

# Systematic evaluation of fMRI data-processing pipelines for consistent functional connectomics

---

Received: 4 October 2022

---

Accepted: 10 May 2024







---

Published online: 04 June 2024

---

 Check for updates

---

Andrea I. Luppi<sup>1,2,3,4</sup> ✉, Helena M. Gellersen<sup>5,6</sup>, Zhen-Qi Liu<sup>4</sup>, Alexander R. D. Peattie<sup>1,2</sup> , Anne E. Manktelow<sup>1,2</sup> , Ram Adapa<sup>1,2</sup>, Adrian M. Owen<sup>7,8</sup>, Lorina Naci<sup>9</sup> , David K. Menon<sup>1</sup>, Stavros I. Dimitriadis<sup>10,11,12,13,14,15,16,17</sup>  & Emmanuel A. Stamatakis<sup>1,2,17</sup> 

Functional interactions between brain regions can be viewed as a network, enabling neuroscientists to investigate brain function through network science. Here, we systematically evaluate 768 data-processing pipelines for network reconstruction from resting-state functional MRI, evaluating the effect of brain parcellation, connectivity definition, and global signal regression. Our criteria seek pipelines that minimise motion confounds and spurious test-retest discrepancies of network topology, while being sensitive to both inter-subject differences and experimental effects of interest. We reveal vast and systematic variability across pipelines' suitability for functional connectomics. Inappropriate choice of data-processing pipeline can produce results that are not only misleading, but systematically so, with the majority of pipelines failing at least one criterion. However, a set of optimal pipelines consistently satisfy all criteria across different datasets, spanning minutes, weeks, and months. We provide a full breakdown of each pipeline's performance across criteria and datasets, to inform future best practices in functional connectomics.

The human brain is a remarkably complex system, comprising a large number of regions interacting over time. To address this challenge and obtain insights about distributed brain function and dysfunction, neuroscientists have turned to network science, whereby different parts of the brain can be viewed as nodes in a network, and the

statistical relationships between them are used to represent connections between nodes<sup>1-5</sup>. This powerful approach uses graph theory to quantify key aspects of brain network organisation in vivo, illuminating the neurobiological underpinnings of healthy and pathological cognition, behaviour and individual differences<sup>6-12</sup>. In

---

<sup>1</sup>Division of Anaesthesia, University of Cambridge, Cambridge, UK. <sup>2</sup>Department of Clinical Neurosciences, University of Cambridge, Cambridge, UK. <sup>3</sup>St John's College, University of Cambridge, Cambridge, UK. <sup>4</sup>Montreal Neurological Institute, McGill University, Montreal, Canada. <sup>5</sup>German Center for Neurodegenerative Diseases, Magdeburg, Germany. <sup>6</sup>Department of Psychology, University of Cambridge, Cambridge, UK. <sup>7</sup>Department of Psychology, Western Institute for Neuroscience (WIN), Western University, London, ON, Canada. <sup>8</sup>Department of Physiology and Pharmacology, Western Institute for Neuroscience (WIN), Western University, London, ON, Canada. <sup>9</sup>Trinity College Institute of Neuroscience, School of Psychology, Trinity College Dublin, Dublin, Ireland. <sup>10</sup>Department of Clinical Psychology and Psychobiology, University of Barcelona, Barcelona, Spain. <sup>11</sup>Institut de Neurociències, University of Barcelona, Barcelona, Spain. <sup>12</sup>Neuroinformatics Group, Cardiff University Brain Research Imaging Centre (CUBRIC), School of Psychology, College of Biomedical and Life Sciences, Cardiff, Wales, UK. <sup>13</sup>Division of Psychological Medicine and Clinical Neurosciences, School of Medicine, College of Biomedical and Life Sciences, Cardiff University, Cardiff, Wales, UK. <sup>14</sup>Neuroscience and Mental Health Research Institute, School of Medicine, College of Biomedical and Life Sciences, Cardiff University, Cardiff, Wales, UK. <sup>15</sup>MRC Centre for Neuropsychiatric Genetics and Genomics, School of Medicine, College of Biomedical and Life Sciences, Cardiff University, Cardiff, Wales, UK. <sup>16</sup>Integrative Neuroimaging Lab, Thessaloniki, Greece. <sup>17</sup>These authors jointly supervised this work: Stavros I. Dimitriadis, Emmanuel A. Stamatakis. ✉ e-mail: [al857@cam.ac.uk](mailto:al857@cam.ac.uk)

particular, resting-state functional MRI (rs-fMRI) is a very popular imaging tool, due to its excellent spatial resolution and wide applicability<sup>13</sup>: being task-free, it can be easily administered even to challenging populations, from in-utero foetuses<sup>14</sup> to severely injured and even unconscious patients<sup>15–17</sup>. Indeed, aberrant functional connectivity patterns have been observed in many neurological and psychiatric conditions<sup>18–22</sup>.

However, recent studies have highlighted how different analysis workflows can lead to sometimes drastically different conclusions about the same neuroimaging dataset<sup>23</sup>, owing to a vast pool of possible methodological choices which effectively constitute a combinatorial explosion problem<sup>24</sup>. Crucially, such a combinatorial explosion also plagues network analyses of the human brain: even beyond substantial differences introduced by data preprocessing and denoising procedures<sup>25,26</sup>, a wide variety of approaches have been proposed to derive brain networks from preprocessed functional neuroimaging data<sup>27,28</sup>. The very definition of nodes in brain networks is controversial: although fMRI voxels have no intrinsic biological meaning, it is well-established based on both functional involvement and lesion studies but also cellular, molecular, and fibre architecture that the brain exhibits biologically meaningful regional organisation, such that voxels can be grouped together into anatomically distinct areas<sup>24,29–31</sup>. However, there is yet no consensus on the most appropriate parcellation of the human brain, or the number and spatial extent of brain regions, or whether they should be discrete or overlapping, spatially contiguous or discontinuous<sup>24,30,32–34</sup>. Similar difficulties arise for the definition of functional connections (edges) between nodes: how to quantify them, which ones to retain for analysis, and whether to emphasise the presence/absence of connections (binary network) or their relative strength (weighted network)<sup>21,27</sup>, highlighting the intricacies of this issue<sup>32</sup>. This challenge has practical consequences: even with high-quality data, a poor choice of network construction pipeline may produce misleading conclusions about neurobiology and functional organisation, and possibly misinform biomarker discovery and clinical practice. Thus, to ensure the value of graph-based estimates as clinical biomarkers, it is of paramount priority to establish what is the most appropriate way to construct a functional brain network from rs-fMRI data.

Reliability of network topology is of fundamental importance for any subsequent analysis of network properties<sup>35</sup>: any pipelines that recover vastly different topologies from two scans of the same individual taken within the same hour, are liable to produce misleading results when used to associate network properties with behavioural traits<sup>13</sup> or clinical outcomes<sup>36</sup>. Thus, identification of reliable network construction pipelines represents a fundamental prerequisite for both network-based investigation of individual differences using functional neuroimaging<sup>37,38</sup> and subsequent efforts aimed at clinical translation<sup>39</sup>. Existing scientific work comparing different network construction steps typically focused on specific global or local network properties (e.g., modularity, small-world character, global or local efficiency, down to individual edges<sup>35</sup>) and evaluated the different alternatives by maximising the intra-class correlation of the adopted global or local network properties<sup>30,35,40–50</sup>.

However, these approaches both have limitations. On the one hand, focusing on local aspects (individual edges, node-level properties) runs the risk of “missing the forest for the trees”<sup>46</sup>, because networks are more than just collections of edges: rather, the way that edges are organised gives rise to micro-, meso- and macro-scale structure, which is precisely what makes network-based approaches so powerful. On the other hand, focusing on specific high-level properties of the network will inevitably limit the generalisability of results, because a vast and ever-growing array of network properties can be defined and used to obtain insights about brain function<sup>51,52</sup>, but there is no guarantee that recommendations pertaining to one will also apply to others.

In the present study, we introduce a framework to explicitly address and tame the combinatorial explosion. First, we evaluate network construction pipelines end-to-end, rather than restricting our attention to individual steps in isolation, as most previous studies have done. Second, we base our evaluation on the network’s topology, that is, the network’s organisation as a whole. For this purpose, we take advantage of the recently introduced “Portrait divergence” (PDIV) measure of dissimilarity between networks<sup>53</sup>. This information-theoretic measure simultaneously takes into account all scales of organisation within a network, from local structure to motifs to large-scale connectivity patterns. Therefore, it incorporates all aspects of network topology, enabling us to go beyond the use of specific and arbitrarily-chosen graph-theoretical properties.

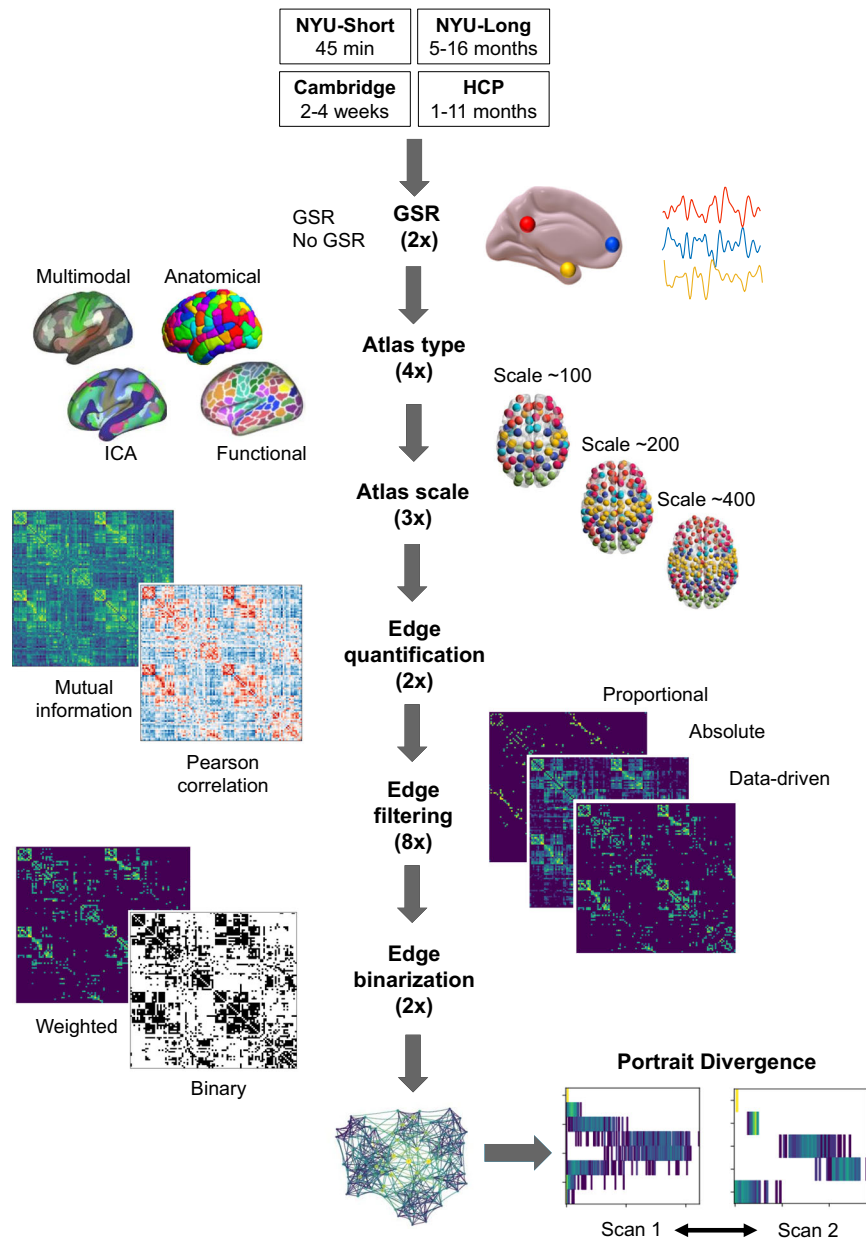
Third, test–retest reliability is a necessary but arguably not sufficient condition for a pipeline to be suitable for functional connectomics<sup>54</sup>. In particular, test–retest reliability would be of limited value if it were driven simply by constant but unimportant features. Rather, a suitable pipeline should also be able to detect meaningful experimental differences, when such exist. Therefore, we seek to identify network construction pipelines that minimise spurious (noise- or motion-induced) differences between brain networks of the same individual across repeated scan sessions, but that also satisfy additional criteria of biological relevance: sensitivity to individual differences, clinical contrasts of interest and experimental manipulations – here operationalised by pharmacological intervention with the general anaesthetic propofol. Fourth, to ensure the generalisability of our results<sup>36</sup>, each pipeline is evaluated across two independent test–retest datasets, spanning short (45 minutes), medium (2–4 weeks) and long-term delays (5–16 months). Our focus here is not on preprocessing/denoising approaches to fMRI data (where a vast literature exists<sup>55–58</sup>), but rather on the workflow that begins with preprocessed fMRI data and results in a brain network. However, to ensure that our recommendations can be further generalised to datasets acquired with different scanning parameters and preprocessed with different methods, we also require that optimal pipelines should meet all the above-mentioned criteria in an additional independent dataset (test–retest dataset from the Human Connectome Project), which was acquired with higher spatial (2 mm) and temporal resolution (TR = 0.72 s) than the other datasets; preprocessed using a surface-based rather than volume-based workflow; and denoised with a different method than the anatomical CompCor used for our main datasets (FIX-ICA, which is designed to affect artefacts specifically and avoid modifying the neural signal of interest)<sup>55,59,60</sup>.

Through this comprehensive, multi-criterion approach, we compare the topologies of functional brain networks obtained from systematic combinations of different options at each step in the network construction process. (i) First, given our interest in robustness and generalisability, we conduct all our analyses on two versions of the same data: with versus without the controversial preprocessing step of global signal regression (GSR)<sup>61</sup>. This allows us to make recommendations that are specific for GSR-processed data, or for non-GSR-processed data, as well as identifying network processing pipelines that are suitable for both. (ii) Definition of network nodes: from discrete parcellations of spatially circumscribed regions-of-interest based on anatomical landmarks, or functional characteristics (combination of local homogeneity and global gradients of connectivity, from combined resting-state and task-based fMRI), or multimodal structural and functional MRI features accounting for cortical myeloarchitecture, functional activation, connectivity and topography; or from continuous, spatially overlapping maps from spatial Independent Components Analysis<sup>24,29</sup>. (iii) Number of nodes: approximately 100, 200 or 300–400, for each type of parcellation. (iv) Two different ways to define network edges from BOLD time-series: Pearson correlation or mutual information. (v) Eight different approaches to filter the network’s edges: by

imposing a pre-specified density (retaining 5%, 10%, or 20% of total edges, or matching the density of the structural connectome), or imposing a pre-specified minimum edge weight (0.3 or 0.5), or using data-driven methods (Efficiency Cost Optimisation, ECO and Orthogonal Minimum Spanning Trees, OMST, two different strategies to define and then optimise the balance between network efficiency and wiring cost)<sup>62–64</sup>. (vi) Use of either binary or weighted networks. Figure 1 illustrates the set of choices across the investigated network construction steps that influence the construction of

a functional brain network, yielding a total set of 768 pipelines ( $2 \times 4 \times 3 \times 2 \times 8 \times 2$ ). We assess each of these pipelines across 18 distinct combinations of criteria and datasets, yielding a total of 13,824 unique evaluations, which we make available to the reader through an interactive Pipeline Selection Tool (Supplementary Data 2).

Overall, a strength of our current study is our ability to make recommendations for the choice of pipelines end-to-end, not only on the basis of theoretical gold standard metrics (test–retest) but also on



**Fig. 1 | Overview of the steps to turn functional MRI data into a network.**

Starting from preprocessed and denoised data, the following steps are involved. (i) Use of data with vs without global signal regression (GSR), in addition to other denoising protocol (aCompCor for NYU-short, NYU-long and Cambridge datasets; FIX-ICA for HCP); (ii) Definition of nodes (based on anatomical features, local and global functional characteristics, or multimodal features; or Independent Components Analysis); (iii) Choice of number of nodes (approximately 100, 200, or 400); (iv) Definition of connectivity measure (from Pearson correlation or mutual information); (v) Choice of edges to retain (8 filtering schemes considered, based on a priori choices of network density, or minimum edge weight, or data-driven strategies to optimise the balance between network efficiency and wiring cost); (vi) Use

of binary or weighted edges. In total, we consider  $2 \times 4 \times 3 \times 2 \times 8 \times 2 = 768$  unique pipelines. For each pipeline, the resulting functional networks are compared for the same subject across different time-spans (minutes, weeks, or months) using the Portrait Divergence. A network portrait for a binary network is a matrix  $B$  whose rows each correspond to a histogram obtained by thresholding the matrix of shortest paths between the networks's constituent nodes, at each path length  $l$  between 0 and the network's diameter  $L$ , such that entry  $B_{l,k}$  encodes the number of nodes that have  $k$  nodes at distance  $l$ . For weighted networks, the histogram is obtained by binning (see Methods). Illustration of parcellations adapted from refs. 34 and 44; illustration of Portrait Divergence adapted from ref. 53.

the basis of practical relevance: meaningful inference about changes in brain functional network topology and individual differences, and robustness and generalisability. To anticipate our main findings, we discovered large and systematic variability among pipelines' ability to recover a reliable network topology, with the majority of pipelines failing to meet at least one criterion. Choice of an inappropriate network construction pipeline can lead to results that are not only misleading (statistically significant in the opposite direction as the true effect), but replicably so (being observed in two independent datasets). While our results show that an uninformed choice of pipeline will likely be suboptimal, we also identified a number of pipelines that satisfy all our criteria, in all four test–retest comparisons, making them suitable candidates for functional connectomics and biomarker discovery. Through this multi-dataset, multi-criteria, multi-scale and multi-step approach, we provide a comprehensive set of benchmarks for trustworthy functional connectomics.

## Results

We used an information-theoretic measure of distance between network topologies across scales, termed Portrait divergence (PDiv), to systematically compare 768 alternative network construction pipelines in terms of their ability to recover similar brain network topologies from functional MRI scans of the same individual across minutes (NYU dataset, same-session scans), weeks (Cambridge dataset), or months (NYU dataset, between-sessions comparison) (see "Methods" section and Figs. S1–2 for examples of network portraits and their divergence). Additionally, we considered an additional dataset (HCP test–retest) that was acquired with higher spatial (2 mm isotropic) and temporal resolution (0.72 s TR); with longer duration (1200 volumes); denoised using FIX-ICA instead of aCompCor; and parcellated on the surface rather than in volumetric space, as for the other datasets<sup>65–67</sup>. Our end-to-end approach allowed us to simultaneously assess the effects of parcellation type and number of nodes; connectivity quantification, thresholding and binarisation; and global signal regression; while ensuring robustness to aspects such as acquisition, time between test and retest and denoising method.

Being grounded in information theory, the Portrait divergence between two networks can be interpreted as measuring how much information is lost when using one network to represent another: it ranges from 0 (no information loss) to 1 (complete information loss)<sup>53</sup>.

To identify suitable pipelines, we required each of the following criteria to be met:

- Criterion (I): Avoiding spurious differences ("PDiv ranking"). Since the two networks that we consider are derived from different scans of the same healthy individuals under conditions in which no experimentally meaningful changes in functional network topology are expected, we aim to identify pipelines that minimise test–retest PDiv. We consider pipelines as candidates for optimal if they are in the top 20% in terms of the average PDiv rank calculated across all test–retest intervals.
- Criterion (II): Detecting true experimental differences ("propofol"). Suitable pipelines should detect a significant effect for propofol, in the right direction, in both propofol datasets, i.e., a pipeline is excluded if it fails to detect the expected effect (greater change between wakefulness and anaesthesia than between two awake scans) in either of the two propofol datasets.
- Criterion (III): Detecting inter-individual differences ("within-between"). A pipeline fails this criterion if the resulting networks are more similar between than within subjects more than 50% of the times, for any of the four test–retest datasets.
- Criterion (IV): Avoiding motion-induced differences ("motion"). A pipeline fails this criterion if its PDiv has a significant correlation with differences in head motion in any of the four test–retest datasets.

- Criterion (V): Non-empty networks. As a final check, we also exclude any pipelines that remove all connections from a network, in any of the four test–retest datasets.

These criteria also incorporate the need for recommendations to be generalisable across datasets and acquisition/preprocessing choices, since we only consider a criterion to be met if it is met in all the relevant datasets.

A summary of all pipeline characteristics can be found in the interactive Pipeline Selection Tool (Supplementary Data 2). We provide an Excel spreadsheet with an interactive table, including filters that allow selection based on multiple criteria at once to identify pipelines that adhere to the specific criteria desired by the reader. We encourage readers to view the interactive table concurrently with the results described below, as this will allow a closer inspection of associations between a pipeline's specific network processing choices and the desirable properties described in each subsection of the Results. A user guide for the interactive Pipeline Selection Tool (Supplementary Data 2) is also included in the Supplementary Material.

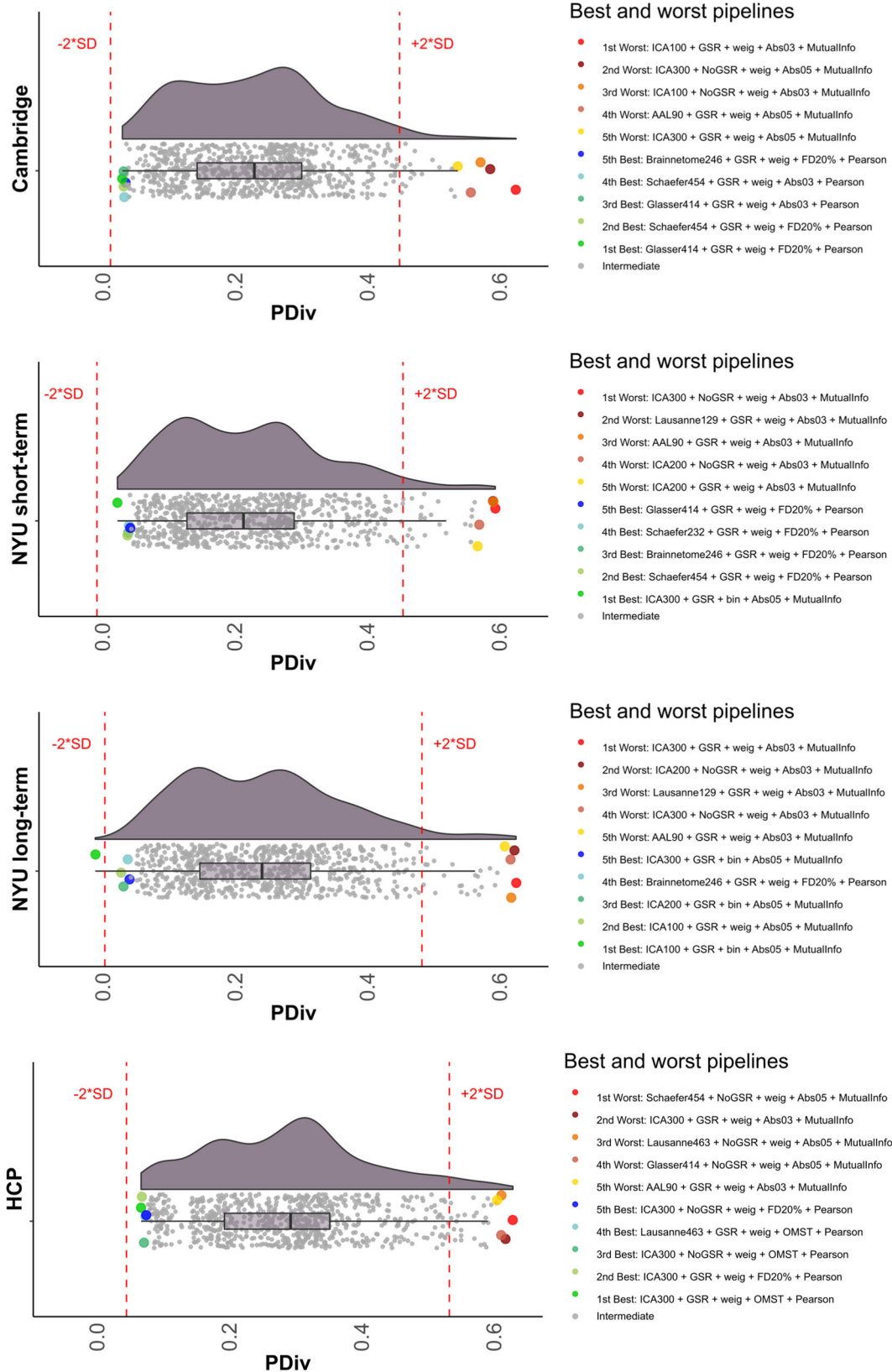
### Portrait Divergence identifies drastic and systematic variability across pipelines' capacity to avoid spurious differences

For each dataset, Fig. 2 illustrates the distributions of group-mean test–retest similarities of network topologies (portrait divergence) across the full set of 768 pipelines (See Fig. S3–30 for the distribution of PDiv across pipelines, broken down by network construction step, for each dataset). Clearly, two patterns can be observed. First, network construction pipelines differ widely in how well they are able to recover the same network topology across different scans of the same individual, on average - whether on a timescale of minutes, weeks, or months. The worst pipelines induce a greater than five-fold increase in topological dissimilarity (PDiv) between functional connectomes of the same individual, compared with the best-performing ones.

Second, our results indicate high consistency across the four test–retest comparisons considered here, in terms of which data-processing steps feature prominently among the pipelines that are best (and worst) at minimising the average within-subject PDiv. Correlation between all pipelines' ranks across time intervals revealed very high consistency between all datasets (Spearman's  $\rho$  ranging from 0.73 to 0.97, all  $p = 2.2 \times 10^{-16}$  (Fig. 3), indicating that pipelines' suitability for network construction is not dataset-specific but rather can generalise to independent groups of individuals - spanning time intervals from hours to months. We view a small PDiv in these datasets as a desirable property: even though learning and plasticity could account for some amount of connectome reorganisation over weeks or months in healthy adults, such factors cannot plausibly be expected to be the cause of any network-wide reorganisation observed within the course of a single hour (in the absence of any intervention), which should instead be treated as unwanted noise.

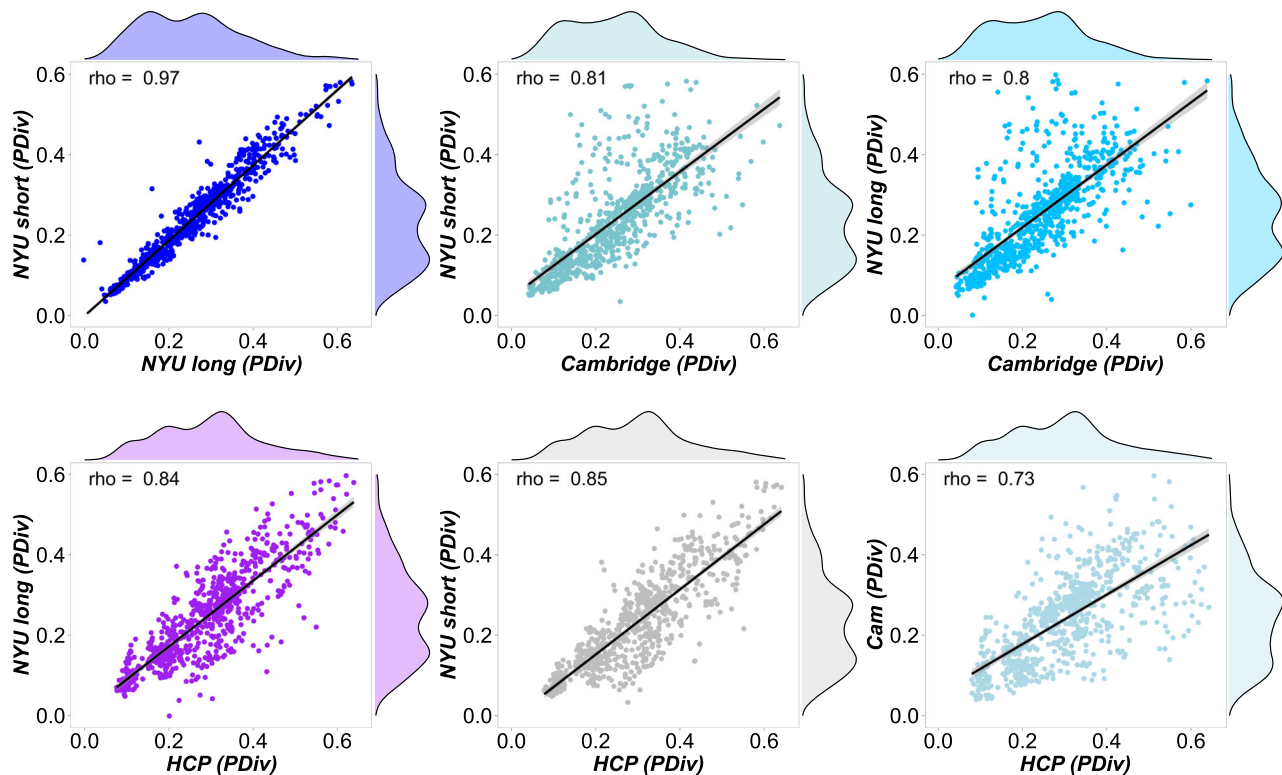
### Sensitivity to experimental differences: Low-PDiv pipelines are more likely to detect pharmacologically-induced connectome reorganisation

We have shown that network construction pipelines vary drastically and systematically in their robustness to noise-induced changes in the functional connectomes of the same individuals scanned multiple times. However, this minimisation of noise-induced differences should not come at the expense of also minimising meaningful changes in network topology, such as control-patient contrasts (an example of this would be a pipeline that never detects any changes). Rather, a good pipeline should simultaneously minimise noise-induced differences, while remaining sensitive to true ones. In other words, test–retest reliability is not the only criterion that neuroscientists need to consider for their choice of network construction pipelines:



**Fig. 2 | Distribution of group-average portrait divergence values for each of 768 alternative network construction pipelines, across different time intervals.** From top to bottom: Cambridge dataset (rescan within 2–4 weeks). NYU short-term dataset (rescan within 45 minutes). NYU long-term dataset (rescan within 16 months; average 11.4); HCP dataset (rescan 1–11 months). Right-side: highlighting the top 5 (lowest PDiv) and bottom 5 performers (highest PDiv). Each data-point

represents one pipeline ( $n = 768$ ). Red lines mark 2 standard deviations from the mean of the distribution. Box plot centre line, median; box limits, upper and lower quartiles; whiskers, 1.5x interquartile range. GSR Global Signal Regression, OMST Orthogonal Minimal Spanning Trees, PDiv Portrait Divergence. Source data are provided as a Source Data file.



**Fig. 3 | Rank-based correlations of the pipelines' performance across datasets.** PDIV portrait divergence, HCP Human Connectome Project data, NYU New York University dataset. All  $p = 2.2 \times 10^{-16}$  (two-sided) from Spearman correlation. Each

data-point represents one pipeline ( $n = 768$ ). Shading indicates standard error of the fitted line to model the linear relationship between the two respective variables. Source data are provided as a Source Data file.

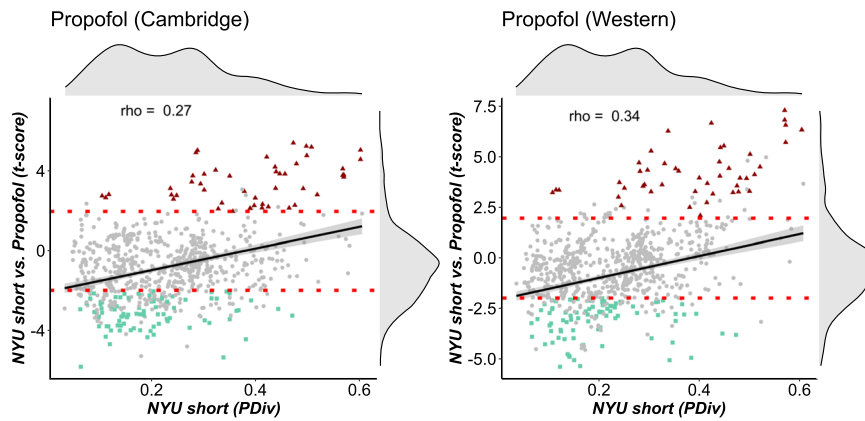
ultimately, the resulting networks need to also demonstrate empirical usefulness by providing neurobiologically meaningful results<sup>43,44</sup>. An ideal pipeline would therefore strike a balance between sensitivity to experimental manipulations or contrasts of interest on the one hand, and low portrait divergence in test–retest over relatively short periods of time in healthy individuals and under the same test conditions on the other hand. Therefore, in addition to identifying pipelines that do not detect differences when we know that there should be none or only minor ones (best exemplified by test–retest scanning within the same hour), we should find pipelines that can also detect a difference, when we know that a difference must be present: we need to combine a low rate of false positives (low test–retest PDIV) with a low rate of false negatives.

Perhaps the most drastic possible difference that can be induced between two scans of the same individual, is that between consciousness and unconsciousness. General anaesthetics such as the intravenous agent propofol can rapidly and reversibly induce a state of unconsciousness, whereby the subject is behaviourally unresponsive and has no subjective experiences. There is arguably no short-term, reversible alteration of the mind that is so all-encompassing, and it cannot be expected to leave the functional connectome unaltered. Therefore, if a pipeline is unable to detect anaesthetic-induced differences in the topology of the functional connectome, we can reasonably conclude that it is not sensitive enough for use in network neuroscience.

Following this rationale, we compared the PDIV from the NYU-short dataset (two scans within the same hour) against the PDIV observed between an awake rs-fMRI scan, and a second scan of the same individuals while under propofol-induced general anaesthesia (also acquired within the same visit). We seek to identify pipelines that produce significantly greater PDIV between an awake and an anaesthetised scan of the same individual, than between two awake scans acquired at a comparable distance in time. To ensure the reliability of

our approach, we repeat this analysis for two independent datasets of propofol anaesthesia to further bolster the reproducibility and generalisability of our findings.

Across both datasets, our results suggest that pipelines with lower PDIV also tend to have t-scores reflective of the expected effect of propofol (Fig. 4), as demonstrated by significant correlations between short-term test–retest PDIV (based on the NYU dataset) and t-scores both for the Western ( $\rho = 0.34$ ,  $p = 2.2 \times 10^{-16}$ ) and the Cambridge propofol datasets ( $\rho = 0.27$ ,  $p = 4.4 \times 10^{-14}$ ). As control test–retest PDIV becomes larger, t-scores also seem to become more variable. Reassuringly, we identified multiple pipelines that provide the expected effect in both datasets (Fig. 4, green dots). Intriguingly, however, we also identified a number of large-PDIV pipelines that detect a statistically significant difference between test–retest and anaesthesia, but in the opposite direction: that is, greater connectome reorganisation between two awake scans, than between an awake and an anaesthetised scan (Fig. 4, red triangles). In other words, these pipelines produce networks that are actively misleading about what we have strong reason to believe must be the ground truth (because there is a very substantial difference introduced by anaesthesia, reflected in the suspension of the brain's input-processing abilities and cognitive function more broadly). These pipelines can be found in the interactive Pipeline Selection Tool (Supplementary Data 2; pipelines labelled “Opposite” in the columns Status Propofol West and Status Propofol Cam). Worryingly, we find that a non-negligible number of pipelines (38) produce the opposite of the expected effect for both propofol datasets - thereby returning results that are systematically misleading, and highlighting the dangers of an inappropriate choice of network construction workflow. Of note, all the consistently misleading pipelines use an absolute threshold; all but three use weighted edges; and 24/47 use mutual information to quantify connectivity. Overall, 85 pipelines show the expected effect for both



**Fig. 4 | Correlation between low PDiv and ability to detect significant difference between anaesthesia and test-retest.** Left: Cambridge anaesthesia dataset (Spearman's  $\rho = 0.27$ ,  $p = 4.4 \times 10^{-14}$ , two-sided). Right: western anaesthesia dataset (Spearman's  $\rho = 0.34$ ,  $p = 2.2 \times 10^{-16}$ , two-sided). The  $t$ -scores are obtained from permutation-based two-sample  $t$ -tests comparing PDiv from test-retest NYU short, against PDiv from awake vs anaesthesia. Horizontal red lines indicate  $t \pm 1.96$  from two-sample  $t$ -test (two-sided), corresponding to a statistically significant difference

between the two groups' mean, with negative  $t$ -scores corresponding to PDiv (anaesthesia) > PDiv (test-retest). Each data-point represents one pipeline ( $n = 768$ ). Green dots indicate pipelines that produce the expected effect in both datasets. Red triangles indicate pipelines that produce a misleading effect in both datasets. PDiv Portrait Divergence. Shading indicates standard error of the fitted line to model the linear relationship between the two respective variables. Source data are provided as a Source Data file.

propofol datasets, thereby satisfying this criterion, whereas 45 pipelines are neutral (failing to detect statistically significant differences in both propofol datasets).

### Sensitivity to inter-individual differences

Another means by which the adequacy of a pipeline may be assessed is by comparing PDiv within subjects (scan 1 vs. scan 2 for subject 1, scan 1 vs. scan 2 for subject 2, etc...) and PDiv between subjects (subject 1 vs. subject 2, etc...). The proportion of participants for whom the within-subjects (WS) PDiv is smaller than between-subjects (BS) PDiv may be used as an additional criterion of pipeline quality, with the rationale that even after accounting for bona fide changes due to plasticity and learning, an individual's functional connectome should not differ from itself at another point in time, more than it differs from the connectomes of other individuals.

Our results suggest that pipelines with smaller PDiv are also better at producing networks that are sensitive to individual differences, such that the same subject's brain network diverges less from the same subject's network than from those of other people. This was the case for the NYU short test-retest data ( $\rho = -0.31$ ,  $p = 2.2 \times 10^{-16}$ ), the medium-term test-retest time interval for the Cambridge dataset ( $\rho = -0.44$ ,  $p = 2.2 \times 10^{-16}$ ), the NYU long test-retest data ( $\rho = -0.28$ ,  $p = 2.2 \times 10^{-16}$ ) and the HCP dataset ( $\rho = -0.51$ ,  $p = 8.08 \times 10^{-9}$ ; Fig. 5).

Passing and failing pipelines on the basis of this within-between criterion can be found in the interactive Pipeline Selection Tool (Supplementary Data 2; column "Criterion within-between all"). The Pipeline Selection Tool also lists the proportion of participants in a given pipeline for which within-subject PDiv is smaller than between-subjects PDiv in columns Within-between Cam (%), Within-between NYU short (%), Within-between NYU long (%) and Within-between HCP (%). In the Cambridge dataset, 72 pipelines were excluded based on this criterion. This was the case for 62 in the NYU short-term test-retest data and for 67 for the NYU long-term data as well as for 39 pipelines in the HCP data. In total, on the basis of the overall within-between criterion across datasets, 159 unique pipelines were excluded and 609 were retained.

### Avoiding motion confound

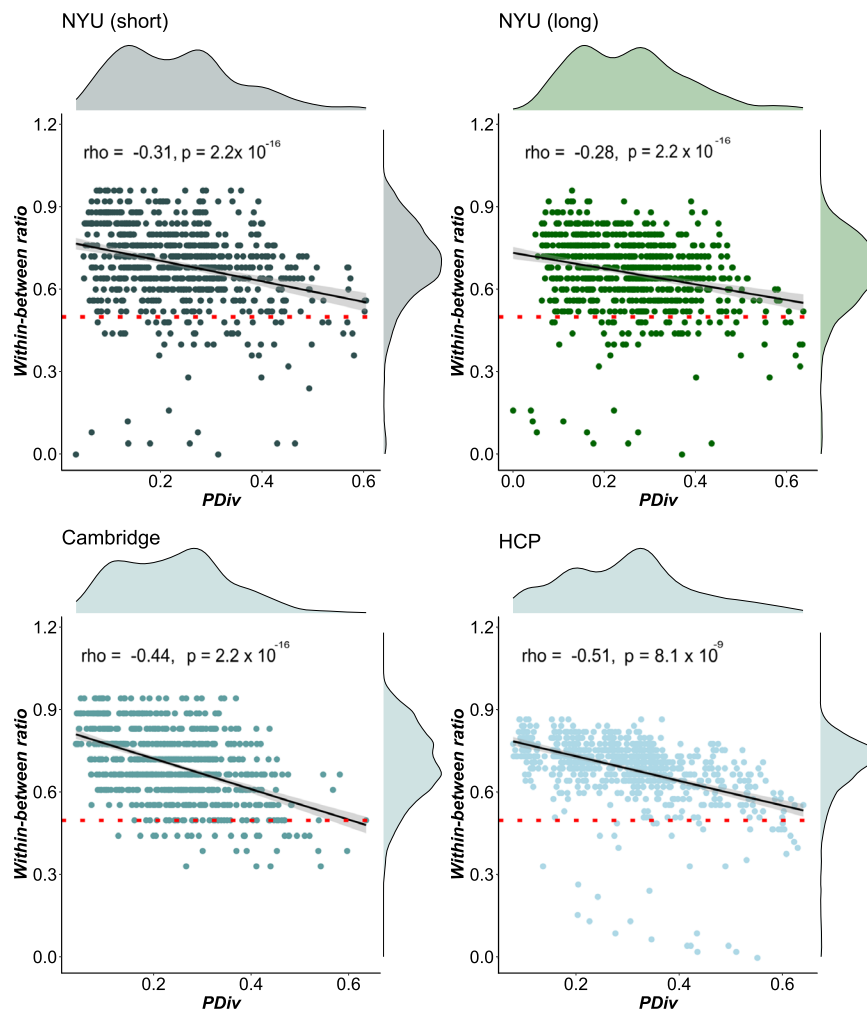
As a further criterion, we sought to identify and exclude pipelines whose PDiv is significantly correlated with differences in subject motion (mean framewise displacement). For the Cambridge dataset,

45 pipelines showed a significant correlation between PDiv and motion (magnitude of the Spearman correlation coefficient  $\rho$  ranging between 0.60 and  $-0.67$ ). For the NYU short-term dataset, 17 pipelines exhibited a significant correlation between PDiv and motion (magnitude of the correlation ranging between 0.48 and  $-0.57$ ). For the NYU long-term dataset, 24 pipelines exhibited a significant correlation between PDiv and motion (magnitude of the correlation ranging between  $-0.53$  and 0.56). Finally, for the HCP dataset we found that PDiv and motion were correlated significantly in 59 pipelines ( $\rho$  between  $-0.38$  and 0.53).

It is argued in the literature that GSR can help to mitigate the noise induced by subject motion<sup>42,45</sup>. When contrasting all pipelines with GSR against those without GSR, no significant difference in the strength of the correlation (absolute  $\rho$ -statistic) between PDiv and motion based on this option was found in the Cambridge ( $t(752) = 0.03$ ,  $p = 0.973$ ,  $d = 0$ ), the NYU short test-retest ( $t(700) = 1.42$ ,  $p = 0.157$ ,  $d = 0.11$ ) or the NYU long-term test-retest ( $t(700) = 0.75$ ,  $p = 0.453$ ,  $d = 0.06$ ). That is, whether GSR was or was not applied, this decision had no bearing on the degree to which test-retest portrait divergence was associated with motion, on average across all pipelines. However, in the HCP data, there was a small but significant effect of GSR on the magnitude of the correlation between motion and PDiv ( $t(684) = -3.89$ ,  $p = 1.09 \times 10^{-4}$ ,  $d = -0.29$ ), showing a stronger association between PDiv and motion in pipelines without GSR than with GSR.

### Avoiding empty networks

Pipelines employing an a priori threshold on the strength of edges, rather than on their density (i.e., removing all edges with weight below a pre-specified value, also known as an "absolute" threshold) run the risk of removing all edges in the network, if none surpass the threshold value. This would be unquestionably incorrect, but it is conceivable that such an occurrence might never materialise in practice. Indeed, we found that this never occurred when edge weights were defined in terms of Pearson correlation. However, empty networks were returned for at least one subject by a total of 68 unique pipelines employing mutual information for edge weight definition (50 occurrences in the NYU short test-retest, 52 occurrences in the NYU long test-retest, 68 in the HCP dataset, eight in the Cambridge dataset). As expected, all of these pipelines used absolute threshold values: mostly with the 0.5 threshold, but for 20 pipelines this was also the case for the more lenient 0.3 threshold (reported in the interactive Pipeline Selection



**Fig. 5 | PDIV within versus between individuals.** Pipeline PDIV (portrait divergence) in a given dataset is plotted against the proportion of participants in the same dataset for whom the within-subject PDIV (baseline vs follow up) is smaller than between-subject PDIV. Each data-point represents one pipeline ( $n = 768$ ). Pipelines above the red line meet the within-between criterion such that portrait

divergence is smaller for within-subject test-retest compared to between subject comparisons. Shading indicates standard error of the fitted line to model the linear relationship between the two respective variables. Spearman correlation coefficient (two-sided) is used to assess the strength of relationship and its statistical significance. Source data are provided as a Source Data file.

Tool [Supplementary Data 2] under the column Criterion edge failure). Therefore, any pipeline which removes all edges in any one dataset is excluded from further consideration as a suitable candidate. However, note that pipelines that fail this check would also be eliminated from consideration based on the other four criteria: only one of those that failed the criterion of avoiding empty networks satisfied both the within-between and propofol criteria (Lausanne129 + No GSR + binarisation + Abs 0.3 + Mutual Info).

### Overall recommendations for network construction pipelines

As a final step, we combined all the criteria identified above:

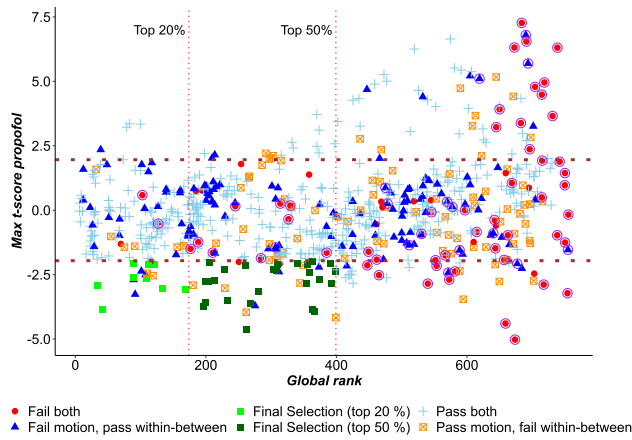
- (I) Avoiding spurious differences: we operationalise this as having low PDIV (pipelines with the average global rank in the top 20%, as calculated from the average of independent rankings within each dataset; 154 pipelines fulfilled this criterion);
- (II) Detecting true experimental differences: ability to correctly identify statistically greater PDIV in anaesthesia than test-retest, across both propofol datasets (85 pipelines passed);
- (III) Sensitivity to inter-individual differences: ability to detect smaller within- than between-subjects PDIV in at least 50% of subjects, in each of the four test-retest datasets (609 pipelines passed);

- (IV) Avoiding motion confounds: no significant correlation between PDIV and subject motion, in any of the four test-retest datasets (566 passed);
- (V) Non-empty networks: we rejected pipelines that produce empty networks for any subject in any of the four test-retest datasets (700 pipelines pass).

Out of the full set of 768 pipelines considered here, we found that only 9 (~1%) jointly satisfied all of our criteria in each of the test-retest datasets that we considered – meaning that the vast majority of pipelines (759 out of 768) may be less than optimal (Fig. 6 and Supplementary Data 1). However, 71 pipelines were excluded from the optimal ones because they each failed one single criterion in one single dataset, such that their failures were neither systematic nor pervasive. In particular, the set of optimal pipelines would expand to 35 (~5% of the total) if a less stringent criterion for the PDIV were adopted, such that all pipelines in the upper 50% were admissible (while still having to satisfy all other criteria in each of the relevant datasets).

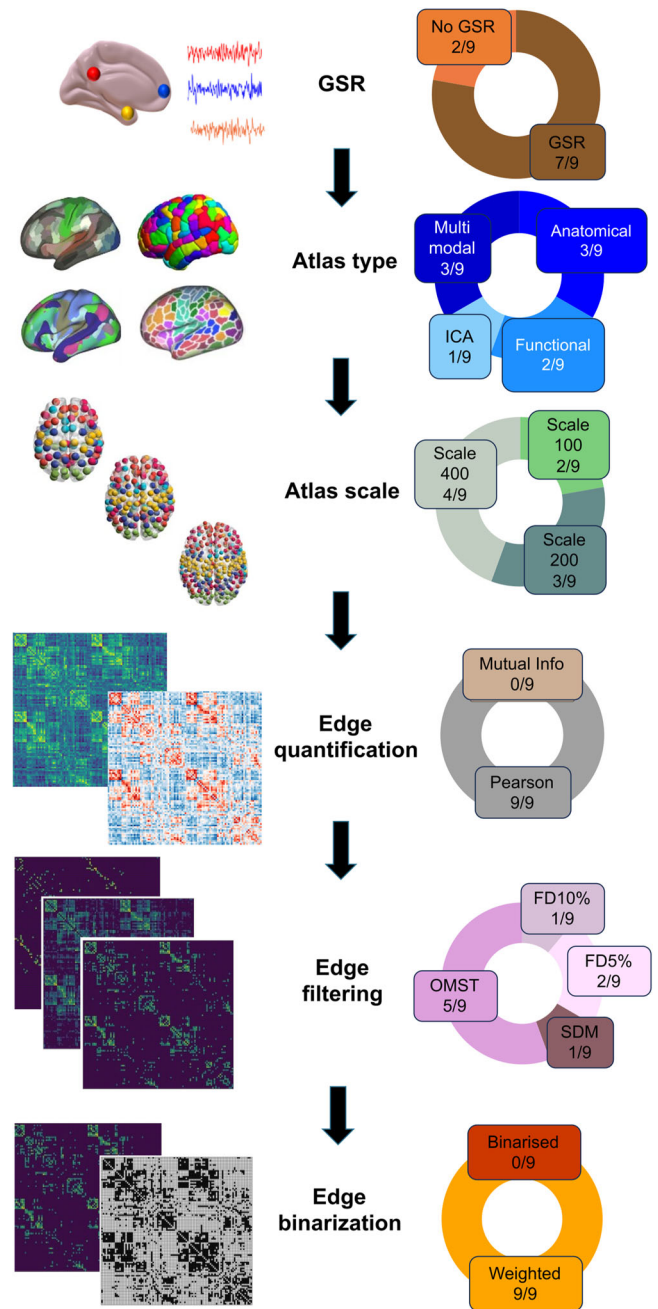
When considering the distribution of individual pipeline steps among the 9 optimal ones, three clear patterns emerge: all pipelines use weighted (rather than binary) edges, and all quantify connectivity





**Fig. 6 | Evaluating pipelines across all criteria.** Each data-point represents one pipeline ( $n = 768$ ), with colour and shape reflecting which criteria are met. Criterion (I): Avoiding spurious differences (“PDiv ranking”). We consider pipelines as optimal if they are in the top 20% in terms of the global rank based on PDiv (Portrait Divergence) calculated as the average rank achieved in each dataset. We further show pipelines which fulfil all other criteria while being among the top 50% in terms of the average global rank. The maximum average PDiv among the top 50% pipelines is 0.169. Criterion (II): Detecting true experimental differences (“propofol”). Suitable pipelines should detect a significant effect for propofol, in the right direction, in both propofol datasets, i.e. a pipeline is excluded if it fails to detect the expected effect in either of the two propofol datasets. The Y axis reports the maximum between the two t-statistics obtained for the two propofol datasets, so pipelines satisfy the sensitivity criterion if they score  $< 1.96$  on this axis (i.e., find a significant effect for propofol, in the *right direction*, in *both* propofol datasets). Criterion (III): Detecting inter-individual differences (“within-between”). A pipeline fails this criterion if the resulting networks are more similar between than within subjects more than 50% of the times, for any of the four test–retest datasets. Criterion (IV): avoiding motion-induced differences (“motion”). A pipeline fails this criterion if its PDiv has a significant correlation with differences in head motion in any of the four test–retest datasets. Criterion (V): non-empty networks. As a final check, we also exclude any pipelines that remove all connections from a network, in any of the four test–retest datasets. “Fail both” refers to pipelines failing in terms of motion and within-between criteria, while “Pass both” refers to pipelines which satisfy both of these criteria. Points circled in purple represent pipelines that produced empty networks. Overall, 9 pipelines satisfy all criteria in all datasets; this number grows to 35 if a more liberal PDiv criterion is adopted (top 50% global rank). Source data are provided as a Source Data file.

in terms of Pearson correlation (rather than mutual information) (Fig. 7). Moreover, the preferred filtering method among the optimal pipelines is the OMST, a method to optimise the balance between network efficiency and wiring cost in a data-driven manner (selected in 5/9 cases). In other words, the single combination of Pearson correlation, weighted edges and OMST accounts for 5 out of 9 optimal pipelines, despite being only one out of  $2 \times 8 \times 2 = 32$  equally likely combinations of edge quantification, thresholding and binarisation. This is highly unlikely to occur just by chance:  $p = 4 \times 10^{-5}$  for the probability of randomly choosing 9 pipelines out of 768 and having 5 or more of them belong to the same group (out of 32 possible groups), assessed with permutation testing. In contrast to the clear importance of edge definition, parcellation choice seems to have less bearing on a pipeline’s performance, but we do observe greater prevalence of pipelines using GSR than not (7 out of 9 optimal ones, and 22/35 near-optimal). We also find that while edges based on Pearson correlation still dominate under the less stringent criterion (28/35), there is now also a number of well-performing pipelines using proportional thresholds (either fixed or SDM) with binarised edges (18/35). Node type and number remain less clearly decisive: among the 35 near-optimal pipelines, every single combination of parcellation type and size is present at least once.



**Fig. 7 | Prevalence of specific network construction steps among the 8 optimal pipelines.** Pie charts demonstrate, for each network construction step, the proportion and absolute number of each option that is found among the optimal pipelines. FD fixed density, GSR global signal regression, OMST orthogonal minimal spanning tree, SDM structural density. See Fig. S31 for a version of this figure with a breakdown of the pipelines under the more liberal PDiv criterion. Illustration of parcellations adapted from refs. 34 and 44.

Overall, inspecting the whole list of optimal pipelines (Supplementary Data 1) clearly reveals that considering each pipeline step in isolation from the others does not provide the full picture. Specifically, we found that a few combinations of options account for most of the optimal pipelines (Fig. 8), with 5 out of 9 pipelines which meet all inclusion criteria using the combination of weighted edges, Pearson correlation and OMST filtering for edge definition and thresholding. These results suggest that a pipeline’s performance is not solely attributable to any specific step: rather, some combinations of steps seem to be especially favourable.



**Fig. 8 | Optimal edge processing combinations.** Pie chart displays the frequency of each combination of edge type definition, filtering, and binarisation among the 9 pipelines which fulfil all criteria for a suitable network construction pipeline. See Fig. S32 for a version of this figure with a breakdown of the pipelines under the more liberal global rank criterion, and Figs. S33–35 for a breakdown of the relationship between PDiv and commonly studied graph properties, in terms of edge quantification, binarisation, and filtering method. FD5%, fixed-density threshold at 5% density; FD10%, fixed-density threshold at 10% density. OMST Orthogonal Minimal Spanning Trees, SDM structural density matching, FD fixed-density.

### Evaluating the full range of pipeline performance

Although we defined pass-fail criteria to be able to provide clear recommendations for the reader, it is important to appreciate that the performance of each pipeline is in fact continuous for all our evaluation metrics (except the presence of empty networks). Therefore, in addition to the optimal-nonoptimal distinction presented above, we can also evaluate the full performance of each pipeline, across all datasets and criteria. This approach allows us to identify patterns of similarity between pipelines, and their main drivers (Fig. 9).

In particular, hierarchical clustering reveals that the main distinction between pipelines, in terms of ranked performance across all datasets and criteria, is between pipelines producing binary versus weighted networks. We confirmed this statistically: the mean rank of weighted pipelines is 330.59 (SD = 73.76), whereas the mean rank for binary pipelines is 400.48 (SD = 59.33),  $t(698) = 13.81$ ,  $p < 0.001$ , Hedge's  $g = 1.04$ , from independent-samples  $t$ -test (Fig. S38). Among the other network construction steps, we also found significant differences for edge type (with Pearson correlation outperforming mutual information) (Fig. S39), filtering choice (with OMST and FD20% performing best, on average; Fig. S40), and number of nodes (with a small preference for more fine-grained parcellations; Fig. S41) In contrast, no significant difference in pipelines' overall rank were found as a result of GSR use, or whether nodes are defined from an anatomical atlas, functional atlas, multimodal atlas, or Independent Components Analysis (Fig. S42–S43).

These results indicate that network construction steps vary in terms of their overall impact on pipeline performance, and are in line with our observation that optimal pipelines tend to share specific steps pertaining to edge type, filtering and binarization. However, not every combination of the individually best-performing steps is optimal, and

therefore it is not sufficient to consider individual steps in isolation: only end-to-end evaluation of full pipelines provides the full picture.

### Discussion

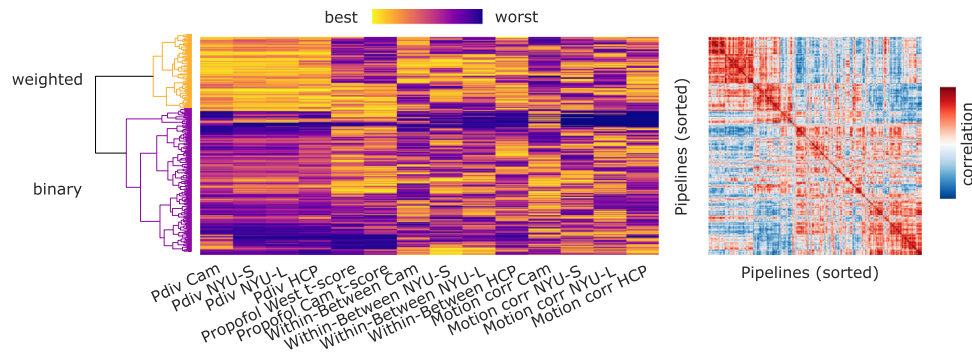
A tremendous amount of neuroimaging research with functional MRI is devoted to finding reliable functional connectomic biomarkers for brain function and its disorders – but this process involves a combinatorial explosion of arbitrary choices<sup>21,22,27</sup>.

Here, we tackled this challenge by systematically investigating 768 unique pipelines that a neuroscientist could adopt to obtain brain networks from resting-state fMRI data, arising from the combination of several key data-processing steps. To do so, we departed from most previous studies in a number of key respects. First, we explicitly addressed the combinatorial explosion, by considering pipelines end-to-end, rather than restricting our attention to specific steps. Second, rather than choosing any arbitrary local or global graph-theoretical property for our comparisons, we focused on the pipelines' ability to recover the networks' overall topology across all scales. Third, we did not focus exclusively on test–retest reliability, but rather we adopted an entire battery of criteria that any appropriate pipeline for functional connectomics should meet, in order to provide practically useful results: these include minimising both random (noise-induced) and systematic (motion-induced) topological distortions, while also being sensitive to differences between individuals and between experimental conditions. Finally, we required all criteria to be consistently met in each of several independent datasets, encompassing short (minutes), medium (weeks) and long timespans (up to 16 months), and using different spatial and temporal resolution, and different pre-processing/denoising approaches, to ensure the generalisability of our recommendations. Through this multi-dataset, multi-criteria, multi-scale and multi-step approach, our goal was to provide a comprehensive set of benchmarks for trustworthy functional connectomics.

Our first finding is that inappropriate network pipelines are ubiquitous and can produce systematically misleading results. The substantial majority of the pipelines that we considered failed to meet at least one of our criteria for consistent functional connectomics. We also observed drastic and systematic variability among pipelines' performance: an inappropriate choice of pipeline can greatly impair one's ability to recover a reliable network topology. Even for scans obtained less than 45 minutes apart, we observed up to a 5-fold increase in topological dissimilarity (PDiv) compared with the best-performing pipelines (Fig. 2), even across several months. Put differently, adoption of an inappropriate pipeline can distort the functional connectome more drastically than the passage of nearly a year – which may have far-reaching repercussions for longitudinal studies of brain network properties.

A recent review of statistical power in network neuroscience suggested that “many real effects may be missed by current studies”<sup>68</sup>. Our results are in line with this observation: we found that the vast majority (approximately 90%) of pipelines considered were unable to reliably detect the effect of general anaesthesia on the functional connectome. Thus, one potential implication of our work is that some true effects may have been missed due to a suboptimal choice of network construction pipelines for functional connectomics.

Even more worryingly, choice of the wrong pipeline can lead to results that are not only misleading (statistically significant in the opposite direction as the true effect), but replicably so (being observed in two independent propofol datasets): we found this to be the case for 38 pipelines. This means that adopting an inappropriate pipeline for network analysis can turn the replicability of results against researchers, boosting their confidence in results that are actively the opposite of the truth. Being consistently wrong rather than randomly so, these results would not be “washed out” by approaches such as meta-analytic aggregation: on the contrary, they would propagate to the meta-analysis itself. Clearly, such a scenario would have



**Fig. 9 | Clustering pipelines based on performance across criteria and datasets.**

Left: hierarchical clustering of pipelines, in terms of similarity (correlation) of their performance across datasets and criteria. The clustering solution highlights the difference between pipelines producing binary versus weighted networks. For the PDiv criterion, best performance refers to the smallest PDiv; for the propofol criterion, best performance is the greatest t-score in the correct direction; for the within-between criterion, best performance means the greatest proportion of

participants for whom the within-subjects PDiv is smaller than between-subjects PDiv; for the motion correlation criterion, best performance is identified as the smallest magnitude of correlation with motion. The empty networks criterion is not included, since it is not continuous. Overall rank is the mean across all columns. Right: correlation between each pair of pipelines in terms of performance, arranged by the same hierarchical clustering. See Fig. S37 for the same figure, but sorted by overall rank.

devastating consequences for the use of functional connectomics for biomarker identification; in the worst-case scenario, a treatment that actually makes the disease worse may be systematically mis-identified as making it better.

Finally, our results show that the above-mentioned concerns cannot be easily dismissed, because suboptimal pipelines are not a rare exception, but rather the rule: the vast majority of pipelines among those considered (759 out of 768) failed to meet at least one of our criteria (or 733 if the criterion of having low PDiv on average is relaxed). In other words, our results clearly demonstrate that even when combining steps for network construction that are individually sensible, it is overwhelmingly likely (over 98%, in our sample of pipelines) that the resulting overall pipeline will *not* be appropriate for functional connectomics – at least not optimal. Indeed, we find that no single step uniquely determines a pipeline’s ability (or inability) to accurately recover the network’s topology: pipelines differing by only one step are largely overlapping in terms of their portrait divergence distributions (Figs. S7–S30). These observations highlight the importance of focusing on entire pipelines as we do here, in contrast to most published approaches that typically consider only one or two steps in isolation. This approach enabled us to identify some pipelines (i.e., specific combinations of steps) that successfully pass all our criteria in every single one of our datasets.

Fortunately, we were able to identify a number of pipelines (9 out of 768) that consistently recover effects in the correct direction, and that additionally satisfy all our other criteria for trustworthy functional connectomics: low PDiv for test–retest scans, indicating that the pipeline minimises spurious differences; greater PDiv across subjects than within the same subject on average, indicating that the pipeline reflects the ground-truth difference between networks; no empty networks; and no correlation between PDiv and motion. We emphasise that each of these criteria had to be met in *all* our datasets, which included both differences in time-span, and also differences in data resolution and preprocessing/denoising.

Additionally, we found that pipelines’ performance on our criteria is far from random, nor does it vary idiosyncratically with each dataset, instead being highly correlated across different independent datasets spanning short, medium and long timespans (with Spearman’s  $\rho$  ranging between 0.73 and 0.97; Fig. 3). Ability to minimise test–retest differences is also correlated with a pipeline’s ability to detect true differences, when they do exist: both between different individuals (Fig. 5), and within the same individual (induced by potent pharmacological intervention; Fig. 4). In other words, there are systematic factors at play. Indeed, patterns of similarity clearly emerge among the

pipelines that satisfy all our criteria. Specifically, 5 out of 9 optimal pipelines employ the same procedure for edge definition (out of 32 possible ones), consisting of Pearson correlation, weighted edges, and the OMST method of optimising the balance between network efficiency and wiring cost. This is a statistically unlikely occurrence, suggesting that there may be something about this combination that makes it especially appropriate. In fact, all 9 (or 28/35 under the less stringent PDiv criterion) employ Pearson correlation for edge definition. More combinations for edge construction become available if pipelines with PDiv rank in the top 50% are included, with fixed-density thresholds at 5% and 20% density also performing well in combination with weighted and binary edges, respectively.

The edge construction part of the pipeline therefore appears as the most crucial choice: once it is fixed, both GSR and NoGSR options are available among the optimal pipelines, and many combinations of parcellation type and size. This observation is corroborated by our assessment of pipelines’ overall ranked performance across all criteria and datasets (Fig. 9): significant predictors of better pipeline rank include edge type (Pearson correlation outperforming mutual information), filtering scheme (OMST and FD20% being the best) and especially the use of weighted networks instead of binary ones (Figs. S38–S40). In contrast, pipelines’ overall performance does not significantly differ as a function of use of GSR, or parcellation type, when these steps are considered in isolation (Figs. S42–S43). In other words, the relevance of these network construction steps for our criteria becomes apparent when they are considered as part of a full pipeline.

It is especially reassuring that our results about pipeline performance are shared across multiple independent datasets. Likewise, our results generalise across different popular methods for functional MRI denoising (aCompCor and FIX-ICA). The Cambridge and NYU datasets were acquired with parameters for spatial and temporal resolution that are widely used in functional neuroimaging studies. Therefore, we expect our results to generalise to other datasets with similar specifications, such as the publicly available and intensely studied CamCAN<sup>69</sup>, Philadelphia Neurodevelopmental Cohort<sup>70</sup>, CENTER-TBI<sup>71</sup>, Harvard Aging Brain Study<sup>72</sup>, Autism Brain Imaging Data Exchange (ABIDE)<sup>73</sup>, and UCLA Neurophenomics<sup>74</sup> datasets among others, enabling the functional connectomics community to make the most of these valuable resources to study development, aging, and disease. Importantly though, our results about pipeline performance and choice of optimal pipelines also replicated in the high-quality HCP data, which have higher temporal and spatial resolution (suitable for surface-based analysis). Therefore, we expect that our recommendations should also be applicable to more recent datasets acquired with

HCP-like specifications, such as UK Biobank<sup>75</sup>. However, our recommendations are intended to complement investigators' domain expertise, not replace it: each study has its own driving hypotheses and unique challenges. For this reason, we have made available our interactive Pipeline Selection Tool (Supplementary Data 2), which provides a full breakdown of each pipeline's performance across each criterion and each dataset, totalling 13,824 unique assessments: to enable readers to engage with our results, and identify pipelines that fit their specific requirements.

Our optimal pipelines are those that pass *all* our tests across *all* datasets: they minimise noise-driven differences, but correctly detect genuine ones, in a way that is consistent across datasets. While this stringency undoubtedly contributed to the exclusion of many pipelines – 71 of which only due to a single failure in a single dataset – it should equally bolster our confidence about the recommended pipelines' suitability to provide sensible results, including across different time-spans and different data acquisition and preprocessing choices. By recommending a select number of network construction pipelines that provide the most replicable and generalisable results, we hope that the present work will facilitate future meta-analyses of functional connectomics studies.

We next discuss consistent features among optimal network construction pipelines. Owing to its ease of application and interpretation, Pearson correlation is a cornerstone of functional connectomics, and remains the most widely used method to quantify connectivity between regions across thousands of published studies (accounting for over 75% of the studies reviewed by ref. 21). It is therefore reassuring for the field that our optimal pipelines overwhelmingly favour Pearson correlation to quantify functional connectivity. Consistent with our results, correlation was also shown in previous work to outperform mutual information and partial correlation in terms of test–retest reliability, but also in terms of fingerprinting accuracy<sup>76</sup>.

At the microscopic level of neurons and circuits, the brain is unquestionably a nonlinear system. However, the good performance of Pearson correlation that we observed in our results dovetails with previous evidence by Hlinka and colleagues, whose extensive modelling led them to conclude that “the practical relevance of nonlinear methods trying to improve over linear correlation might be limited by the fact that the data are indeed almost Gaussian”<sup>77</sup>. More broadly, although nonlinear aspects of fMRI timeseries can be identified, capturing variability related to patient-control differences<sup>78,79</sup>, multiple studies have provided quantifiable evidence that at the macroscale level observed by functional MRI, signals may be suitably accounted for as linear<sup>80,81</sup>. This has been shown in terms of obtaining limited or no additional benefit when using more complex nonlinear methods to relate structural and functional connectivity<sup>82,83</sup>, or to predict demographic variables from functional MRI<sup>84</sup>, or when comparing the ability of linear versus nonlinear models to fit high-resolution BOLD timeseries<sup>85</sup>.

Crucially, the observed predominance of linear dynamics in macroscale brain signals cannot be dismissed as a mere artifact of functional MRI<sup>85</sup>. Although fMRI's low temporal resolution does contribute to linearising the signal due to both temporal averaging and the limited number of samples, linear models were also recently shown by Nozari and colleagues (2023) to outperform nonlinear ones in terms of their ability to fit intracranial EEG (iEEG) time-series<sup>85</sup>, which are electrodynamic rather than haemodynamic in origin, and have much higher temporal resolution. Thus, empirical results from diverse neuroimaging modalities converge with both simulations<sup>85</sup> and theoretical analysis<sup>86</sup>, showing that the dynamics of nonlinear stochastic populations converge to linear dynamics at the macroscale, as a result of spatial averaging. In other words, observing good performance of linear methods at the macroscale should not be viewed as un-physiological, or a mere artifact of a specific imaging modality, or a denial of

the brain's microscale nonlinearity. Rather, linearisation is an inherent consequence of observing brain activity at the macroscale as afforded by modern neuroimaging, and this phenomenon contributes to explaining why Pearson correlation is suitable for quantifying functional connectivity – at least according to the criteria that we adopted. Nonetheless, we note that although none of the 9 optimal pipelines employ MI, 7 out of the 35 near-optimal ones do. In particular, some of the near-optimal pipelines include versions of the same pipeline that only differ in the use of Pearson correlation or MI: ICA-300 GSR-Top20%-binary, ICA-200 NoGSR-Top20%-binary, Lausanne-463 NoGSR-Top20%-binary, Lausanne-463 GSR-Top20%-binary.

Pertaining to edge filtering, the OMST (Orthogonal Minimum Spanning Trees algorithm), our main recommended approach, is a data-driven method that optimises the balance between efficiency and wiring cost of the network. OMST is unique among the filtering schemes considered here, for multiple reasons. First, because it guarantees that the resulting network is not fragmented into disconnected components (Fig. S35), which we know should not be the case in the brain. This feature makes OMST analogous to percolation-based filtering schemes, whereby the weakest edges are iteratively removed from the network, up to the point where further removal would make the network disconnected, which corresponds to the percolation threshold<sup>87–89</sup>. Thus, OMST and percolation thresholding both ensure that global connectivity is not impacted by removal of a few weak but topologically important edges. Unlike percolation, however, OMST is not restricted to preserving only the strongest edges. Rather, weaker edges can be preferred to stronger ones and be included in the OMST-filtered network, if they contribute to an optimal balance of efficiency and cost. Because of this ability to include weaker edges over stronger ones based on their role in the overall topology, OMST avoids a pitfall of percolation thresholding, whereby the presence of a single node whose edges are all relatively weak, can result in a network that is potentially very dense (because the percolation threshold is determined by the weakest edge whose removal would make the network disconnected, and if this edge is very weak, many other edges may survive the threshold).

In other words, the second feature that makes OMST unique among the filtering schemes considered here is that OMST takes into account not only the strength of connections, but also their more general topological role in the network. Therefore, connectomes obtained through OMST can include edges that both absolute and proportional thresholding methods would simply disregard as too weak, regardless of any further role they may play in network organisation. The key role of weak connections acting as shortcuts between segregated modules, often referred to as the “strength of weak ties”<sup>89–91</sup>, has been increasingly recognised across artificial and biological networks, including the human brain – a clear argument in favour of OMST's ability to reconstruct biologically plausible networks, especially in combination with weighted (rather than binary) edges, which is consistent with our optimal pipelines.

It is essential to note that despite the similar name, OMST is very different both in theory and in practice from simple Minimum Spanning Tree (MST) filtering. Reducing the network to its minimum spanning tree will enforce every individual's network to have the same number of edges: namely, the minimum number possible. In other words, with Minimum Spanning Tree filtering the density of the reconstructed network is known a priori, given its number of nodes. This is not the case for OMST, which is a data-driven method. Specifically, while the OMST does ensure the desirable property of not leaving any node disconnected from the rest of the brain, it determines the final number of edges in a data-driven manner by optimising the network's balance of efficiency and wiring cost. This approach therefore produces plausibly sparse networks, but not maximally sparse, and without imposing the same a-priori level across all individuals (arguably a biologically implausible feature of fixed-density methods).

This difference has important practical consequences for the physiological plausibility of the reconstructed networks: simple Minimum Spanning Tree filtering would lead to brain networks with the biologically implausible property of a null clustering coefficient<sup>64</sup>, whereas this is not the case for the OMST method that we use (as clearly demonstrated in Fig. S34). We further illustrate this distinction in Fig. S36, which shows that for all datasets and individuals, pipelines using OMST filtering produce plausible values of mean clustering coefficient. In contrast, as expected the Minimum Spanning Tree filtering always results in a mean clustering coefficient of zero, regardless of any other steps in the network construction pipeline. This clearly illustrates that despite similar names, OMST (which we used) and the Minimum Spanning Tree filtering (which we *did not* use) lead to very different results and must not be confused.

The good performance of OMST is arguably due to this method being data-driven based on each individual connectome, rather than a one-size-fits-all. Indeed, although OMST is a relatively recent method, its use has already been recommended by several studies on multiple grounds. OMST filtering was shown to minimise topological differences between pipelines<sup>62</sup>; it has outperformed alternative thresholding schemes for functional networks in terms of recognition accuracy and reliability<sup>37,63,92</sup>; and it has also been recommended for use with alternative neuroimaging modalities such as electro- and magnetoencephalography<sup>37,63,92</sup>, suggesting that its applicability may generalise beyond rs-fMRI. Finally, the use of OMST (as well as 20% fixed-density thresholding) was also recommended by another recent study<sup>44</sup> that evaluated a large number of individual options (though without combining them, and using as criterion the ICC of specific network properties instead of our topological approach). Therefore, our results suggest a convergence of recommendations for brain network construction across different criteria and different studies – possibly heralding the emergence of consistent analytic practices in the field. This convergence may in part be helped by our choice to use the Portrait divergence (PDiv), which enabled us to take into account both local and global aspects of network organisation across scales<sup>53</sup>: by considering the network's topology as a whole, our results are inherently more general than results based on any specific graph-theoretical metric.

In this study, we endeavoured to systematically sample and combine many of the most common options across each step in the process of constructing a functional brain network from rs-fMRI data – resulting in 768 unique pipelines. However, due to combinatorial explosion, it would be unfeasible to consider every single option that has been proposed in the literature, and this inevitable limitation should be borne in mind when interpreting our results.

Pertaining to node definition, we considered both atlas-based parcellations, and Independent Components Analysis. Atlas-based methods are the most widely used approach for defining nodes in functional connectomics<sup>21,33</sup>, with their prevalence enjoying an exponential growth in recent years<sup>33</sup>. This enduring popularity is due to multiple reasons. First, biological interpretability and dimensionality reduction. Additionally, parcellations enable integration between data from neuroimaging modalities whose different acquisition methods and resolution are intrinsically different<sup>93</sup>, such as fMRI and PET<sup>94</sup>, cortical morphometry<sup>95</sup> or post-mortem transcriptomics<sup>96</sup>. In the words of Revell and colleagues: “The atlases defining anatomical structures (whether they are functionally, histologically, genetically, procedurally, multi-modally, or randomly defined) are the link between structural connectivity and functional connectivity measurements of the brain”<sup>33</sup>. Reflecting the key role of parcellations in today's human neuroscience, there have been recent calls to adopt a “standard set of atlases”<sup>33</sup> – which includes all the atlases considered here. Indeed, some open neuroimaging resources provide data only in the form of specific parcellations: often precisely those that we evaluated here<sup>97–99</sup>. Therefore, an evaluation of the role of such popular

parcellations in the network reconstruction process is especially pressing for informing best practices in the field.

We considered some of the most widely used atlas-based parcellation schemes for defining nodes in the brain, which vary along some of the most relevant dimensions for network construction<sup>24,30</sup>. Our parcellations range from traditional anatomical atlases based on structural landmarks from one (AAL) or multiple individuals (Lausanne/Desikan-Killiany), to functional parcellations combining task-based and resting-state fMRI in over 1400 individuals (Schaefer), to multimodal parcellations accounting for cortical myeloarchitecture, functional activation, connectivity and topography (Glasser, Brainnetome). However, despite the diversity of approaches covered, encompassing the most common range of network sizes used in the field, we inevitably could not include all the possible atlases in existence<sup>24,30,31,33,100–104</sup>, and we chose to focus on some of the most widely adopted. In fact, the parcellations that we considered are all among those that an exhaustive multi-criterion evaluation of 55 different brain atlases, recently recommended as the “standard set of atlases” for inclusion in neuroimaging studies, due to their prevalence and coverage of the space of parcellation types and sizes<sup>33</sup>.

Dividing the brain into discrete, spatially circumscribed regions provides both dimensionality reduction and ease of interpretation. However, atlas-based parcellations also come with implicit assumptions about spatial localisation (e.g., by imposing the constraints that parcels should be spatially contiguous and non-overlapping) and about what should be regarded as the true functional units of the brain<sup>31</sup>. An alternative to atlas-based methods for node definition is the use of Independent Components Analysis (or analogous approaches such as functional modes<sup>105</sup>), which can provide spatially overlapping, “soft” parcels without hard boundaries, possibly exhibiting multiple spatially discontinuous peaks, which may be better able to reflect the complexity of brain organisation<sup>106–110</sup>. Simulations previously suggested that defining nodes based on ICA may outperform the use of regions-of-interest (e.g., atlas-based parcellations)<sup>108</sup>, and ICA also performed well at behavioural prediction<sup>11</sup>.

Here, therefore, we also included nodes defined as continuous, spatially overlapping independent components from ICA performed at different dimensionalities. However, we did not find significant differences across parcellation types in terms of overall performance (Fig. S42). Indeed, each type of node definition (anatomical atlas, functional atlas, multimodal atlas and ICA-based) features among the optimal pipelines – suggesting that parcellation type may be of limited relevance on its own (at least among the 768 pipelines that we evaluate), and what matters is more the combination of specific node definition choices as part of specific pipelines.

Although we endeavoured to provide a representative sampling of widely used node definition schemes (including all those in the “standard set of atlases” recently proposed<sup>33</sup>), many additional atlases exist in the literature, as well as parcellation-free methods<sup>24,30</sup>. The combinatorial explosion prevented us from extending our investigation to alternatives, but we briefly outline them below. Voxelwise/vertexwise networks with thousands of nodes<sup>43,111</sup> provide maximal spatial resolution, but “volumetric pixels” have no more intrinsic biological meaning than the pixels of a 2D image – and the computational burden can be considerable. Additionally, by failing to aggregate biologically meaningful units (since cognitive functions are known to cover cortical areas larger than single voxels<sup>27</sup>), this approach can incur a loss of both physiological interpretability and statistical power, due to lower signal-to-noise ratio and higher rate of multiple comparisons. Recent gradient-based and eigenmode-based approaches provide alternative representations of the brain that are spatially extended, overlapping, and continuous rather than discrete (analogously to ICA), offering a complementary perspective on the constituent elements of the brain's functional organisation<sup>32,112–116</sup>. It is also worth mentioning that, whatever the macroscopic units of brain function may be in terms

of space, they need not be temporally invariant: future extensions of the present work may consider node definition schemes that allow for node boundaries to vary, and nodes themselves to merge or split dynamically in time, or as a function of task<sup>27,117–120</sup>. Finally, future approaches may enrich nodes with biological annotations such as microstructure, chemoarchitecture, and heterogeneities reflecting additional biological properties<sup>93–95,121,122</sup>, thereby providing a path towards a more integrative neuroscience.

Overall, there are a myriad ways to define nodes for functional connectomics, both using parcellations of different types and sizes, and using parcellation-free methods entirely. Selection criteria specific for this choice have also been proposed in the literature<sup>33</sup>. Future work may reveal whether some of these alternative approaches to node definition perform consistently better – or consistently worse – across our own criteria, than the parcellation-based node definitions adopted in the present work. However, we reiterate that our results point towards a more prominent role of edge definition than node definition, for determining the success of a pipeline – at least among the 768 pipelines considered here. Joining previous authors, we caution the reader that there may simply not be a “one size fits all” atlas<sup>33,119</sup>: our results suggest that the remaining steps in the pipeline provide essential context for determining whether a given atlas is suitable or not.

Pertaining to edge definition, future work may adopt more sophisticated methods of quantifying connectivity<sup>123</sup>: for instance, by adopting multivariate connectivity estimators<sup>124</sup> or methods from information decomposition capable of recovering different kinds of information sharing between regions<sup>125–128</sup> or the directionality of connections (transfer entropy, Granger causality, Dynamic Causal Modelling<sup>123,129,130</sup>), or disambiguating between direct and indirect connections (e.g. partial correlation, regularised partial correlation<sup>11,107</sup>). In particular, recent studies have suggested that although partial correlation tends to exhibit lower reliability than the commonly used Pearson correlation<sup>76,131</sup>, this shortcoming may be compensated by higher validity<sup>76</sup> and discriminability<sup>132</sup>.

More broadly, many alternative thresholding methods also exist, whether based on statistical significance<sup>133</sup>, percolation<sup>87–89</sup>, or shrinkage methods<sup>107,134</sup> – or avoiding thresholding entirely, by using analytic methods that can deal with fully connected and signed networks<sup>52</sup>. Additionally, it remains to be determined how our results will generalise to the case of frequency-specific or even multilayer networks obtained from EEG or MEG<sup>135</sup> (although see Jiang et al.<sup>44</sup> and Dimitriadis et al.<sup>136,137</sup> for recent investigations of frequency bands for fMRI network construction); and time-varying (“dynamic”) networks, an increasingly popular approach in fMRI functional connectivity, whereby edges change over time<sup>138–141</sup>.

It is also known that different motion correction strategies can influence the validity of BOLD signals and subsequent network characteristics; however, no correction strategy offered perfect motion correction<sup>26</sup>. Here, we adopted a widely used denoising strategy (anatomical CompCor), and required our results to also replicate in a dataset denoised with FIX-ICA instead, which unlike aCompCor is designed to affect artifacts specifically and avoid modifying the neural signal of interest<sup>59,60</sup>. Additionally, we also considered two versions of each dataset, preprocessed with versus without the additional step of global signal regression, due to ongoing controversy about the effect of GSR on functional connectivity<sup>61,142</sup>. Finally, to further mitigate the potential impact of motion on our recommendations, we also explicitly included as one of our criteria that pipelines should not produce a PDiv distribution that is significantly correlated with the distribution of differences in subject motion, across any of the four test–retest datasets.

Notably, our final recommendations include pipelines both with and without GSR – although the latter is somewhat more prevalent among the very best-performing ones. In particular, we even found

that the set of optimal pipelines includes versions of the same pipeline both with and without GSR: Brainnetome-246 for Pearson-OMST-weighted (with GSR and no-GSR versions both featuring among the 9 optimal pipelines); and in the expanded set, Schaefer454 Top20%-binary-Pearson, Lausanne-463 Top20%-binary-Pearson, Lausanne-463 Top20%-binary-MutualInfo, ICA-300 Top20%-binary-Pearson and ICA-200 Top20%-binary-MutualInfo. Therefore, our results suggest that investigators may have some discretion in the choice of using GSR, depending on their specific datasets and hypotheses. As an example, GSR may remove physiological and motion-induced noise<sup>56,58</sup>, and it may strengthen brain-behaviour associations<sup>142</sup>, but it can also remove signal of interest pertaining to some pharmacological and pathological conditions<sup>39,143</sup>, or distort group<sup>46</sup> and individual differences<sup>144</sup>. Likewise, a recent study observed reduced generalisability of graph-theoretical properties across sites, sessions, and paradigms when GSR was used<sup>42</sup>, although Tozzi et al.<sup>48</sup> delineated a more intricate picture, whereby GSR decreases reliability for networks and most edges, but increases it for some others, and GSR appeared overall beneficial in a recent evaluation of multiple denoising strategies for fMRI<sup>145</sup>. A comprehensive evaluation of the relative advantages and drawbacks of GSR is beyond the scope of this paper, and the reader is referred to Fox & Murphy<sup>61</sup> and Liu et al.<sup>142</sup> for extensive discussions. Finally, we did not explore potential differences between resting-state conditions (eyes-open vs eyes-closed vs naturalistic viewing)<sup>49,146</sup>, or the impact of scan duration and spontaneous fluctuations in arousal state – although we did include datasets with different scan duration, up to 1200 volumes<sup>21,147</sup>.

There are also additional considerations for functional connectomics that deserve to be mentioned. In addition to increasing the number of options and pipelines considered, future work may further expand on the present results in several ways: it remains to be determined to what extent our results apply to task-based rather than resting-state fMRI<sup>148,149</sup>. The generalisability of the proposed framework beyond healthy individuals is also worthy of future exploration. Compared to healthy controls, some clinical populations have demonstrated lower test–retest reliability<sup>150,151</sup>. Reliability across the lifespan should be also considered by comparing age groups, as early evidence untangled age-related differences in test–retest reliability of rs-fMRI<sup>152</sup>. The choice of the optimal pipeline for functional connectomics may, therefore, vary by clinical characteristics, which still remains to be ascertained and may benefit from topology-based approaches such as the one adopted here. This is an important next step following the present work. It is also possible that a different proportion of optimal pipelines would be found when alternative reconstruction methods are included, or different neuroimaging modalities. Additionally, although here we considered numerous criteria for trustworthy functional connectomics, other criteria could also be of relevance for specific research questions, such as fingerprinting accuracy<sup>76</sup>, discriminating cognitive states<sup>132</sup>, or structure-function correspondence<sup>33,99,153</sup>, among others.

The time-spans that we considered here range from less than an hour to nearly a year between scans. Certainly, in addition to measurement noise, some degree of change in the topology of the functional connectome over the course of weeks or months is to be expected, due to learning and plasticity which may even change the underlying structural connectome. However, such physiological phenomena cannot be expected to appreciably reorganise the entire functional connectome within the span of less than an hour (in the absence of experimental interventions). Therefore, any test–retest PDiv observed within the same hour is most plausibly attributable to noise, and an appropriate pipeline should simply minimise it, as per our test–retest criterion. Additionally, plasticity and learning should not make the functional connectome so different that it becomes indistinguishable from the connectomes of other individuals: rather, such an occurrence should be minimised, as it indicates measurement

noise. Our results clearly show a convergence of these criteria: pipelines that produce small test–retest PDIV over weeks or months are also those that minimise within-hour PDIV, and that minimise the misidentification of individuals. Thus, across datasets and time-spans we observed an encouraging convergence of criteria for reliable functional connectomics.

At a broader level, our approach has been to identify which pipelines produce sensible functional connectomes, so that researchers may have a guide to orient their choice among the “forking paths” of analytical possibilities. However, alternative approaches exist. For instance, pipelines can be evaluated in terms of classification and prediction of behavioural and demographic variables and individual differences from functional connectomics<sup>11,154</sup>; atlas-specific criteria have also been proposed by Revell and colleagues<sup>33</sup>. Alternatively, researchers may perform a “multiverse” analysis, adopting not one but many pipelines and then finding suitable ways to aggregate the results – or using machine learning tools to characterise a low-dimensional space of pipelines<sup>155</sup>. These approaches are not mutually exclusive, but rather complementary: our criteria and our final recommendations could be used to prune the number of branching options to a manageable number of optimal pipelines, and a multiverse analysis could then be carried out in parallel across them, with the confidence that the overall picture will not be contaminated by inappropriate choices.

Ultimately, brain function is extremely complex. Studying the brain as an organised network, rather than a mere collection of areas, has been a tremendous step forward in neuroscientists’ ability to tame this complexity<sup>6</sup>. In fact, it has been argued that there is no a priori reason to assume that the space of brain function should coincide with the space of brain anatomy<sup>28</sup>. Perhaps if a network representation of brain function is desired, the best way will turn out to be one where the nodes represent different temporal frequencies, or different spatial frequencies/eigenmodes<sup>28,113–116,156</sup>, or even the recent “edge-centric” functional connectivity, whereby the edges of the functional connectome are themselves treated as nodes of an edge-to-edge network<sup>157</sup>. However, even the most sophisticated way of representing the functional connectome, would still be a major simplification of the true underlying complexity. Our purpose here was not – and could not be – to identify the one perfect combination of network construction steps that reflects true brain function. Rather, among a large set of ways to simplify the brain into a network, we sought to identify which (if any) respect a set of criteria that we deem sensible and of wide applicability across the field of functional connectomics. However, it is clear that sometimes additional complexity may be required to address specific questions. In the words of Korhonen, Zanin and Papo: “network neuroscientists should choose those elements that yield the representation that best serves their specific goals”<sup>27</sup>.

Overall, among the pipelines that we considered, most fail to meet our threshold for optimality (meeting all five criteria in each of four datasets), and this should raise some measure of concern: an unprincipled choice of pipeline is very likely to be suboptimal. This pervasiveness of suboptimal pipelines makes it all the more urgent to know which pipelines are optimal. Although each of our pipelines represents a unique combination of network construction steps, they all also share some features, such as the range of network sizes considered, producing undirected and sparse networks, and being based on resting-state fMRI (to name just a few). It is conceivable that any of these shared features may be responsible for the non-optimality of most pipelines considered here. However, the present work cannot address whether alternative pipelines would perform better or worse when using different steps than those considered here (such as any of the many parcellation-free approaches in existence, or directed measures of connectivity, or a number of nodes outside our range, or a different imaging modality, and so on). We welcome the ongoing development and evaluation of new methods and pipelines.

However, the fact that most of the pipelines considered here fail at least one of our criteria, does not mean that we should abandon functional connectomics. The failure of some pipelines does not negate the success of others. By considering pipelines end-to-end rather than focusing on individual steps, we were able to identify a small but consistent set of pipelines that successfully meet all our criteria in all our datasets, and which are not random but rather share some consistent features (especially pertaining to edge definition). Therefore, our key message is that although an uninformed choice of pipeline will likely be suboptimal, researchers may still confidently use popular methods for network reconstruction, so long as they rely on an appropriate combination of steps.

In conclusion, our study provides a principled framework to search for the best network construction pipelines across hundreds of candidates, with the aim of recovering brain networks that satisfy multiple criteria for scientific accuracy and practical utility. We revealed drastic differences across pipelines in terms of their ability to recover similar network topologies across different scans of the same individual – even within the same hour – and to recover the true directionality of experimental effects of interest: pipelines vary widely in their ability to detect true effects while mitigating spurious ones. The present demonstration of the existence and prevalence of systematically misleading pipelines further enhances the importance of identifying suitable network construction pipelines. Thus, our results indicate that researchers should pay careful consideration to their choice of network processing pipeline: since most pipelines fail at least one criterion, an uninformed choice of pipeline will likely produce suboptimal (and possibly misleading) results. Nevertheless, the present work also identifies a number of optimal pipelines that may be used with confidence, since they reliably satisfy all our criteria across all our datasets.

Our findings further indicate that no single step in the network construction workflow can single-handedly guarantee that all criteria will be met. Fortunately, however, we also show that by carefully combining different steps in the network construction workflow, neuroscientists can obtain functional brain networks that satisfy all our criteria, across datasets covering different time-spans and different acquisition and preprocessing procedures, and may be used with confidence. These recommendations can inform future studies, to help investigators make principled choices and minimise the chance that an inappropriate choice of network construction will lead to unreliable or false negatives results. Overall, by enabling systematic evaluation of network processing steps in a way that does not require the arbitrary selection of specific network properties of interest, we hope that the topology-based, multi-criteria framework proposed here will lead towards an objective consensus and more consistent practices in functional connectomics.

## Methods

### NYU test–retest dataset

This is an open dataset from the International Neuroimaging Data-Sharing Initiative (INDI) ([http://www.nitrc.org/projects/nyu\\_trt](http://www.nitrc.org/projects/nyu_trt)), originally described in Shehzad et al.<sup>158</sup>. Briefly, this dataset includes 25 participants (mean age  $30.7 \pm 8.8$  years, 16 females) with no history of psychiatric or neurological illness. The study was approved by the institutional review boards of the New York University School of Medicine and New York University, and participants provided written informed consent and were compensated for their participation.

For each participant, 3 resting-state scans were acquired. Scans 2 and 3 were conducted in a single scan session, 45 min apart, which took place on average 11 months (range 5–16 months) after scan 1. Each scan was acquired using a 3T Siemens (Allegra) scanner, and consisted of 197 contiguous EPI functional volumes (TR = 2000 ms; TE = 25 ms; flip angle = 90°; 39 axial slices; field of view (FOV) = 192 × 192 mm<sup>2</sup>; matrix = 64 × 64; acquisition voxel size = 3 × 3 × 3 mm<sup>3</sup>). Participants were

instructed to relax and remain still with their eyes open during the scan. For spatial normalisation and localisation, a high-resolution T1-weighted magnetisation prepared gradient echo sequence was also obtained (MPRAGE, TR = 2500 ms; TE = 4.35 ms; TI = 900 ms; flip angle = 8°; 176 slices, FOV = 256 mm).

### Cambridge test–retest dataset

Right-handed healthy participants ( $N=22$ , age range, 19–57 years; mean age, 35.0 years; SD 11.2; female-to-male ratio, 9/13) were recruited via advertisements in the Cambridge area and were paid for their participation. Cambridgeshire 2 Research Ethics Committee approved the study (LREC 08/H0308/246) and all volunteers gave written informed consent before participating and were compensated for their participation. Exclusion criteria included National Adult Reading Test (NART) < 70, Mini Mental State Examination (MMSE) < 23, left-handedness, history of drug/alcohol abuse, history of psychiatric or neurological disorders, contraindications for MRI scanning, medication that may affect cognitive performance or prescribed for depression, and any physical handicap that could prevent the completion of testing.

The study consisted of two visits (separated by 2–4 weeks). For each visit, resting-state fMRI was acquired for 5:20 minutes using a Siemens Trio 3 T scanner (Erlangen, Germany). Functional imaging data were acquired using an echo-planar imaging (EPI) sequence with parameters TR 2000 ms, TE 30 ms, Flip Angle 78°, FOV  $192 \times 192 \text{ mm}^2$ , in-plane resolution  $3.0 \times 3.0 \text{ mm}$ , 32 slices  $3.0 \text{ mm}$  thick with a gap of  $0.75 \text{ mm}$  between slices. A 3D high-resolution MPRAGE structural image was also acquired, with the following parameters: TR 2300 ms, TE 2.98 ms, Flip Angle 9°, FOV  $256 \times 256 \text{ mm}^2$ . Task-based data were also collected, and have been analysed before to investigate separate experimental questions<sup>159,160</sup>. A final set of 18 participants had usable data for both resting-state fMRI scans and were included in the present analysis.

### Human Connectome Project test–retest data

This dataset is a subset of the 1200 Human Connectome Project (HCP) subjects<sup>65,66</sup>. It includes resting-state functional MRI (and accompanying structural MRI) scans for 46 healthy individuals (13 male, age 22–35 years), who were each scanned twice at 3 T, at intervals ranging between 1 month and 11 months). All HCP scanning protocols were approved by the local Institutional Review Board at Washington University in St. Louis, and participants provided written informed consent. Detailed information about the acquisition and imaging is provided in the dedicated HCP publications. Briefly: anatomical (T1-weighted) images were acquired in axial orientation, with FOV =  $224 \times 224 \text{ mm}$ , voxel size  $0.7 \text{ mm}^3$  (isotropic), TR 2400 ms, TE 2.14 ms, flip angle 8°. Functional MRI data (1200 volumes) were acquired with EPI sequence, 2 mm isotropic voxel size, TR 720 ms, TE 33.1 ms, flip angle 52°, 72 slices.

### Cambridge propofol dataset

The Cambridge University (“Cambridge”) propofol dataset has been published before<sup>161–163</sup>; we refer the reader to the original study for a detailed description<sup>161</sup>. As previously reported, 16 healthy volunteers were initially recruited for scanning. In addition to the original 16 volunteers, data were acquired for nine participants using the same procedures, bringing the total number of participants in this dataset to 25 (11 males, 14 females; mean age 34.7 years, SD = 9.0 years). Ethical approval for these studies was obtained from the Cambridgeshire 2 Regional Ethics Committee, and all subjects gave informed consent to participate in the study and were compensated for their participation. Volunteers were informed of the risks of propofol administration, such as loss of consciousness, respiratory and cardiovascular depression. They were also informed about more minor effects of propofol such as pain on injection, sedation and amnesia. In addition, standard

information about intravenous cannulation, blood sampling and MRI scanning was provided.

Three target plasma levels of propofol were used: no drug (Awake), 0.6 mg/ml (Mild sedation) and 1.2 mg/ml (Moderate sedation). Scanning (rs-fMRI) was acquired at each stage, and also at Recovery; anatomical images were also acquired. The level of sedation was assessed verbally immediately before and after each of the scanning runs. Propofol was administered intravenously as a “target controlled infusion” (plasma concentration mode), using an Alaris PK infusion pump (Carefusion, Basingstoke, UK). A period of 10 min was allowed for equilibration of plasma and effect-site propofol concentrations. Blood samples were drawn towards the end of each titration period and before the plasma target was altered, to assess plasma propofol levels. In total, 6 blood samples were drawn during the study. The mean (SD) measured plasma propofol concentration was 304.8 (141.1) ng/ml during mild sedation, 723.3 (320.5) ng/ml during moderate sedation and 275.8 (75.42) ng/ml during recovery. Mean (SD) total mass of propofol administered was 210.15 (33.17) mg, equivalent to 3.0 (0.47) mg/kg. Two senior anaesthetists were present during scanning sessions and observed the subjects throughout the study from the MRI control room and on a video link that showed the subject in the scanner. Electrocardiography and pulse oximetry were performed continuously, and measurements of heart rate, non-invasive blood pressure, and oxygen saturation were recorded at regular intervals.

The acquisition procedures are described in detail in the original study<sup>161</sup>. As previously reported, MRI data were acquired on a Siemens Trio 3T scanner (WBIC, Cambridge). For each level of sedation, 150 rs-fMRI volumes (5 min scanning) were acquired. Each functional BOLD volume consisted of 32 interleaved, descending, oblique axial slices, 3 mm thick with interslice gap of 0.75 mm and in-plane resolution of 3 mm, field of view =  $192 \times 192 \text{ mm}$ , TR = 2000 ms, acquisition time = 2000 ms, time echo = 30 ms, and flip angle 78. T1-weighted structural images at 1 mm isotropic resolution were also acquired in the sagittal plane, using an MPRAGE sequence with TR = 2250 ms, TI = 900 ms, TE = 2.99 ms and flip angle = 9 degrees, for localisation purposes. During scanning, volunteers were instructed to close their eyes and think about nothing in particular throughout the acquisition of the resting-state BOLD data. Of the 25 healthy subjects, 15 were ultimately retained (7 males, 8 females): 10 were excluded, either because of missing scans ( $n=2$ ), or due of excessive motion in the scanner ( $n=8$ , 5 mm maximum motion threshold). Here, we only use data from the Awake and Moderate anaesthesia resting-state scanning.

### Western propofol dataset

The Western University (“Western”) propofol data have been published before<sup>17,164–166</sup> and we refer the reader to the original study for a detailed description. Briefly, data were collected between May and November 2014 at the Robarts Research Institute, Western University, London, Ontario (Canada). The study received ethical approval from the Health Sciences Research Ethics Board and Psychology Research Ethics Board of Western University (Ontario, Canada). Healthy volunteers ( $n=19$ ) were recruited (18–40 years; 13 males). Volunteers were right-handed, native English speakers, and had no history of neurological disorders. In accordance with relevant ethical guidelines, each volunteer provided written informed consent, and received monetary compensation for their time. Due to equipment malfunction or physiological impediments to anaesthesia in the scanner, data from  $n=3$  participants (1 male) were excluded from analyses, leaving a total  $n=16$  for analysis<sup>17</sup>.

Resting-state fMRI data were acquired at different propofol levels: no sedation (Awake), Deep anaesthesia (corresponding to Ramsay score of 5) and also during post-anaesthetic recovery. As previously reported<sup>17</sup>, for each condition fMRI acquisition began after two anaesthesiologists and one anaesthesia nurse independently assessed



Ramsay level in the scanning room. The anaesthesiologists and the anaesthesia nurse could not be blinded to experimental condition, since part of their role involved determining the participants' level of anaesthesia. Note that the Ramsay score is designed for critical care patients, and therefore participants did not receive a score during the Awake condition before propofol administration: rather, they were required to be fully awake, alert and communicating appropriately. To provide a further, independent evaluation of participants' level of responsiveness, they were asked to perform two tasks: a test of verbal memory recall, and a computer-based auditory target-detection task. Wakefulness was also monitored using an infrared camera placed inside the scanner.

Propofol was administered intravenously using an AS50 auto syringe infusion pump (Baxter Healthcare, Singapore); an effect-site/plasma steering algorithm combined with the computer-controlled infusion pump was used to achieve step-wise sedation increments, followed by manual adjustments as required to reach the desired target concentrations of propofol according to the TIVA Trainer (European Society for Intravenous Anaesthesia, [eurosva.eu](http://eurosva.eu)) pharmacokinetic simulation programme. This software also specified the blood concentrations of propofol, following the Marsh 3-compartment model, which were used as targets for the pharmacokinetic model providing target-controlled infusion. After an initial propofol target effect-site concentration of  $0.6 \mu\text{g mL}^{-1}$ , concentration was gradually increased by increments of  $0.3 \mu\text{g mL}^{-1}$ , and Ramsay score was assessed after each increment: a further increment occurred if the Ramsay score was  $<5$ . The mean estimated effect-site and plasma propofol concentrations were kept stable by the pharmacokinetic model delivered via the TIVA Trainer infusion pump. Ramsay level 5 was achieved when participants stopped responding to verbal commands, were unable to engage in conversation, and were rousable only to physical stimulation. Once both anaesthesiologists and the anaesthesia nurse all agreed that Ramsay sedation level 5 had been reached, and participants stopped responding to both tasks, data acquisition was initiated. The mean estimated effect-site propofol concentration was  $2.48 (1.82\text{--}3.14) \mu\text{g mL}^{-1}$ , and the mean estimated plasma propofol concentration was  $2.68 (1.92\text{--}3.44) \mu\text{g mL}^{-1}$ . Mean total mass of propofol administered was  $486.58 (373.30\text{--}599.86) \text{ mg}$ . These values of variability are typical for the pharmacokinetics and pharmacodynamics of propofol. Oxygen was titrated to maintain SpO<sub>2</sub> above 96%.

At Ramsay 5 level, participants remained capable of spontaneous cardiovascular function and ventilation. However, the sedation procedure did not take place in a hospital setting; therefore, intubation during scanning could not be used to ensure airway security during scanning. Consequently, although two anaesthesiologists closely monitored each participant, scanner time was minimised to ensure return to normal breathing following deep sedation. No state changes or movement were noted during the deep sedation scanning for any of the participants included in the study<sup>17</sup>. Propofol was discontinued following the deep anaesthesia scan, and participants reached level 2 of the Ramsey scale approximately eleven minutes afterwards, as indicated by clear and rapid responses to verbal commands.

As previously reported<sup>17</sup>, once in the scanner participants were instructed to relax with closed eyes, without falling asleep. Resting-state functional MRI in the absence of any tasks was acquired for 8 minutes for each participant. A further scan was also acquired during auditory presentation of a plot-driven story through headphones (5 min long). Participants were instructed to listen while keeping their eyes closed. The present analysis focuses on the resting-state data only, from the Awake and Deep scanning; the story scan data have been published separately<sup>165</sup> and will not be discussed further here.

As previously reported<sup>17</sup>, MRI scanning was performed using a 3-Tesla Siemens Tim Trio scanner (32-channel coil), and 256 functional volumes (echo-planar images, EPI) were collected from each participant, with the following parameters: slices = 33, with 25% inter-slice

gap; resolution = 3 mm isotropic; TR = 2000ms; TE = 30 ms; flip angle = 75°; matrix size =  $64 \times 64$ . The order of acquisition was interleaved, bottom-up. Anatomical scanning was also performed, acquiring a high-resolution T1-weighted volume (32-channel coil, 1 mm isotropic voxel size) with a 3D MPRAGE sequence, using the following parameters: TA = 5 min, TE = 4.25 ms,  $240 \times 256$  matrix size, 9° flip angle<sup>17</sup>.

### Functional MRI preprocessing and denoising

Preprocessing of the functional MRI data for all datasets except HCP followed the same standard workflow as in our previous studies<sup>62</sup>, and was implemented in the CONN toolbox (<http://www.nitrc.org/projects/conn>), version 17f<sup>67</sup>. The following steps were performed: removal of the first 5 volumes to allow for steady-state magnetisation; functional realignment, motion correction and spatial normalisation to Montreal Neurological Institute (MNI-152) standard space with  $2 \times 2 \times 2 \text{ mm}$  isotropic resolution. Denoising followed the anatomical CompCor (aCompCor) method of removing cardiac and motion artifacts, by regressing out of each individual's functional data the first 5 principal components corresponding to white matter signal, and the first 5 components corresponding to cerebrospinal fluid signal, as well as six subject-specific realignment parameters (three translations and three rotations) and their first-order temporal derivatives, and nuisance regressors identified by the artifact detection software *art*<sup>68</sup>. The subject-specific denoised BOLD signal time-series were linearly detrended and band-pass filtered between 0.008 and 0.09 Hz to eliminate both low-frequency drift effects and high-frequency noise. No spatial smoothing was applied, since all analyses were performed on parcellated data, whereby the signal was averaged across voxels belonging to the same ROI (see below, section *Node definition*).

For the HCP test-retest dataset, we instead used the minimally preprocessed functional data made available by HCP, which were further denoised with FIX-ICA<sup>59,60</sup>. This popular approach is intended to remove non-BOLD noise arising from multiple known sources, including spatially specific noise from head motion, cardiac pulsation, breathing and scanner artifacts. Using different denoising methods enables us to ensure that our final results are not specific to a particular way of denoising rs-fMRI data, thereby ensuring their robustness and generalisability.

A further, particularly controversial denoising step is global signal regression (GSR): although some authors suggest that GSR may improve subsequent construction of functional brain networks<sup>41,50</sup>, others did not find such an effect<sup>40,43</sup> or even reported GSR as deleterious<sup>42,48</sup>. Here, we therefore evaluated the performance of different network construction pipelines on two versions of each dataset: with the application of GSR, and without the application of GSR.

### Node definition

When deciding on how to turn preprocessed and denoised fMRI data into a brain network, the first decision that needs to be made is: what are the elements of the network? Different approaches exist in the literature, from the use of each voxel as a node to maximise spatial resolution, to the use of Independent Components Analysis and similar data-driven techniques to obtain study- or even subject-specific clusterings of brain signals, which may be spatially extended or even nested within each other, coalescing and splitting over time. Although each of these approaches has unquestionable merits, perhaps the most common approach for defining nodes in human network neuroscience is the use of parcellations: pre-defined assignments of spatially contiguous voxels into regions-of-interest (ROIs) – typically on the ground of neuroanatomical/cytoarchitectonic considerations, or shared function, or some combination thereof. A wide variety of parcellations exist<sup>24</sup>, and recent work reported how the choice of parcellation scheme can affect aspects such as structure-function similarity estimation<sup>153</sup> but also the intra-subject and inter-subject variability of the functional connectome and whole-brain resting-state

modelling<sup>48,169</sup>. Parcellation schemes vary on two main dimensions: the criterion based on which clusters are identified (e.g., based on neuroanatomy, or functional considerations, or a combination thereof from multiple modalities) and the number of ROIs – ranging from a few tens to thousands. The number of ROIs involves a trade-off between the superior spatial resolution of finer-grained parcellations, and the robustness and increased signal-to-noise ratio that derive from spatial averaging of many neighbouring voxels.

Here, we considered both of these dimensions: we employed parcellations spanning three scales (approximately 100, 200 and 300–400 nodes; see Table S1 for specific details) and obtained based on anatomical, functional, or multimodal considerations, or from spatial Independent Components Analysis, across one or multiple scales (summarised in Table S1).

**Lausanne multi-scale atlas.** We consider the multi-scale anatomical Lausanne (also known as “Cammoun”) atlas with 129, 234 and 463 cortical and subcortical nodes obtained by subdividing the sulcus-based Desikan-Killiany atlas<sup>170,171</sup>.

**Schaefer+Tian multi-scale atlas.** We also consider the functional multi-scale atlas developed by Schaefer and colleagues<sup>172</sup> which combines local gradients and global similarity across task-based and resting-state functional connectivity. Following our previous work, we included versions with 100, 200 and 400 cortical regions, respectively supplemented with 16, 32 or 54 subcortical regions from the recent subcortical functional atlas developed by Tian and colleagues<sup>173</sup>.

**Single-scale atlases.** We also include three widely used single-scale atlases: (i) the Automated Anatomical Labelling (AAL) atlas, an anatomical parcellation with 90 cortical and subcortical regions<sup>174</sup>; (ii) the Brainnetome atlas, which comprises 210 cortical and 36 subcortical regions, identified by combining anatomical, functional and meta-analytic information<sup>175</sup>; (iii) and the Glasser atlas comprising 360 cortical regions identified by combining multi-modal information about cortical architecture, function, connectivity and topography<sup>176</sup>. The volumetric Glasser parcellation in MNI-152 space made available by Preti and Van de Ville<sup>177</sup> was used. Since the Glasser atlas is cortical-only, it was also supplemented with the 54-region version of the Tian atlas, in order to include a comparable number of subcortical regions, resulting in 414 ROIs.

**Spatial independent components analysis.** Finally, we include parcellations obtained from group spatial Independent Components Analysis (ICA) at different dimensionality. We used the independent components generated and made available by the Human Connectome Project consortium, by performing spatial-ICA on the combined cortical (vertexwise) and subcortical (voxelwise) rs-fMRI timeseries from 820 HCP participants using FSL MELODIC tool<sup>178,179</sup>. These authors performed spatial-ICA at several different dimensionalities, from 15 to 300. Dimensionality corresponds to the number of independent spatial components to be identified by MELODIC, which is analogous to the number of discrete parcels in an atlas-based parcellation. Here we used dimensionalities 100, 200 and 300, as the most compatible with the dimensionality of the other parcellations included in the present study. ICA differs from atlas-based methods in several key respects. Unlike atlas parcels, ICA maps are not binary but rather consist of continuous weights: some parts of the brain will contribute more or less to a given spatial component. As a result, ICA maps are spatially overlapping, and need not be spatially contiguous.

For atlas-based parcellations, the average denoised BOLD timeseries across all voxels belonging to a given ROI were extracted, for each individual. For the ICA parcellations, we applied FSL Dual Regression to obtain the time-course associated with each spatial independent component map of each individual<sup>180</sup>. Therefore, for the

ICA parcellations, nodes do not represent discrete spatial locations, but rather different weightings of spatially extended and possibly overlapping independent components.

For all but the HCP dataset, we used parcellations made available in volumetric MNI-152 space. For the HCP test–retest dataset, given the higher spatial resolution, we opted to use a surface-based parcellation approach instead – thereby enabling us to verify that our final results are not specific to a given parcellation approach.

### Functional connectivity

We considered two alternative ways of quantifying the interactions between regional BOLD signal timeseries. First, we used Pearson correlation, whereby for each pair of nodes  $i$  and  $j$ , their functional connectivity  $F_{ij}$  was given by the Pearson correlation coefficient between the timecourses of  $i$  and  $j$ , over the full scanning length. Second, we also used the mutual information  $I$ , which quantifies the interdependence between two random variables  $X$  and  $Y$ , and is defined as the average reduction in uncertainty about  $X$  when  $Y$  is given (or vice versa, since this quantity is symmetric):

$$I(X;Y) = H(X) + H(Y) - H(X,Y) = H(X) - H(X|Y) \quad (1)$$

With  $H(X)$  being the Shannon entropy of a variable  $X$ . Unlike Pearson correlation, mutual information does not provide negative values: both strong positive and strong negative correlations will be mapped (non-linearly) onto high values of MI. For consistency with previous work<sup>62</sup>, the values in each individual matrix of mutual information were divided by the maximum value in the matrix, thereby rescaling them to lie between zero and unity.

### Filtering schemes

Both Pearson correlation and MI provide continuous values for the statistical association between pairs of nodes, resulting in a dense matrix of functional connections. Therefore, some form of filtering is typically employed to remove spurious connections that are likely to be driven by noise, and obtain a sparse network of functional connectivity. However, there is no gold standard approach to decide which connections to retain, and different filtering schemes have emerged in the literature. Here, we considered 8 different edge filtering schemes (Table S2), described below. The Brain Connectivity Toolbox<sup>51,52</sup> was used to implement absolute and proportional thresholds and quantify network density, as well as the networks' mean clustering coefficient and characteristic path length (Fig. S33–S35).

**Absolute thresholding.** The simplest approach to decide which edges to retain is to accept or reject edges based on a pre-determined minimum acceptable weight. However, there is no consensus in the literature about which threshold one should adopt. Here, we considered absolute threshold values of 0.3 or 0.5 (for Pearson correlation, only positively-valued edges were considered).

**Proportional thresholding.** Absolute thresholding can produce networks with very different densities, which can introduce confounds in subsequent network analyses. Therefore, a popular approach simply retains a fixed proportion of the strongest edges. However, there is once again no consensus in the literature on the correct proportion of edges to retain. We therefore employed three different density levels, in the range commonly reported in the literature: fixed density (FD) of 5%, 10% and 20% of the strongest edges.

**Structural density matching.** The main problem with proportional thresholding is the selection of an appropriate target density – especially since this may vary depending on the number of nodes in the network. To address this issue in a principled manner, we recently introduced a method termed structural density matching (SDM)<sup>62</sup>,

whereby the proportion of functional edges to retain corresponds to the density  $s$  of the corresponding structural connectome (the network of anatomical connectivity obtained from the group-averaged diffusion-weighted MRI data from the Human Connectome Project<sup>181</sup>). In other words, SDM ensures that functional and structural networks obtained using the same parcellation have the same density, instead of using an arbitrary target density. For the ICA parcellations, since each component map is spatially extended over the entire cortex and sub-cortex, we used instead the density of the structural connectome obtained from the Schaefer parcellation at the same number of parcels (100, 200 or 300), in order to obtain results that are comparable to the other parcellations.

**Efficiency cost optimisation.** The efficiency cost optimisation (ECO) is designed to optimise the trade-off between the network's overall efficiency (sum of global and average local efficiency) and its wiring cost (number of edges)<sup>64</sup>, by ensuring that the network maximises the following target function  $J$ :

$$J = \frac{E_g + E_l}{\rho} \quad (2)$$

With  $E_g$  and  $E_l$  being the global and mean local efficiency of the network, respectively. This filtering scheme produces sparse graphs while still preserving their structure, as demonstrated by its empirical success at discriminating between different graph topologies<sup>64</sup>. Here, we obtained ECO-thresholded graphs by setting a proportional threshold such that the average node degree would be 3, since previous analytic and empirical results indicate that the optimal density corresponds to enforcing an average node degree approximately equal to 3<sup>64</sup>.

**Orthogonal minimum spanning trees.** OMST<sup>63,92</sup> is another data-driven approach intended to optimise the balance between efficiency and density of the network, while also ensuring that the network is fully connected. Specifically, the method involves three steps: (1) identifying the minimum set of edges such that each node can be reached from each other node – known as the minimum spanning tree (MST); (2) identifying an alternative (orthogonal) MST, and combining it with the previous one; (3) repeating steps (1) and (2) until the network formed by the progressive addition of orthogonal MSTs optimises a global cost function defined as  $E_g - \text{Cost}$  (with Cost corresponding to the ratio of the total weight of the selected edges, divided by the total strength of the original fully weighted graph). This approach produces plausibly sparse networks without imposing an a-priori level across all subjects, and it has been shown that the resulting networks provide higher recognition accuracy and reliability than many alternative filtering schemes<sup>63,92</sup>.

**Binarisation.** For all filtering schemes considered here, edges that were not selected were set to zero. However, edges that were included in the network could be weighted or unweighted. In the case of unweighted (binary) networks, we set all non-zero edges to unity. Otherwise, their original weight was retained.

### Topological distance as Portrait Divergence

To quantify the difference between network topologies, we used the recently developed Portrait Divergence (PDIV). The Portrait Divergence between two graphs  $G_1$  and  $G_2$  is the Jensen-Shannon divergence between their “network portraits”, which encode the distribution of shortest paths of the two networks<sup>53</sup>. Specifically, the network portrait is a matrix  $B$  whose entry  $B_{lk}$ ,  $l = 0, 1, \dots, d$  (with  $d$  being the graph diameter),  $k = 0, 1, \dots, N - 1$ , is the number of nodes having  $k$  nodes at shortest-path distance  $l$ .

Thus, to compute the Portrait Divergence one needs to compute the probability  $P(k, l)$  (and similarly  $Q(k, l)$  for the second graph) of randomly choosing two nodes at distance  $l$  and, for one of the two nodes, to have  $k$  nodes at distance  $l$ :

$$P(k, l) = P(k|l)P(l) = \frac{1}{N} B_{lk} \frac{1}{\sum_c n_c^2} \sum_{k'=0}^N k' B_{lk'} \quad (3)$$

where  $n_c$  is the number of nodes in the connected component  $c$ . Then, the Portrait Divergence distance is defined using the Jensen-Shannon divergence (an information-theoretic notion of distance):

$$D(G_1, G_2) = \frac{1}{2} KL(P||M) + \frac{1}{2} KL(Q||M) \quad (4)$$

where  $M = (P + Q)/2$  is the mixture distribution of  $P$  and  $Q$ , and  $KL(\cdot||\cdot)$  is the Kullback-Leibler divergence.

The Portrait Divergence offers three key advantages that make it well suited for the present investigation. First, it is based on network portraits, which do not change depending on how a graph is represented. Comparing network topologies based on such “graph invariants” is highly desirable, because it removes the potential confound of encoding format. Second, the Portrait Divergence does not require the networks in question to have the same number of nodes or edges, and it can be applied to both binary and weighted networks – making it ideally suited for the applications of the present study. And finally, the Portrait Divergence is not predicated on a single specific network property, but rather it takes into account all scales of structure within networks, from local structure to motifs to large scale connectivity patterns: that is, it considers the topology of the network as a whole<sup>53</sup>.

For each subject, at each timepoint, we obtained one brain network following each of the possible combinations of steps above (768 distinct pipelines in total).

For each pipeline, we then computed the Portrait Divergence between networks obtained from the same subject at different points in time, and subsequently obtained a group-average value of Portrait Divergence for each pipeline.

### Reporting summary

Further information on research design is available in the Nature Portfolio Reporting Summary linked to this article.

### Data availability

Source data are provided with this paper. The NYU dataset is freely available from the International Neuroimaging Data-Sharing Initiative (INDI) ([http://www.nitrc.org/projects/nyu\\_trt](http://www.nitrc.org/projects/nyu_trt)). The Cambridge datasets are available upon request from author EAS (email: eas46@cam.ac.uk). The Western propofol dataset is available on the OpenNeuro data repository (doi: 10.18112/openneuro.ds003171.v2.0.1). The HCP data are available from <https://www.humanconnectome.org/>. The AAL atlas is available online at <https://www.gin.cnrs.fr/en/tools/aal/>. The Brainnetome atlas is available online at <https://atlas.brainnetome.org/download.html>. The Glasser parcellation is available online at <https://balsa.wustl.edu/study/show/RVVG>. The Lausanne multi-scale atlas can be obtained from [https://github.com/mattcieslak/easy\\_lausanne](https://github.com/mattcieslak/easy_lausanne). The Schaefer multi-scale atlas is available at [https://github.com/ThomasYeoLab/CBIG/tree/master/stable\\_projects/brain\\_parcellation/Schaefer2018\\_LocalGlobal](https://github.com/ThomasYeoLab/CBIG/tree/master/stable_projects/brain_parcellation/Schaefer2018_LocalGlobal). The Tian subcortical multi-scale atlas is available at <https://github.com/yetianned/subcortex>. The Group-ICA parcellations are available from <https://www.humanconnectome.org/>. Source data are provided with this paper.

### Code availability

The CONN toolbox is freely available online (<http://www.nitrc.org/projects/conn>). Python (v3.6) code for the portrait divergence is freely

available online (<https://github.com/bagrow/network-portrait-divergence>). MATLAB (v2019a) code for the Orthogonal Minimum Spanning Tree thresholding is freely available online ([https://github.com/stdimitr/topological\\_filtering\\_networks](https://github.com/stdimitr/topological_filtering_networks)). The Brain Connectivity Toolbox code used for graph-theoretical analyses is freely available online (<https://sites.google.com/site/bctnet/>). FSL (FMRIB Software Library) v6 toolbox is freely available for academic research at <https://fsl.fmrib.ox.ac.uk/>.

## References

- Bullmore, E. & Sporns, O. Complex brain networks: graph theoretical analysis of structural and functional systems. *Nat. Rev. Neurosci.* **10**, 312–312 (2009).
- Fox, M. D. et al. The human brain is intrinsically organized into dynamic, anticorrelated functional networks. *Proc. Natl Acad. Sci. USA* **102**, 9673–9678 (2005).
- Park, H. J. & Friston, K. Structural and functional brain networks: From connections to cognition. *Science* **342**, 1238411 (2013).
- Sporns, Olaf. *Networks of the Brain* (MIT Press, 2011).
- Petersen, S. E. & Sporns, O. Brain networks and cognitive architectures. *Neuron* **88**, 207–219 (2015).
- Bassett, D. S. & Sporns, O. Network neuroscience. *Nat. Neurosci.* **20**, 353–364 (2017).
- Fornito, A., Zalesky, A. & Breakspear, M. The connectomics of brain disorders. *Nat. Rev. Neurosci.* **16**, 159–172 (2015).
- Smith, S. M. et al. A positive-negative mode of population covariation links brain connectivity, demographics and behavior. *Nat. Neurosci.* **18**, 1–7 (2015).
- Perovnik, M., Rus, T., Schindlbeck, K. A. & Eidelberg, D. Functional brain networks in the evaluation of patients with neurodegenerative disorders. *Nat. Rev. Neurol.* **19**, 73–90 (2023).
- Miller, K. L. et al. Multimodal population brain imaging in the UK Biobank prospective epidemiological study. *Nat. Neurosci.* **19**, 1523–1536 (2016).
- Pervaiz, U., Vidaurre, D., Woolrich, M. W. & Smith, S. M. Optimising network modelling methods for fMRI. *Neuroimage* **211**, 116604 (2020).
- Finn, E. S. et al. Functional connectome fingerprinting: identifying individuals using patterns of brain connectivity. *Nat. Neurosci.* **18**, 1664–1671 (2015).
- Smith, S. M. et al. Functional connectomics from resting-state fMRI. *Trends Cogn. Sci.* **17**, 666–82 (2013).
- Turk, E. et al. Functional connectome of the fetal brain. *J. Neurosci.* **39**, 9716–9724 (2019).
- Demertzi, A. et al. Human consciousness is supported by dynamic complex patterns of brain signal coordination. *Sci. Adv.* **5**, 1–12 (2019).
- Huang, Z., Zhang, J., Wu, J., Mashour, G. A. & Hudetz, A. G. Temporal circuit of macroscale dynamic brain activity supports human consciousness. *Sci. Adv.* **6**, 87–98 (2020).
- Luppi, A. I. et al. Consciousness-specific dynamic interactions of brain integration and functional diversity. *Nat. Commun.* **10**, 4616 (2019).
- Burianová, H. et al. Altered functional connectivity in mesial temporal lobe epilepsy. *Epilepsy Res.* **137**, 45–52 (2017).
- Holiga, Š. et al. Patients with autism spectrum disorders display reproducible functional connectivity alterations. *Sci. Transl. Med.* **11**, eaat9223 (2019).
- Karbasforoushan, H. & Woodward, N. D. Resting-state networks in schizophrenia. *Curr. Top. Med. Chem.* **12**, 2404–2414 (2012).
- Hallquist, M. N. & Hillary, F. G. Graph theory approaches to functional network organization in brain disorders: a critique for a brave new small-world. *Netw. Neurosci.* **3**, 1–26 (2019).
- Petrella, J. R. Use of graph theory to evaluate brain networks: a clinical tool for a small world? *Radiology* **259**, 317–320 (2011).
- Botvinik-Nezer, R. et al. Variability in the analysis of a single neuroimaging dataset by many teams. *Nature* **582**, 84–88 (2020).
- Eickhoff, S. B., Yeo, B. T. T. & Genon, S. Imaging-based parcellations of the human brain. *Nat. Rev. Neurosci.* **19**, 672–686 (2018).
- Carp, J. On the plurality of (methodological) worlds: estimating the analytic flexibility of fmri experiments. *Front. Neurosci.* <https://doi.org/10.3389/fnins.2012.00149> (2012).
- Parkes, L., Fulcher, B., Yücel, M. & Fornito, A. An evaluation of the efficacy, reliability, and sensitivity of motion correction strategies for resting-state functional MRI. *NeuroImage* **171**, 415–436 (2018).
- Korhonen, O., Zanin, M. & Papo, D. Principles and open questions in functional brain network reconstruction. *Human Brain Mapp.* <https://doi.org/10.1002/hbm.25462> (2021).
- Papo, D., Zanin, M. & Buldú, J. M. Reconstructing functional brain networks: have we got the basics right? *Front. Hum. Neurosci.* **8**, 8–11 (2014).
- Toga, A. W., Thompson, P. M., Mori, S., Amunts, K. & Zilles, K. Towards multimodal atlases of the human brain. *Nat. Rev. Neurosci.* **7**, 952–966 (2006).
- Arslan, S. et al. Human brain mapping: a systematic comparison of parcellation methods for the human cerebral cortex. *NeuroImage* **170**, 5–30 (2018).
- Genon, S., Reid, A., Langner, R., Amunts, K. & Eickhoff, S. B. How to characterize the function of a brain region. *Trends Cogn. Sci.* **22**, 350–364 (2018).
- Bijsterbosch, J. et al. Challenges and future directions for representations of functional brain organization. *Nat. Neurosci.* **23**, 1484–1495 (2020).
- Revell, A. Y. et al. A framework For brain atlases: lessons from seizure dynamics. *NeuroImage* **254**, 118986 (2022).
- Zhi, D., King, M., Hernandez-Castillo, C. R. & Diedrichsen, J. Evaluating brain parcellations using the distance-controlled boundary coefficient. *Hum. Brain Mapp.* **43**, 3706–3720 (2022).
- Noble, S., Scheinost, D. & Constable, R. T. A decade of test-retest reliability of functional connectivity: a systematic review and meta-analysis. *NeuroImage* **203**, 116157 (2019).
- Woo, C.-W., Chang, L. J., Lindquist, M. A. & Wager, T. D. Building better biomarkers: brain models in translational neuroimaging. *Nat. Neurosci.* **20**, 365–377 (2017).
- Messaritaki, E., Dimitriadis, S. I. & Jones, D. K. Optimization of graph construction can significantly increase the power of structural brain network studies. *NeuroImage* **199**, 495–511 (2019).
- Nichols, T. E. et al. Best practices in data analysis and sharing in neuroimaging using MRI. *Nat. Neurosci.* **20**, 299–303 (2017).
- Chen, X. et al. Topological analyses of functional connectomics: a crucial role of global signal removal, brain parcellation, and null models. *Hum. Brain Mapp.* **39**, 4545–4564 (2018).
- Andellini, M., Cannatà, V., Gazzellini, S., Bernardi, B. & Napolitano, A. Test-retest reliability of graph metrics of resting state MRI functional brain networks: a review. *J. Neurosci. Methods* **253**, 183–192 (2015).
- Braun, U. et al. Test-retest reliability of resting-state connectivity network characteristics using fMRI and graph theoretical measures. *NeuroImage* **59**, 1404–1412 (2012).
- Cao, H. et al. Test-retest reliability of fMRI-based graph theoretical properties during working memory, emotion processing, and resting state. *NeuroImage* **84**, 888–900 (2014).
- Du, H.-X. et al. Test-retest reliability of graph metrics in high-resolution functional connectomics: a resting-state functional MRI study. *CNS Neurosci. Ther.* **21**, 802–816 (2015).
- Jiang, C. et al. Optimizing network neuroscience computation of individual differences in human spontaneous brain activity for test-retest reliability. *Netw. Neurosci.* **7**, 1080–1108 (2023).

45. Romero-Garcia, R., Atienza, M., Clemmensen, L. H. & Cantero, J. L. Effects of network resolution on topological properties of human neocortex. *NeuroImage* **59**, 3522–3532 (2012).
46. Saad, Z. S. et al. Trouble at rest: how correlation patterns and group differences become distorted after global signal regression. *Brain Connect.* **2**, 25–32 (2012).
47. Termenon, M., Jaillard, A., Delon-Martin, C. & Achard, S. Reliability of graph analysis of resting state fMRI using test-retest dataset from the Human Connectome Project. *NeuroImage* **142**, 172–187 (2016).
48. Tozzi, L., Fleming, S. L., Taylor, Z. D., Raterink, C. D. & Williams, L. M. Test-retest reliability of the human functional connectome over consecutive days: identifying highly reliable portions and assessing the impact of methodological choices. *Netw. Neurosci.* **4**, 925–945 (2020).
49. Wang, J.-H. et al. Graph theoretical analysis of functional brain networks: test-retest evaluation on short- and long-term resting-state functional MRI data. *PLoS ONE* **6**, e21976 (2011).
50. Welton, T., Kent, D. A., Auer, D. P. & Dineen, R. A. Reproducibility of graph-theoretic brain network metrics: a systematic review. *Brain Connect.* **5**, 193–202 (2015).
51. Rubinov, M. & Sporns, O. Complex network measures of brain connectivity: uses and interpretations. *NeuroImage* **52**, 1059–1069 (2010).
52. Rubinov, M. & Sporns, O. Weight-conserving characterization of complex functional brain networks. *NeuroImage* **56**, 2068–2079 (2011).
53. Bagrow, J. P. & Boltt, E. M. An information-theoretic, all-scales approach to comparing networks. *Appl. Netw. Sci.* **4**, 45 (2019).
54. Noble, S., Scheinost, D. & Constable, R. T. A guide to the measurement and interpretation of fMRI test-retest reliability. *Curr. Opin. Behav. Sci.* **40**, 27–32 (2021).
55. Burgess, G. C. et al. Evaluation of denoising strategies to address motion-correlated artifacts in resting-state functional magnetic resonance imaging data from the Human Connectome Project. *Brain Connect.* **6**, 669–680 (2016).
56. Power, J. D., Schlaggar, B. L. & Petersen, S. E. Recent progress and outstanding issues in motion correction in resting state fMRI. *NeuroImage* <https://doi.org/10.1016/j.neuroimage.2014.10.044> (2015).
57. Power, J. D. et al. Methods to detect, characterize, and remove motion artifact in resting state fMRI. *NeuroImage* **84**, 320–341 (2014).
58. Ciric, R. et al. Benchmarking of participant-level confound regression strategies for the control of motion artifact in studies of functional connectivity. *NeuroImage* **154**, 174–187 (2017).
59. Salimi-Khorshidi, G. et al. Automatic denoising of functional MRI data: combining independent component analysis and hierarchical fusion of classifiers. *NeuroImage* **90**, 449–468 (2014).
60. Griffanti, L. et al. ICA-based artefact removal and accelerated fMRI acquisition for improved resting state network imaging. *NeuroImage* **95**, 232–247 (2014).
61. Murphy, K. & Fox, M. D. Towards a consensus regarding global signal regression for resting state functional connectivity MRI. *NeuroImage* **154**, 169–173 (2017).
62. Luppi, A. I. & Stamatakis, E. A. Combining network topology and information theory to construct representative brain networks. *Netw. Neurosci.* **5**, 96–124 (2021).
63. Dimitriadis, S. I., Antonakakis, M., Simos, P., Fletcher, J. M. & Papanicolaou, A. C. Data-driven topological filtering based on orthogonal minimal spanning trees: application to multigroup magnetoencephalography resting-state connectivity. *Brain Connect.* **7**, 661–670 (2017).
64. De Vico Fallani, F., Latora, V. & Chavez, M. A topological criterion for filtering information in complex brain networks. *PLoS Comput. Biol.* **13**, e1005305 (2017).
65. Glasser, M. F. et al. The minimal preprocessing pipelines for the Human Connectome Project. *NeuroImage* **80**, 105–124 (2013).
66. Van Essen, D. C. et al. The WU-Minn Human Connectome Project: an overview. *NeuroImage* **80**, 62–79 (2013).
67. Smith, S. M. et al. Resting-state fMRI in the Human Connectome Project. *NeuroImage* **80**, 144–168 (2013).
68. Helweggen, K., Libedinsky, I. & van den Heuvel, M. P. Statistical power in network neuroscience. *Trends Cogn. Sci.* **27**, 282–301 (2023).
69. Shafto, M. A. et al. The Cambridge Centre for Ageing and Neuroscience (Cam-CAN) study protocol: a cross-sectional, lifespan, multidisciplinary examination of healthy cognitive ageing. *BMC Neurol.* **14**, 1–25 (2014).
70. Satterthwaite, T. D. et al. The Philadelphia Neurodevelopmental Cohort: a publicly available resource for the study of normal and abnormal brain development in youth. *NeuroImage* **124**, 1115–1119 (2016).
71. Maas, A. I. R. et al. Collaborative European NeuroTrauma Effectiveness Research in Traumatic Brain Injury (CENTER-TBI): a prospective longitudinal observational study. *Neurosurgery* **76**, 67–80 (2014).
72. Dagle, A. et al. Harvard Aging Brain Study: dataset and accessibility. *NeuroImage* **144**, 255–258 (2017).
73. Di Martino, A. et al. The autism brain imaging data exchange: towards a large-scale evaluation of the intrinsic brain architecture in autism. *Mol. Psychiatry* **19**, 659–667 (2014).
74. Poldrack, R. A. et al. A phenome-wide examination of neural and cognitive function. *Sci. Data* **3**, 1–12 (2016).
75. Littlejohns, T. J. et al. The UK Biobank imaging enhancement of 100,000 participants: rationale, data collection, management and future directions. *Nat. Commun.* **11**, 2624 (2020).
76. Mahadevan, A. S., Tooley, U. A., Bertolero, M. A., Mackey, A. P. & Bassett, D. S. Evaluating the sensitivity of functional connectivity measures to motion artifact in resting-state fMRI data. *NeuroImage* **241**, 118408 (2021).
77. Hlinka, J., Palus, M., Vejmelka, M., Mantini, D. & Corbetta, M. Functional connectivity in resting-state fMRI: is linear correlation sufficient? *NeuroImage* **54**, 2218–2225 (2011).
78. Motlaghian, S. M. et al. A method for estimating and characterizing explicitly nonlinear dynamic functional network connectivity in resting-state fMRI data. *J. Neurosci. Methods* <https://doi.org/10.1016/j.jneumeth.2023.109794> (2023).
79. Motlaghian, S. M. et al. Nonlinear functional network connectivity in resting functional magnetic resonance imaging data. *Hum. Brain Mapp.* **43**, 4556–4566 (2022).
80. Friston, K. J., Harrison, L. & Penny, W. Dynamic causal modelling. *NeuroImage* **19**, 1273–1302 (2003).
81. Deco, G. et al. How local excitation-inhibition ratio impacts the whole brain dynamics. *J. Neurosci.* **34**, 7886–7898 (2014).
82. Honey, C. J. et al. Predicting human resting-state functional connectivity from structural connectivity. *Proc. Natl Acad. Sci. USA* **106**, 2035–2040 (2009).
83. Saggio, M. L., Ritter, P. & Jirsa, V. K. Analytical operations relate structural and functional connectivity in the brain. *PLoS ONE* **11**, e0157292 (2016).
84. Schulz, M. A. et al. Different scaling of linear models and deep learning in UK Biobank brain images vs. machine-learning datasets. *Nat. Commun.* **11**, 4238 (2020).
85. Nozari, E. et al. Macroscopic resting-state brain dynamics are best described by linear models. *Nat. Biomed. Eng.* <https://doi.org/10.1038/s41551-023-01117-y> (2023).
86. Ahmed, S. & Nozari, E. On the linearizing effect of spatial averaging in large-scale populations of homogeneous nonlinear systems. in *2022 IEEE 61st Conference on Decision and Control (CDC)* 641–648 <https://doi.org/10.1109/CDC51059.2022.9993260> (2022).

87. Esfahlani, F. Z. & Sayama, H. A percolation-based thresholding method with applications in functional connectivity analysis. in *Complex Networks IX* (eds. Cornelius, S., Coronges, K., Gonçalves, B., Sinatra, R. & Vespignani, A.) 221–231 (Springer International Publishing, Cham, 2018).
88. Bordier, C., Nicolini, C. & Bifone, A. Graph analysis and modularity of brain functional connectivity networks: searching for the optimal threshold. *Front. Neurosci.* **11**, 1–9 (2017).
89. Gallos, L. K., Makse, H. A. & Sigman, M. A small world of weak ties provides optimal global integration of self-similar modules in functional brain networks. *Proc. Natl. Acad. Sci. USA* **109**, 2825–2830 (2012).
90. Gallos, L. K., Sigman, M. & Makse, H. A. The conundrum of functional brain networks: small-world efficiency or fractal modularity. *Front. Physiol.* **3**, 123 (2012).
91. Granovetter, M. S. The strength of weak ties. *Am. J. Sociol.* **78**, 1360–1380 (1973).
92. Dimitriadis, S. I., Salis, C., Tarnanas, I. & Linden, D. E. Topological filtering of dynamic functional brain networks unfolds informative chronnectomics: A novel data-driven thresholding scheme based on orthogonal minimal spanning trees (OMSTs). *Front. Neuroinform.* **11**, 28 (2017).
93. Markello, R. D. et al. neuromaps: structural and functional interpretation of brain maps. *Nat. Methods* **19**, 1472–1479 (2022).
94. Hansen, J. Y. et al. Mapping neurotransmitter systems to the structural and functional organization of the human neocortex. *Nat. Neurosci.* **25**, 1569–1581 (2022).
95. Hansen, J. Y. et al. Local molecular and global connectomic contributions to cross-disorder cortical abnormalities. *Nat. Commun.* **13**, 4682 (2022).
96. Markello, R. D. et al. Standardizing workflows in imaging transcriptomics with the Abagen toolbox. *eLife* **10**, e72129 (2021).
97. Royer, J. et al. An open MRI dataset for multiscale neuroscience. *Sci. Data* **9**, 569 (2022).
98. Larivière, S. et al. The ENIGMA Toolbox: multiscale neural contextualization of multisite neuroimaging datasets. *Nat. Methods* **18**, 698–700 (2021).
99. Mansour, L. S., Di Biase, M. A., Smith, R. E., Zalesky, A. & Seguin, C. Connectomes for 40,000 UK Biobank participants: a multi-modal, multi-scale brain network resource. *Neuroimage* **283**, 120407 (2023).
100. Craddock, R. C., James, G. A., Holtzheimer, P. E., Hu, X. P. & Mayberg, H. S. A whole brain fMRI atlas generated via spatially constrained spectral clustering. *Hum. Brain Mapp.* **33**, 1914–1928 (2012).
101. Craddock, R. C. et al. Imaging human connectomes at the macroscale. *Nat. Methods* **10**, 524–539 (2013).
102. Amunts, K., Mohlberg, H., Bludau, S. & Zilles, K. Julich-Brain: a 3D probabilistic atlas of the human brain's cytoarchitecture. *Science* **369**, 988–992 (2020).
103. Gordon, E. M. et al. Generation and evaluation of a cortical area parcellation from resting-state correlations. *Cereb. Cortex* **26**, 288–303 (2016).
104. Shen, X., Tokoglu, F., Papademetris, X. & Constable, R. T. Group-wise whole-brain parcellation from resting-state fMRI data for network node identification. *Neuroimage* **0**, 403–415 (2013).
105. Dadi, K. et al. Fine-grain atlases of functional modes for fMRI analysis. *NeuroImage* **221**, 117126 (2020).
106. Kiviniemi, V. et al. Functional segmentation of the brain cortex using high model order group PICA. *Hum. Brain Mapp.* **30**, 3865–3886 (2009).
107. Smith, S. M. et al. Network modelling methods for FMRI. *NeuroImage* **54**, 875–891 (2011).
108. Yu, Q. et al. Comparing brain graphs in which nodes are regions of interest or independent components: a simulation study. *J. Neurosci. Methods* **291**, 61–68 (2017).
109. Iraj, A. et al. Identifying canonical and replicable multi-scale intrinsic connectivity networks in 100k+ resting-state fMRI datasets. *Hum. Brain Mapp.* **44**, 5729–5748 (2023).
110. Duda, M. et al. Reliability and clinical utility of spatially constrained estimates of intrinsic functional networks from very short fMRI scans. *Hum. Brain Mapp.* **44**, 2620–2635 (2023).
111. Thirion, B., Varoquaux, G., Dohmatob, E. & Poline, J.-B. Which fMRI clustering gives good brain parcellations? *Front. Neurosci.* **8**, 167 (2014).
112. Margulies, D. S. et al. Situating the default-mode network along a principal gradient of macroscale cortical organization. *Proc. Natl. Acad. Sci. USA* **113**, 12574–12579 (2016).
113. Atasoy, S., Donnelly, I. & Pearson, J. Human brain networks function in connectome-specific harmonic waves. *Nat. Commun.* **7**, 1–10 (2016).
114. Glomb, K. et al. Functional harmonics reveal multi-dimensional basis functions underlying cortical organization. *Cell Rep.* **36**, 109554 (2021).
115. Lioi, G., Gripon, V., Brahim, A., Rousseau, F. & Farrugia, N. Gradients of connectivity as graph Fourier bases of brain activity. *Netw. Neurosci.* **5**, 322–336 (2021).
116. Pang, J. C. et al. Geometric constraints on human brain function. *Nature* <https://doi.org/10.1038/s41586-023-06098-1> (2023).
117. Boukhdhir, A., Zhang, Y., Mignotte, M. & Bellec, P. Unraveling reproducible dynamic states of individual brain functional parcellation. *Netw. Neurosci.* **5**, 28–55 (2021).
118. Iraj, A. et al. The spatial chronnectome reveals a dynamic interplay between functional segregation and integration. *Hum. Brain Mapp.* **40**, 3058–3077 (2019).
119. Salehi, M. et al. There is no single functional atlas even for a single individual: Functional parcel definitions change with task. *NeuroImage* **208**, NeuroImage (2020).
120. Iraj, A., Miller, R., Adali, T. & Calhoun, V. D. Space: a missing piece of the dynamic puzzle. *Trends Cogn. Sci.* **24**, 135–149 (2020).
121. Bazinet, V. et al. Assortative mixing in micro-architecturally annotated brain connectomes. *Nat. Commun.* **14**, 2850 (2023).
122. Luppi, A. I. et al. From abstract networks to biological realities. *Phys Life Rev.* **49**, 12–14 (2024).
123. Reid, A. T. et al. Advancing functional connectivity research from association to causation. *Nat. Neurosci.* **22**, 1751–1760 (2019).
124. Yoo, K. et al. Multivariate approaches improve the reliability and validity of functional connectivity and prediction of individual behaviors. *NeuroImage* **197**, 212–223 (2019).
125. Mediano, P. A. M. et al. Towards an extended taxonomy of information dynamics via Integrated Information Decomposition. *arXiv* <https://doi.org/10.48550/arXiv.2109.13186> (2021).
126. Luppi, A. I. et al. A synergistic core for human brain evolution and cognition. *Nat. Neurosci.* **25**, 771–782 (2022).
127. Varley, T. F., Sporns, O., Schaffelhofer, S., Scherberger, H. & Dann, B. Information-processing dynamics in neural networks of macaque cerebral cortex reflect cognitive state and behavior. *Proc. Natl. Acad. Sci. USA* **120**, e2207677120 (2023).
128. Luppi, A. I., Rosas, F. E., Mediano, P. A. M., Menon, D. K. & Stamatakis, E. A. Information decomposition and the informational architecture of the brain. *Trends Cogn. Sci.* **28**, 352–368 (2024).
129. Friston, K. J. et al. Dynamic causal modelling revisited. *NeuroImage* **199**, 730–744 (2019).
130. Novelli, L. & Lizier, J. T. Inferring network properties from time series using transfer entropy and mutual information: validation of multivariate versus bivariate approaches. *Netw. Neurosci.* **5**, 373–404 (2021).

131. Fiecas, M. et al. Quantifying temporal correlations: a test-retest evaluation of functional connectivity in resting-state fMRI. *Neuroimage* **65**, 231–241 (2013).
132. Sala-Llonch, R., Smith, S. M., Woolrich, M. & Duff, E. P. Spatial parcellations, spectral filtering, and connectivity measures in fMRI: optimizing for discrimination. *Hum. Brain Mapp.* **40**, 407–419 (2018).
133. Váša, F., Bullmore, E. T. & Patel, A. X. Probabilistic thresholding of functional connectomes: application to schizophrenia. *NeuroImage* **172**, 326–340 (2018).
134. Wang, M. B., Owen, J. P., Mukherjee, P. & Raj, A. Brain network eigenmodes provide a robust and compact representation of the structural connectome in health and disease. *PLoS Comput. Biol.* **13**, e1005550 (2017).
135. Pourmotabbed, H., de Jongh Curry, A. L., Clarke, D. F., Tyler-Kabara, E. C. & Babajani-Feremi, A. Reproducibility of graph measures derived from resting-state MEG functional connectivity metrics in sensor and source spaces. *Hum. Brain Mapp.* **43**, 1342–1357 (2022).
136. Dimitriadis, S. I. Complexity of brain activity and connectivity in functional neuroimaging. *J. Neuro Res.* **96**, 1741–1757 (2018).
137. Dimitriadis, S. I. Assessing the repeatability of multi-frequency multi-layer brain network topologies across alternative researcher's choice paths. *bioRxiv* <https://doi.org/10.1101/2021.10.10.463799> (2021).
138. Hutchison, R. M. et al. Dynamic functional connectivity: promise, issues, and interpretations. *Neuroimage* **80**, 5–79 (2013).
139. Preti, M. G., Bolton, T. A. & Van De Ville, D. The dynamic functional connectome: state-of-the-art and perspectives. *NeuroImage* **160**, 41–54 (2017).
140. Lurie, D. J. et al. Questions and controversies in the study of time-varying functional connectivity in resting fMRI. *Netw. Neurosci.* **4**, 30–69 (2020).
141. Holme, P. & Saramäki, J. Temporal networks. *Phys. Rep.* **519**, 97–125 (2012).
142. Liu, T. T., Nalci, A. & Falahpour, M. The Global Signal in fMRI: nuisance or information? *Neuroimage* **150**, 213–229 (2017).
143. Tanabe, S. et al. Altered global brain signal during physiologic, pharmacologic, and pathologic states of unconsciousness in humans and rats. *Anesthesiology* <https://doi.org/10.1097/ALN.0000000000003197> (2020).
144. Li, J. et al. Topography and behavioral relevance of the global signal in the human brain. *Sci. Rep.* **9**, 1–10 (2019).
145. Wang, H.-T. et al. Continuous evaluation of denoising strategies in resting-state fmri connectivity using fMRIPrep and Nilearn. *PLoS Comput. Biol.* **20**, e1011942 (2024).
146. Van Dijk, K. R. A. et al. Intrinsic functional connectivity as a tool for human connectomics: theory, properties, and optimization. *J. Neurophysiol.* **103**, 297–321 (2010).
147. Laumann, T. O. et al. On the stability of BOLD fMRI correlations. *Cereb. Cortex* **27**, 4719–4732 (2017).
148. Shah, L. M., Cramer, J. A., Ferguson, M. A., Birn, R. M. & Anderson, J. S. Reliability and reproducibility of individual differences in functional connectivity acquired during task and resting state. *Brain Behav.* **6**, e00456 (2016).
149. Kristo, G. et al. Task and task-free fMRI reproducibility comparison for motor network identification. *Hum. Brain Mapp.* **35**, 340–352 (2012).
150. Somandepalli, K. et al. Short-term test-retest reliability of resting state fMRI metrics in children with and without attention-deficit/hyperactivity disorder. *Dev. Cogn. Neurosci.* **15**, 83–93 (2015).
151. Conwell, K. et al. Test-retest variability of resting-state networks in healthy aging and prodromal Alzheimer's disease. *Neuroimage Clin.* **19**, 948–962 (2018).
152. Song, J. et al. Age-related differences in test-retest reliability in resting-state brain functional connectivity. *PLoS ONE* **7**, e49847 (2012).
153. Messé, A. Parcellation influence on the connectivity-based structure–function relationship in the human brain. *Hum. Brain Mapp.* **41**, 1167–1180 (2020).
154. Finn, E. S. & Rosenberg, M. D. Beyond fingerprinting: choosing predictive connectomes over reliable connectomes. *NeuroImage* **239**, 118254 (2021).
155. Dafflon, J. et al. A guided multiverse study of neuroimaging analyses. *Nat. Commun.* **13**, 3758 (2022).
156. Robinson, P. A. et al. Eigenmodes of brain activity: neural field theory predictions and comparison with experiment. *NeuroImage* **142**, 79–98 (2016).
157. Faskowitz, J., Zamani Esfahlani, F., Jo, Y., Sporns, O. & Betzel, R. F. Edge-centric functional network representations of human cerebral cortex reveal overlapping system-level architecture. *Nat. Neurosci.* **23**, 1644–1654 (2020).
158. Shehzad, Z. et al. The resting brain: unconstrained yet reliable. *Cereb. Cortex* **19**, 2209–2229 (2009).
159. Vatansever, D., Menon, X. D. K., Manktelow, A. E., Sahakian, B. J. & Stamatakis, E. A. Default mode dynamics for global functional integration. *J. Neurosci.* **35**, 15254–15262 (2015).
160. Manktelow, A. E., Menon, D. K., Sahakian, B. J. & Stamatakis, E. A. Working memory after traumatic brain injury: the neural basis of improved performance with methylphenidate. *Front. Behav. Neurosci.* **11**, 58 (2017).
161. Stamatakis, E. A., Adapa, R. M., Absalom, A. R. & Menon, D. K. Changes in resting neural connectivity during propofol sedation. *PLoS ONE* **5**, e14224 (2010).
162. Adapa, R. M., Davis, M. H., Stamatakis, E. A., Absalom, A. R. & Menon, D. K. Neural correlates of successful semantic processing during propofol sedation. *Hum. Brain Mapp.* <https://doi.org/10.1002/hbm.22375> (2013).
163. Varley, T. F. et al. Consciousness & brain functional complexity in propofol anaesthesia. *Sci. Rep.* **10**, 1–13 (2020).
164. Naci, L., Sinai, L. & Owen, A. M. Detecting and interpreting conscious experiences in behaviorally non-responsive patients. *NeuroImage* **145**, 304–313 (2017).
165. Kandeepan, S. et al. Modeling an auditory stimulated brain under altered states of consciousness using the generalized ising model. *NeuroImage* **223**, 117367 (2020).
166. Naci, L. et al. Functional diversity of brain networks supports consciousness and verbal intelligence. *Sci. Rep.* **8**, 1–15 (2018).
167. Whitfield-Gabrieli, S. & Nieto-Castanon, A. Conn: a functional connectivity toolbox for correlated and anticorrelated brain networks. *Brain Connect.* **2**, 125–141 (2012).
168. Behzadi, Y., Restom, K., Liao, J. & Liu, T. T. A component based noise correction method (CompCor) for BOLD and perfusion based fMRI. *NeuroImage* **37**, 90–101 (2007).
169. Popovych, O. V. et al. Inter-subject and inter-parcellation variability of resting-state whole-brain dynamical modeling. *NeuroImage* <https://doi.org/10.1016/j.neuroimage.2021.118201> (2021).
170. Cammoun, L. et al. Mapping the human connectome at multiple scales with diffusion spectrum MRI. *J. Neurosci. Methods* **203**, 386–397 (2012).
171. Daducci, A. et al. The connectome mapper: an open-source processing pipeline to map connectomes with MRI. *PLoS ONE* **7**, e48121 (2012).
172. Schaefer, A. et al. Local-global parcellation of the human cerebral cortex from intrinsic functional connectivity MRI. *Cereb. Cortex* **28**, 3095–3114 (2018).
173. Tian, Y., Margulies, D., Breakspear, M. & Zalesky, A. Topographic organization of the human subcortex unveiled with functional connectivity gradients. *Nat. Neurosci.* **23**, 1421–1432 (2020).

174. Tzourio-Mazoyer, N. et al. Automated anatomical labeling of activations in SPM using a macroscopic anatomical parcellation of the MNI MRI single-subject brain. *NeuroImage* **15**, 273–289 (2002).
175. Fan, L. et al. The human brainnetome atlas: a new brain atlas based on connectonal architecture. *Cereb. Cortex* **26**, 3508–3526 (2016).
176. Glasser, M. F. et al. A multi-modal parcellation of human cerebral cortex. *Nature* **536**, 171–178 (2016).
177. Preti, M. G. & Van De Ville, D. Decoupling of brain function from structure reveals regional behavioral specialization in humans. *Nat. Commun.* **10**, 4747 (2019).
178. Hyvarinen, A. Fast and robust fixed-point algorithm for independent component analysis. *IEEE Trans. Neural Netw. Learn. Syst.* **10**, 626–634 (1999).
179. Beckmann, C. F. & Smith, S. M. Probabilistic independent component analysis for functional magnetic resonance imaging. *IEEE Trans. Med. Imaging* **23**, 137–152 (2004).
180. Nickerson, L. D., Smith, S. M., Öngür, D. & Beckmann, C. F. Using dual regression to investigate network shape and amplitude in functional connectivity analyses. *Front. Neurosci.* **11**, 115 (2017).
181. Yeh, F.-C. et al. Population-averaged atlas of the macroscale human structural connectome and its network topology. *NeuroImage* **178**, 57–68 (2018).

## Acknowledgements

This work was supported by the Gates Cambridge Trust (OPP 1144) [A.I.L.]; the Canadian Institute for Advanced Research (CIFAR; grant RCZB/072 RG93193) [to D.K.M. and E.A.S.]; The National Institute for Health Research (NIHR, UK), Cambridge Biomedical Research Centre and NIHR Senior Investigator Awards [D.K.M.]; The British Oxygen Professorship of the Royal College of Anaesthetists [D.K.M.]; The Stephen Erskine Fellowship, Queens' College, University of Cambridge [E.A.S.]; the Wellcome Trust Research Training Fellowship (grant no. 083660/Z/07/Z), Raymond and Beverly Sackler Studentship, and the Cambridge Commonwealth Trust [to R.A.]; the Medical Research Council Doctoral Training Grant (#RG86932) [H.M.G.]; Pinsent Darwin Award [H.M.G.]; Joachim Herz Foundation Add-on Fellowship in Interdisciplinary Life Sciences [H.M.G.]; MRC grant MR/K004360/1 [S.I.D.], a Marie Skłodowska-Curie COFUND EU-UK Research Fellowship [S.I.D.], a Beatriu de Pinós fellowship (2020 BP 00116) [S.I.D.]; AMO acknowledges support by the Canada Excellence Research Chairs program (215063); LN acknowledges support by the L'Oréal-Unesco for Women in Science Excellence Research Fellowship; ZQL is supported by the Fonds de Recherche du Québec - Nature et Technologies (FRQNT). Acquisition of the NYU test-retest dataset was funded by Stavros S. Niarchos Foundation, the Leon Lowenstein Foundation, NARSAD (The Mental Health Research Association) grants to F.Xavier Castellanos; and Linda and Richard Schaps, Jill and Bob Smith, and the Taubman Foundation gifts to F. Xavier Castellanos. This work was performed using resources provided by the Cambridge Service for Data Driven Discovery (CSD3) operated by the University of Cambridge Research Computing Service ([www.csd3.cam.ac.uk](http://www.csd3.cam.ac.uk)), provided by Dell EMC and Intel using Tier-2

funding from the Engineering and Physical Sciences Research Council (capital grant EP/T022159/1), and DiRAC funding from the Science and Technology Facilities Council ([www.dirac.ac.uk](http://www.dirac.ac.uk)). For the purpose of open access, the authors have applied a Creative Commons Attribution (CC BY) licence to any Author Accepted Manuscript version arising from this submission.

## Author contributions

A.I.L., E.A.S. and S.I.D. conceived the study. A.I.L., H.M.G., S.I.D. and E.A.S. designed the methodology and the analysis. Z.Q.L. and A.R.D.P. contributed to data analysis. A.I.L. and H.M.G. analysed the data. E.A.S., S.I.D. and D.K.M. supervised the work. E.A.S., A.E.M., R.A., A.M.O. and L.N. participated in data collection. A.I.L., S.I.D. and H.M.G. wrote the manuscript with feedback from all co-authors.

## Competing interests

The authors declare no competing interests.

## Additional information

**Supplementary information** The online version contains supplementary material available at <https://doi.org/10.1038/s41467-024-48781-5>.

**Correspondence** and requests for materials should be addressed to Andrea I. Luppi.

**Peer review information** *Nature Communications* thanks Onerva Korhonen and the other, anonymous, reviewer(s) for their contribution to the peer review of this work. A peer review file is available.

**Reprints and permissions information** is available at <http://www.nature.com/reprints>

**Publisher's note** Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.

**Open Access** This article is licensed under a Creative Commons Attribution 4.0 International License, which permits use, sharing, adaptation, distribution and reproduction in any medium or format, as long as you give appropriate credit to the original author(s) and the source, provide a link to the Creative Commons licence, and indicate if changes were made. The images or other third party material in this article are included in the article's Creative Commons licence, unless indicated otherwise in a credit line to the material. If material is not included in the article's Creative Commons licence and your intended use is not permitted by statutory regulation or exceeds the permitted use, you will need to obtain permission directly from the copyright holder. To view a copy of this licence, visit <http://creativecommons.org/licenses/by/4.0/>.

© The Author(s) 2024