spatial and topological embedding. In section Scaling of Information-Theoretic Measures, we discuss the scaling of our measures as brain subsystems of different sizes are considered, their distribution across the cortical surface, and their relation to long-range efficiency. In section Subsystem Predictability and Integration vs. Segregation Trade-Off, we discuss the fact that Connectome subsystems tend to be highly internally integrated while also being coupled to the rest of the brain, and that the subsystems that optimize this trade-off cluster into communities that resemble previously identified resting state networks. In section Methodological Considerations we discuss some important methodological considerations and

assumptions involved in this work. In the last section Future Directions, we suggest some possible directions for further work.

### INFORMATION-THEORETIC MEASURES FOR STUDYING THE ORGANIZATION OF THE BRAIN

As mentioned in the Introduction, much recent research has been devoted to characterizing the structural and functional roles of different brain regions and networks. Many of these characterizations have identified certain regions as structural and functional hubs, decomposed the brain into weakly-coupled modules and networks, and investigated the role of large-scale integrative backbones.

Information theory provides a natural language for talking about systemic aspects of the organization of functional brain activity, including the presence of quasi-independent modular subsystems and the integrative properties of functional hubs and networks. Several information-theoretic measures for studying brain organization have been proposed in the literature. One measure of particular interest is TSE complexity (Tononi et al., 1994), which is based on the idea that low-level processing is performed in localized, segregated brain regions that operate in parallel and interconnect along hierarchical lines, while high-level association and integration is performed in large-scale distributed networks (Felleman and Essen, 1991; Yeo et al., 2011). TSE complexity quantifies this notion by looking at the scaling of total correlation as increasingly large subsystems are considered. The degree to which the total correlation of large-scale regions (having many components) exceeds that of small-scale regions (containing few components) is a quantitative signature of integration at large scales.

Importantly, the activity of the brain unfolds across physical space and structural connectivity networks. For this reason, it can be expected to qualitatively follow Tobler's *first law of geography*: "Everything is related to everything else, but near things are more related than distant things" (Tobler, 1970). In fact, as previously reported (Salvador et al., 2005; Honey et al., 2009; Power et al., 2013) and as also shown here, functional interactions are stronger between spatially proximate regions. TSE complexity, however, considers integration at a given scale by looking at all possible subsets of component of a given size. Thus, while it represents a promising step toward an information-theoretic treatment of large-scale integration, it disregards spatial and connectivity information and the fact that the organization of functional activity is often dominated by physically localized interactions.

In this work, we looked at the scaling of information-theoretic measures across both physical space and the Connectome, and compared it to scaling over the Randomized metric (which, like TSE, disregards actual spatial and topological organization). As we will discuss, we found that the Randomized metric poorly represents the functional organization of our brain data, and this weakness may also be expected of TSE. In fact, underlying spatial and connectome structure must be taken into account in order to properly quantify the amount of large-scale integration in the brain. In addition, our methodology, which captures systematic relationships between the size and the strength of functional constraints in spatially-compact subsystems, allows us to compute localized information-theoretic measures of scaling. This

allows for the characterization of the variation of integration and predictability across the cortical surface.

### SCALING OF INFORMATION-THEORETIC MEASURES

We measured the amount of functional coupling between each ROI (the "seed") and the set of most proximate neighbors (the local "subsystem") as well as the set of most distant neighbors (the distant "environment"). We also computed how the strength of functional coupling scales as increasing numbers of neighbors are chosen according to one of the three different distance metrics—Euclidean, Connectome, or Randomized.

As discussed in section Information-Theoretic Measures for Studying the Organization of the Brain, functional activity organized according to an underlying metric will display stronger functional coupling between nearby locations vs. more distant ones. Thus, information-theoretic measures computed on sets of ROIs chosen according to a more "representative" space are expected to give rise to higher values of Seed-Subsystem MI and Seed-Neighbor MI for close neighbors (i.e., more integration within local regions) as well as lower values of Seed-Environment MI (i.e., more segregation between local regions and the rest of the system).

According to these criteria, both Euclidean and Connectome metrics better represent the functional organization of resting state activity than the Randomized metric (Figure 3). On average, for small scales, the Euclidean metric captures more strong functional couplings than does the Connectome metric, as shown by higher values of Seed-Neighbor MI (Figure 3A) and Seed-Subsystem MI (Figure 3B) measures for Euclidean vs. Connectome subsystems. Generally, for the range of scales considered, ROIs chosen according to the Connectome metric display an amount of functional coupling with neighbors between those chosen according to the Euclidean metric on one hand and Randomized on the other. We discuss some possible reasons for the intermediate role played by Connectome subsystems below.

The strength of functional coupling between seeds and Euclidean subsystems (Figure 3B), as well as the fact that the environments chosen in terms of distant Euclidean neighbors display the least functional coupling (Figure 3C), demonstrates that resting-state brain data is highly spatial, in that it exhibits strong correlations over small Euclidean scales (some reasons for this are discussed below in section Methodological Considerations). However, while short-Euclidean range interactions are strong, the brain also integrates information globally and exhibits functional coupling over large spatial scales. Because Seed-Environment MI quantifies the functional coupling between seed ROIs and remote locations, we defined Euclidean Coupling Range as the number of Euclidean neighbors at which the Seed-Environment MI drops below a threshold of 0.3 bits, and looked at the distribution of this measure across the cortical surface (Figure 4).

This measure was found to have a highly heterogeneous distribution across the brain. Low values of Euclidean Coupling Range—indicating that only short-scale correlations present—are found in unimodal sensorimotor cortices, including locations corresponding to V1, motor areas in the precentral gyrus, somatosensory areas in postcentral gyrus, paracentral areas corresponding to the supplementary motor area, and



superior temporal areas corresponding to auditory cortex. On the other hand, locations in the brain having high Euclidean Coupling Range—indicating the presence of long-range functional couplings—include recognized high-level hub areas (van den Heuvel and Sporns, 2013), such as the precuneus, inferior parietal, superior frontal gyrus, anterior cingulate, temporoparietal junction and ventral frontal cortex. In addition, regions thought to have functional roles at intermediate levels of the cortical hierarchy, such as higher-order visual and auditory cortices as well as somatosensory association cortices, tend to display intermediate values of Euclidean Coupling Range.

Importantly, variation in Euclidean Coupling Range arises due to variation in the range of spatial coupling of different ROIs and is not simply due to differences in their inherent level of predictability. We compared Euclidean Coupling Range with Total MI, a measure of mutual information between each ROI and the rest of the ROIs in its hemisphere. Total MI does not rely on any underlying metric and quantifies the inherent predictability of different regions. This measure also displayed a heterogeneous distribution across the brain, indicating that during resting-state some ROIs are much more predictable than others. Regions with the highest predictability included large areas of the occipital lobe, primarily corresponding to the primary and higher-order visual cortices, as well as some regions of the parietal lobe such as the inferior parietal lobule. Notably, many regions high in Euclidean Coupling Range—such as those in the frontal lobe—did not have exceptionally high Total MI, nor did many regions with high Total MI—such as visual cortices—have high Euclidean Coupling Range (**Figure 4C**). Thus, Euclidean Coupling Range is a continuous measure that separates regions having spatial segregation (low values) from those having spatial integration (high values) and identifies functional hubs at multiple scales of the cortical hierarchy. Our results are in agreement with previous research showing a connection between functional hubs and long-spatial-range functional coupling (Sepulcre et al., 2010).

We also evaluated whether functional coupling between spatially distant regions may be driven by long-range efficiency. Hence we correlated Seed-Environment MI and long-range efficiency (over the three metrics) between ROIs and their Euclidean environments for a wide range of scales. For most scales, long-range efficiency over the Connectome was positively correlated with the Seed-Environment MI, while correlations were much smaller with long-range efficiency over Euclidean and Randomized metrics. Thus, the presence of Connectome shortest-paths between the seed and spatially distant ROIs was the best predictors of strong functional coupling between them, reflecting a possible fingerprint of structural connections in driving functional coupling over large spatial scales.

### SUBSYSTEM PREDICTABILITY AND INTEGRATION vs. SEGREGATION TRADE-OFF

Our information-theoretic approach measured not only the predictability of seed ROIs, but also the multivariate predictability of sets of ROIs in subsystems (**Figure 6**). We investigated two complementary measures: Subsystem-Environment MI and Subsystem Integration. On average Connectome subsystems displayed nearly as much Subsystem Integration as Euclidean

subsystems up to subsystem size 90, and more integration for larger sizes. Across many scales, Euclidean subsystems located in the occipital lobe (corresponding to the visual cortices) displayed the highest amounts of integration (these subsystems, for example, are the cluster of points with very high integration shown in the scatter plot of **Figure 7A**). On the other hand, in comparison to Euclidean subsystems, Connectome subsystems had a higher Subsystem-Environment MI, indicating that they were less functionally segregated from the rest of the hemisphere. Randomized subsystems were much less internally integrated and much less segregated from their environments than either Euclidean or Connectome subsystems.

At first glance, subsystems with high functional integration are also expected to display high functional segregation. The fact that Connectome subsystems have relatively high values of both Subsystem Integration and Subsystem-Environment MI suggests that they may balance a trade-off between two important information-processing functions: accessing information from large areas of the brain and integrating it efficiently across a network of hub regions (Zamora-López et al., 2010). We investigated this question by looking at particular values of Subsystem Integration and Subsystem-Environment MI for subsystems of size of 11 (~5% of one hemisphere) (**Figure 7A**). We chose Connectome subsystems with high values on both Subsystem Integration and Subsystem-Environment MI and found that they are distributed into four minimally-overlapping subsystem communities (**Figure 7B**). Interestingly, these communities can be interpreted in terms of neural anatomy as well as in terms of previous work on functional resting state networks. The yellow communities in the left and right hemispheres occupy areas corresponding to primary and secondary visual and auditory cortices, the light blue communities roughly correspond to locations in the default mode network, while the dark red and blue communities contain regions reported to be part of the ventral attention, dorsal attention, and fronto-parietal control resting state networks (Yeo et al., 2011). The anatomical regions (**Figure 7C**) most represented in the light blue, dark blue and dark red communities are known to include many functional hub regions, such as superior frontal gyrus, inferior and superior parietal lobules, supra-marginal gyrus and insula. Interestingly, in both hemispheres, the superior frontal gyrus included ROIs corresponding to all three of these communities, suggesting that it may be a location where these separate high-level integrative networks intersect.

Overall, this shows that our multivariate information-theoretic measures provide useful characterization of integration and coupling in subsystems. Furthermore, we found that they identify regions that display large values of integration and coupling, some of which are similar to previously reported resting-state networks.

### METHODOLOGICAL CONSIDERATIONS

The Randomized metric was used as a control for comparison and was not expected to correspond closely to the functional organization. On the other hand, the fact that nearby Connectome neighbors exhibited increased functional coupling (**Figure 2B**) suggests that connections captured in the DSI data do correspond to actual anatomical connections that drive neural interactions and

produce correlations in the multivariate BOLD signal. However, proximity in physical space, as captured by the Euclidean metric, corresponded to even higher correlations. The strong correlations between physical neighbors is driven in part by the overlap between structural and Euclidean neighbors, in that anatomical connections are enriched in spatially proximate regions (Honey et al., 2009). However, other causes may also be responsible, including undetected connections (such as local cortico-cortical connectivity and subcortical-mediated circuitry) as well as spatial smoothing due to BOLD-signal blurring due to vasculature effects, head motion artifacts, and MRI preprocessing (Honey et al., 2009; Power et al., 2012, 2013).

The framework proposed here looks at the scaling of information-theoretic measures. It is not tied to any particular way of estimating information-theoretic measures from empirical data and can be applied both to continuous and discrete data. However, as discussed in section Information-Theoretic Measures and Efficiency, for practical purposes in this work we assumed that the activity of ROIs was distributed as a multivariate Gaussian. Due to the Gaussian assumption, the covariance matrix of each hemisphere's multivariate fMRI time series served as a sufficient statistic for all of our measures of predictability and integration. In addition to the Gaussian assumption, we also combined the time series from all 37 subjects into a single "pooled" subject possessing ~10,000 time points (see Section MRI Data). Because of the subject pooling, enough time points were acquired to get a reasonable estimate of the entries in this covariance matrix (defined by nearly ~20,000 parameters). We thus could estimate information-theoretic measures for high dimensional spaces, such as for the entropies of the joint activity of the ~220 ROIs present in each hemisphere.

Computing predictability using the Gaussian assumption is equivalent to predicting the activity of seed ROIs and subsystems by linear regression. The drawback of using the covariance matrix for estimating information-theoretic quantities is that it disregards non-linear interactions between ROI activities, as well as interactions of higher-order than pairwise. Though it has been suggested that bivariate fMRI time series are sufficiently Gaussian to not warrant the estimation of non-linear effects in functional connectivity (Hlinka et al., 2011), there are a number of estimators that could be used that do take into account such effects, such as for example nearest-neighbor estimators (Kozachenko and Leonenko, 1987; Singh et al., 2003; Kraskov et al., 2004; Lizier et al., 2011). However, these estimators require of a large number of samples for reliable estimates and in our case gave unstable entropy estimates (data not shown). Overall, questions about the importance of non-linear and higher-order interactions in describing the functional organization of the brain present great interest for future investigation using our framework.

For similar reasons, it was not feasible to accurately estimate our multivariate information-theoretic measures using individual subjects' time series, which included only 276 samples per ROI per subject. Our method of subject pooling, which was performed for reasons of statistical estimation, is defensible because resting state functional activity is known to be fairly similar across healthy subjects (Damoiseaux et al., 2006). In addition,

structural connectivity is also similar enough across healthy subjects so that connectome pooling can be used to reduce the effect of DSI-tractography false negatives (i.e., undetected fibers) (Hagmann et al., 2008; de Reus and van den Heuvel, 2013). However, this approach prohibits us from investigating questions of inter-subject variation in information-theoretic measures as well as their relation to individual-subject structural measures. Questions of inter-subject variability of information-theoretic measures also present great interest for future investigation, which may become feasible given the availability of datasets containing longer fMRI time series.

## FUTURE DIRECTIONS

As mentioned, with longer recordings it may be possible to investigate the role of non-linear coupling and higher-order interactions in the functional organization of the human brain, as well as the inter-subject variability of information-theoretic measures. In addition, it may be possible to apply these measures in a time-dependent manner in order to look for evidence of dynamic re-organization of the integrative properties of different regions. Another interesting avenue of development would be to apply our methodology to task-dependent datasets in order to test differences in information-theoretic measures exhibited under different cognitive loads and tasks. Finally, recent work on using entropy measures for diagnostic purposes (Mäki-Marttunen et al., 2013) suggests that the kinds of measures developed here may hold promise as possible sources of diagnostic markers.

Generally, the idea of using information-theory to study the functional organization of the brain draws connections to fields of statistical learning, coding theory, statistical physics, complex systems and other fields that are playing a central part in modern computational and systems neuroscience. It may also be relevant to recent ideas regarding the criticality of brain functional activity. Criticality is a concept closely tied to long-range scaling of correlations, and it has been shown in models that information-theoretic measures of integration (Erb and Ay, 2004; Feldman et al., 2008; DeDeo and Krakauer, 2012) are maximized at critical parameter values. As we have argued, however, properly measuring the scaling of integration should take into account underlying topologies on which system constraints are organized. This suggests that our approach may be useful for investigating the hypothesis that brain is poised at or nearby a critical state (Haimovici et al., 2013; Marinazzo et al., 2013).

## FUNDING

Artemy Kolchinsky received summer research support from the Indiana University School of Informatics through the NSF IGERT program in Brain-Body-Environment Systems. Martijn P. van den Heuvel was supported by Netherlands Organization for Scientific Research Grant VENI-451-12-001 and a fellowship of the Brain Center Rudolf Magnus. Alessandra Griffo was supported by the Swiss National Science Foundation (Schweizerische Nationalfonds Grant 320030-130090). Luis M. Rocha was supported by Intelligence Advanced Research Projects Activity (Open Source Indicators) and Indiana University Collaborative Research

Grant. Olaf Sporns and Joaquín Goñi were supported by the J. S. McDonnell Foundation.

## ACKNOWLEDGMENT

We thank Richard F. Betzel and Andrea Avena-Koenigsberger for useful comments and discussions.

## SUPPLEMENTARY MATERIAL

The Supplementary Material for this article can be found online at: <http://www.frontiersin.org/journal/10.3389/fninf.2014.00066/abstract>

## REFERENCES

- Abdelnour, F., Voss, H. U., and Raj, A. (2014). Network diffusion accurately models the relationship between structural and functional brain connectivity networks. *Neuroimage* 90, 335–347. doi: 10.1016/j.neuroimage.2013.12.039
- Achard, S., Salvador, R., Whitcher, B., Suckling, J., and Bullmore, E. (2006). A resilient, low-frequency, small-world human brain functional network with highly connected association cortical hubs. *J. Neurosci.* 26, 63–72. doi: 10.1523/JNEUROSCI.3874-05.2006
- Ahmed, N. A., and Gokhale, D. V. (1989). Entropy expressions and their estimators for multivariate distributions. *Inf. Theory IEEE Trans.* 35, 688–692. doi: 10.1109/18.30996
- Ay, N., Olbrich, E., Bertschinger, N., and Jost, J. (2006). “A unifying framework for complexity measures of finite systems,” in *Proceedings of ECCS06, European Complex Systems Society* (Oxford, UK).
- Beckmann, C. F., DeLuca, M., Devlin, J. T., and Smith, S. M. (2005). Investigations into resting-state connectivity using independent component analysis. *Philos. Trans. R. Soc. B Biol. Sci.* 360, 1001–1013. doi: 10.1098/rstb.2005.1634
- Betz, R. F., Griffa, A., Avena-Koenigsberger, A., Goñi, J., Thiran, J.-P., Hagmann, P., et al. (2014). Multi-scale community organization of the human structural connectome and its relationship with resting-state functional connectivity. *Netw. Sci.* 1, 353–373. doi: 10.1017/nws.2013.19
- Bialek, W., Nemenman, I., and Tishby, N. (2001). Predictability, complexity, and learning. *Neural Comput.* 13, 2409–2463. doi: 10.1162/089976601753195969
- Blondel, V. D., Guillaume, J. L., Lambiotte, R., and Lefebvre, E. (2008). Fast unfolding of communities in large networks. *J. Stat. Mech. Theory Exp.* 2008:P10008. doi: 10.1088/1742-5468/2008/10/P10008
- Cammoun, L., Gigandet, X., Meskaldji, D., Thiran, J. P., Sporns, O., Do, K. Q., et al. (2012). Mapping the human connectome at multiple scales with diffusion spectrum MRI. *J. Neurosci. Methods* 203, 386–397. doi: 10.1016/j.jneumeth.2011.09.031
- Cover, T. M., and Thomas, J. A. (2012). *Elements of Information Theory*. New York, NY: John Wiley & Sons.
- Daducci, A., Gerhard, S., Griffa, A., Lemkaddem, A., Cammoun, L., Gigandet, X., et al. (2012). The connectome mapper: an open-source processing pipeline to map connectomes with MRI. *PLoS ONE* 7:e48121. doi: 10.1371/journal.pone.0048121
- Damoiseaux, J. S., Rombouts, S., Barkhof, F., Scheltens, P., Stam, C. J., Smith, S. M., et al. (2006). Consistent resting-state networks across healthy subjects. *Proc. Natl. Acad. Sci. U.S.A.* 103, 13848–13853. doi: 10.1073/pnas.0601417103
- Deco, G., Ponce-Alvarez, A., Mantini, D., Romani, G. L., Hagmann, P., and Corbetta, M. (2013). Resting-state functional connectivity emerges from structurally and dynamically shaped slow linear fluctuations. *J. Neurosci.* 33, 11239–11252. doi: 10.1523/JNEUROSCI.1091-13.2013
- DeDeo, S., and Krakauer, D. C. (2012). Dynamics and processing in finite self-similar networks. *J. R. Soc. Interface* 9, 2131–2144. doi: 10.1098/rsif.2011.0840
- de Reus, M. A., and van den Heuvel, M. P. (2013). Estimating false positives and negatives in brain networks. *Neuroimage* 70, 402–409. doi: 10.1016/j.neuroimage.2012.12.066
- Desikan, R. S., Ségonne, F., Fischl, B., Quinn, B. T., Dickerson, B. C., Blacker, D., et al. (2006). An automated labeling system for subdividing the human cerebral cortex on MRI scans into gyral based regions of interest. *Neuroimage* 31, 968–980. doi: 10.1016/j.neuroimage.2006.01.021
- Erb, I., and Ay, N. (2004). Multi-information in the thermodynamic limit. *J. Stat. Phys.* 115, 949–976. doi: 10.1023/B:JOSS.0000022375.49904.a
- Feldman, D. P., McTague, C. S., and Crutchfield, J. P. (2008). The organization of intrinsic computation: complexity-entropy diagrams and the diversity of natural information processing. *Chaos Interdiscip. J. Nonlin. Sci.* 18, 043106. doi: 10.1063/1.2991106
- Felleman, D. J., and Essen, D. C. V. (1991). Distributed hierarchical processing in the primate. *Cereb. Cortex* 1, 1–47. doi: 10.1093/cercor/1.1.1
- Fox, M. D., and Raichle, M. E. (2007). Spontaneous fluctuations in brain activity observed with functional magnetic resonance imaging. *Nat. Rev. Neurosci.* 8, 700–711. doi: 10.1038/nrn2201
- Fox, M. D., Zhang, D., Snyder, A. Z., and Raichle, M. E. (2009). The global signal and observed anticorrelated resting state brain networks. *J. Neurophysiol.* 101, 3270–3283. doi: 10.1152/jn.90777.2008
- Gong, G., He, Y., Concha, L., Lebel, C., Gross, D. W., Evans, A. C., et al. (2009). Mapping anatomical connectivity patterns of human cerebral cortex using *in vivo* diffusion tensor imaging tractography. *Cereb. Cortex* 19, 524–536. doi: 10.1093/cercor/bhn102
- Goñi, J., van den Heuvel, M., Avena-Koenigsberger, A., Velez de Mendizabal, N., Betzel, R., Griffa, A., et al. (2014). Resting brain functional connectivity predicted by analytic measures of network communication. *Proc. Natl. Acad. Sci. U.S.A.* 111, 833–838. doi: 10.1073/pnas.1315529111
- Grassberger, P. (1986). Toward a quantitative theory of self-generated complexity. *Int. J. Theor. Phys.* 25, 907–938. doi: 10.1007/BF00668821
- Hagmann, P., Cammoun, L., Gigandet, X., Meuli, R., Honey, C. J., Wedeen, V. J., et al. (2008). Mapping the structural core of human cerebral cortex. *PLoS Biol.* 6:e159. doi: 10.1371/journal.pbio.0060159
- Haimovici, A., Tagliazucchi, E., Balenzuela, P., and Chialvo, D. R. (2013). Brain organization into resting state networks emerges at criticality on a model of the human connectome. *Phys. Rev. Lett.* 110:178101. doi: 10.1103/PhysRevLett.110.178101
- Hlinka, J., Paluš, M., Vejmelka, M., Mantini, D., and Corbetta, M. (2011). Functional connectivity in resting-state fMRI: is linear correlation sufficient? *Neuroimage* 54, 2218–2225. doi: 10.1016/j.neuroimage.2010.08.042
- Honey, C. J., Sporns, O., Cammoun, L., Gigandet, X., Thiran, J.-P., Meuli, R., et al. (2009). Predicting human resting-state functional connectivity from structural connectivity. *Proc. Natl. Acad. Sci. U.S.A.* 106, 2035–2040. doi: 10.1073/pnas.0811168106
- Honey, C. J., Thivierge, J.-P., and Sporns, O. (2010). Can structure predict function in the human brain? *Neuroimage* 52, 766–776. doi: 10.1016/j.neuroimage.2010.01.071
- Kozachenko, L. F., and Leonenko, N. N. (1987). Sample estimate of the entropy of a random vector. *Probl. Peredachi Inf.* 23, 9–16.
- Kraskov, A., Stögbauer, H., and Grassberger, P. (2004). Estimating mutual information. *Phys. Rev. E* 69:066138. doi: 10.1103/PhysRevE.69.066138
- Latora, V., and Marchiori, M. (2001). Efficient behavior of small-world networks. *Phys. Rev. Lett.* 87:198701. doi: 10.1103/PhysRevLett.87.198701
- Lizier, J. T., Heinze, J., Horstmann, A., Haynes, J. D., and Prokopenko, M. (2011). Multivariate information-theoretic measures reveal directed information structure and task relevant changes in fMRI connectivity. *J. Comput. Neurosci.* 30, 85–107. doi: 10.1007/s10827-010-0271-2
- Mäki-Marttunen, V., Cortes, J. M., Villarreal, M. F., and Chialvo, D. R. (2013). Disruption of transfer entropy and inter-hemispheric brain functional connectivity in patients with disorder of consciousness. *BMC Neurosci.* 14(Suppl. 1):P83. doi: 10.1186/1471-2202-14-S1-P83
- Marinazzo, D., Pellicoro, M., Wu, G.-R., Angelini, L., Cortes, J. M., and Stramaglia, S. (2013). Information transfer of an ising model on a brain network. *BMC Neurosci.* 14(Suppl. 1):P376. doi: 10.1186/1471-2202-14-S1-P376
- Maslov, S., and Sneppen, K. (2002). Specificity and stability in topology of protein networks. *Science* 296, 910–913. doi: 10.1126/science.1065103
- Moussa, M. N., Steen, M. R., Laurienti, P. J., and Hayasaka, S. (2012). Consistency of network modules in resting-state fMRI connectome data. *PLoS ONE* 7:e44428. doi: 10.1371/journal.pone.0044428
- Power, J. D., Barnes, K. A., Snyder, A. Z., Schlaggar, B. L., and Petersen, S. E. (2012). Spurious but systematic correlations in functional connectivity MRI networks arise from subject motion. *Neuroimage* 59, 2142–2154. doi: 10.1016/j.neuroimage.2011.10.018
- Power, J. D., Schlaggar, B. L., Lessov-Schlaggar, C. N., and Petersen, S. E. (2013). Evidence for hubs in human functional brain networks. *Neuron* 79, 798–813. doi: 10.1016/j.neuron.2013.07.035

- Salvador, R., Suckling, J., Coleman, M. R., Pickard, J. D., Menon, D., and Bullmore, E. D. (2005). Neurophysiological architecture of functional magnetic resonance images of human brain. *Cereb. Cortex* 15, 1332–1342. doi: 10.1093/cercor/bhi016
- Sepulcre, J., Liu, H., Talukdar, T., Martincorena, I., Yeo, B. T. T., and Buckner, R. L. (2010). The organization of local and distant functional connectivity in the human brain. *PLoS Comput. Biol.* 6:e1000808. doi: 10.1371/journal.pcbi.1000808
- Singh, H., Misra, N., Hnizdo, V., Fedorowicz, A., and Demchuk, E. (2003). Nearest neighbor estimates of entropy. *Am. J. Math. Manag. Sci.* 23, 301–321. doi: 10.1080/01966324.2003.10737616
- Smith, S. M., Fox, P. T., Miller, K. L., Glahn, D. C., Fox, P. M., Mackay, C. E., et al. (2009). Correspondence of the brain's functional architecture during activation and rest. *Proc. Natl. Acad. Sci. U.S.A.* 106, 13040–13045. doi: 10.1073/pnas.0905267106
- Sporns, O. (2013). The human connectome: origins and challenges. *Neuroimage* 15, 53–61. doi: 10.1016/j.neuroimage.2013.03.023
- Sporns, O., Tononi, G., and Kötter, R. (2005). The human connectome: a structural description of the human brain. *PLoS Comput. Biol.* 1:e42. doi: 10.1371/journal.pcbi.0010042
- Tobler, W. R. (1970). A computer movie simulating urban growth in the detroit region. *Econ. Geogr.* 46, 234–240. doi: 10.2307/143141
- Tononi, G., Sporns, O., and Edelman, G. M. (1994). A measure for brain complexity: relating functional segregation and integration in the nervous system. *Proc. Natl. Acad. Sci. U.S.A.* 91, 5033–5037. doi: 10.1073/pnas.91.11.5033
- van den Heuvel, M. P., Kahn, R. S., Goñi, J., and Sporns, O. (2012). High-cost, high-capacity backbone for global brain communication. *Proc. Natl. Acad. Sci. U.S.A.* 109, 11372–11377. doi: 10.1073/pnas.1203593109
- van den Heuvel, M. P., and Sporns, O. (2011). Rich-club organization of the human connectome. *J. Neurosci.* 31, 15775–15786. doi: 10.1523/JNEUROSCI.3539-11.2011
- van den Heuvel, M. P., and Sporns, O. (2013). Network hubs in the human brain. *Trends Cogn. Sci.* 17, 683–696. doi: 10.1016/j.tics.2013.09.012
- Wedeen, V. J., Wang, R. P., Schmahmann, J. D., Benner, T., Tseng, W. Y. I., Dai, G., et al. (2008). Diffusion spectrum magnetic resonance imaging (DSI) tractography of crossing fibers. *Neuroimage* 41, 1267–1277. doi: 10.1016/j.neuroimage.2008.03.036
- Yeo, B. T. T., Krienen, F. M., Sepulcre, J., Sabuncu, M. R., Lashkari, D., Hollinshead, M., et al. (2011). The organization of the human cerebral cortex estimated by intrinsic functional connectivity. *J. Neurophysiol.* 106, 1125–1165. doi: 10.1152/jn.00338.2011
- Zamora-López, G., Zhou, C., and Kurths, J. (2010). Cortical hubs form a module for multisensory integration on top of the hierarchy of cortical networks. *Front. Neuroinform.* 4:1. doi: 10.3389/neuro.11.001.2010

**Conflict of Interest Statement:** The authors declare that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.

Received: 17 January 2014; accepted: 26 June 2014; published online: 24 July 2014.

Citation: Kolchinsky A, van den Heuvel MP, Griffa A, Hagmann P, Rocha LM, Sporns O and Goñi J (2014) Multi-scale integration and predictability in resting state brain activity. *Front. Neuroinform.* 8:66. doi: 10.3389/fninf.2014.00066

This article was submitted to the journal *Frontiers in Neuroinformatics*.

Copyright © 2014 Kolchinsky, van den Heuvel, Griffa, Hagmann, Rocha, Sporns and Goñi. This is an open-access article distributed under the terms of the Creative Commons Attribution License (CC BY). The use, distribution or reproduction in other forums is permitted, provided the original author(s) or licensor are credited and that the original publication in this journal is cited, in accordance with accepted academic practice. No use, distribution or reproduction is permitted which does not comply with these terms.