



**HAL**  
open science

## **MULAN: Evaluation and ensemble statistical inference for functional connectivity**

Huifang Wang, Karl J. Friston, Christian Bénar, Marmaduke Woodman,  
Patrick Chauvel, Viktor Jirsa, Christophe Bernard

► **To cite this version:**

Huifang Wang, Karl J. Friston, Christian Bénar, Marmaduke Woodman, Patrick Chauvel, et al..  
MULAN: Evaluation and ensemble statistical inference for functional connectivity. *NeuroImage*, 2018,  
166, pp.167-184. 10.1016/j.neuroimage.2017.10.036 . hal-02562257

**HAL Id: hal-02562257**

**<https://hal.science/hal-02562257v1>**

Submitted on 4 May 2020

**HAL** is a multi-disciplinary open access archive for the deposit and dissemination of scientific research documents, whether they are published or not. The documents may come from teaching and research institutions in France or abroad, or from public or private research centers.

L'archive ouverte pluridisciplinaire **HAL**, est destinée au dépôt et à la diffusion de documents scientifiques de niveau recherche, publiés ou non, émanant des établissements d'enseignement et de recherche français ou étrangers, des laboratoires publics ou privés.



Contents lists available at ScienceDirect

NeuroImage

journal homepage: [www.elsevier.com/locate/neuroimage](http://www.elsevier.com/locate/neuroimage)

## MULAN: Evaluation and ensemble statistical inference for functional connectivity



Huifang E. Wang<sup>a,\*\*</sup>, Karl J. Friston<sup>b</sup>, Christian G. Bénar<sup>a</sup>, Marmaduke M. Woodman<sup>a</sup>,  
Patrick Chauvel<sup>a,c</sup>, Viktor Jirsa<sup>a</sup>, Christophe Bernard<sup>a,\*</sup>

<sup>a</sup> Aix Marseille Univ, INSERM, INS, Inst Neurosci Syst, Marseille, France

<sup>b</sup> The Wellcome Trust Centre for Neuroimaging, University College London, Queen Square, London WC1N 3BG, UK

<sup>c</sup> Epilepsy Center, Neurological Institute, Cleveland Clinic, Cleveland, OH 44195, USA

### ABSTRACT

Many analysis methods exist to extract graphs of functional connectivity from neuronal networks. Confidence in the results is limited because, (i) different methods give different results, (ii) parameter setting directly influences the final result, and (iii) systematic evaluation of the results is not always performed. Here, we introduce MULAN (Multiple method ANalysis), which assumes an ensemble based approach combining multiple analysis methods and fuzzy logic to extract graphs with the most probable structure. In order to reduce the dependency on parameter settings, we determine the best set of parameters using a genetic algorithm on simulated datasets, whose temporal structure is similar to the experimental one. After a validation step, the selected set of parameters is used to analyze experimental data. The final step cross-validates experimental subsets of data and provides a direct estimate of the most likely graph and our confidence in the proposed connectivity. A systematic evaluation validates our strategy against empirical stereotactic electroencephalography (SEEG) and functional magnetic resonance imaging (fMRI) data.

### Introduction

Recordings from a large number of network nodes (e.g. from high density multisite electrodes, whole brain fMRI, or calcium imaging) are nowadays used routinely to study network properties and its information processing capacities. Functional networks are usually represented in the form of a graph, and various analysis methods are available to assess functional connectivity between two network nodes. Choosing a particular method represents a real challenge for the following reasons. Each method is based upon a hypothesis on how network nodes influence each other. This choice is critical, since different methods may produce different graphs. Some methods perform better as a function of the internal structure of the data (e.g. linear versus non-linear coupling) (Wang et al., 2014). The final graph is thresholded empirically to accept/reject links, the threshold value having a strong influence on the number of false positives and false negatives. When analyzing empirical data, it is impossible to know *a priori*, which methods and threshold values to use. Since the final graph cannot be validated (the ground truth structure is not known), it is important to provide a measure of confidence in the final graph. Ensemble based systems provide a way to solve the problem of decision making (Polikar, 2006). Applied to functional connectivity, it means combining several methods to decrease the dependency of the

results on the choice of only one method and on the choice of parameters, in order to increase confidence in the result. Ensemble approaches and statistical inference have been successfully applied to the extraction of structural connectomes from diffusion-weighted imaging (Pestilli et al., 2014; Takemura et al., 2016).

The conceptual contribution offered in this paper is the application of ensemble learning to pressing issues in systems neuroscience; namely, structural and functional connectomes. The key idea behind ensemble learning is to use the plurality or diversity of classification models to ensure that all data features (or their characterizations) can contribute to the best model of connectivity that underlies observed data. The key aspect of our approach rests upon using a diversity of models that are brought to bear upon the problem. In brief, we exploit the many characterizations of connectivity that have been proposed in the literature to assemble a repertoire or ensemble of models that can be fitted to time series (or possibly structural) data. Using cross validation accuracy as a proxy for model evidence or marginal likelihood, one then scores the evidence for each model in the ensemble to perform model averaging. This approach rests upon the convergence of three recent developments in this field. First the availability of diverse methods for characterizing connectivity, second the machine learning procedures for assessing model evidence in terms of cross validation accuracy and, finally, ways of

\* Corresponding author.

\*\* Corresponding author.

E-mail addresses: [Elizabethhw@gmail.com](mailto:Elizabethhw@gmail.com) (H.E. Wang), [christophe.bernard@univ-amu.fr](mailto:christophe.bernard@univ-amu.fr) (C. Bernard).

<https://doi.org/10.1016/j.neuroimage.2017.10.036>

Received 7 July 2017; Received in revised form 27 September 2017; Accepted 17 October 2017

Available online 27 October 2017

1053-8119/© 2017 The Authors. Published by Elsevier Inc. This is an open access article under the CC BY license (<http://creativecommons.org/licenses/by/4.0/>).

integrating different models using fuzzy logic. In contrast to current approaches to functional (and effective) connectivity in the brain, we do not search for the best model. Rather, we maximize the diversity of models that are applied to the same data and score their relative evidence. This approach is guaranteed to be better than using any single model (provided the single model is contained within the ensemble). In short, our ensemble learning for network analysis rests upon one single imperative; namely, to include as many models for data features as possible.

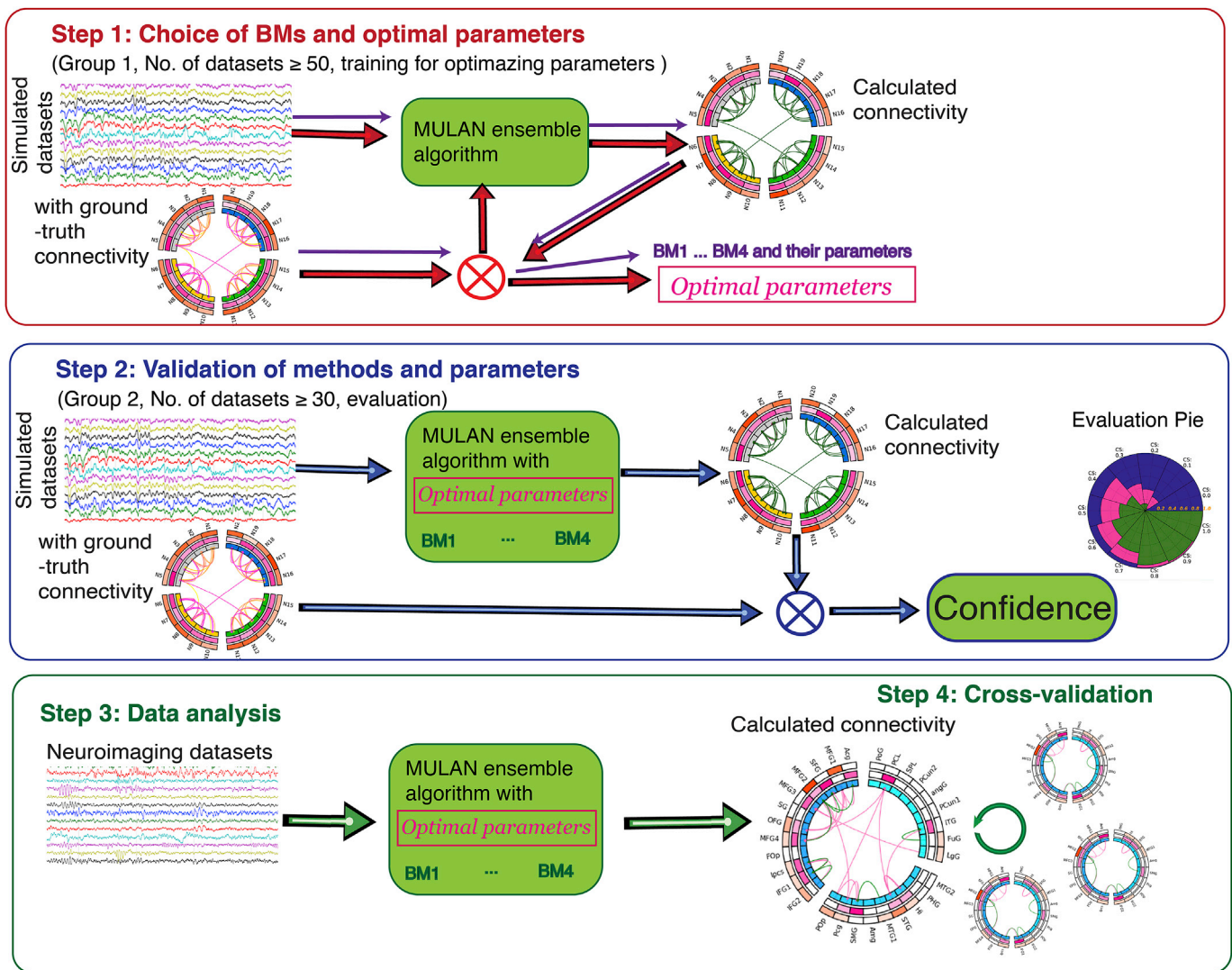
In what follows, we will assume that the forward or generative model offers a ground truth for the data it generates. However, it is useful to appreciate that there is no real ground ‘truth’. In other words, the evidence for any model of data is just a statement of its ability to explain those data in a parsimonious way. One will never know what the true model is – one can only identify the best sort of model. This is particularly prescient in the current context because both directed functional connectivity, i.e. afforded by measures like Granger causality (Barrett and Barnett, 2013) and effective connectivity, i.e. based on concepts like DCM (Friston et al., 2003) can involve a multiplicity of synaptic mechanisms and indeed polysynaptic influences among many routes. In this sense, there is no ‘true’ connectivity – just different ‘best’ characterizations of connectivity (e.g., structural, functional or effective). The underlying motivation behind MULAN is to use ensemble procedures to

provide the best model or description that elude the idiosyncrasies of any particular characterization. In this instance, the characterization of architectures of directed functional connectivity that best explain the data under a known generative process.

In this study, we introduce MULAN (MULTiple method ANalysis), a method, which uses an ensemble approach to extract functional connectivity and a statistical evaluation to directly estimate confidence in the final graph. The MULAN ensemble approach combines several methods to calculate functional connectivity between nodes. A genetic algorithm is used to set parameter and threshold values using simulated datasets. A statistical evaluation compares the calculated ensemble connectivity with a large number of simulated datasets to estimate confidence in the final graph. Finally, we show how partitioning the data can be used to cross-validate the final graph.

**General strategy**

When building a functional connectivity graph from experimental time series, it is not possible to verify the validity of the results, since, by definition; the ground truth structure is not available. We propose the general strategy to start with simulated datasets with known ground structures for an *a priori* selection of existing basic methods (BMs) and



**Fig. 1.** Flowing chart for MULAN strategy. In order to obtain the optimal parameters (step 1) and estimate the confidence in this choice (step 2), we use more than 80 simulated datasets, which are split into 2 groups (one for training and one for validation) with the same generative parameters as for the experimental data to analyze, such as number of nodes and sampling frequency. Step 1: We generate a set of BMs and optimal parameter values using training simulated datasets with different ground truth structures. Step 2: We validate these choices with the evaluation dataset. Step 3: Once validated we used the chosen BMs and optimal parameters to analyze experimental data. Step 4: We cross-validate the results using sub-datasets.

parameters (step 1, Fig. 1). After validation (step 2), the methods and parameters thus selected are used to analyze the experimental dataset (step 3), followed by a cross-validation phase (step 4). Each step uses MULAN ensemble algorithm, which takes the time series as inputs, and which outputs a directed graph. MULAN is an open platform, i.e. it can accommodate any user-designed procedures (e.g., cross-validation and genetic algorithms) to optimize not only the choice of methods and parameters, but also the strategies used in the MULAN algorithm itself. Since MULAN ensemble algorithm is used at each of the previously described steps (Fig. 1), we start describing the algorithm. Then, we show how to choose the BMs and parameter values (step 1), and how to validate these choices (step 2). Finally, we apply these to experimental electrophysiological datasets (step 3) and show how to cross-validate the results (step 4). We also show how to analyze graphs with a large number of nodes using the human brain as a prototypical large network, the sensitivity of the analysis to artefacts and the computational cost. Finally, we apply MULAN to human fMRI data.

**Methods**

*MULAN ensemble algorithm*

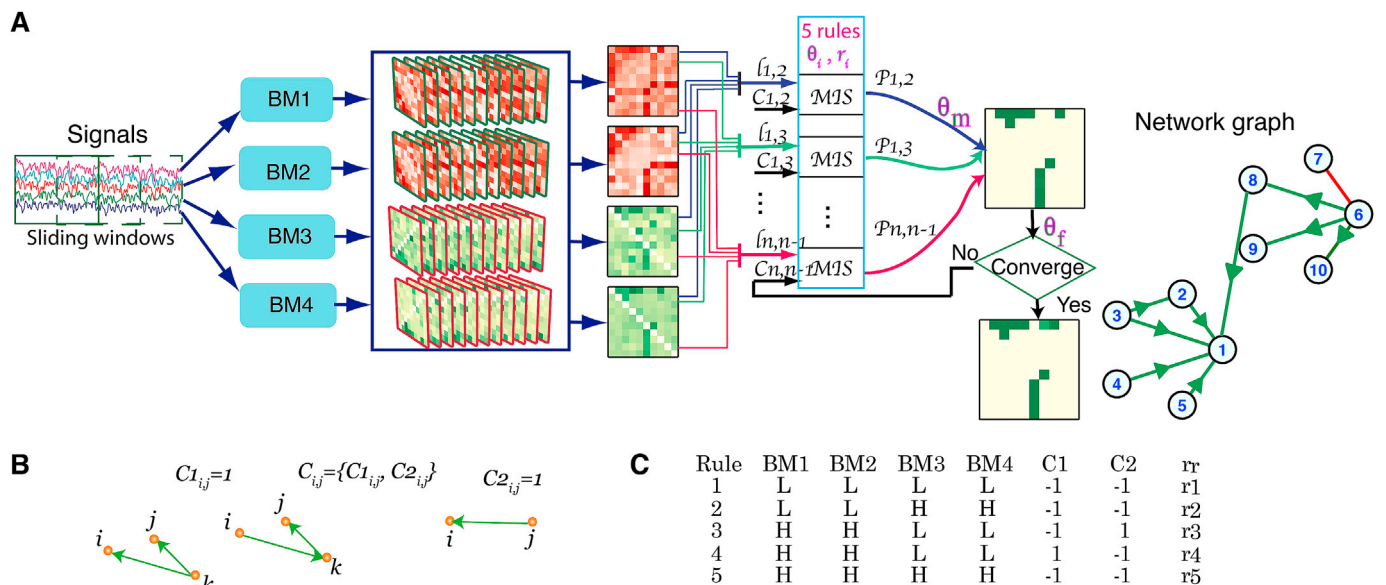
In a previous study, we have performed a systematic evaluation of 40 directed and undirected basic methods (BMs) designed to analyze functional connectivity (Wang et al., 2014). We use this library (Supplementary Fig. 1A) for our ensemble approach. We take two directed and two undirected BMs from this library as examples. MULAN ensemble algorithm (Fig. 2) uses fuzzy inference to combine the information from the four BMs to obtain the directed graph. MULAN can provide direction information, if one of BMs has direction information. Each BM returns a connectivity matrix quantifying the functional connectivity between any two nodes with a normalized strength between 0 and 1 (Appendix A and Supplementary Fig. 1B). Then, we assess each individual link  $l_{i,j}$  using MULAN fuzzy inference scheme to evaluate the plausibility of existence of the link (Appendix A and Supplementary Fig. 2). Each link is

characterized by four values, one for each BM. Each value is then assigned two fuzzy values (between 0 and 1): L (for low probability of existence) and H (for high probability) (Supplementary Fig. 1C). The algorithm is then an iterative process, generating graphs until convergence is achieved. Let us consider the  $n$ th iteration. For each link  $l_{i,j}$  between two nodes ( $i,j$ ), we apply five rules based on five different L/H combinations for the four BMs and two conditions (Appendix A and Fig. 2C). The first condition C1 tests whether a third node contacts the two nodes, or if there is an indirect path between them via a third node, based on the structure of the graph obtained at the  $(n-1)$ th iteration (Fig. 2B). The second condition C2 tests the directionality of the link if the two directed methods output that the link has a low probability of existence while the two undirected methods output that the link has a high probability of existence, based on the structure of the  $(n-1)$ th graph (Fig. 2B). Each of the five rules returns a value (between 0 and 1), and their weighted average gives the probability  $P_{i,j}$  that the link exists (Supplementary Fig. 2, Fig. 2A). After performing the same procedure for all links, we apply a threshold  $\theta_m$  to all  $P_{i,j}$  to obtain a directed graph that is binary (link/no link). If the structure of the  $n$ th graph is not the same as the previous graph, we perform a further iteration until convergence. After convergence, we use two final thresholds  $\theta_{f1}, \theta_{f2}$  to distinguish between strong (green) and weak (pink) links, respectively (Appendix A). Note that for the first iteration, there is no prior graph, and C1 and C2 values are set to  $-1$ .

*Choice of basic methods*

Since the ground truth structure is usually not known when analyzing empirical datasets, it is not possible to validate *a posteriori* the choice of BMs and parameter values. Rather than using a heuristic approach, MULAN strategy is *a priori* strategy. We simulate the empirical data using the same mechanics, number of nodes, duration, and sampling frequency, but with different ground truth graph architecture to generate multiple datasets for training (Step 1 in Fig. 1).

MULAN allows modeling any type of activity (EEG, MEG, functional

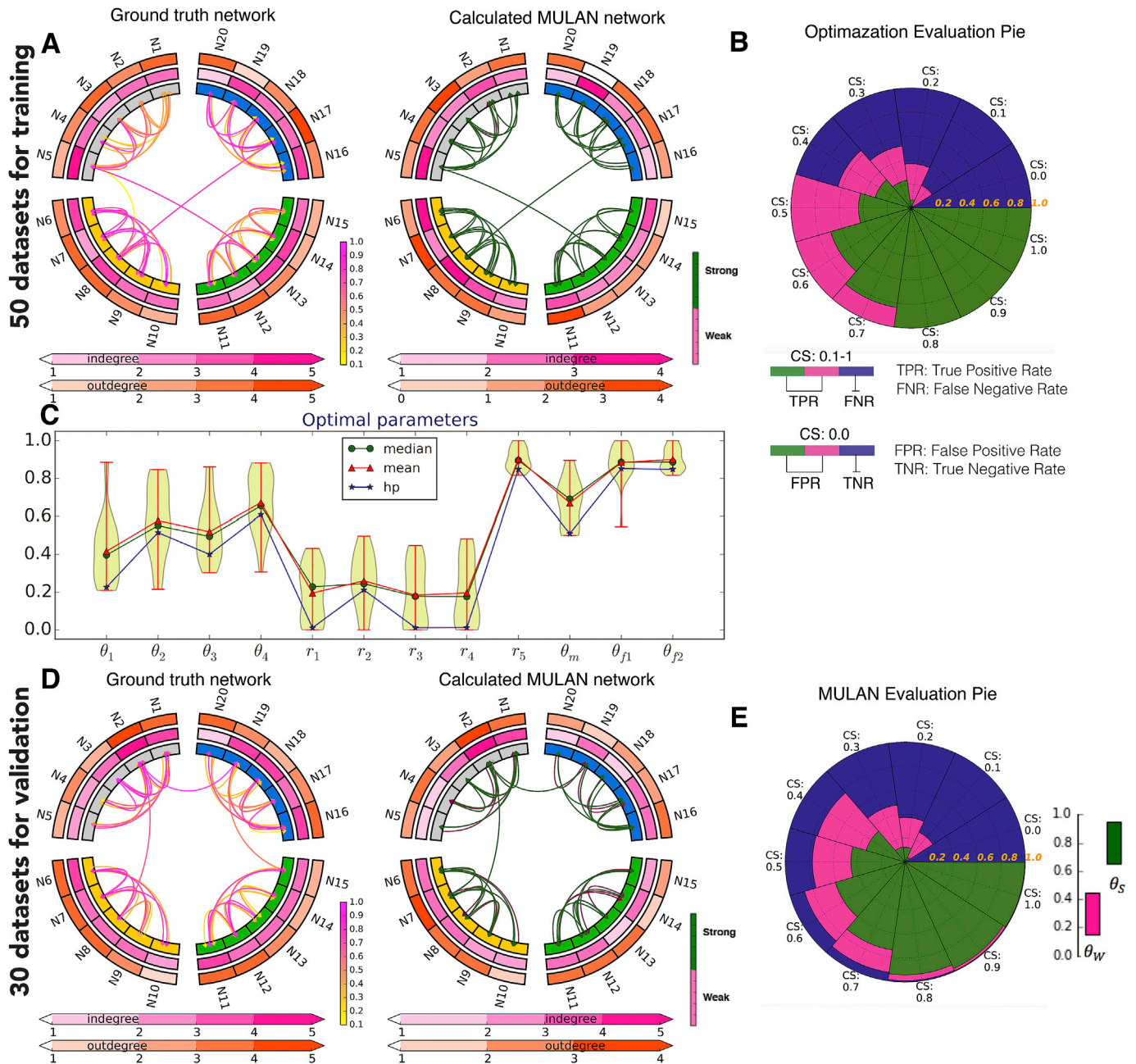


**Fig. 2.** (A) Flowchart of MULAN ensemble algorithm. A dataset is analyzed with four basic methods (BM1-BM4), using  $N$  sliding windows, thus generating  $4 \times N$  connection matrices. For each BM, we average the  $N$  connection matrices to obtain 4 averaged matrices. After bootstrapping and normalization (Methods), each link  $l_{i,j}$  between nodes  $i$  and  $j$  is characterized by four values (between 0 and 1); one value for each BM. We assess each link using MULAN inference system (MIS) (Supplementary Figs. 2A–B) to obtain  $P_{i,j}$  the probability of existence of the link.  $\theta_i$  and  $r_i$  are two important sets of parameters for MIS (Methods). A link is considered if  $P_{i,j}$  is larger than a threshold  $\theta_m$ . After obtaining the first graph, we apply conditions  $C_{ij}$ , combined with  $l_{i,j}$ , as input for the next iteration. If the new graph thus generated differs from the previous graph, we perform another iteration, using the new  $P_{i,j}$  values to change the rule conditions, until convergence is achieved (i.e. the graph of the  $(n+1)$ th iteration is the same as the  $n$ th). (B) Condition  $C_{ij}$  considers two cases:  $C1_{ij} = 1$  if there is a common source  $k$  for  $i$  and  $j$ , OR if there is an indirect path from  $i$  to  $j$  through  $k$ ; and  $C2_{ij} = 1$  if there is a link in the reverse direction i.e.  $j \rightarrow i$ . (C) We define five If-Then MULAN rules. For each considered rule,  $-1$  means an unnecessary condition and 1 a necessary one.

MRI etc.) in any type of network architecture, which can be user-defined. As a prototypical example, let us consider the analysis of a dataset of field EEGs obtained from 30 recording sites. We use a convolution-based neural mass model (NMM) for EEG (David et al., 2006; Moran et al., 2013) to generate field recordings at the millisecond timescale (Appendix C and Supplementary Fig. 3). Brain networks typically reveal a small-world topology, which is characterized by dense local clustering, while keeping a short path length between any pair of nodes (Bassett and Bullmore, 2006; Bullmore and Sporns, 2009, 2012; He et al., 2007). We consider 50 network architectures composed of subnetworks with local connectivity of variable density (each sub-network has 5 nodes and 5, 10

or 15 links), where links between subnetworks mediate short cuts in the graph (Supplementary Fig. 4A). In each graph, the connection strengths (CSs) of the links are normalized and can vary between 0 and 1. We obtain 50 different datasets.

We first perform a systematic evaluation of 40 BMs (Wang et al., 2014) on the 50 datasets (Supplementary Fig. 4). For the sake of simplicity, we consider here two cases in which all CSs have the same value: 0.8 or 1. We use the AUC value, i.e. the area under the receiver operating characteristic (ROC) curve, to evaluate the performance of each BM (Methods and Supplementary Fig. 5). Then, we select 18 different quadruplets of BMs (Supplementary Fig. 4d) from the



**Fig. 3.** Choosing optimal values for MULAN parameters for graphs comprising links with connection strengths varying from 0.1 to 1, and evaluation results. (A) Left, one example out of 50 training dataset to obtain the distribution of optimal parameter values. The ground truth network consists of links of varying strengths (color-coded from 0.1 to 1). Right: resulting MULAN graph showing green (strong) and pink (weak) links. (B) The combined results for all 50 datasets demonstrate the parameter optimization objectives: using two thresholds (strong – green and weak – pink) MULAN always identifies the links with ground truth  $CS \geq 0.5$ , and all strong links (green) with ground truth  $CS \geq 0.8$ . We ignore the cases for ground truth  $CS \in [0.1, 0.4]$ . MULAN never produces false positives when a link does not exist ( $CS = 0$ ). (C) Distribution of optimal parameter values from 50 training datasets. (D) Left, example of one of 30 networks used for evaluation. Right, corresponding MULAN graph by choosing the optimal parameter values. All links with high CS values are identified, but many low CS links are not identified. (E) Distribution of MULAN values as a function of ground truth CS values for 30 validation datasets, in which all the strong links ( $CS \geq 0.9$ ) are identified. There is no false positive ( $CS = 0$ ). The rate of false negatives (percentage of blue in each slice for  $CS > 0$ ) increases as CS decreases. There is also a negligible false positives rate of 0.6% for pink links.

correlation (Lopes da Silva et al., 1989),  $h^2$  (Ansari-Asl et al., 2006; Lopes da Silva et al., 1989; Wendling et al., 2001), transfer entropy (Barnett et al., 2009; Chicharro, 2011; Schreiber, 2000), Granger (Barnett and Seth, 2014; Granger, 1969; Seth, 2010) and  $\overline{\mathcal{N}}/\mathcal{N}$  family (Baccalá, 2007; Kaminski and Blinowska, 1991; Wang et al., 2014) (Supplementary Figs. 4B–C). We consider 12 quadruplets with 4 good performing BMs and 6 quadruplets with one badly performing BM (to test robustness). Half of the best performing BMs already generate a significant number of false positives (AUC < 0.99) for strong connection strengths (CS = 1). All methods start to fail for CS = 0.8 (Supplementary Fig. 4C). In contrast, MULAN ensemble algorithm outperforms individual BMs, in particular for CS = 0.8 (Supplementary Fig. 4D). MULAN is robust even if a poorly performing BM is included (G13 to G18). MULAN is agnostic regarding BMs and their number. Users can include any non-listed method in the algorithm, and change the number of BMs. In our experience, using 4 BMs already provides good results. In the worked example of experimental data described later (Supplementary Fig. 12G), we compare the results obtained using 6 BMs instead of 4 BMs. Using 6 BMs slightly decreases the number of false negatives for links with CSs < 0.9, but slightly increases the false positive rate of weak link for CS = 0 from 0.0011 to 0.0012. This is the reason why, we are using 4 BMs.

### Selection of the optimal parameters

We have previously analyzed the performance of individual BMs and defined the optimal set of parameters for each BM, which depends on the structure of the data (Wang et al., 2014). We use these parameter values (Appendix E). In addition to these parameters, MULAN ensemble algorithm requires setting 12 specific parameters. Four parameters  $\theta_i$ ,  $i = 1, 2, 3, 4$  determine the values of L and H for the 4 BMs (Fig. 2 and Supplementary Figs. 1–2). Five parameters  $r_r$ ,  $r = 1, 2, 3, 4, 5$ , the output from rule  $r$ , characterize the confidence that a link exists. Three parameters  $\theta_m$ ,  $\theta_{f1}$  and  $\theta_{f2}$  set the threshold for accepting a link within the MULAN inference loop, and the thresholds to distinguish between strong and weak links, respectively. We apply a genetic algorithm (Mitchell, 1998) to determine the optimal values for MULAN parameters. For each simulated dataset, the genetic algorithm (Appendix E and Supplementary Fig. 6) returns an optimal set of 12 values. For each of the 50 training datasets, we consider that a set of parameters is optimal if, in the MULAN graph (right panel of Fig. 3A): (i) there are no false positives for CS = 0, (ii) all strong links (CS ≥ 0.8) are identified as such (green), and (iii) if all CSs between 0.5 and 0.8 are correctly identified as existing, although they can be labeled as strong (green) or weak (pink). The evaluation pie chart for 50 datasets (Fig. 3B) indicates that all links are correctly identified for CS ≥ 0.5. Following this training period, we obtain a distribution,  $\mathcal{S}_s$ , of 50 values for each parameter (Fig. 3C).

## Results

### Evaluation

The training session is designed to obtain the best match between the ground truth structure and the computed graph. To validate the choice of BMs and parameter values, we generate 30 new datasets (step 2 in Fig. 1). For each dataset, we randomly pick 20 groups of 12 parameters from  $\mathcal{S}_s$  (Appendix F,G and Supplementary Figs. 7 and 8). We thus obtain 20 different graphs for each dataset. We then calculate the median value for each link. In order to assess the validity of the final averaged graph, we need to include two new thresholds (used only for validation purposes, i.e. they are not part of the general algorithm, hence they do not need to be optimized). A strong (green) link is assigned if the median value of the link is larger than  $\theta_s$  and a weak (pink) link for values larger than  $\theta_w$  but smaller than  $\theta_s$  (Appendix H, Fig. 3D). The evaluation from all 30 datasets (Fig. 3E) demonstrates the absence of false positives and negatives for CS = 0 and CS ≥ 0.9 respectively. MULAN correctly identifies strong

links for CS larger than 0.8, but below this value false negatives start to appear. The rate of false negatives is 10% for CS = 0.6, which is acceptable. The number of missed links increases as the connection strength decreases. Importantly, a weak (pink) link is reliably identified as an existing link (with a 0.0006 false positive rate: 9 links among 12995 links).

For comparison, the same evaluation procedure performed on two well performing directed basic methods, PDC and BCorrD, led to much less robust results (Supplementary Fig. 9).

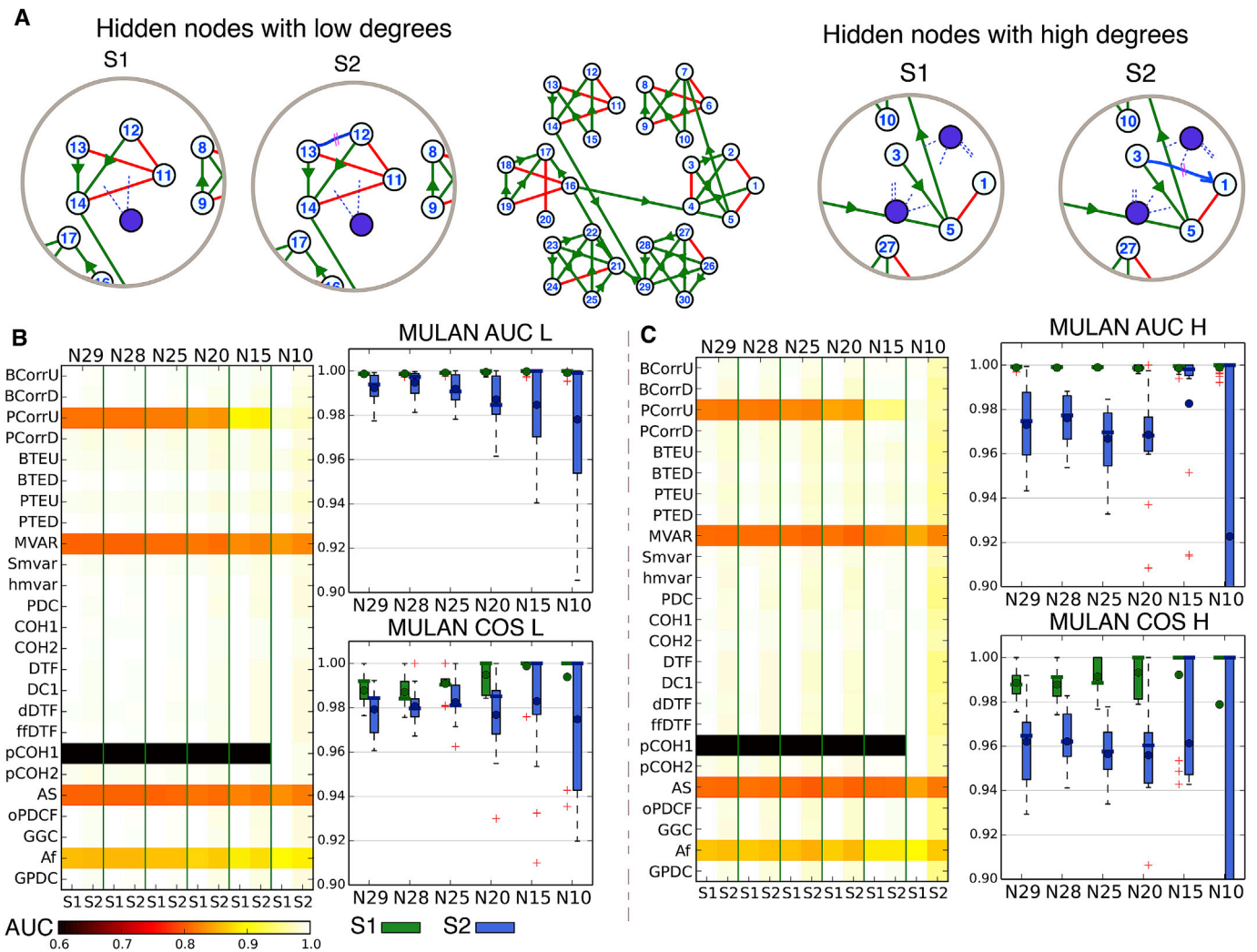
We conclude that the training procedure allowed finding optimal sets of BMs and parameter values, enabling MULAN to identify the graphs of test datasets.

### Cross-validation

In order to validate the graph, we have designed a cross-validation technique consisting in partitioning the original network to analyze into complementary sub-networks. Each sub-network will thus receive information from many hidden nodes (from those removed from the original network). We first determine the influence of the presence of hidden nodes on the graph structure. Indeed, a hidden node may act as a common source for two nodes, generating functional connectivity in the absence of effective connectivity between these two nodes. We used graphs with 30 nodes and randomly removed from 1 to 20 nodes (making them hidden nodes). We consider two cases: hidden nodes with high or low numbers of links (Appendix I and Fig. 4). We find that MULAN can infer the underlying connectivity graph, irrespective of the number of hidden nodes (Fig. 4B–C and Supplementary Fig. 10). Importantly, the presence of a hidden common source or the presence of indirect paths through hidden nodes does not influence the final graph (no false link). The implicit robustness to hidden nodes allows us to use a cross-validation for empirical data (Fig. 5A) and calculate graphs containing a large number of nodes (Fig. 5B), as described in the next sections.

### Analysis of experimental SEEG datasets

As a practical illustration of MULAN strategy, we analyzed intracranial stereo-electroencephalography (SEEG) recordings (20 s) from 30 channels obtained from a patient with drug-resistant epilepsy during an idle awake state (Fig. 6A–C and Supplementary Fig. 11A). When using BMs, we find that the functional connectivity matrices they return are very different from one another (Supplementary Fig. 11B), supporting the necessity of an ensemble approach. We follow the strategy described in Fig. 1. In step 1. Instead of using custom-designed small world architectures (designed to evaluate MULAN in the previous section), we now use realistic networks based on human connectomes made of 84 nodes (Appendix K). Although tractography-based connectomes have limitations, they provide the closest estimate of human structural connectivity, which is sufficient for our evaluation (i.e. it does not affect the generality of our conclusions). We randomly choose 30 nodes out of 84 to avoid overfitting. We use a neuronal mass model generating local field potentials, and we select sampling and dimensionality to match experimental conditions. We down-sample the signals to 250 Hz to reduce computational cost and to avoid having too high autoregressive model orders, and we choose 6 s sliding windows with a 3 s overlap. We select the 4 best performing BMs: BCorrU, COH1, BCorrD and PDC based on their AUC values (Supplementary Fig. 12A). Applying the genetic algorithm we obtain the distribution of parameter values  $\mathcal{S}_e^{30}$  (Supplementary Fig. 12D). For validation (step 2), we use 50 other connectomes from the database (Appendix K) and the parameter set  $\mathcal{S}_e^{30}$  determined previously. The evaluation results (Fig. 6D) for all 50 datasets show that for green strong links output of MULAN there are no false positives for CS = 0 and no false negative for CS > 0.9. In step 3, we use the same group of BMs and parameter values, we analyze SEEG signals and obtain the connectivity graph (Fig. 6E).



**Fig. 4.** Sensitivity to hidden nodes (A) An example of a designed ground-truth structure is shown in the top middle panel. We remove nodes from the graph (filled blue circle), considering low ( $< 3$ ) and high ( $\geq 3$ ) degrees in the right and left panels, respectively. S1 is the new ground truth structure to be recovered after removing the hidden node and its incident edges. S2 corresponds to S1, with the addition of false positive links (indicated in blue), which may have a common source from hidden nodes (left panel) and an indirect path through hidden nodes (right panel). (B) and (C) show the results obtained for hidden nodes with low and high degrees, respectively. Left, we show the AUC values for 25 BMs if either S1 or S2 is considered as the ground truth structure. We varied the number of hidden nodes (1, 2, 5, 10, 15 and 20), with Nx indicating that we analyze a network comprised of x nodes (i.e. we removed 30-x nodes, which become hidden nodes). Right, AUC values (top) and COS values (bottom) when using MULAN. Green and blue boxes test the hypothesis that S1 and S2 are the ground truth structures, respectively. Note that MULAN always finds S1, demonstrating that the algorithm is robust to hidden nodes. We used two sets of quadruplet: BCorrU, PCorrU, BCorrD, Hmvar, and BTEU, PTEU, BCorrD, ffDTF and  $\theta_1 \in [0.5, 0.6]$  and  $\theta_2 \in [0.6, 0.7]$  with  $N_{bc} = 20$ .

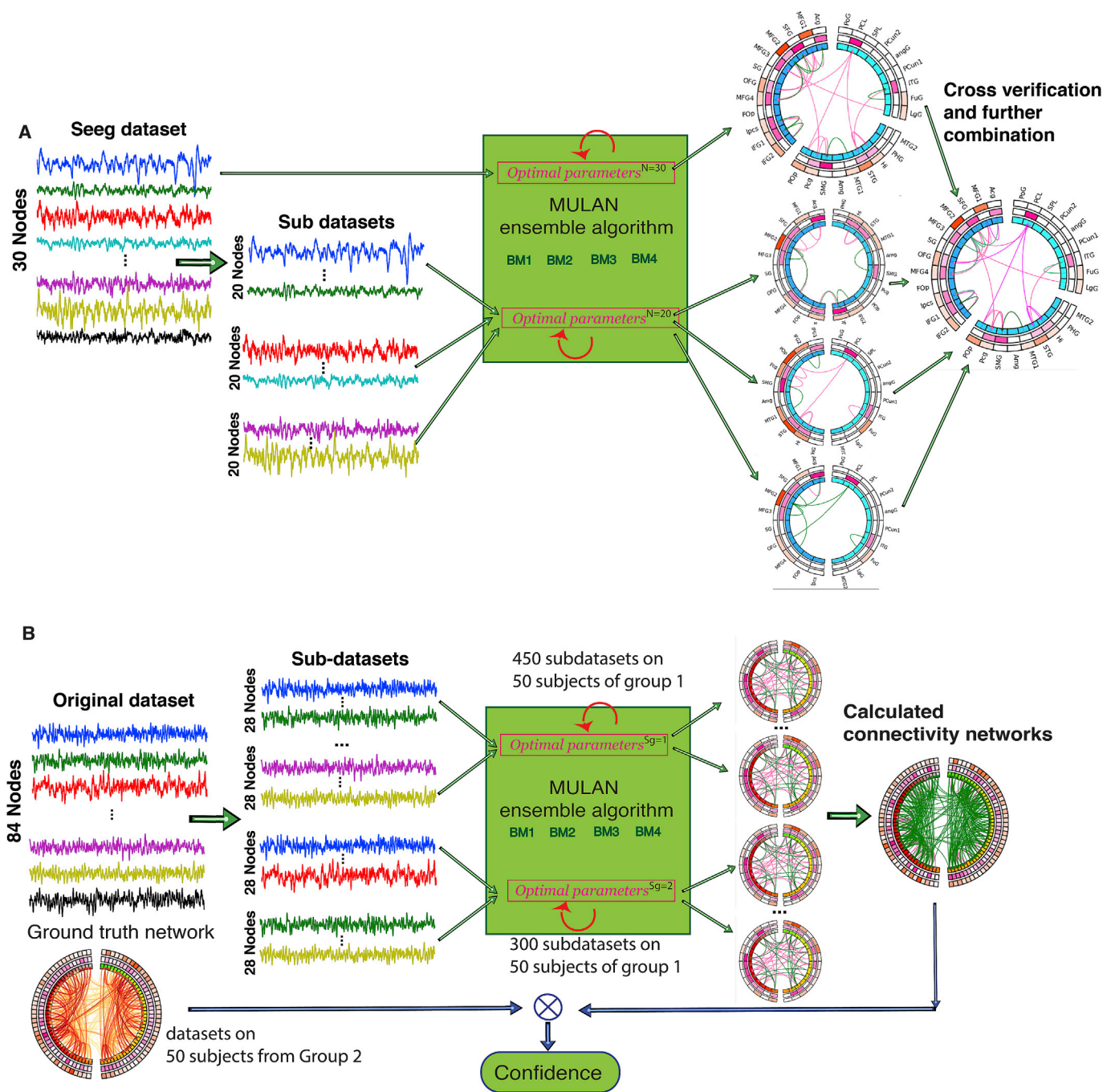
**In step 4,** in order to cross-validate our results, we use the sub-dataset strategy described above (Fig. 5A). We start with simulated datasets from the first 50 connectomes. From each 30-node dataset, we extract three sub-datasets comprising 20 nodes, each link of the 30 nodes being considered at least once in the 3 sub-datasets. We thus obtain  $3 \times 50$  training datasets, which allow us to calculate the distribution of optimal parameter values  $\mathcal{L}_e^{20h}$  (Supplementary Fig. 13A). We then use these optimal parameters to analyze the remaining 50 simulated datasets for validation. For each of these 50 datasets, we extract three 20-node sub-datasets and obtain 4 MULAN matrices: one for the 30 node time-series and three for the 20 node sub-datasets (Supplementary Figs. 13C–D). We average each link value to obtain the final MULAN matrix. After averaging, the CSs are distributed between 0 and 1, which allows us to consider several categories of CSs, from weak to strong. We consider five thresholds [0.1, 0.3, 0.4, 0.5, 0.7] coded with five colors to obtain the final MULAN graph (Supplementary Fig. 13B). With the procedure using sub-datasets (i.e. introducing hidden nodes), some of these weaker links are better classified as such (Fig. 6F). Also this procedure decreases the number of false negatives for links with CSs $<0.9$ . Therefore, the sub-

dataset method can be used to obtain a more precise graph in terms of quantitative connection strengths. The same strategy is applied to the real SEEG dataset (Supplementary Fig. 13D) to obtain a new graph (Fig. 6G). The difference with the first analysis, (Fig. 6E) is the addition of one link and a finer grain distribution of CSs.

The previous results validate our ensemble strategy. We now discuss the computational cost of the algorithm, how to handle datasets with a large number of nodes, and how the presence of experimental artefacts may affect the results.

*Handling networks with a large number of nodes (application to human connectomes)*

Human neuroimaging data typically involve large-scale networks. A typical parcellation of the human brain includes 84 nodes (Fischl et al., 2004). As ground-truth structures, we use the published connectomes of 100 unrelated subjects (Appendix K). We divide the dataset into two 50 datasets, one for training, one for validation. When the number of nodes is large (e.g. 84), the training phase could not converge to achieve 0 false links for 50 datasets. However, we always achieve convergence to 0 false

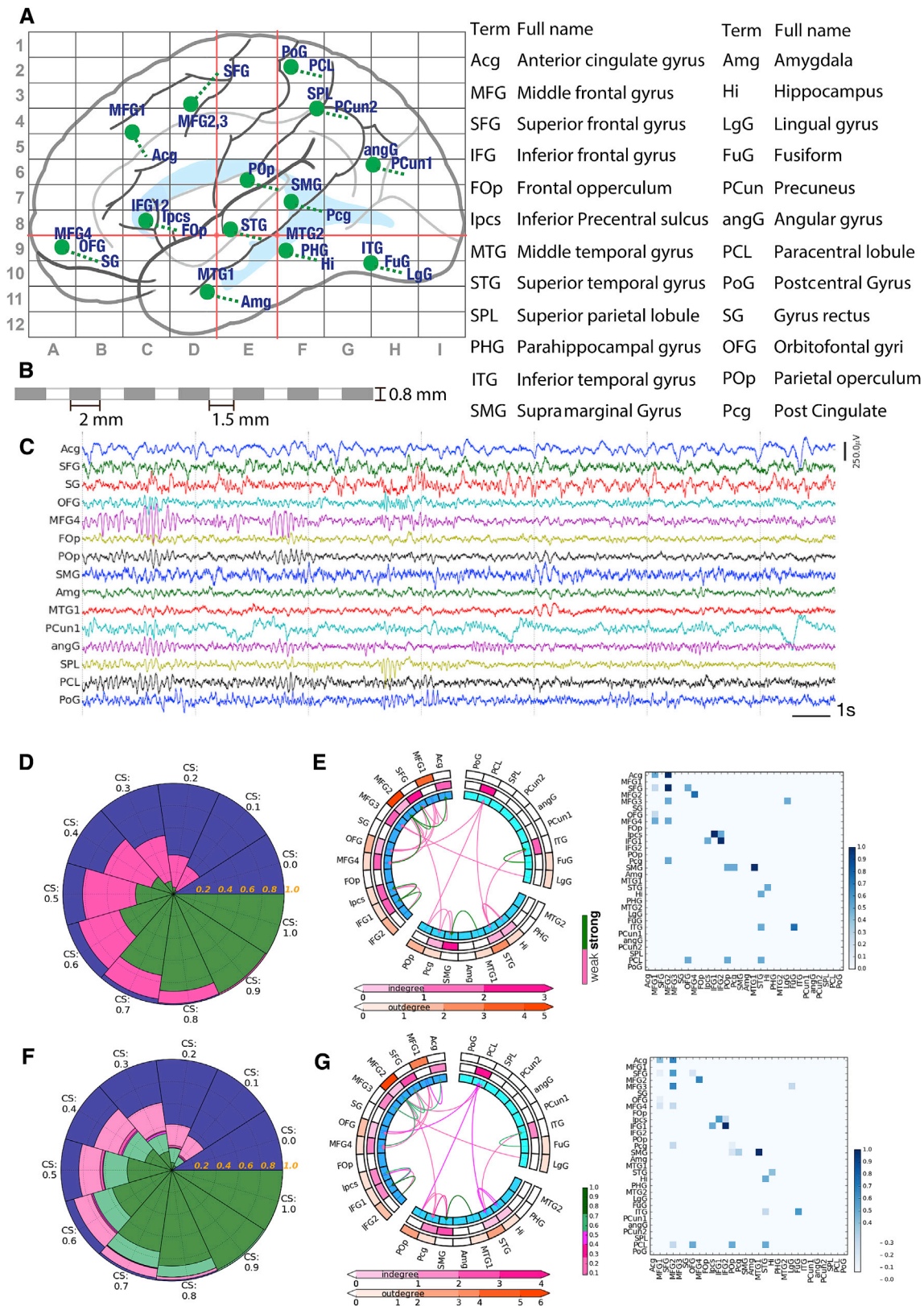


**Fig. 5.** Two flowcharts for MULAN strategy using sub-datasets on (A) cross-verification and (B) datasets with a large number of nodes. (A) For example, if the original dataset contains 30 nodes, we extract 3 sub-networks made of 20 nodes, we calculate the corresponding graphs, and combine these to verify and refine the results (examples in Fig. 6 and Supplementary Fig. 13). (b) For example, a dataset with 84 nodes (example in Fig. 7) is first divided into fifteen 28-node sub-datasets. We divide them further into 2 sub-groups according to the localization of the channels. By comparing the calculated MULAN graphs on other 50 subjects, we obtain the two optimal parameter sets for the two sub-groups. For each sub-datasets we obtain a sub-network, which we combine to obtain the final MULAN network. Confidence is assessed using datasets of 50 other subjects (Fig. 7G).

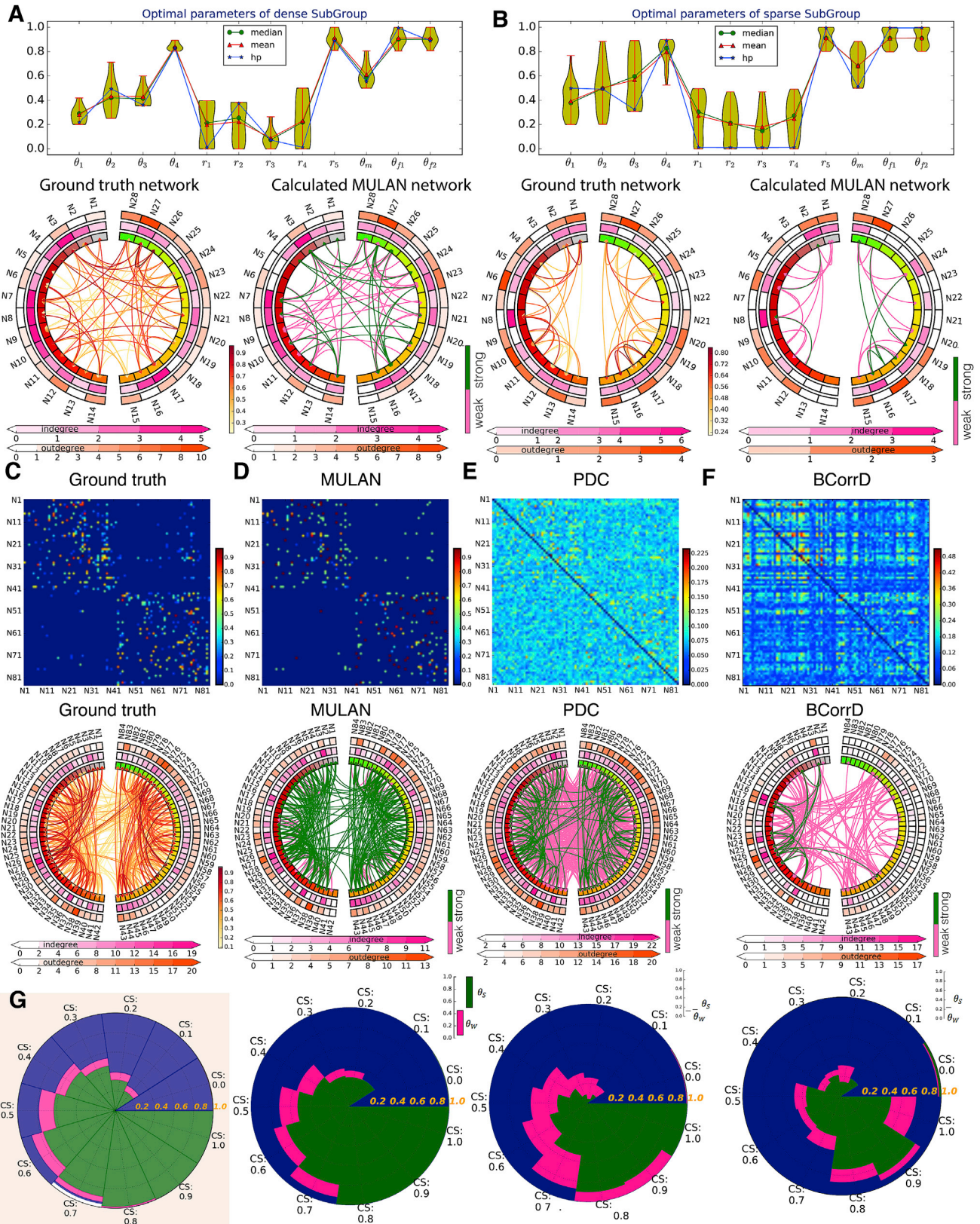
links when using 30 nodes networks. For each connectome of the training dataset, we use 15 28-node sub-datasets, which is the minimal combination of sub-datasets necessary to account for all links. According to structural connectivity of any connectome (Fig. 7C), denser links are intrahemispheric, whilst interhemispheric links are less dense. Based on this observation, we use 2 subgroups, one for dense sub-networks (all nodes within one hemisphere, Fig. 7A) and one for sparse sub-networks (links in both hemispheres, Fig. 7B). During the training phase, we obtain the distribution of optimal parameters for each subgroup (Top Fig. 7A–B) and then use these optimal parameters to calculate MULAN graphs (bottom Fig. 7A–B). Using these parameters, we perform the validation

step. One example of validation connectome is shown on Fig. 7C. The calculated MULAN graph is shown on Fig. 7D, showing no false negatives for  $CS > 0.8$ ; and a 0.0016 false positive rate for  $CS = 0$  (11 false positive links out of 6487). For comparison, we report the results obtained from two BMs, PDC (Fig. 7E) and BCorrD (Fig. 7F), which belong to the quadruplet of BMs used by MULAN for the ensemble approach. Both BMs fail to recover the structure as compared to MULAN (Top of Fig. 7D–F). MULAN is robust for a large range of evaluation threshold values, whilst thresholds need to be set in very narrow ranges for the two BMs (Fig. 7E–F). These results confirm that the ensemble approach of MULAN greatly improves the reliability of the results as compared to the best BMs





**Fig. 6.** Application of MULAN strategy to intracerebral Stereo-Electro-Encephalographic (SEEG) signals recorded from a patient with drug-resistant epilepsy. (A) Location of the 30 recording sites on a Talairach map of the left hemisphere – and list of recorded region. (B) Schematic diagram of the SEEG electrode annular leads: length 2 mm, diameter 0.86 mm and 1.5 mm inter-contact distance. (C) Time-series of original SEEG signals from the 30 recording sites. For the analysis, we use 6 s-long sliding windows with 3 s overlap. (D) Ensemble connectivity graph generated by MULAN using BCorrU, COH1, BCorrD, PDC as BMs (Supplementary Fig. 11B) and optimal parameter values from 50 training datasets with human connectome ground truth networks (Supplementary Fig. 12D). (E) Distributions of true/false positive/negative rates from MULAN results as a function of ground truth CS values for 50 validation datasets with human connectome structures. (F) After cross-validation from hidden nodes, the network is refined by introducing 5 CS levels (Supplementary Fig. 13) and is evaluated on 50 simulated datasets in (G).



**Fig. 7.** Application of MULAN to 84-node graphs obtained from a database of 100 human connectomes. After dividing an 84-node graph into fifteen 28 node graphs, themselves divided into dense (A) and sparse (B) graphs, we obtain the corresponding distribution of optimal parameters and MULAN graphs based on 50 training datasets. The matrices, graphs with green and pink links using two thresholds and evaluation results according to the graphs from three methods: MULAN, PDC and BCorrD are shown (C to F). Note that MULAN finds the original connectivity graph, when the best performing methods, PDC and BCorrD, fail. (G) Evaluation results for MULAN on simulated datasets obtained from the remaining 50 evaluation subjects.

and that MULAN is more robust for parameter setting.

If we now combine the results of the analysis of the 50 connectomes from the validation dataset, we find no false negatives for  $CS \geq 0.8$ ; and a 0.002 rate of false positives for  $CS = 0$  (Fig. 7G). Therefore, the sub-dataset procedure enables handling networks with a large number of nodes.

#### Computational cost

MULAN computational cost mainly depends on the numbers of sliding windows and the number of nodes. We use 20-s-long datasets and 6 sliding windows. The computational cost of BMs increases exponentially with the number of nodes. However, the computational cost of MULAN algorithm itself is much lower than BMs (Supplementary Fig. 14A). The cost of 20 runs as a function of nodes is shown in Supplementary Fig. 14B. The computational cost of BMs increases linearly with the numbers of sliding windows, whereas MULAN ensemble algorithm itself is not influenced by the number of sliding windows. MULAN ensemble algorithm works on the mean values of the sliding windows, so that the number of sliding windows does not really increase the computational cost. When using subnetworks for cross-verification, there is a near linear increase of the computational time with the number of nodes (Supplementary Fig. 14C). This means that the dataset can be divided in subnetworks without increasing the computational cost and without affecting the final results, as demonstrated above. This result is particularly important for the analysis of networks containing large number of nodes, as shown below.

#### Artefact sensitivity

To assess MULAN sensitivity to artefacts, we use two types of artefacts common to SEEG signals: muscle artefacts and jump artefacts (Fig. 8 A,C). For each type of artefact (around 1 s long), we systematically add the artefacts to a 30-channel 20-s SEEG dataset. We generate 270 datasets with  $A_t$  artefacts every 2 s on each  $A_c$  channels, where  $A_t$  ranges from 1 to 9 (from the beginning until ends of signals) and  $A_c$  ranges from 1 to 30. We compare artefact-free and artefact-contaminated results with cosine similarity (Appendix D, Fig. 8B,D). The algorithm is more robust for muscle artefacts than for jump artefacts. For both types of artefacts, the algorithm is more sensitive on the time axis (number of artefacts) than the space axis (number of channels with artefacts). If we take 0.96 as a threshold, MULAN starts to fail when half of signals contain muscle artefacts or 1/6 of jump artefacts. Note that experimentalists would generally reject such artefact-ridden signals from analysis.

#### Analysis of fMRI data

Functional connectivity is commonly used to analyze fMRI datasets, and construct graphs (David et al., 2008). As a worked example, we use a 6-min fMRI recording in a child with autism spectrum disorder (Di Martino et al., 2017) (Fig. 9H). In step 1, we start with 50 simulated fMRI datasets (Appendix C) using 50 different human connectomes as ground truth. We use 180-s sliding windows with a 90-s overlap and select 2 best performing BMs: BCohF, BCorrD from 10 basic methods based on their AUC values (Supplementary Fig. 15A). Note that, as reported before (Smith et al., 2011), none of the BMs is able to extract the ground truth graph even with an *a priori* knowledge of the most appropriate parameters. Using the training dataset and 2 BMs instead of 4, we obtain the distribution of parameter values for both dense and sparse sub-networks (Fig. 9A and B). Note that fMRI datasets, having very low sampling rates, do not include enough information for extracting the underlying ground truth. We obtain less than 21 false links. In step 2, we evaluate the parameter choice using 50 other connectomes. An example of a connectome from the validation set is shown on Fig. 9C, and the computed MULAN graph is shown on Fig. 9D). There is a 20% false negative rate for strong links and a 10% false positive rate for  $CS = 0$ . MULAN improves

the results of individual BMs: BCohF (Fig. 9E) and (Fig. 9F). However, the evaluation results for all 50 datasets (Fig. 9G) shows that, although MULAN does better than individual BMs, it is more difficult to extract ground truth graphs when analyzing fMRI data as compared to electro-physiological signals. Adding the third best BM (BCorrU) on these 50 datasets (Supplementary Fig. 15B) did not improve the performance. In step 3, we use 2 BMs in MULAN and the parameter values determined previously to analyze experimental fMRI data (Fig. 9H) and obtain the connectivity graph (Fig. 9I).

## Discussion

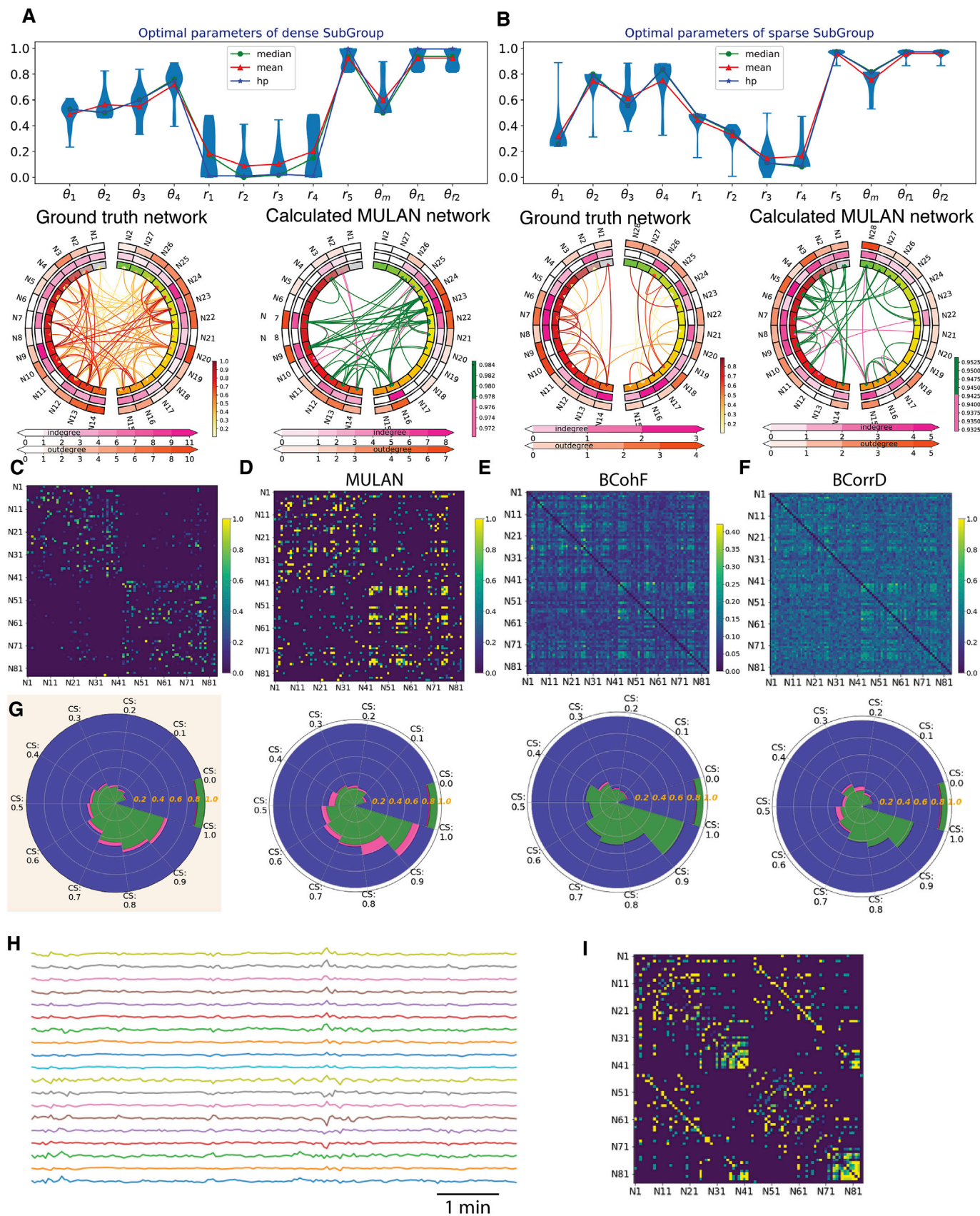
### The ensemble approach

Structural connectomes are derived from diffusion-weighted imaging coupled with tractography to map human white matter fascicles. Functional connectivity is derived from multiple functional neuroimaging data to map the relationships between different regions. Many methods have been proposed to extract structural connectomes and functional connectivity graphs, but the final graph is method-dependent and parameter-dependent. A solution has been recently proposed to extract structural connectomes based on statistical evaluation (Pestilli, 2015; Pestilli et al., 2014) and ensemble inference (Takemura et al., 2016).

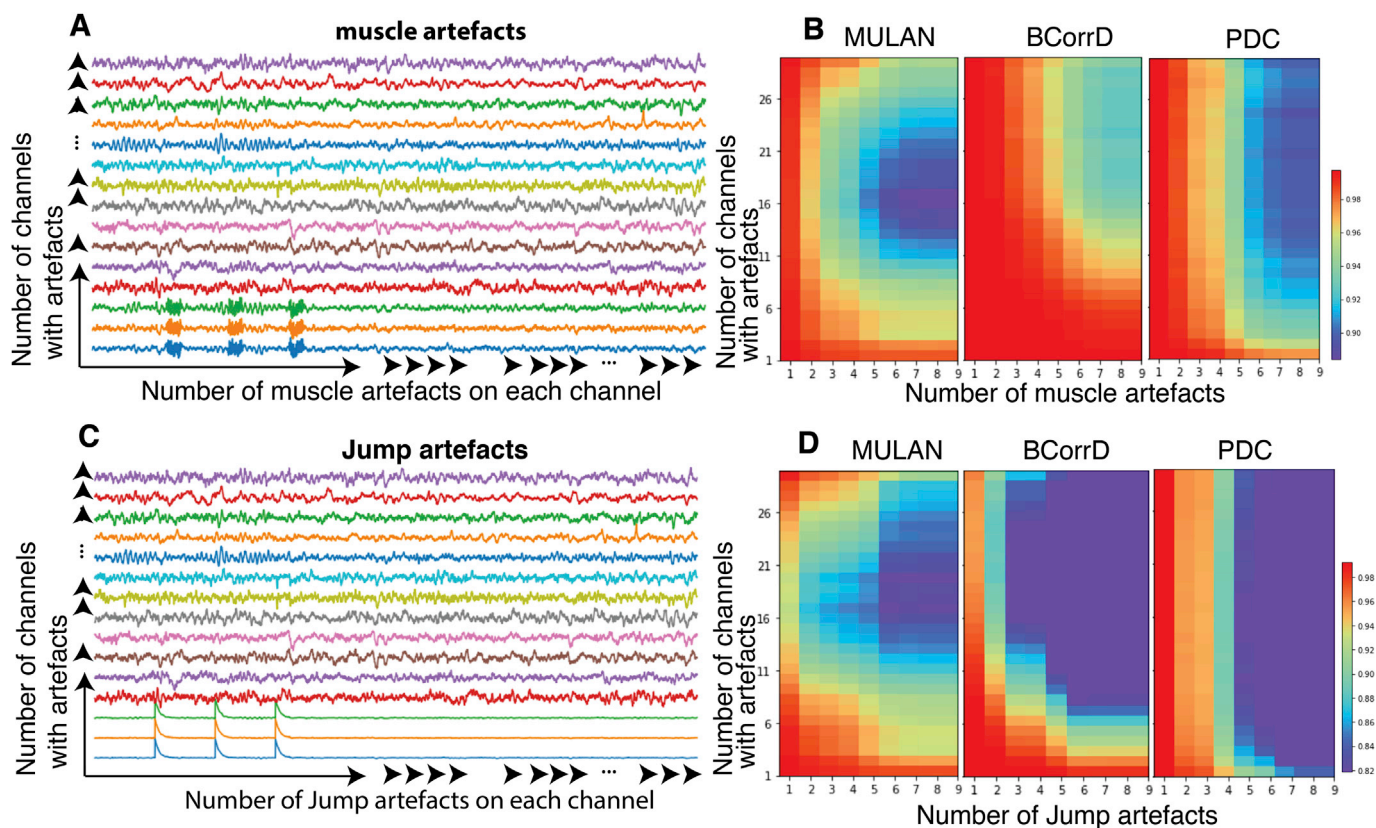
The present work builds on these studies to extend the concept to functional connectivity. The importance is twofold: (1) structural connectomes are related to the underlying neuroimaging data through diffusion processes captured by the Stejskal-Tanner equation (Pestilli et al., 2014; Stejskal and Tanner, 1965); to make the equivalent link for functional connectivity is more difficult, for a large part due to our insufficient understanding of how local neuro-electric and -chemical processes organize themselves across multiple scales in space and time. An ensemble approach, based on statistical validation methods, is perhaps an alternative for accurate network identification. (2) To date, we understand that no single connectome mapping method will be optimal in all situations (Takemura et al., 2016) for this reason, statistical approaches will become fundamental in identifying connectomes with high degree of sensitivity and specificity (Zalesky et al., 2016). A growing literature focuses on the generative models of the human connectome that yield synthetic networks combining geometric and topological factors in order to better understand the human connectome (Betzel et al., 2016), which could be further used for the functional connectivity.

### Functional, effective and ensemble connectivity

MULAN infers ensemble connectivity by pooling information from various functional connectivity measures. Different functional connectivity methods provide complementary information about statistical dependencies in time-series. MULAN essentially tests the hypothesis that the underlying coupling can be inferred from the information provided by different aspects of functional connectivity. The systematic evaluation of MULAN using thousands of datasets supports our hypothesis. Dynamic causal modeling (DCM) (Friston et al., 2003; Friston et al., 2013) is the most established approach to estimate effective connectivity (a model of interaction or coupling). DCM is based on a neuronal model that describes causal interactions, which are mediated by unobservable neuronal dynamics – and an observation model that describes the mapping from neural activity to observed responses. Thus, an interesting issue for future work is a validation of our ensemble measures of coupling in relation to explicit models of effective connectivity. MULAN and DCM have complementary aims for inferring underlying coupling. Crucially, MULAN can be directly applied to any given time series to detect the coupling among empirically sampled nodes, independently of any specific (or explicit) causal model. Interestingly, MULAN could be used as a prior in DCM-based approaches.



**Fig. 8.** Sensitivity to two types of artefacts: (A) muscle artefacts and (C) jump artefacts with an example with 3 artefacts on 3 channels in a 30-channel 20-s SEEG dataset. Cosine similarity values to compare artefact-free and artefact-containing data as a function of the number of artefacts in time and their distribution across channels (B and D). From left to right, we show the results when using MULAN and two directed BMs: BCoRD and PDC.



**Fig. 9.** Application of MULAN to fMRI datasets with 84 nodes. Step 1: After dividing an 84 node graph into fifteen 28 node graphs, themselves divided into dense (A) and sparse (B) graphs, we obtain the corresponding distribution of optimal parameters and MULAN graphs based on 50 training datasets. We apply the optimal parameters on the example dataset with ground truth (C) and obtain the connectivity matrices and evaluation results from three methods: (D) MULAN, (E) BCohF and (F) BCorrD. Note that BCohF can not provide direction information. (G) Evaluation results for MULAN on 50 evaluation subjects. Step 3: We apply MULAN (H) a fMRI dataset with 84 channel and 6 minutes of a child with autism and obtain (I) the MULAN matrix.

### Adaptive MULAN

The present version of MULAN uses fuzzy logic and a genetic algorithm for the following reasons. Fuzzy logic is a consensus of multiple inputs and rules, which fits well with the present question. Fuzzy logic is more robust by providing degree of the truths rather than classic crisp logic whose conclusions are either true or false. We used a genetic algorithm because it can handle the high degree of nonlinearity imposed by our strategy (traditional linear optimization tools cannot be used). Genetic algorithms have demonstrated the ability, efficiency and robustness in handling complex search spaces (Bancaud et al., 1970; Gürocak, 1999; Herrera et al., 1995) and achieving the complex objectives we set (the calculated connectivity graph needs to be similar to the ground-truth). However, other algorithms could be used and implemented.

In principle, any connectivity method, which outputs weighted connection matrices, can be incorporated as a basic method into MULAN. MULAN affords the opportunity to include, evaluate and use any new method. We suggest using basic methods, which have low computational cost and which do not require prior knowledge of the underlying generative model; such as correlation functions, coherence, transfer entropy and  $\bar{A}H$  families. We also suggest using at least two classes of BMs (that would usually report instantaneous or undirected functional connectivity and directed functional connectivity based on temporal dependencies). In this respect, it is useful to note that our results show that MULAN is robust to the inclusion of a poorly performing basic method.

MULAN parameters can be adjusted to different types of data. Importantly, the five fuzzy logic rules used in MULAN depend on the combination of BMs. Although these five rules are robust and widely applicable, users still can add or replace rules to optimize results from their chosen BMs. Because no ground-truth connectivity network is

available for empirical datasets, we suggest starting the analysis procedure with a systematic evaluation on simulated data to optimize the BMs and MULAN parameters. This involves generating multiple simulated datasets, with the same number of nodes, signal length and sampling frequency as the empirical dataset. First we perform a systematic evaluation (Wang et al., 2014) to select the best performing basic methods, and their associated parameters. Then we use a genetic method to obtain MULAN parameters – as illustrated in this paper. The results thus obtained enable an informed choice of parameters that can then be used to analyze empirical data. An overview of parameters entailed by MULAN in provided in [Supplementary Table 2](#).

Basic methods generally require a threshold, which is impossible to know *a priori*. Our analyses show that results can be very sensitive to a slight change in threshold when using a given BM. In contrast, a major strength of MULAN is that it is relatively insensitive to such small variations. Finally, the links identified by MULAN can be characterized by their strength, the classification (e.g. null, weak, strong in the paper) is mainly for evaluation purpose.

### Hidden nodes

The MULAN scheme is robust to the influence of hidden nodes. This result is important, as some network nodes are not observable or measurable in many empirical data (e.g., deep sources in MEG/EEG). Hence, the presence of hidden nodes or indirect pathways between nodes does not influence the ability of MULAN to detect a link between two observed nodes. We found this behavior particularly useful when using hidden nodes to cross-validate MULAN. Importantly, its low sensitivity to the influence of hidden nodes provides an alternative way to analyze datasets with a very large number of nodes (>50); namely, by simply

combining the results obtained from the analysis of sub-networks with a smaller number of nodes. Such a procedure could obtain the satisfied connectivity results by using the subnetwork optimal parameters, because when large networks are characterized optimal parameters are hardly obtained with 0 false links.

#### Application and limitations of MULAN

Although an ensemble approach provides a better estimate of connected graphs as compared to single methods, the accuracy depends upon the individual performance of the BMs. The analysis of fMRI data illustrates this. Even the best performing BMs are not able to extract the correct graph, perhaps because the neuronal responses are convolved with a hemodynamic response function with a very low sample frequency. As a result, MULAN cannot output the correct graph, even if its estimation is better than that provided by individual BMs. In contrast to electrophysiology data for which the convergence criterion can be set to 0 false positive link, in the case of fMRI data, the threshold had to be empirically set to 21 false links for the algorithm to converge. Better performing BMs need to be designed to improve the accuracy of fMRI functional graphs. Our simulated data corresponds to electrophysiological (MEG or EEG) time series, with inherent spectral structure and correlations induced by (simulated) neuronal processes. For SEEG datasets, we can get confident connectivity results based on the given locations and high time resolutions. Whereas from fMRI time series where the neuronal responses are convolved with a haemodynamic response function and with very low sample frequency, it is more difficult to find the ground-truth graph (because all BMs do not perform well enough). Future work may include other algorithms such as DCM, and then work on the connectivity of hidden states (Friston et al., 2003; Friston et al., 2013).

Finally, it is important to note that we assume that the network is in a

steady state regime. However, functional links evolve in time, a concept coined as Functional Connectivity Dynamics (FCD) (Hansen et al., 2015). FCD has been observed in task and resting conditions of the human brain (Allen et al., 2014; Hansen et al., 2015). Subsequent work may include cases when functional connectivity fluctuates over time.

All MULAN procedures are written in Matlab/Python code and are available freely with submitted code, which will be pushed on github once this paper is published.

#### Competing interests

All authors in this paper have non-financial competing interests.

#### Acknowledgement

We are very grateful to Amanda Eason for preprocessing fMRI datasets, Michael Wibral, Pedro A. Valdés-Sosa, Fabrice Wendling, Timothée Proix and Mathieu Golos for helpful discussions, and Andrea Brovelli for useful discussions and technical help. This work was supported by grants from the Fondation pour la Recherche Médicale en France FRM:ING20121226265, ANR-2010-BLAN-1520-01, ANR-13-NEUC-0005, and Ligue Française Contre l'Epilepsie. K.J.F is funded by a Wellcome Trust Principal Research Fellowship (Ref: 088130/Z/09/Z).

Fibre-tracking was performed using the MRtrix package (J-D Tournier, Brain Research Institute, Melbourne, Australia, <https://github.com/MRtrix3/mrtrix3>). 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.

## APPENDICES.

### A. MULAN ensemble algorithm

#### 1 Normalization of the connections strengths generated by the Basic Methods

For MULAN ensemble algorithm, for each basic method (BM), we average the  $N$  connection matrices obtained for each analysis window. The value of each matrix element corresponds to the connection strength between two nodes. However, the range of values depends upon each BM. For example, the calculated values from BCorrU (Bivariate correlation undirected) range from 0 to 0.4, whereas the values from Hmvar (H\_matrix based on MVAR) range from 0 to 0.003 (Supplementary Fig. 1B). In order to combine the results from all BMs, we normalize the values according to their probability distribution function. After bootstrapping 500 times, on  $N$  connection matrices from  $N$  sliding windows with replacement, we obtain 500 averaged connection matrices. Then, we build the histogram (100 bins) of all connection strengths thus obtained (blue bars in Supplementary Fig. 1B). The histogram is used to compute the cumulative distribution function (in green in Supplementary Fig. 1B), defined by:  $R_i = (\sum_{j=1}^i m_j) / (\sum_{j=1}^{N_{bin}} m_j)$ , with  $m_j$  the histogram value of the  $j^{th}$  bin. All connection strengths are normalized to 1. The four normalized values from the four BMs form the vector  $l_{ij}$ , which is used as the input to the MULAN fuzzy inference system (MIS).

#### 2 Initial conditions

MULAN algorithm is an iterative scheme. We start with an empty graph, and the first iteration produces a graph with links.

#### 3 MULAN Fuzzy Inference System (MIS)

MULAN fuzzy inference system (MIS) works on each link  $i, j$  and uses fuzzy set theory (Tang, 1993; Zadeh, 1965) to map normalized values  $l_{ij}$  from the basic methods and current network conditions  $C_{ij}$  to a MULAN score  $P_{ij}$  (Supplementary Fig. 2). MIS comprises four steps: designing fuzzy rules, defining membership functions, performing the inference operations on the rules and obtaining MULAN scores by combining the rule strength and output. The detailed description for each step is as follows. For simplicity, we omit the subscripts  $ij$  in this section.

##### Step 1: Fuzzy if-then rules with conditions

Fuzzy if-then rules are expressions of the form IF A THEN B, where A and B are labels of fuzzy sets characterized by membership functions. MIS uses

the fuzzy if-then rules with conditions, i.e. in case of C, IF A THEN B. The definition of MIS rules depends on the chosen BMs. Here we give an example of 5 MIS rules, which we use in this paper (Supplementary Fig. 2D). BM1 and BM2 are two undirected methods, whereas BM3 and BM4 are two directed methods. H (high) corresponds to high likelihood for existence of a link, whereas L (low) corresponds to low likelihood. The value “-1” represents an unnecessary condition, and “1” a necessary one. Output  $r_r$  is 0 if the link unlikely exists and 1 if the link most likely exists.

The five rules:

- Rule 1: IF all the basic methods predict that the link unlikely exists (L), THEN the link unlikely exists (output:  $r_1 \in [0, 0.5]$ ).
- Rule 2: IF undirected methods output L AND directed methods predict it likely exists (H), THEN it unlikely exists (output:  $r_2 \in [0, 0.5]$ ).
- Rule 3: In the case of C2: existence of a link in the opposite direction, IF undirected methods output H AND directed methods output L, THEN the link unlikely exists (output:  $r_3 \in [0, 0.5]$ ).
- Rule 4: In the case of C1: existence of a common source or an indirect path, IF undirected methods output H AND directed methods output L, THEN the link unlikely exists (output:  $r_4 \in [0, 0.5]$ ).
- Rule 5: IF all basic methods output H, THEN it likely exists (output:  $r_5 \in [0.8, 1]$ ).

We use these conservative ranges of  $r_r$  values (instead of just 0 or 1) to cover all possibilities. When analyzing data, the  $r_r$  values are determined during the optimization process, as detailed below. These values can vary each time when a new dataset is analyzed.

### Step 2: Membership functions

Membership functions map the normalized input value from each BM into a degree of non-existence/existence of this link; i.e., L or H in our case. We use generalized bell-shaped membership function  $\mu(x) = 1 / \left( 1 + \left| \frac{x-c}{a} \right|^{2b} \right)$ , where  $a = \theta_i$  for L and  $1 - \theta_i$  for H;  $b = 2k-a$ ,  $c = 0$  for L and 1 for H (Supplementary Fig. 1C). Here  $\theta_i$ ,  $i = 1, 2, 3, 4$  determines the half width of the bell shape of the membership functions for each of the four BMs. The examples in Supplementary Fig. 1C show that  $\theta_i$  strongly influences the membership functions; in particular, for values close to the steep parts of the sigmoid functions. The slope of the function is also controlled by  $k$ . Here, we used  $k = 5$  as our simulations show no major effect of varying  $k$  between 3 and 7.

### Step 3: Inference operators

We need to obtain the output value  $r_r$  of a given rule  $r$ . The first step is to verify the conditions  $\prod_{c=1}^2 C_c^r$ ,  $c = 1, 2$ ; i.e., condition C1 or C2 for the  $r^{\text{th}}$  rule.  $C_c^r = 1$  in two situations: a condition  $C_c$ ,  $c = 1, 2$  is unnecessary for a rule  $r$ ; Or a condition  $C_c$  is necessary AND  $C_c = 1$  (the condition is satisfied). Otherwise  $C_c^r = 0$ , i.e. if a condition is necessary AND  $C_c = 0$ . If  $C_c^r = 0$ , we do not use this rule. Let us consider the example shown in Supplementary Fig. 2B: here, during one iteration, there is neither a common source nor an indirect path for two nodes under consideration. Hence, C1 is not satisfied (False) and Rule 4 is not used. If a rule is used, we multiply the membership values from the 4 BMs to obtain the weight of rule  $r$ , i.e.  $w_r = \prod_{b=1}^4 \mu_b^r$  ( $w_r$  and  $r_r$  are shown as dark and light blue bars, respectively, in Supplementary Fig. 2B). For each rule  $r$ , the inference outputs a weighted  $w_r$  value for a given  $r_r$  value (Supplementary Figs. 2A–B).

### Step 4: MULAN score

For each link, MULAN outputs the value  $P = \frac{1}{N_r} \sum_{r=1}^{N_r} w_r r_r$ , where  $N_r = 5 - \text{Number of unused rules}$ . The unused rules fail on the conditions,  $\prod_{c=1}^2 C_c^r$ ,  $c = 1, 2$ .

## B. Basic methods – functional connectivity measures

In this study, we consider 40 functional connectivity measures from 7 families as candidates of basic methods: correlation,  $h^2$ , mutual information, coherence, Granger, transfer entropy and  $\overline{\mathcal{A}}\mathcal{H}$  (MVAR-frequency domain based techniques, see below for details). The mathematical form of these measures of functional connectivity can be found in Supplementary Table 1 and note 1. We select these methods because they have low computational cost and do not require prior knowledge of the underlying generative model.

We have demonstrated the way to select the best sets of parameters for the basic methods according to their AUC values on the simulation datasets with known ground truth and the number of nodes is 5 (Wang et al., 2014). The important parameters are length of windows, the length of signals for all basic methods; the frequency ranges for the frequency domain methods and time delay for the time-domain methods. In this paper we extended to the cases with more number of nodes, detailed the parameters in Supplementary Table 2. The type and property of datasets are important when simulation datasets considered. In this paper, we use neural mass models and fMRI models. In addition, three mathematical models such as linear and nonlinear models with linear/nonlinear couples can refer to (Wang et al., 2014). Users are encouraged to find the suitable models for their datasets. Many models are available from open source software toolkits such as Statistical Parametric Mapping and The Virtual Brain.

## C. Simulated data:

### 1 Neuronal Mass Model (NMM)

We consider synaptic convolution-based neural mass models (NMM) to test the methods at the millisecond timescale, which is typical for electrophysiological signals. These models (David et al., 2006; Moran et al., 2013) consider cortical columns with three subpopulations: spiny stellate cells in granular layer IV, pyramidal cells and inhibitory interneurons in extra granular layers (II and III, V and VI) (Supplementary Fig. 3). Thirteen neural states include currents  $i_0 - i_5$  and membrane potentials  $v_1 - v_7$  for these three cell subpopulations. A connection from node  $j$  to  $i$  corresponds to a link from the pyramidal subpopulation of node  $j$  to the spiny stellate subpopulation of node  $i$ . We use  $C_{ij}$  to denote the connection strength from node  $j$  to node  $i$ .  $C_{ij}$  varies between 0 (no effect from  $j$  to  $i$ ) to 1 (the input from pyramidal column  $i$  has the same strength as the input from node  $i$  itself).

The details of the mathematical definition of the NMM model can be found in Supplementary Figs. 3A–C. For simplicity, time  $t$  and node  $i$  are omitted. The sigmoid function is defined as  $S(x) = 1/(1 + e^{-x}) - 1/(1 + e^2)$ . The main parameters are:  $K_e = -4 \text{ ms}^{-1}$ ,  $K_i = -16 \text{ ms}^{-1}$ ,  $H_e = 8 \text{ mV}$ ,  $H_j = 32 \text{ mV}$ ,  $\gamma_{1,2,3,4,5} = \{128, 128, 64, 64, 4\}$ . The input  $U$  comprises signals containing both white and pink noise, with amplitude 0.0865. The original codes are from <http://www.fil.ion.ucl.ac.uk/spm>. We use 125, 250 and 1000 Hz sampling frequencies in this study. For visualization and comparison, we plot the time-series and time frequency plot for simulated signals (Supplementary Fig. 3D) and experimental data from gyrus rectus (Supplementary Fig. 3E). Note that the similarities in frequency content.

## 2 fMRI

We used the standard generative model used in DCM to generate simulated BOLD signals. The neuronal signals were generated using linear differential equations with one neuronal state per node (K. J. Friston et al., 2003; Stephan et al., 2007). This neural activity is transformed into a BOLD response using a non-linear haemodynamic model (incorporating the empirically validated balloon model). The associated haemodynamic response function effectively acts as a non-linear convolution operator to remove high-frequency fluctuations in neuronal signals. This means simulated BOLD time series have a slower timescale as compared to EEG signals. We used the BOLD signal to analyze the underlying connectivity between the neuronal states of nodes  $i$  and  $j$ . We also added system noise to the neuronal state and observation noise to the BOLD signals.

## D. Evaluation criteria

### 1 AUC value

For each simulated dataset, a particular basic method computes a connectivity matrix, which assigns a connection value between 0 and 1 between any two nodes. The connection strength can be thresholded to produce a network graph. For example, in the connection matrix 2 in Supplementary Fig. 5A, choosing a connection threshold of 0.4 will produce a graph with many more links (those in orange color) than the ground truth structure. However, with the threshold of 0.5, the method identifies the ground truth structure. Based on these ground truth structures, we can calculate the number of false positive and true positive edges as a function of the threshold. This enables us to compute receiver operating characteristic (ROC) curves (Zweig and Campbell, 1993) and evaluate the area under the curve (AUC) (Supplementary Fig. 5A). Note that, for the undirected methods, the computed connection matrices are symmetric, thus we take the symmetrical adjacency matrix as a ground truth for calculating their ROC and AUC (Supplementary Fig. 5B). The closer the AUC is to 1, the closer the computed graph is to the ground truth structure. Note that, if  $\text{AUC} = 1$ , then there exists a threshold that can identify the correct structure. But this threshold cannot be empirically determined when analyzing experimental data, because the ground truth of the graph is not known.

### 2 COS value

We use the COS value (Singhal, 2001) to measure the similarity between two connectivity graphs. First, we use a binary vector  $v$  of  $N(N - 1)$  dimension to characterize a graph with  $N$  nodes, in which 1 indicates the presence of a link and 0 the absence of a link. Then, to compare two graphs, we calculate  $\text{COS}(v_1, v_2) = \frac{v_1 \cdot v_2}{\|v_1\| \|v_2\|} = \frac{\sum v_{1,i} v_{2,i}}{\sqrt{\sum (v_{1,i})^2} \sqrt{\sum (v_{2,i})^2}}$ . The values thus range from 1 (identical graphs) and 0 (totally different graphs).

### 3 Pie chart

We use a multi-level pie chart to compare the computed graphs with ground-truth (such as Fig. 3B,E, Fig. 6D,F, Fig. 7D-G and Fig. 9D-G). MULAN outputs a network graph with green and pink lines to indicate strong and weak links, respectively. To characterize the accuracy of the graph, we use a pie chart. Each slice corresponds to a given ground-truth connection strength from 0 to 1 with 0.1 steps. For a given ground-truth connection strength, we calculate the number of times MULAN outputs a green, pink link or no link (blue), noted as  $N_c^{cs}$ , where  $c \in \{\text{green, pink, blue}\}$  and  $cs \in \{0, 0.1, \dots, 0.9, 1\}$ . Then, we calculate the rate  $\frac{N_c^{cs}}{\sum_{c \in \{\text{green, pink, blue}\}} N_c^{cs}}$ .

## E. Genetic algorithm for parameter optimization

The genetic algorithm is a method for solving optimization problems based on natural evolution, such as inheritance, mutation, selection and crossover. This type of algorithm is well suited for optimization problems in which the objective function is discontinuous, non-differentiable, or highly nonlinear. We use this algorithm for optimizing 10 MULAN parameters  $[\theta_1, \theta_2, \theta_3, \theta_4, r_1, r_2, r_3, r_4, r_5, \theta_m]$  (Supplementary Fig. 6). Using training datasets of known structure containing the same number of nodes as the empirical data to analyze, we start with 16 sets of MULAN randomly generated parameters within a given range,  $\theta_{1,2} \in [0.2, 0.9]$ ,  $\theta_{3,4} \in [0.3, 0.9]$ ,  $r_{1,2,3,4} \in [0, 0.5]$ ,  $r_5 \in [0.8, 1]$ , and  $\theta_m \in [0.5, 0.9]$  (we used a large range that was empirically determined during preliminary simulations). The first 16 sets are called the first generation. MULAN is used to generate 16 graphs (one per set). From these, we extract the 2 sets with the lowest number of false positives and negatives. We retain these two sets. Then, we randomly pick two sets (parents) from the 16, and from these we generate 2 children by crossover. We repeat the operation (randomly picking 2 parents) 6 times, to obtain 12 children. Finally, we randomly choose two sets from the 16, and we apply a mutation to 4 parameters. This procedure gives the next generation of 16 sets of parameters. Note that the probability of selecting one set is proportional to the accuracy of the corresponding graph. This process is repeated until one of three following termination conditions is reached: the generated connective graph has 0 false link, there are 20 generations with the same false



links or 40 generations have been performed. Notice that using this termination conditions, the process dose not always converge to 0 false positive link as SEEG datasets do. For example, in the case of fMRI data, the optimization process terminates around 21 false positive links.

The last parameter group is for generating the MULAN graph. We consider two scenarios. When the MULAN graph has one parameter,  $\theta_f$  – for strong connections here shown in green – or to parameters  $\theta_{f1}, \theta_{f2}$  for two-color mapping (green links for the strong connection and pink ones for the weak). These parameters can be determined easily when comparing the MULAN output with ground truth structures (using 50 training datasets).

### F. Statistical results

Although MULAN returns connectivity matrices with AUC values close to 1, some links may be incorrectly assigned. To further improve the performance, we use the bootstrapping procedure to define the final connection graph based on combination of  $N_{bt}$  matrices, each from on bootstrapping on  $N$  sliding windows (Supplementary Fig. 7A). We vary the number of nodes (20, 30 and 50) and the density of links (5\*, 10\* and 15\*; a \* indicating the number of links within each sub-network), using 10 datasets for each case. We use the cosine similarity (COS) to compare the final connection graphs with the ground truth. After testing  $N_{bt}$  from 10 to 300, we find that  $N_{bt} = 20$  is sufficient to recover the ground truth structure (Supplementary Fig. 7B).

### G. MULAN application on strong connection strengths

Here we demonstrate the examples with only strong connection strengths for better understanding MULAN algorithm and evaluation system. We simulate the data with 30 number of nodes with 50 different small-world ground truth structures (i.e. links with connection strengths varying between 0.8 and 1.0 to simplicity). We then apply a genetic algorithm (Mitchell, 1998) to determine the optimal values for MULAN parameters. For each simulated dataset, the genetic algorithm (methods and Supplementary Fig. 7) returns an optimal set of 11 values. Supplementary Fig. 8A shows one example from 50 simulated datasets with the ground truth network in the left and the MULAN graph in the right. The performance based on all 50 simulated datasets (Supplementary Fig. 8B) illustrates that MULAN produces neither false positives nor false negatives when using optimal parameters. Since 50 simulated datasets generate a set of optimal parameters  $\mathcal{S}_s$ , we obtain a distribution of values for each parameter (Supplementary Fig. 8C).

To evaluate MULAN using this optimal distribution, we generate 30 new datasets with the same generative parameters (numbers of nodes etc.) but with ground truth structures differing from the 50 training datasets. We randomly pick 20 groups of parameters from the set  $\mathcal{S}_s$  of optimal parameters. For each dataset, we obtain a matrix as the median value of 20 MULAN matrices (Supplementary Fig. 8D). We thus obtain the final MULAN graph in which strong and weak links are shown in green and pink, respectively. When we evaluated MULAN on the 30 test datasets, we find that all links were correctly identified, although some strong  $CS = 0.8$  links in the original graphs are labeled as “weak” (pink) in the final MULAN graph (Supplementary Figs. 8E–F). In short, MULAN can always identify strong links ( $CS \geq 0.8$ ), producing neither false positives nor false negatives.

#### 1 Selection of optimal parameters

To select a particular value from each distribution (Supplementary Fig. 8C) of optimal parameters obtained from 50 training datasets, we generate other 30 new datasets with the same generative parameters (the numbers of nodes etc.) but with ground truth structures differing from the 50 training datasets. We evaluate 4 alternative ways to choose the optimal parameters using bootstrapping ( $N_{bt} = 20$ ) (Supplementary Fig. 8D). In the first three, we use 20 sets of optimal parameters randomly generated from normal distributions having their expectation as ‘mean’, ‘median’ or ‘highest peak’ (hp) of the optimal parameter set  $\mathcal{S}_s$ , and half standard deviation. The fourth approach is to randomly pick 20 groups of parameters from the set  $\mathcal{S}_s$  of optimal parameters. In each case, we obtain 20 MULAN matrices. Since MULAN decides whether a link exists (green) or not, we assign the value of 1 to each green link. Each link is thus characterized by a distribution of 20 binary values. We represent each link by either the mean or the median value of the distribution, thus generating two matrices. We apply a final threshold  $\theta_s = 0.9$  (to select strong links only) and calculate the distribution of false negatives/positives (Supplementary Fig. 8D). We find that using the fourth option (randomly selecting from  $\mathcal{S}_s$ ) and using the median value of the distribution produces the smallest number of false positives (Supplementary Fig. 8D). In order to remove false negatives, we introduce another threshold for weak links  $\theta_w$ ; i.e., for links between  $[0.4, 0.9]$ . We thus obtain the final MULAN graph in which strong and weak links are shown in green and pink, respectively. When we evaluate MULAN on the test 30 datasets, we find that all links are correctly identified, although some strong  $CS = 0.8$  links in the original graphs are labeled weak (pink) in the final MULAN graph (Supplementary Fig. 8F). In short, MULAN can always identify strong links ( $CS \geq 0.8$ ), producing neither false positives nor false negatives.

### H. Evaluation thresholds

In order to assess the validity of the final averaged graph from 20 sets of parameters, we include two new thresholds (used only for validation purposes, i.e. they are not part of the general algorithm, hence they do not need to be optimized). A strong (green) link is assigned if the median value of the link is larger than  $\theta_s$  and a weak (pink) link for values larger than  $\theta_w$  but smaller than  $\theta_s$ . For the example of 30 nodes NMM datasets, MULAN is robust for a wide range  $\theta_w \in [0.15, 0.45]$  and  $\theta_s \in [0.65, 0.95]$  to obtain zero false negative for  $CS \geq 0.9$ . Hence, any value can be used within these intervals. Because for each MULAN graph from one set of parameters, we set 0.5 for weak links and 1 for strong links. We search  $\theta_w$  less than 0.5 and  $\theta_s \in [0.5, 1]$ . For comparison, the same evaluation procedure performed on two well performing directed basic methods, PDC and BCorrD. For both BMs, we have to fix  $\theta_w$  (0.1 for PDC and 0.16 for BCorrD) and  $\theta_s$  (0.14 for PDC and 0.24 for BCorrD) to obtain zero false negative for  $CS \geq 0.9$ . While the MULAN algorithm leads 9 false positive pink link, PDC has 402 false links and BCorrD, has 364. Thus MULAN provides a better estimate of the connectivity as compared to individual BMs. Importantly, MULAN is more robust regarding threshold values as compared to BMs applied alone.

For the evaluation on datasets based on 100 human connectome with 84 nodes, we apply MULAN algorithm on 50 subjects using the obtained optimal parameters from other 50 subjects. Then, we use two thresholds  $\theta_w$  and  $\theta_s$  to identify weak (pink) and strong (green) links (middle of Fig. 7D–F.) For one example in the text, MULAN is robust for a large range of threshold values ( $\theta_w = [0.1, 0.49]$ ;  $\theta_s = [0.5, 1]$ ) to the absence of false negatives for  $CS \geq 0.8$ ; the false positive rate for  $CS = 0$  is 0.0016 (11 false positive links out of 6487). Whilst thresholds need to be set in very narrow ranges for the two BMs (PDC:  $\theta_w = 0.12$  and  $\theta_s = [0.14, 0.15]$ ; BCorrD:  $\theta_w = 0.26$  and  $\theta_s = 0.28$ ) (Fig. 7E–F, bottom). PDC finds no false negatives for  $CS \geq 0.8$  and the rate of false positive is 0.005 for  $CS = 0$  (33 links out of 6487), performing better than BCorrD (Fig. 7E–F, bottom).

## I. Hidden nodes

Here, we examine how the presence of hidden nodes influences the resulting graph. We use structures with 30 nodes and a link density of  $10^*$  (10 links with 5 node-sub-structure middle panel of Fig. 4A). We then randomly remove 1, 2, 5, 10, 15 or 20 nodes (making them hidden nodes) and analyze the data from the remaining 29, 28, 25, 20, 15 or 10 nodes, respectively. We consider two cases: hidden nodes with high ( $\geq 3$ , right panel of Fig. 4A) or low ( $<3$ , left panel of Fig. 4A) numbers of links. Finally, we consider two structures S1 and S2. S1 corresponds to the ground truth structure without the hidden nodes. S2 is similar to S1, with the addition of a false positive link between two nodes, if they are jointly influenced by a hidden node (left panel of Fig. 4A), or if they are connected vicariously via hidden nodes (right panel of Fig. 4A).

We use AUC values to evaluate the sensitivity of 25 BMs to hidden nodes (left panels of Fig. 4B and C). We consider two cases: the method discovers S1 (true) or S2 (false positive) structures. The AUC values for the BMs show that most methods are more likely to identify S1 than S2, for both low and high degree hidden nodes, which provides sufficient information for MULAN (Supplementary Fig. 10). Using both AUC and COS values, we find that MULAN can infer S1 as the underlying connectivity graph, irrespective of the number of hidden nodes (Fig. 4B and C). Crucially, MULAN recovers S1 rather than S2; i.e., the presence of a hidden common sources or the presence of indirect paths through hidden nodes does not influence the final inference (no false link). The implicit robustness to hidden nodes allows us to calculate connectivity with a large number of nodes (Fig. 5A) and to design a cross-validation for empirical data (Fig. 5B), as described in the next two sections.

## J. Empirical example – stereo-electroencephalography (SEEG)

A male patient with drug resistant epilepsy (28 years old) was recorded with SEEG electrodes during pre-surgical evaluation at the Cleveland Clinic Epilepsy Center. SEEG recordings were performed using 14 intra-cerebral multiple contact electrodes in the left cerebral hemisphere and according to the Talairach stereotactic method (Bancaud et al., 1970; Talairach et al., 1992). Each electrode had 10–15 contacts with length of 2 mm, 1.5 mm interspace and a diameter of 0.8 mm) (Fig. 6B).

We choose 20-second recordings from 30 contacts of 14 electrodes, located in different brain regions: cingulate gyrus, superior/middle/inferior frontal gyrus, frontal operculum, inferior precentral sulcus, superior/middle/inferior temporal gyrus, superior parietal lobule, supramarginal gyrus, hippocampus, parahippocampal gyrus, amygdala, lingual gyrus, fusiform, precuneus, angular gyrus, paracentral lobule, postcentral gyrus, gyrus rectus, orbitofrontal gyri, parietal operculum (Fig. 6A). A fusion of MRI and CT scans provides the exact location of each electrode contact. We use the original SEEG recordings (common reference montage) and a sampling frequency of 1000 Hz.

## K. Human connectome

We use 100 structural connectomes from the human connectome project (<https://db.humanconnectome.org/>) based on preprocessed structural and diffusion MRI from 100 unrelated subjects. We use the software MRtrix3 (Tournier et al., 2012) and FreeSurfer (Desikan et al., 2006). We first convert the diffusion images into a non-compressed format, embed the diffusion gradient encoding information, make volume data contiguous for each voxel and convert to floating point representations. Then, we generate a whole brain mask from a DWI image. Diffusion weighted and structural volumes are used to obtain a mask that includes both brain tissue and Cerebrospinal fluid. Next, we estimate the response function as the kernel during the deconvolution step. For the white matter, this is the signal expected for a voxel containing a single, coherently oriented fiber bundle. We also estimate fiber orientation distributions from diffusion data using spherical deconvolution. For structural image processing, we first generate a mask image appropriate for seeding streamlines on the grey matter-white matter interface.

After processing the structural and diffusion images, we generate the initial tractogram by setting the maximum length to 250 mm. Then, we apply the spherical-deconvolution informed filtering of tractograms (SIFT) to reduce the overall streamline count and provide more biologically meaningful estimates of structural connection density. Finally, we map streamlines to the parcellated image based on a FreeSurfer Desikan-Killiany atlas (Fischl et al., 2004) segmentation to produce a connectome with 84 nodes. Based on the log-normal distribution of values (Buzsáki and Mizuseki, 2014), we normalize  $\log(c_{ij}^d/32)$  to  $[0, 1]$  as the ground truth structure, where  $c_{ij}^d$  is the connection density for area  $i$  to area  $j$  from the structural connectome. To avoid overfitting when we optimize the parameters and evaluate the final results, we randomly assign the direction to  $i \rightarrow j$  or  $j \rightarrow i$ . We chose this procedure since, in this study, we focus more on connection patterns than on the real human connectome.

## Appendix A. Supplementary data

Supplementary data related to this article can be found at <https://doi.org/10.1016/j.neuroimage.2017.10.036>.

## References

- Allen, E.A., Damaraju, E., Plis, S.M., Erhardt, E.B., Eichele, T., Calhoun, V.D., 2014. Tracking whole-brain connectivity dynamics in the resting state. *Cereb. Cortex* (New York, N.Y. 1991) 24 (3), 663–676.
- Ansari-Asl, K., Senhadji, L., Bellanger, J.J., Wendling, F., 2006. Quantitative evaluation of linear and nonlinear methods characterizing interdependencies between brain signals. *Phys. Rev. E* 74 (3), 31916.
- Baccalá, L., 2007. Generalized partial directed coherence. In: *Digital Signal Processing, 2007 15th International Conference on*. IEEE, Baccald, LA, pp. 163–166.
- Bancaud, J., Angelergues, R., Bernouilli, C., Bonis, A., Bordas-Ferrer, M., Bresson, M., Talairach, J., 1970. Functional stereotaxic exploration (SEEG) of epilepsy. *Electroencephalogr. Clin. Neurophysiology* 28, 85.
- Barnett, L., Barrett, A.B., Seth, A.K., 2009. Granger causality and transfer entropy are equivalent for gaussian variables. *Phys. Rev. Lett.* 103 (23), 238701.
- Barnett, L., Seth, A.K., 2014. The MVGC multivariate Granger causality toolbox: a new approach to Granger-causal inference. *J. Neurosci. Methods* 223, 50–68.
- Barrett, A.B., Barnett, L., 2013. Granger causality is designed to measure effect, not mechanism. *Front. Neuroinformatics* 7.
- Bassett, D.S., Bullmore, E., 2006. Small-world brain networks. *The Neuroscientist: a review journal bringing neurobiology. Neurology Psychiatry* 12 (6), 512–523.
- Betz, R.F., Avena-Koenigsberger, A., Goñi, J., He, Y., de Reus, M.A., Griffa, A., Sporns, O., 2016. Generative models of the human connectome. *NeuroImage* 124, 1054–1064.
- Bullmore, E., Sporns, O., 2009. Complex brain networks: graph theoretical analysis of structural and functional systems. *Nat. Rev. Neurosci.* 10 (3), 186–198.
- Bullmore, E., Sporns, O., 2012. The economy of brain network organization. *Nat. Rev. Neurosci.* 13 (5), 336–349.
- Buzsáki, G., Mizuseki, K., 2014. The log-dynamic brain: how skewed distributions affect network operations. *Nat. Rev. Neurosci.* 15 (4), 264–278.
- Chicharro, D., 2011. On the spectral formulation of Granger causality. *Biol. Cybern.* 105 (5–6), 331–347.

- David, O., Guillemain, I., Sallet, S., Reyt, S., Deransart, C., Segebarth, C., Depaulis, A., 2008. Identifying neural drivers with functional MRI: an electrophysiological validation. *PLoS Biol.* 6 (12), e315.
- David, O., Kiebel, S.J., Harrison, L.M., Mattout, J., Kilner, J.M., Friston, K.J., 2006. Dynamic causal modeling of evoked responses in EEG and MEG. *NeuroImage* 30 (4), 1255–1272.
- Desikan, R.S., Ségonne, F., Fischl, B., Quinn, B.T., Dickerson, B.C., Blacker, D., Killiany, R.J., 2006. An automated labeling system for subdividing the human cerebral cortex on MRI scans into gyral based regions of interest. *NeuroImage* 31 (3), 968–980.
- Di Martino, A., O'Connor, D., Chen, B., Alaerts, K., Anderson, J.S., Assaf, M., Milham, M.P., 2017. Enhancing studies of the connectome in autism using the autism brain imaging data exchange II. *Sci. Data* 4, 170010.
- Fischl, B., van der Kouwe, A., Destrieux, C., Halgren, E., Ségonne, F., Salat, D.H., Dale, A.M., 2004. Automatically parcellating the human cerebral cortex. *Cereb. Cortex* (New York, N.Y. 1991) 14 (1), 11–22.
- Friston, K.J., Harrison, L., Penny, W., 2003. Dynamic causal modelling. *Neuroimage* 19 (4), 1273–1302.
- Friston, K., Moran, R., Seth, A.K., 2013. Analysing connectivity with Granger causality and dynamic causal modelling. *Curr. Opin. Neurobiol.* 23 (2), 172–178.
- Granger, C.W.J., 1969. Investigating causal relations by econometric models and cross-spectral methods. *Econ. J. Econ. Soc.* 424–438.
- Gürocak, H.B., 1999. A genetic-algorithm-based method for tuning fuzzy logic controllers. *Fuzzy Sets Syst.* 108 (1), 39–47.
- Hansen, E.C.A., Battaglia, D., Spiegler, A., Deco, G., Jirsa, V.K., 2015. Functional connectivity dynamics: modeling the switching behavior of the resting state. *NeuroImage* 105, 525–535.
- He, Y., Chen, Z.J., Evans, A.C., 2007. Small-world anatomical networks in the human brain revealed by cortical thickness from MRI. *Cereb. Cortex* (New York, N.Y. 1991) 17 (10), 2407–2419.
- Herrera, F., Lozano, M., Verdegay, J.L., 1995. Tuning fuzzy logic controllers by genetic algorithms. *Int. J. Approx. Reason.* 12 (3–4), 299–315.
- Kaminski, M.J., Blinowska, K.J., 1991. A new method of the description of the information flow in the brain structures. *Biol. Cybern.* 65 (3), 203–210.
- Lopes da Silva, F., Pijn, J.P., Boeijinga, P., 1989. Interdependence of EEG signals: linear vs. nonlinear associations and the significance of time delays and phase shifts. *Brain Topogr.* 2 (1–2), 9–18.
- Mitchell, M., 1998. *An Introduction to Genetic Algorithms*. MIT Press.
- Moran, R., Pinotsis, D., Friston, K., 2013. Neural masses and fields in dynamic causal modeling. *Front. Comput. Neurosci.* 7, 57. May.
- Pestilli, F., 2015. Test-retest measurements and digital validation for in vivo neuroscience. *Sci. Data* 2, 140057.
- Pestilli, F., Yeatman, J.D., Rokem, A., Kay, K.N., Wandell, B.A., 2014. Evaluation and statistical inference for human connectomes. *Nat. Methods* 11 (10), 1058–1063.
- Polikar, R., 2006. Ensemble based systems in decision making. *IEEE Circuits Syst. Mag.* 6 (3), 21–45.
- Schreiber, T., 2000. Measuring information transfer. *Phys. Rev. Lett.* 85 (2), 461–464.
- Seth, A.K., 2010. A MATLAB toolbox for Granger causal connectivity analysis. *J. Neurosci. Methods* 186 (2), 262–273.
- Singhal, A., 2001. Modern information retrieval: a brief overview. *IEEE Data Eng. Bull.* 24 (4), 35–43.
- Smith, S.M., Miller, K.L., Salimi-Khorshidi, G., Webster, M., Beckmann, C.F., Nichols, T.E., Woolrich, M.W., 2011. Network modelling methods for FMRI. *NeuroImage* 54 (2), 875–891.
- Stejskal, E.O., Tanner, J.E., 1965. Spin diffusion measurements: spin echoes in the presence of a time-dependent field gradient. *J. Chem. Phys.* 42 (1), 288.
- Stephan, K.E., Weiskopf, N., Drysdale, P.M., Robinson, Pa, Friston, K.J., 2007. Comparing hemodynamic models with DCM. *NeuroImage* 38 (3), 387–401.
- Takemura, H., Caiafa, C.F., Wandell, B.A., Pestilli, F., 2016. Ensemble tractography. *PLOS Comput. Biol.* 12 (2), e1004692.
- Talairach, J., Bancaud, J., Bonis, A., Szikla, G., Trottier, S., Vignal, J.P., Chodkiewicz, J.P., 1992. Surgical therapy for frontal epilepsies. *Adv. Neurology* 57, 707–732.
- Tang, J.R., 1993. ANFIS: adaptive network based fuzzy inference systems. *IEEE Trans. Syst. Cybern.* 23, 515–520.
- Tournier, J.-D., Calamante, F., Connelly, A., 2012. MRtrix: diffusion tractography in crossing fiber regions. *Int. J. Imaging Syst. Technol.* 22 (1), 53–66.
- Wang, H.E., Bénar, C.G., Quilichini, P.P., Friston, K.J., Jirsa, V.K., Bernard, C., 2014. A systematic framework for functional connectivity measures. *Front. Neurosci.* 8, 1–22 (December).
- Wendling, F., Bartolomei, F., Bellanger, J.J., Chauvel, P., 2001. Interpretation of interdependencies in epileptic signals using a macroscopic physiological model of the EEG. *Clin. Neurophysiol.* 112 (7), 1201–1218.
- Zadeh, L.A., 1965. Fuzzy sets. *Inf. Control* 8 (3), 338–353.
- Zalesky, A., Fornito, A., Cocchi, L., Gollo, L.L., van den Heuvel, M.P., Breakspear, M., 2016. Connectome sensitivity or specificity: which is more important? *NeuroImage* 142, 407–420.
- Zweig, M., Campbell, G., 1993. Receiver-operating characteristic (ROC) plots: a fundamental evaluation tool in clinical medicine. *Clin. Chem.* 39 (4), 561–577.