A data-driven approach to identifying and evaluating connectivity-based neural correlates of consciousness

Annie G. Bryant [1,2,✉] **and Christopher J. Whyte** [2,3]

[1]School of Physics, The University of Sydney, Camperdown, NSW, Australia

[2]Centre for Complex Systems, The University of Sydney, Camperdown, NSW, Australia

[3]Brain and Mind Centre, The University of Sydney, Camperdown, NSW, Australia

complex systems | consciousness | brain dynamics | visual perception | functional connectivity

Correspondence: *annie.bryant@sydney.edu.au*

Abstract

Identifying the neural correlates of consciousness remains a major challenge in neuroscience, requiring theories that bridge between subjective experience and measurable neural correlates. However, theoretical interpretation of empirical evidence is often post hoc and susceptible to confirmation bias. Building upon the adversarial collaboration mediated by the COGITATE Consortium, we present a generalizable approach for the data-driven identification, evaluation, and theoretical modeling of connectivity-based neural correlates of consciousness. Using the same magnetoencephalography (MEG) dataset and accompanying pre-registered hypotheses from the COGITATE Consortium, we systematically compared 246 functional connectivity (FC) measures between regions predicted to underlie conscious vision by Integrated Information Theory (IIT) and/or Global Neuronal Workspace Theory (GNWT). We identified a family of FC measures based on the barycenter—tracking the 'center of mass' between two signals—as the top-performing stimulus decoding measures that generalize across regions central to predictions of both IIT and GNWT. To interpret these findings within a theoretical framework, we developed neural mass models that recapitulate the neural dynamics hypothesized to underlie conscious perception by each theory. Comparing simulated barycenter values from these models against empirically measured MEG data revealed that the GNWT-based model, featuring delayed ignition dynamics, better captured observed connectivity patterns than the IIT-based model, which relied on highly synchronized sensory dynamics. Beyond dataset-specific conclusions and limitations, we introduce a framework for systematically identifying and testing candidate neural correlates of consciousness in an unbiased and interpretable manner.

1 Introduction

Characterizing the neural correlates of consciousness remains one of the great open challenges in modern neuroscience. At the heart of this challenge is the search for a theory of consciousness capable of mapping between subjective experiences and their observable neural correlates [1]. Although the field sees an increasing number of publications each year, theoretical interpretation is primarily performed posthoc; moreover, design choices of the methodology itself are highly predictive of which theory the study's findings will support [2]. These biases have cultivated a fragmented literature that lacks a cohesive methodology, in terms of both experimental paradigms and theoretical interpretation.

In a collective endeavor to move beyond current theoretical silos and confirmation bias, researchers

from competing theoretical camps have jointly entered into adversarial collaborations that directly test their preregistered predictions in shared empirical paradigms, conducted by theory-neutral third party experimental groups. This adversarial collaboration, led by the COGITATE Consortium, focused on the predictions of two prominent theories of consciousness: integrated information theory (IIT) and global neuronal workspace theory (GNWT). IIT identifies consciousness with the emergent cause–effect structure of physical systems [3]. According to IIT, a system is conscious to the degree that the system as a whole exhibits greater informational content than the sum of information contained within all sub-partitions of the system. In contrast, GNWT [4, 5] identifies consciousness with the availability of information in the 'global workspace' [6], a recurrent network with long-range excitatory projections connecting prefrontal cortex (PFC) and parietal cortices.

Within the domain of conscious vision in particular, IIT and GNWT yield contrasting predictions regarding stimulus decodability, the location and temporal dynamics of conscious vision, the role of pre-stimulus activity, and the nature of inter-areal coupling. IIT postulates that consciousness is a persistent structure, with conscious perception requiring phase synchronization between low-level visual areas and category-selective visual areas (lateral occipital cortex [7] and fusiform face area [8]). Together with regions in the parietal cortex, this network is often referred to as the 'posterior hot zone'—and notably does not include the PFC, which IIT predicts is only necessary for cognitive aspects of consciousness, such as introspection and self-report [9]. By contrast, GNWT posits that stimulus information becomes conscious when it enters the global workspace through a late wave of neural signal propagation from sensory cortices to PFC. This wave comprises a nonlinear ignition event, independent of task demands, through which the global workspace broadcasts information between otherwise isolated processes throughout the brain. In the experiments conducted by the COGITATE Consortium [10], the results of which were recently made available [11], these predictions were tested through a shared visual perception paradigm. Specifically, participants were shown a stream of images belonging to one of four categories. Across multiple functional neuroimaging modalities, the Consortium reported varying degrees of evidence supporting these two theories based on region-specific decoding and the timing of conscious percepts. However, neither theory's predictions about the relationship between inter-areal coupling (assessed via functional connectivity) and stimulus decodability were borne out—raising the question of whether there exists a functional connectivity (FC) measure that can discriminate between the neural correlates of visual experience predicted by IIT and/or GNWT.

Of note, the Consortium only examined two measures of inter-areal connectivity: pairwise phase consistency [12] and dynamic FC ([13]; using the Pearson correlation coefficient) between respective pairs of theory-predicted brain regions. Many other measures of inter-areal coupling could, in principle, distinguish between the predictions of each theory in the relevant region–region pairs. For example, in previous work, we systematically compared diverse coupling types—for example, synchronous vs. asynchronous, directed vs. undirected, and time vs. frequency domain-based [14, 15]. The results validated commonly used measures like the Pearson correlation coefficient across neuroimaging modalities, while simultaneously highlighting the potential utility of previously unexplored types of measures of pairwise interaction like directed information. Moreover, the theories put forward by both IIT and GNWT were intrinsically qualitative, making it difficult to precisely quantify the evidence for either theory's predictions.

In this work, we address these limitations in the quantitative testing of competing theories by presenting a unified and data-driven approach for identifying, evaluating, and theoretically interpreting candidate neural correlates of consciousness. Based on the aims of the COGITATE Consortium [11] pertaining to inter-areal coupling, we present a highly comparative approach to search for inter-areal coupling (i.e., FC) measures that reliably quantify the neural correlates of conscious vision from magnetoencephalography (MEG) imaging data. This comprehensive suite of 246 pairwise coupling measures [14] includes those

typically employed in functional neuroimaging, including the Pearson correlation coefficient, pairwise phase consistency, power envelop correlation [16, 17], and ϕ^* [18]—enabling direct comparison to under-explored alternatives. To interpret the FC results within the theoretical frameworks of IIT and GNWT, respectively, we then develop two neural mass models [19]— implementing the stimulus-evoked dynamics hypothesized by each theory to be necessary for conscious perception. Crucially, implementing the hypothesized dynamics for each brain region in computational models allows us to explicitly compute exemplar time series and quantitatively compare the predictions of each theory based on the top-performing FC measures. In sum, we present a data-driven approach for identifying the FC measures that maximally encode stimulus-specific information, and then implement simple and interpretable neural models to derive theoretical insight from the data-driven FC measures.

2 Materials and methods

2.1 Neuroimaging data acquisition and task paradigm

All data was obtained from the COGITATE Consortium [11] as prepared for a functional connectivity data science challenge (cf. https://www.arc-cogitate.com/biomag-2024) [20]. MEG data were downloaded in BIDS format for 100 participants, $N = 94$ of which had the requisite anatomical scans to perform source localization. This subset of $N = 94$ participants included $N = 54$ females (23.1 $\pm$ 3.5 years) and $N = 40$ males (22.3 $\pm$ 3.6 years) with no known neurological or psychological conditions and no age differences between males and females (Wilcox rank-sum test, P=0.19). All imaging acquisition and behavioral task details for this dataset are provided in Consortium et al. [11]. Briefly, participants viewed supra-threshold visual stimuli comprised of faces, objects, letters, or false fonts, for variable durations (0.5 s, 1.0 s, or 1.5 s) at different rotation angles. Participants were instructed to press a button when either of two specific stimuli were presented, such that stimulus events were classified as 'Relevant target' (i.e., the correct stimulus to respond to), 'Relevant non-target' (i.e., all other stimuli belonging to the same category as one of the target stimuli), or 'Irrelevant' (i.e., a stimulus belonging to a different category to both target stimuli). We included all stimulus types in our analysis, though we excluded the 'Relevant target' setting due to potential motor confounds. Additionally, for simplicity, we focused exclusively on the 1.0 s ('intermediate') stimulus duration to compare neural dynamics during one second of stimulus presentation ('onset') and the subsequent post-stimulus second ('offset'). Participants were conscious of the presented stimuli in both task-relevant and irrelevant trials, and they demonstrated high hit rates and low false alarm rates, eliminating behavioral accuracy as a potential confound [11].

2.2 Neuroimaging preprocessing

2.2.1 Anatomical preprocessing

High-resolution T1-weighted structural volumes were preprocessed with the same methods and accompanying code outlined in Consortium et al. [11]. Briefly, cortical surfaces were reconstructed using FreeSurfer's `recon-all` pipeline (cf. https://surfer.nmr.mgh.harvard.edu/). Scalp surfaces were also reconstructed and combined with an inner skull surface to create a single-shell boundary elements model (using the *MNE* module in Python [21]), from which the solution was used to create a volumetric forward model spanning the full brain volume with a 5 mm grid. We selected four brain regions of interest that are listed in Table 1 and are depicted schematically in a brain map in Fig. 1A. To delineate these brain areas in subject space, we resampled the Schaefer et al. [22] 100-parcel atlas (mapping to the 7-Network atlas from Yeo et al. [23]) to each participant's reconstructed cortical surfaces using the `mri_surf2surf` function from FreeSurfer.

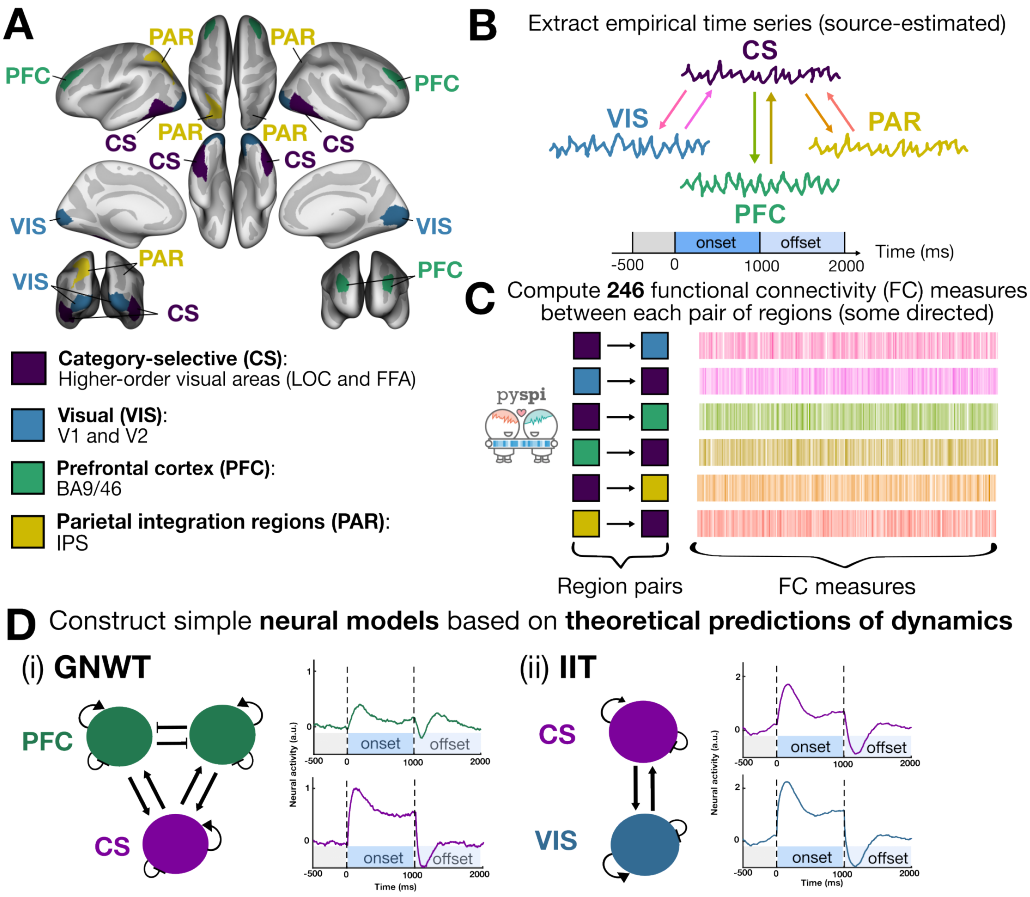

Figure 1. In this study, we systematically compared different ways of quantifying functional coupling from MEG time series derived from regions hypothesized to support the IIT and/or GNWT theories of visual perception and consciousness. A. We selected four regions of interest that are depicted on the inflated cortical surface in `fsaverage` space: Category-selective (CS), comprised of the higher-order visual lateral occipital cortex (LOC) and fusiform face area (FFA) regions; Visual (VIS), comprised of V1 and V2; Prefrontal cortex (PFC), comprised of Brodmann areas 9 and 46; and Parietal integration regions (PAR), comprised of the intraparietal sulcus (IPS). **B.** We extracted the average source-localized MEG time series across all epochs per event-related field type in each region, considering both stimulus onset (first 1000ms) and the subsequent offset period (second 1000ms). **C.** For each functional pathway (i.e., each directed region–region pair), we computed 246 statistics of pairwise interactions (SPIs) using the *pyspi* library. **D.** Neural mass models designed to recapitulate the dynamics of each theoretically relevant region based upon the preregistered predictions of GNWT and IIT set out in Consortium et al. [11]. The dynamics of the GNWT-based model are governed by two key factors: (1) winner-take-all competition between the two PFC neural masses leading to transient 'ignition' events at stimulus onset and offset; and (2) PFC masses have a slower intrinsic timescale leading to delayed peak in PFC activity. The dynamics of the IIT-based model are likewise governed by two key factors: (1) strong recurrent connectivity between CS and VIS masses; and (2) similar intrinsic timescales, which together yield highly synchronous activity at stimulus onset that is maintained for the duration of the stimulus presentation period which then rapidly and synchronously decays at stimulus offset.

2.2.2 MEG preprocessing

MEG data were preprocessed using the same methods previously employed by the COGITATE Consortium for MEG data [11, 24], which included signal-space separation with a Maxwell filter to minimize environment artifacts [25] and FAST-ICA to remove cardiac muscle artifacts [26]. Data were partitioned into 3.5 s epochs, comprised of 1 s pre-stimulus to 2.5 s post-stimulus onset, and epochs were filtered according to gradiometer, magnetometer, and/or artifact properties as described in Consortium et al. [11]. Source modeling was performed using the dynamic statistical parametric mapping method [27]. A common spatial filter was generated from the sum of a noise covariance matrix for the baseline period (-0.5 to 0 s) and a covariance matrix for the active time window period (0 to 2 s). The baseline covariance

matrix was used to spatially pre-whiten the data to combine magnetometer and gradiometer data. For each epoch, we fit and applied an inverse solution to extract the time series from the relevant regions (cf. Sec 2.2.1), averaging across all channels within each parcel. To enhance the signal-to-noise ratio [28], we averaged across trials for each event-related field type (i.e., for each stimulus type and task relevance condition) to yield one average time series per brain region for the given trial type in each participant. As we focused on the 1.0 s stimulus duration, we compared the 1 s window of stimulus presentation ('onset') with the following 1 s post-stimulus period ('offset') for each trial type.

2.3 Data-driven quantification of inter-areal coupling

	Left hemisphere	**Right hemisphere**
Category-Selective (CS)	7Networks_LH_DorsAttn_Post_1-lh	7Networks_RH_Vis_3-rh
Primary visual (VIS)	7Networks_LH_Vis_4-lh	7Networks_RH_Vis_4-rh
	7Networks_LH_Vis_5-lh	7Networks_RH_Vis_5-rh
Prefrontal Cortex (PFC)	7Networks_LH_SalVentAttn_PFCl_1-lh	7Networks_RH_Cont_PFCl_3-rh
Parietal integration (PAR)	7Networks_LH_DorsAttn_Post_3-lh	7Networks_RH_DorsAttn_Post_4-rh

Table 1. The individual brain regions from the 100-parcellation atlas from Schaefer et al. [22] comprising our regions of interest.

We sought to comprehensively quantify diverse types of functional coupling between our four regions. We employed the novel framework for highly comparative analysis of pairwise interactions, `pyspi` (v1.1.0) [14], computing a total of 262 FC measures from the MEG time series between each pair of regions (e.g., CS $\rightarrow$ PAR). As part of the standard `pyspi` workflow, all time series were z-scored along the time axis after which linear detrending was applied with the `detrend` function from the `scipy.signals` module [29]. Some of the FC measures are undirected, such that values are symmetric in both directions between region–region pairs, whereas others are directed (and therefore asymmetric). For quality control, we filtered the 262 FC measures down to those that yielded non-zero variance and real, non-NaN values across region–region pairs for at least 90% (i.e., $N = 85$) of participants, yielding a final set of 246 FC measures. As described in Cliff et al. [14], these FC measures comprise diverse types of pairwise interactions, from basic covariance estimates to multi-metric distances to frequency-based coupling.

2.4 Fitting classifiers

For our primary classification aim ('stimulus type'), we fit a separate model for each region–region pair, task relevance condition, and stimulus period ('on' during the first 1000 ms from stimulus onset and 'off' during the second 1000 ms). For our supplementary aims ('task relevance' and 'cross-task' classification), we still partitioned the data by region–region pair (e.g., VIS $\rightarrow$ CS) and stimulus presentation period ('on' or 'off') into separate classifiers. In all cases, for each FC measure (e.g., power envelope correlation), we fit a binary logistic regression classifier using *scikit-learn* (v1.6.1) [30] to predict either the stimulus type or task relevance from the associated FC measure values (from the given region–region pair). For the stimulus pair classification (primary aim) and task relevance classification (first supplementary aim), we evaluated out-of-sample performance using stratified 10-fold cross-validation with the `StratifiedGroupKFold` function from *scikit-learn*. This allowed us to train each model on a pair of stimuli (or pair of relevance conditions) from one subset of participants and evaluate the model on that pair of stimuli or relevance conditions from another (unseen) subset of participants. For the cross-task classification, each model was trained on the pair of stimuli from all $N = 94$ participants in e.g., 'Relevant non-target' context, and tested on the same pair of stimuli from all $N = 94$ participants in the 'Irrelevant' context (and vice versa). Classification performance was assessed as the mean accuracy across the test subsets per model and is presented throughout this report as the mean cross-validated accuracy $\pm$ 1 SD. As a robustness check,

we also fit a linear support vector machine (SVM) classifier for each FC measure and region–pair and evaluated the cross-validated accuracy to confirm that similar performances were achieved with the logistic regression classifiers.

2.5 Theory-driven neural modeling

To quantitatively link each theory with the empirically observed neural activity (measured with MEG), we constructed two simple neural mass/rate-based models [19, 31, 32] based on the neural activity predicted for each core region (PFC, CS, and VIS) defined by the COGITATE Consortium [11]. The dynamics of each model were governed by the following system of rate-based equations

$$\begin{aligned} \tau_r \frac{dr}{dt} &= -r + f(Wr + u_{stim} - \gamma a + b) \\ \tau_a \frac{da}{dt} &= -a + r, \end{aligned} \tag{1}$$

with state variables $r, a \in \mathbb{R}^{N\times 1}$ representing the aggregate (i.e., coarse-grained) neural activity and (hyperpolarizing) adaptation current (respectively), of each cortical region hypothesized to play a role in conscious perception (for similar models applied to the visual system, see Wilson [33]). For simplicity, the transfer function $f(x)$ is a rectified linear unit. The matrix $W \in \mathbb{R}^{N\times N}$ parameterized the strength of inter-areal and self(within-mass) connectivity; the visual stimulus was encoded in the one-hot vector $u^{stim} \in \mathbb{R}^{N\times 1}$ which, as in the experimental paradigm, lasted 1000 ms of simulation time; τ_r and τ_a are scalar time constants; γ is a scalar constant controlling the strength of the adaptation current; and b is a scalar bias parameter representing the background level of arousal.

Parameter values (supplied in **Table 2**) were manually selected to recapitulate—as precisely as possible—-the time series predictions for each region that were pre-registered by IIT and GNWT theorists [11]. Model simulations and analysis were implemented in custom python (version 3.9.21) code. We solved Eq. (1) numerically using an Euler–Maruyama integration scheme ($\Delta t = 1$ ms), adding zero mean Gaussian noise ($\sigma = 0.005$) to both the neural activity and adaptation current.

	IIT	**GNWT**
W	$\begin{array}{c c} & \begin{array}{cc} VIS & CS \end{array} \\ \begin{array}{c} VIS \\ CS \end{array} & \begin{pmatrix} 0.1 & 0.75 \\ 0.75 & 0.1 \end{pmatrix} \end{array}$	$\begin{array}{c c} & \begin{array}{ccc} CS & PFC_1 & PFC_2 \end{array} \\ \begin{array}{c} CS \\ PFC_1 \\ PFC_2 \end{array} & \begin{pmatrix} 0.05 & 0.75 & 0.75 \\ 0.25 & 0.25 & -2.75 \\ 0.25 & -2.75 & 0.25 \end{pmatrix} \end{array}$
γ	0.5	1
τ_r	τ_{VIS} = 20, τ_{CS} = 40 (ms)	τ_{CS} = 40, τ_{PFC} = 80 (ms)
τ_a	400 (ms)	400 (ms)
b	0.85	0.85

Table 2. Parameter values for the simple IIT- and GNWT-based neural mass models. Parameter values are in arbitrary units unless stated otherwise. These parameters were used construct models as schematically depicted in Fig. 1D.

GNWT predicts that conscious perception should be marked by an 'ignition' event shortly after stimulus onset, as information from high-level visual areas (i.e., CS) updates the contents of the global workspace in PFC through long-range recurrent projections. After this initial ignition event, activity in PFC returns close to baseline. At stimulus offset, when the contents of consciousness change again, GNWT predicts that there should be another 'ignition' event in PFC—updating the contents of the workspace. The GNWT-based model, therefore, consists of three strongly adapting neural masses, one representing CS

and two representing the competitive interactions in PFC responsible for the update dynamics at stimulus onset and offset. The stimulus is delivered to CS, which then projects to the two PFC masses—which compete in a winner-take-all mode to inhibit one another. After stimulus onset, strong adaptation reduces the activity of the winning (i.e., dominant) PFC mass. At stimulus offset, the previously suppressed mass (which is not subject to activity-dependent adaptation) then supersedes and inhibits the dominant mass, generating the hypothesized ignition update dynamics in PFC (cf. Fig. 1D[i]). To generate a single PFC time series, analogous to what would be observed in the MEG data, we averaged the simulated time series of the two PFC masses.

In contrast, according to IIT, conscious contents are maintained in posterior cortical structures throughout the duration of a conscious event via short-range reciprocal projections between low-level visual cortices (i.e., VIS) and higher-level category-selective (CS) cortices generating synchronous activity between sensory regions, maintained for the duration stimulus presentation period (i.e. the duration of the visual experience). The IIT-based model, therefore, consists of two weakly adapting neural masses representing VIS and CS respectively, with strong reciprocal projections. The stimulus is delivered to VIS and is maintained in a high-activity attractor state jointly constituted by VIS and CS throughout stimulus presentation, as depicted schematically in Fig. 1D(ii).

We note that these models should not be regarded as neuronal implementations of IIT or GNWT. Rather, they are simple phenomenological models (i.e., quantitative summaries) of the neural dynamics hypothesized to underlie conscious perception by each theory. Such models allow us to derive quantitative predictions about inter-areal coupling by analyzing the simulated neural data using the FC measures identified in our data-driven analysis of MEG data. Specifically, we compared time-series features computed from the model-simulated time series with those computed from empirically observed MEG signals using two measures: (1) the Kullback–Leibler (K-L) divergence [34] and (2) the Wasserstein distance [35].

The K-L divergence, denoted $D_{\mathrm{KL}}(P \,||\, Q)$, captures the amount of additional information (in units of bits) needed to reconstruct probability distribution P when it is approximated by another distribution Q.

$$D_{\mathrm{KL}}(P \,||\, Q) = \sum_{x \in X} P(x) \log\left(\frac{P(x)}{Q(x)}\right), \tag{2}$$

where X is the set of possible outcomes, and $P(x)$ and $Q(x)$ represent the probabilities assigned to each outcome x by distributions P and Q, respectively. Here, we use the K-L divergence to compute the information 'cost' of approximating the empirical distribution of FC values with simulated distributions derived from each model. Specifically, we compared empirical (P) and model-derived (Q) distributions of time-series features comprised of vectors of discrete observations. These discrete distributions were converted into 100-bin probability distributions with equal bin widths. K-L divergence was computed using custom code in python (version 3.9.21).

To complement the K-L divergence (which is not a formal distance metric and may be biased by outliers), we additionally computed the Wasserstein distance [35]—also known as the 'earth mover's distance'—which, in the discrete case, minimizes both the number of samples which must be moved and the distance those samples are moved for probability distributions P and Q to match (in units of total work). The Wasserstein distance was computed using the `wasserstein_distance` function from the *scipy* package (version 1.13.1) [29].

2.6 Code and data availability

All MEG data analyzed in this study is openly available upon registration at https://doi.org/10.17617/1.wqa3-wk71 [20]. All code needed to reproduce our analyses and visuals is freely available in our GitHub

repository (https://github.com/anniegbryant/MEG_functional_connectivity) [36].

3 Results

As shown in Fig. 1A, we focus here on regions predicted by one or both theories to mediate stimulus- and/or task-specific FC: category-selective visual areas (CS; common to both IIT and GNWT), early visual areas (VIS, comprised of V1 and V2; exclusive to IIT), and lateral prefrontal cortex (PFC, comprised of Brodmann areas 9 and 46; exclusive to GNWT). Due to the combinatorial explosion that can arise from highly comparative analysis, we included an additional parietal region (PAR, comprised of the intraparietal sulcus), as the parietal cortex plays an important role in consciousness in both theories (i.e., the parietal cortex is a part of the posterior substrate of consciousness according to IIT [3], and a part of the 'global workspace' according to GNWT [5]). By including PAR as an additional region, we gained a further constraint on high-performing functional connectivity measures without biasing our search toward either theory. The label names for each area in the Schaefer et al. [22] 100-parcellation atlas are listed in Table 1. We then comprehensively quantify MEG-derived FC between CS and each of the three other regions (taking the average time series across all epochs per event type and participant) using 246 bivariate time-series measures that span an interdisciplinary literature—from spectral properties to information-theoretic measures, using the `pyspi` ('Python toolkit of statistics for pairwise interactions') library [14]. Finally, we used the top-performing FC measures to analyze simulated time series from two theory-derived neural models, allowing us to interpret the empirical results in light of the pre-registered predictions of IIT and GNWT.

Based on the original predictions for inter-areal coupling delineated by the COGITATE Consortium [11], we defined one core classification aim and two supplementary classification aims. In our main analysis targeting the core classification aim, we evaluate whether any FC measure in `pyspi` could distinguish between stimulus types (e.g., 'face' vs. 'object') in either the 'Relevant non-target' and/or 'Irrelevant' settings, in any region–region pair and during either stimulus presentation ('on') or offset ('off'). Our secondary analyses focused on the supplementary classification aims to explore whether FC measures distinguish domain-independent task relevance ('Relevant' vs. 'Irrelevant') and generalize from one relevance type to another in the same stimulus comparison. In all cases, we employed logistic regression classifiers with a grouped cross-validation strategy, reporting the mean $\pm$ standard deviation (SD) out-of-sample accuracy (see Methods, Sec 2.4 for more details).

3.1 Multiple properties of inter-areal coupling can distinguish between stimulus types across region–region pairs

For the primary stimulus-decoding analysis, we first considered the overall range of classification accuracy values across all FC measures and experimental paradigm combinations (e.g., stimulus onset vs. offset, relevant non-target vs. irrelevant category). As shown in Fig. 2, many FC measures classified stimuli with $60 - 67\%$ cross-validated accuracy across region–region pairs and category relevance settings. This range sits within the context of prior visual stimulus-decoding analyses based on inter-areal connectivity derived from electroencephalography (EEG) time series [37] or combined EEG and functional magnetic resonance imaging (fMRI) data [38]. The presence of relatively high-performing FC measures in all stimulus pairs suggests there may be conserved signatures of neural activity in the functional pathways of interest while processing different visual stimuli. However, we note the marked variance in classification performance exhibited among FC measures across both region–region pairs and stimulus pairs (as shown in Fig. 2)—underscoring the heterogeneity in FC patterns across individuals, region–region pairs, and stimulus types.

In the first supplementary classification aim, we probed the neural signatures of task relevance in a

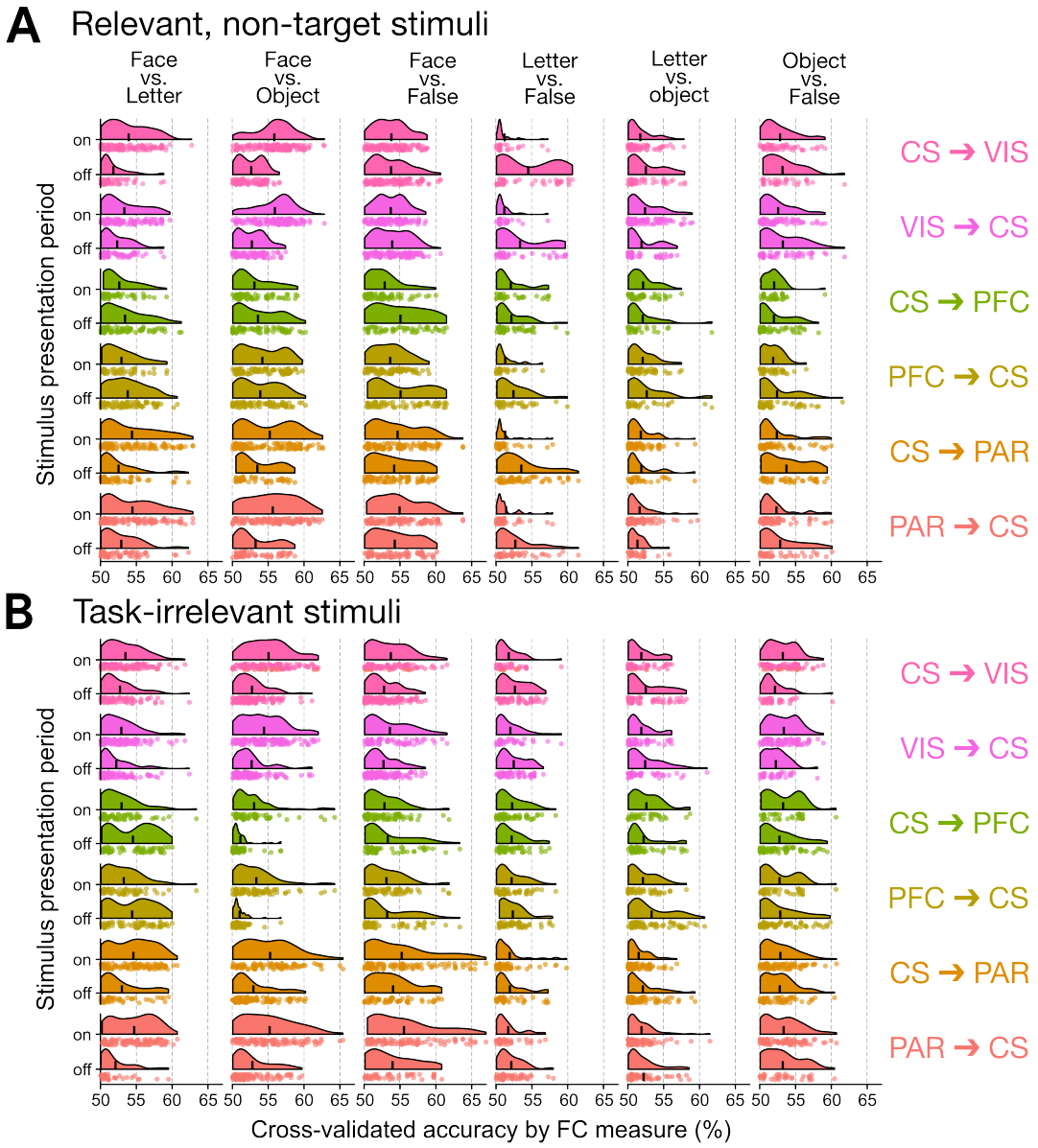

Figure 2. Many FC measures distinguish stimuli with $\geq 60\%$ accuracy across region–region pairs, particularly for face stimuli. A. Each raincloud plot shows the distribution of classification accuracy values for each SPI in the corresponding stimulus pair (column) and region–region pair (row), partitioned by relevant non-target stimulus presentation period ('on' vs. 'off'). Each dot represents one SPI, and the vertical lines represent the mean cross-validated accuracy across all FC measures in the corresponding distribution. The x-axis was truncated at 50% accuracy, such that FC measures yielding lower than 50% accuracy are not shown here. **B.** Same as in **A**, for task-irrelevant stimuli.

domain-general manner, testing whether any FC measure could distinguish 'Relevant non-target' versus 'Irrelevant' trials that combined all stimulus types. Surprisingly, none of the FC measures surpassed 56% accuracy with any region–region pair (cf. Fig. S1), indicating that inter-areal coupling is less distinctive between domain-general task relevance settings. Our second supplementary classification aim was to test 'cross-task' transfer learning (as in Consortium et al. [11]), in which we trained each classifier on FC measure values for a given stimulus pair (e.g., 'face' vs. 'object') during the 'Relevant non-target' setting (i.e., training on all $N = 94$ participants) and tested the classifier in the 'Irrelevant' setting—and vice versa. In line with the predictions of GNWT, cross-task classification yielded similar accuracy distributions across FC measures and region–region pairs (including CS $\leftrightarrow$ PFC) as in our primary stimulus type classification analysis (cf. Fig. S2); therefore, we focus on the primary stimulus decoding classification

analysis hereafter.

3.2 Statistics derived from the barycenter best distinguish stimuli, particularly faces

Having confirmed that multiple FC measures between the examined regions can distinguish between stimulus categories, we now zoom in to focus on the highest-performing measures. This allows us to make general inferences about the types of pairwise dynamics that contain the most information about visual stimulus categories. As shown in Fig. 3A, we identified the maximum classification accuracy across all FC measures, stimulus presentation periods ('on' or 'off'), and category relevance settings ('Relevant non-target' or 'Irrelevant') for each stimulus pair (e.g., 'Face' vs. 'Object') and region–region pair (e.g., VIS $\rightarrow$ CS). Across region–region pairs, the maximum classification accuracy values ranged between 62-67%, with all of the top values corresponding to face stimuli and another category. However, there is marked variability in classification performance across cross-validation folds (indicated by the shaded ribbons in Fig. 3A); therefore, while we report mean accuracy $\pm$ 1 SD, we draw inferences mainly from relative performance trends rather than absolute accuracy values.

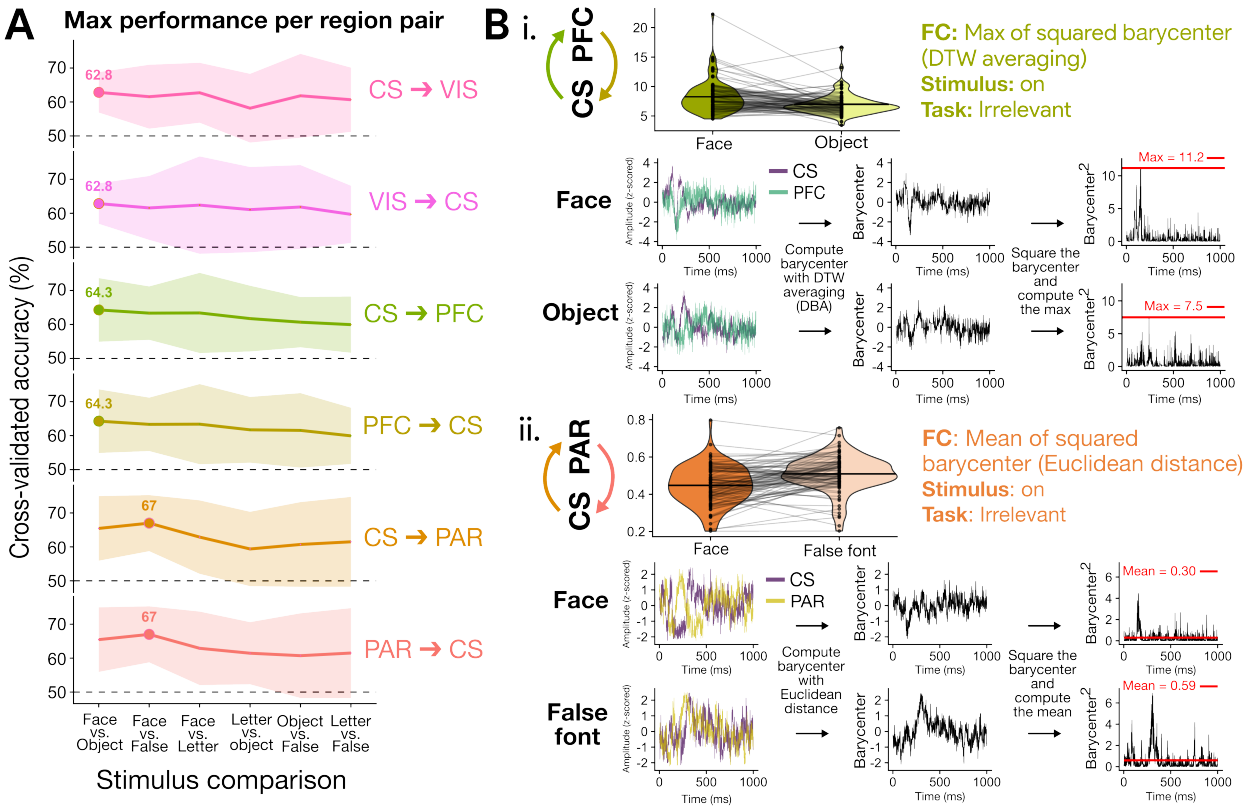

Figure 3. Examining the top-performing SPI per stimulus comparison highlights strong performance of barycenter statistics. A. For each region–region pair (rows) and stimulus comparison (columns), we plot the mean cross-validated classification accuracy for the top-performing SPI (lines), $\pm$ 1 SD (shaded ribbons) across test folds, out of all 246 FC measures, relevance conditions ('Relevant non-target' or 'Irrelevant'), and stimulus presentation periods ('on' or 'off'). The peak classification accuracy is annotated per region–region pair. **B.** (i) The top-performing SPI for CS $\leftrightarrow$ PFC—the maximum of the squared barycenter with DTW averaging—is schematically depicted. The violin plots show the SPI value distributions for task-irrelevant 'face' versus 'object' trials during stimulus onset across all $N = 94$ participants. Below, line plots schematically depict the computation of this SPI from 'face' or 'object' trial-averaged MEG time series in an example participant. (ii) As in (i), the top-performing SPI for CS $\leftrightarrow$ PAR—the mean of the squared barycenter with Euclidean geometry—is schematically depicted. The violin plots show the SPI value distributions for task-irrelevant 'face' versus 'false font' trials during stimulus onset across all $N = 94$ participants. Below, line plots schematically depict the computation of this SPI from 'face' or 'false font' trial-averaged MEG time series in an example participant (n.b., a different participant than in (i)).

To better understand the types of dynamics captured by the top-performing FC measure classes, we focused on the top two classification measures highlighted in Fig. 3A, corresponding to 'Face' vs. 'False font' in the CS $\leftrightarrow$ PAR and 'Face' vs. 'Object' in the CS $\leftrightarrow$ PFC regions, respectively (n.b., the bidirectional arrows indicate symmetric, undirected FC measure connections). In Fig. 3, we compare the feature distributions among all $N = 94$ participants for the top-performing FC measure per region–region pair. Of note, coupling between the two region–region pairs maximally distinguished stimulus categories based on different geometric formulations of the barycenter—a property originally developed in astrophysics to measure the center of mass between two orbiting celestial bodies [39], which extends here to capture the center of the pair of time series using a given distance minimization metric [40]. For CS $\leftrightarrow$ PAR, the top-performing FC measure is the mean of the squared barycenter, derived from Euclidean geometry during stimulus onset in the task-irrelevant setting ($67.0 \pm 8.3\%$ accuracy). We schematically depict the computation of this FC measure using empirical CS and PAR time series from an example participant in Fig. 3B(i). The time average of the squared Euclidean barycenter was nearly twice as high during 'False font' stimulus processing (0.59) than during 'Face' stimulus processing (0.30) for this example participant, which reflects the greater MEG fluctuation synchrony between the CS and PAR regions throughout the 1000 ms 'False font' stimulus period than during the 'Face' period.

In a similar vein, the maximum (over time) of the squared barycenter derived using dynamic time warping (DTW) averaging [41] yielded peak 'Face' vs. 'Object' classification performance for the CS $\leftrightarrow$ PFC connection ($64.3 \pm 9.4\%$ accuracy). We depict the computation of this FC measure using the empirical CS and PFC time series from an example participant in Fig. 3B(ii). Unlike in Euclidean space (as depicted in 3B[i]), DTW geometry permits dilations and/or shifts across time for distance minimization between the two time series. The maximum of the squared DTW-based barycenter provides insight into the variability in the underlying regions' local activity, as periods of higher MEG amplitude will be more heavily weighted in the temporal alignment and therefore exert greater influence on the barycenter. Indeed, the squared DTW-based barycenter peaked at a magnitude of 11.2 during 'Face' stimuli—attributable to a sharp downward spike in both CS and PFC MEG amplitude at approximately 150 ms that is archetypal for face processing [42, 43]—compared to 7.5 during 'Object' stimuli, in which MEG amplitudes were less variable and fluctuated on a slower timescale between the two regions.

3.3 Barycenter statistics dominate the top stimulus decoding results, along with basic covariance and spectral properties

In order to better understand the regional specificity of the top-performing FC measures, we next investigated the overlap in the number and type of FC measures that distinguished stimulus types well between region–region pairs. Since connectivity with the CS region is shared between IIT and GNWT, we focus on FC measures directed to and from the CS region, respectively—noting that undirected FC measures will yield bidirectionally symmetric values. We filtered the same classification results presented in Fig. 2A down to those with $\geq 60\%$ mean cross-validated accuracy and tabulated the number of resulting FC measures projecting either from or to CS (e.g., from VIS, PFC, and PAR to CS). As shown in Fig. 4A-B, some FC measures uniquely distinguished task stimuli in only one functional pathway (e.g., 13 FC measures were unique to PFC $\rightarrow$ CS connectivity). However, we found three FC measures that distinguished at least one stimulus-pair to and from CS to all three other regions, all corresponding to the maximum of the squared barycenter derived with different geometries (Euclidean, DTW, and a variation on DTW geometry known as soft-DTW [44]). This suggests that the maximum of the squared barycenter captures stimulus-relevant information (for specific stimulus types) in all functional pathways of interest to IIT and/or GNWT, poising it as a useful measure with which the theories may be quantitatively evaluated and contrasted.

All FC measures can be categorized into one of six literature-based domains ('basic', 'distance', 'spectral',

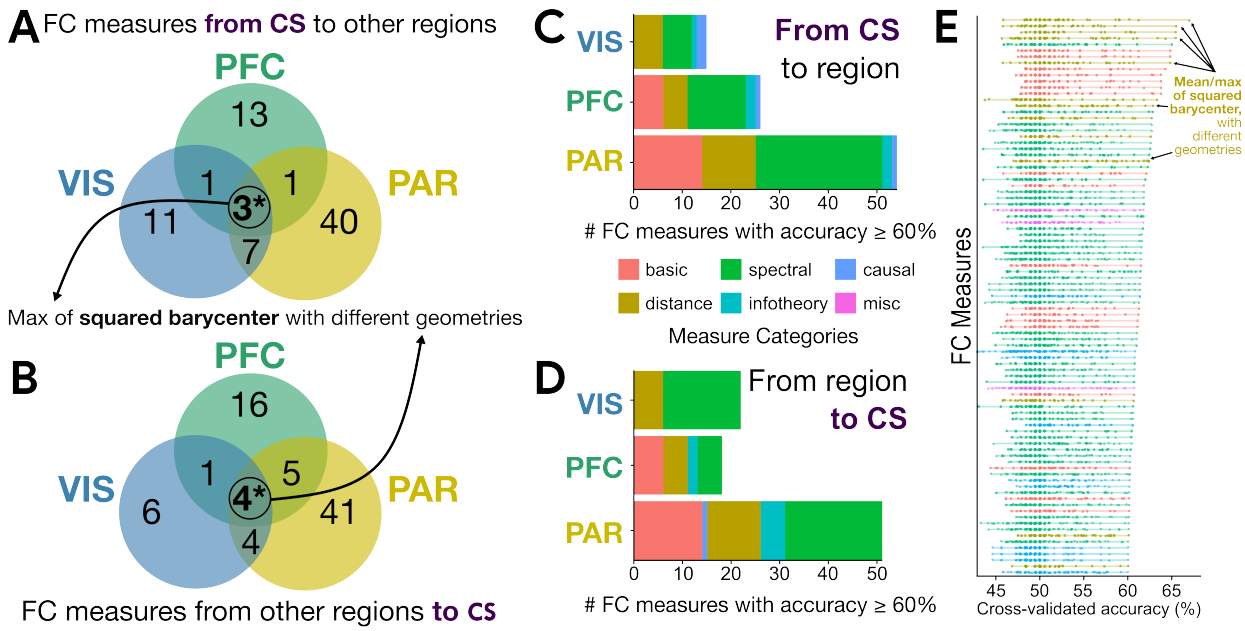

Figure 4. FC measures that detect distance-based, information-theoretic, and frequency-focused coupling distinguish between stimulus types across all region–region pairs. A,B. Here, the number of FC measures yielding at least 62% cross-validated classification accuracy in one or more stimulus pairs are tabulated for each region as either the source (**A**) or target (**B**) region the corresponding interaction. Ellipses within the Venn diagram are color-coded to indicate the corresponding region, following the brain maps depicted in Fig. 1A. The circled '1' at the center indicates that one SPI distinguished between at least one stimulus pair with $\geq 60\%$ accuracy both from and to all regions: the maximum of the squared barycenter, derived using soft-DTW geometry. **C,D**. For the FC measures tabulated in **A** and **B**, here we show the number of FC measures by literature category per region as the source (**C**) and target (**D**). **E**. For each of the 95 FC measures summarized in **A-B**, the cross-validated classification accuracy values are plotted for all 144 parameter combinations (region–region pairs, stimulus comparisons, onset vs. offset, and relevance settings). FC measures are ranked in descending order of the maximum accuracy and colored by literature category, as in **C** and **D**. The mean or maximum of the squared barycenter with different geometries are annotated for reference.

'info-theory', 'causal', or 'miscellaneous'; cf. Cliff et al. [14]), so we next compared the FC measure type composition for the top-performing measures shown in Fig. 4A-B. As shown in Fig. 4C-D, undirected FC measures corresponding to distance measures (including the barycenter) and frequency-based coupling (e.g., power envelop correlation) comprised the top-performing statistics for all three functional pathways both to and from the CS region. In addition, statistics capturing basic covariance structure (including the Pearson correlation coefficient) comprised many of the top-performing FC measures between CS $\leftrightarrow$ PFC and CS $\leftrightarrow$ PAR, but not CS $\leftrightarrow$ VIS. Although there is high-level similarity among FC measures that distinguish stimuli across these functional pathways, the relatively low overlap shown in Fig. 4 indicates that the particular FC measures differ between functional connections. The overlap between CS $\leftrightarrow$ VIS and CS $\leftrightarrow$ PFC corresponds to the mean barycenter computed with soft-DTW geometry, while the FC measure overlap between CS $\leftrightarrow$ PFC and CS $\leftrightarrow$ PAR is driven predominantly by spectral properties like power envelop correlation and imaginary coherence (more detail on these FC measures is shown in Fig. S3).

In addition to characterizing the maximum decoding accuracy for each FC measure across region–region pairs (which represents a 'best-case' stimulus discrimination scenario), we also examined the full performance range for each of the 95 FC measures contained in Figs 4A-B (i.e., all FC measures with at least $\geq 60\%$ cross-validated accuracy result). Fig. 4E shows all decoding accuracy values per FC measure, tabulated across 144 combinations of stimulus contrasts, region–region pairs, task relevance settings, and presentation periods (onset vs offset). The highest maximum accuracy values correspond

to the mean and maximum of the squared barycenter derived from different geometries, as annotated in Fig. 4E—though even these FC measures fail to surpass chance classification performance (i.e., 50%) in many settings. This visualization demonstrates the variability in decoding accuracy exhibited by all 95 examined FC measures, none of which outperformed chance in all 144 experimental parameter combinations.

3.4 Testing predictions via data-driven analysis of theory-driven neural modeling

In the previous sections, we identified a multitude of FC measures that met our core stimulus decoding classification aim across region–region pairs. In particular, FC measures derived from the barycenter generalized across all of the relevant region–region pairs relevant to both IIT and GNWT. Barycenter-based measures are fast to compute, simple to interpret, and novel to functional neuroimaging—prompting further investigation with modeling. To derive theoretical insight into the neural processes predicted to underlie conscious vision from our novel FC measure, we constructed two neural mass models [19] tailored to recapitulate each theory's pre-registered predictions (see Consortium et al. [11]) for neural activity in each region (for details see Sec. 2.5 and schematic depiction in Fig. 1B).

Importantly, because each theory generates predictions about coupling between different region–region pairs (CS $\leftrightarrow$ VIS for IIT, and CS $\leftrightarrow$ PFC for GNWT), our goal in this section is to evaluate the relative success (rather than the overall variance explained) of each theory-derived model in successfully approximating the empirical data. The models are stimulus-agnostic and are built to reproduce the empirical Euclidean barycenter values averaged over all stimulus types. We focused on the stimulus-averaged dynamics rather than the stimulus-specific dynamics, as neither GNWT nor IIT provided stimulus-specific predictions. Rather, they specified predictions about the qualitative features of inter-areal coupling (e.g. synchronized activity between CS and VIS) hypothesized to underlie conscious perception. In addition, to avoid a mismatch between simulated and empirical barycenter values caused by changes in dipole polarity across regions (cf. Fig. S4A-B), we specifically computed the absolute value (i.e., magnitude) of both the empirical and simulated time series before computing the Euclidean barycenter in this section. As a robustness test, we show in Fig. S5 that the stimulus classification results based upon the barycenter are not systematically biased by this extra preprocessing step.

Our model simulation process is schematically depicted in Fig. 5A. Using the IIT- and GNWT-based models, we simulated 1000 trials with dynamical noise and then subtracted the average neural activity before stimulus onset (i.e., the 'baseline') from each trial (Fig. 5A[i-ii]). We approximated noise in the measurement process by adding zero-mean independent Gaussian noise to each trial (Fig. 5A[iii]). Each trial was divided into 1000 ms segments corresponding to stimulus-on and stimulus-off periods for analysis. We show results obtained using $\sigma = 1.0$ added noise in the main text, with results from the full range of $0.5 \leq \sigma \leq 1.0$ depicted in Fig. S7. Within each stimulus-on and stimulus-off epoch, we computed the maximum of the squared barycenter across Euclidean and non-Euclidean geometries (Fig. 5A[iv]). For clarity, we show the results for the maximum of the squared barycenter derived from Euclidean geometry. Results are qualitatively comparable for barycenter measures computed using DTW, soft-DTW, and subgradient descent- (SGD-)DTW geometries (cf. Fig. S6).

Briefly, as schematically represented in Fig. 5B(i), GNWT predicts that visual consciousness is dependent on neural signal propagation from CS to PFC regions, leading to a late 'ignition' of the global workspace after stimulus onset and again at stimulus offset. The GNWT-based model has two key features responsible for the ignition-like updating dynamics in PFC at stimulus onset and offset: (1) the winner-take-all competition between the two PFC masses, and (2) the slow timescale of the PFC neural masses ($\tau_{PFC} = 80$ ms) relative to the comparably fast CS neural mass ($\tau_{CS} = 40$ ms). This leads to a peak in PFC activity ~ 200 ms after stimulus onset—approximately 150 ms after the peak in CS activity,

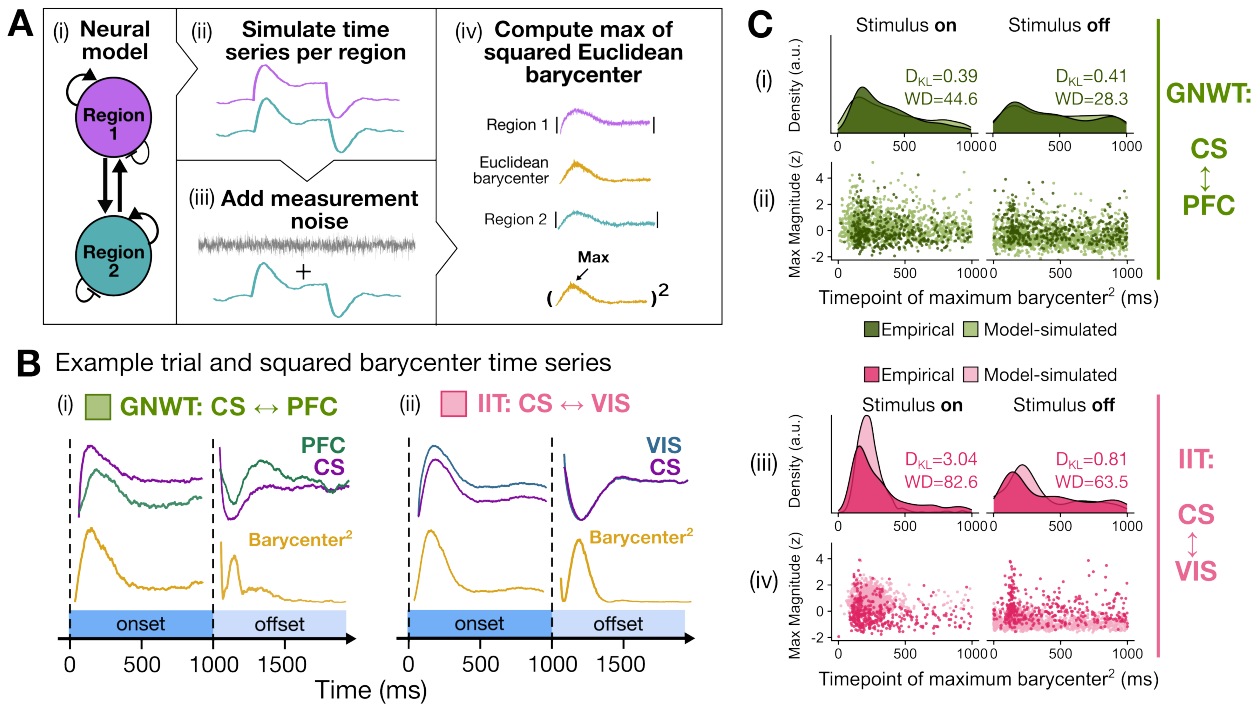

Figure 5. Comparing model-based predictions for barycenter with results from empirical COGITATE MEG data. A Graphical illustration of model analysis pipeline. For both models we simulated 1000-trials with noise, and after baseline subtraction and added zero mean Gaussian noise to each time point of the simulation to mimic the effect of measurement noise before computing the max of the squared barycenter for each trial. For comparability with the empirical data the barycenter was computed after taking the absolute value of the time series avoiding issues of dipole polarity reversal across regions. **B.** Illustration of single trial neural activity and barycenter times series for GNWT- and IIT-based neural mass models. Because the of offset between CS and PFC masses, and the winner-take all dynamics, the timing of the peak barycenter values is more variable across trials in the GNWT-based model than in the IIT-based model. **C.** Here, we examine the timing and magnitude of the squared Euclidean barycenter maximum in both empirical and model-simulated time series (with noise $\sigma = 1$) for both IIT and GNWT during stimulus onset and offset. First, the timing of the barycenter maximum for each (i) CS↔PFC and (iii) CS↔VIS bivariate time series is summarized via kernel density estimation for each stimulus presentation period and theory. The empirical density distribution is dark-shaded and the model-simulated density distribution is light-shaded in each case. The Kullback-Leibler divergence [34], $D_{KL}(\text{Empirical} \,||\, \text{Model})$, is shown for each theory and stimulus presentation period, in units of bits, along with the Wasserstein distance (WD) in units of total work. (ii, iv) The same maximum squared Euclidean barycenter values comprising the density plots in (i) and (iii) are plotted, with the maximum timing on the x-axis and the magnitude of the corresponding squared Euclidean barycenter on the y-axis. The same color scheme is used as in (i) and (iii).

which peaks ~ 50 ms after stimulus onset.

In terms of the (squared) barycenter, which tracks the center of mass of the two time series, the offset in peak activity between the two masses reduces the amplitude of the (squared) barycenter. The relatively low amplitude of the peak PFC signal also means that in the presence of measurement noise, the maximum of the barycenter is broadly distributed throughout the stimulus-on epoch. At stimulus offset, both the CS and PFC masses show a brief dip in activity due to the accumulated adaptation current. The winner-take-all competition between the two PFC masses then generates a secondary bump of activity approximately 300 ms after stimulus offset, at which point the previously suppressed mass takes over. Similar to the stimulus-on period, the offset in the peak activity between the CS and PFC masses, combined with measurement noise, yields a broad distribution of maximum (squared) barycenter values across the stimulus-off period. In contrast, as schematically represented in Fig. 5B(ii), IIT predicts that the conscious experience of a visual stimulus involves synchronized and sustained activity between CS and VIS (i.e., 'posterior hot zone') throughout stimulus presentation. The key features driving the dynamics of the IIT-based model are the strong reciprocal connectivity between the VIS and CS masses, and the

comparably fast and similar timescale of the two masses ($\tau_{VIS} = 20$ ms, $\tau_{CS} = 40$ ms). Because of the similar timescale and strong reciprocal connectivity, both masses increase (and decrease) synchronously following stimulus onset and offset. During the stimulus presentation period, activity is maintained in a stable high-activity attractor state. The synchronized peaks and troughs in activity at stimulus onset and offset generate large peaks in the (squared) barycenter amplitude, yielding a high-density concentration of maximum timing values at the beginning of the epochs across trials.

Although qualitative comparison of the empirical and simulated distributions shows that both the GNWT- and IIT-based roughly approximate the empirical data, quantitative comparison of the simulated and empirical distribution of maximum squared Euclidean barycenter timings shows that the GNWT-based model provides a better approximation to the empirical data than the IIT-based model with lower K-L divergence values in both stimulus-onset ($D_{\mathrm{KL}}[\mathrm{GNWT}](P \,||\, Q)$ = 0.39 bits vs. $D_{\mathrm{KL}}[\mathrm{IIT}](P \,||\, Q)$ = 3.04 bits) and offset ($D_{\mathrm{KL}}[\mathrm{GNWT}](P \,||\, Q)$ = 0.41 bits vs. $D_{\mathrm{KL}}[\mathrm{IIT}](P \,||\, Q)$ = 0.81 bits) periods. This result is also reflected in the Wasserstein distance at stimulus-onset ($W_{\mathrm{GNWT}} = 44.6$ vs. $W_{\mathrm{IIT}} = 82.6$) and stimulus-offset ($W_{\mathrm{GNWT}} = 28.3$ vs. $W_{\mathrm{IIT}} = 63.5$). In other words, simulated maximum squared barycenter values from the GNWT-based model provided a better approximation to the empirical data than those derived from the IIT-based model in terms of both information cost (lower K-L divergence values) and Wasserstein distance. Importantly, this pattern is robust across various noise levels (cf. Fig. S7) and task relevance settings (cf. Fig. S8). We show in Fig. S4F(ii) that the PFC does indeed show a secondary bump in activity after stimulus onset, as GNWT predicts. Based on these results, we argue that the delayed ignition dynamics implemented in the winner-take-all connectivity of the GNWT-based model provide a better approximation to the empirical data than the tightly synchronized sensory dynamics implemented in the IIT-based model. This conclusion is, of course, dependent on the model-based implementation of the theoretical predictions and may be subject to revision.

4 Discussion

Here, we presented an approach for data-driven identification, evaluation, and theoretical interpretation of connectivity-based neural correlates of consciousness. Our work is built upon the datasets and preregistered hypotheses generated by the COGITATE Consortium adversarial collaboration that tested the predictions of IIT and GNWT [11]. We focused here on functional connectivity with the aim of finding a measure(s) that provided reliable stimulus-type decoding—reasoning that if a neural process underlies the content of consciousness, variation in that process should encode information relevant to the contents of conscious perception. Through the most comprehensive comparison performed to date, we systematically evaluated 246 different measures of pairwise coupling, derived from an interdisciplinary time series literature [14], computed between regions hypothesized to drive conscious vision in IIT and/or GNWT.

For each region–region pair, we identified multiple FC measures that decoded stimulus type with $> 60\%$ cross-validated (out-of-sample) accuracy. Importantly, in line with the predictions of GNWT, task relevance did not generally modulate stimulus-type decoding accuracy across FC measures and region–region pairs (including CS $\leftrightarrow$ PFC). Classifiers trained on a pair of stimuli in the task-relevant setting generalized to the same stimuli in the task-irrelevant setting, and vice versa, through a cross-validation paradigm. Top-performing measures included those commonly used to quantify activity from MEG data, such as the power envelop correlation [16, 17], as well as those novel to functional neuroimaging applications, such as the barycenter derived from Euclidean and non-Euclidean geometry. Indeed, the top-performing time series features that provided $> 60\%$ stimulus decoding accuracy across region–region pairs common to both GNWT (CS $\leftrightarrow$ PFC) and IIT (CS $\leftrightarrow$ VIS), were based upon the maximum of the squared barycenter. The barycenter tracks the center of mass of two time series, and is sensitive to both the amplitude and

relative phase of each time series. The maximum of the squared barycenter, therefore, identifies the time point of maximum covariance of both amplitude and phase. Since this class of features maximally decoded stimulus types, we reasoned that statistics derived from the barycenter form strong candidates for encoding information relevant to conscious visual perception.

We next sought to derive theoretical insight into the neural processes hypothesized to underlie conscious vision by comparing empirical barycenter values to simulated barycenter values derived from two simple neural mass models [19] tailored to recapitulate the pre-registered predictions (see Consortium et al. [11]) for what neural activity should be observed in each region by IIT and GNWT. The winner-take-all competition between the two PFC masses, coupled with their slow timescale, yielded a broad distribution of maximum (squared) barycenter timings across the stimulus onset period and stimulus offset period. In contrast, in the IIT-based model, the CS and VIS masses exhibit a similar timescale and strong reciprocal connectivity, such that both masses increased (and decreased) synchronously at stimulus onset and offset. This generated tightly peaked distributions of maximum squared barycenter values at the beginning of each onset and offset period. Quantitatively, simulated barycenter values from the GNWT-based model provided a better approximation for the empirical data than the IIT-based model, in terms of both information cost (lower K-L divergence values [34]) and the Wasserstein distance (also known as 'earth mover's distance' [35]). It can therefore be argued that the delayed ignition dynamics implemented in the GNWT-based model provide a better approximation to the neural processes generating the observed data than the highly synchronized sensory dynamics implemented in the IIT-based model.

When combined with the result that stimulus decoding from CS $\leftrightarrow$ PFC FC measures generalized between task-relevant and task-irrelevant conditions, our findings lend more credence to the predictions of GNWT than the predictions of IIT. Whereas GNWT is agnostic about the role of synchrony in sensory cortex, IIT predicts that ignition-like dynamics in PFC reflect task demands rather than conscious perception. The success of the GNWT model in quantitatively accounting for the data across task-relevance conditions is therefore contrary to the predictions of IIT. Again, however, we note that this result is conditional upon the model-based implementation of hypotheses from GNWT and IIT, and may be subject to revision should another model—still consistent with the hypotheses—be put forward. In practice, theory testing (and refinement) is an iterative, rather than absolute, process [45, 46], so we regard this as a strength, rather than weakness, of our approach. In addition, the predictions set out by COGITATE Consortium [11] are not derived from the core of IIT and GNWT, and therefore may be revised whilst leaving the core tenets of each theory intact [45]. This is not to say that the accuracy of the predictions should not increase or decrease our credence in each theory; rather, this is just to point out that we should not expect to falsify [47] IIT or GNWT outright on the basis of the results.

Before turning to a more general discussion of the approach taken in this paper, we highlight technical limitations that should be considered when interpreting this work and our findings. Even among the top-performing FC measures (including statistics derived from the barycenter), there is variability in stimulus decoding performance across different experimental parameters. Much like the 'no free lunch' theorem in machine learning [48, 49], our findings do not suggest a singular FC measure that will universally perform well across all stimulus decoding contexts. It remains to be seen, therefore, whether statistics based upon the barycenter generalize as top performers beyond the current context. Additionally, we note that our models are phenomenological and should be thought of as quantitative summaries of the neural dynamics hypothesized to underlie conscious perception by IIT and GNWT—rather than full implementations of the theories themselves.

There are also potential limitations specific to our MEG-based connectivity analyses. For example, we averaged across all epochs within a given event-related field type for each participant, such that each of the $N = 94$ participants was represented by a single data point per FC measure in a given

classification analysis. While the temporal averaging was performed to enhance the signal-to-noise ratio and reduce computational load [28], this design might also obscure relevant inter-trial differences and/or inter-individual heterogeneity in neural responses to the same stimulus type and experimental setting. Additionally, we employed the same preprocessing pipeline as in Consortium et al. [11], which employed the widely-used MNE-Python workflow [50]. However, issues remain regarding proper source localization, environmental/biological artifact removal, and latency variability, the latter of which poses issues for interpreting asynchronous (e.g., lagged) FC measures [51]. Field spread is also a concern due to the difference in spatial proximity between regions—for example, VIS and CS regions are much closer together than VIS and PFC—which also means that VIS and CS regions are more likely to exhibit similar MEG time series due to spatial autocorrelation that decays with distance [52]. We do note, however, that this would bias VIS and CS towards greater synchrony (rather than less) such that this limitation does not affect the theoretical conclusions we draw above.

Beyond the conclusions and limitations of the current study, we present a roadmap for identifying and evaluating candidate neural correlates of consciousness in a data-driven manner that integrates with theory-driven modeling to derive quantitative and interpretable insights. Importantly, the data-driven nature of this approach means that one need not make *a priori* assumptions about which time series feature(s) will be best suited for the process of interest. Once informative time-series features are identified, modeling plays an indispensable role in mapping between the neural processes hypothesized to underlie the observed dynamics and the empirically measured time series features themselves. The models we specified here are deliberately as simple as possible, ensuring interpretability, and are also specifically constructed to simulate barycenter values, which are sensitive to the relative amplitude and phase of the two time series. Of note, given the rich literature in quantitative neural theory [31, 32], it will almost always be possible to build a computational model of the neural process of interest that has a sufficient level of complexity to capture the features picked out by a highly comparative data-driven approach. For example, had our data-driven evaluation of FC measures identified a frequency-based measure as the top performing feature (such as the power envelop correlation), a more complex modeling approach that explicitly incorporates frequency-based interactions between neural masses (e.g., Mejias et al. [53]) would have been more appropriate.

One key advantage of formal modeling is that even if the predicted time-series features are not identified as informative, if the model includes the full suite of neural processes hypothesized by a theory to underlie a cognitive/perceptual process of interest, the model can still be used to inform the interpretation of the time series features that are identified as informative. Although it is not always possible, models should ideally be built before data is collected. Indeed, a key limitation of the current study is that we built two models based upon preregistered *qualitative* hypotheses. Because these hypotheses are specified in words rather than equations, there will always be an imprecise mapping between formal models and the hypotheses themselves. It is therefore critical for the progression of the field that we move towards formulating hypotheses in quantitative terms that can directly link to formal models—which provide a concrete starting point for revisions that may better explain the observed empirical data. Here we fit the parameters of each model by hand, an important next step in the analysis pipeline will therefore involve leveraging previous work on model fitting and model comparison to fit and evaluate families of theory-derived models [54, 55].

To our knowledge, this is the first study to systematically compare hundreds of FC measures for decoding visual stimuli across task relevance settings. We focused here on bivariate coupling measures using the `pyspi` toolbox [14], though this approach is readily extendable to highly comparative analysis of both univariate time series [56–58] (to focus on local dynamics of a single region) or higher-order interactions comprising three or more regions, a rapidly growing area of research [59, 60]. Combining across different scales of dynamics, from local to higher-order, could paint a fuller picture of the rich neural

complexity driving a biological phenomenon like conscious visual perception [15]. In addition, an important avenue for future work will be to quantitatively compare measures that distinguish between conscious and unconscious stimuli, allowing us to isolate the processes that make a stimulus conscious—rather than focusing on the neural process underlying consciously perceived stimuli, as we did here. Importantly, our data-driven approach also generalizes across neuroimaging modalities and, when combined with neural modeling, offers clear insights into the processes hypothesized to generate empirically observed brain dynamics. Large openly available datasets are becoming more and more common in consciousness science, while the field is concurrently moving towards more quantitative neural theories [61–66]. There is, therefore, a unique opportunity for collaborations between experimentalists, data analysts, and neural modelers to systematically build, and quantitatively compare, measures and theories of consciousness.

5 Acknowledgments

We are grateful to the COGITATE Consortium for their Herculean effort in setting up the adversarial collaboration, and for making the data openly available. In addition, we would like to thank our supervisors Ben Fulcher and Mac Shine for their guidance and encouragement throughout this project.

References

1. C. Koch, M. Massimini, M. Boly, and G. Tononi. Neural correlates of consciousness: Progress and problems. *Nature reviews neuroscience* **17**, 307 (2016).

2. I. Yaron, L. Melloni, M. Pitts, and L. Mudrik. The ConTraSt database for analysing and comparing empirical studies of consciousness theories. *Nature Human Behaviour* **6**, 593 (2022).

3. L. Albantakis, L. Barbosa, G. Findlay, et al. Integrated information theory (IIT) 4.0: Formulating the properties of phenomenal existence in physical terms. *PLoS computational biology* **19**, e1011465 (2023).

4. S. Dehaene and L. Naccache. Towards a cognitive neuroscience of consciousness: Basic evidence and a workspace framework. *Cognition* **79**, 1 (2001).

5. G. A. Mashour, P. Roelfsema, J.-P. Changeux, and S. Dehaene. Conscious processing and the global neuronal workspace hypothesis. *Neuron* **105**, 776 (2020).

6. B. J. Baars. *A cognitive theory of consciousness*. Cambridge University Press (1993).

7. S. Edelman, K. Grill-Spector, T. Kushnir, and R. Malach. Toward direct visualization of the internal shape representation space by fMRI. *Psychobiology* **26**, 309 (1998).

8. N. Kanwisher, J. McDermott, and M. M. Chun. The fusiform face area: A module in human extrastriate cortex specialized for face perception. *Journal of neuroscience* **17**, 4302 (1997).

9. M. Boly, M. Massimini, N. Tsuchiya, et al. Are the neural correlates of consciousness in the front or in the back of the cerebral cortex? Clinical and neuroimaging evidence. *Journal of Neuroscience* **37**, 9603 (2017).

10. L. Melloni, L. Mudrik, M. Pitts, et al. An adversarial collaboration protocol for testing contrasting predictions of global neuronal workspace and integrated information theory. *Plos one* **18**, e0268577 (2023).

11. C. Consortium, O. Ferrante, U. Gorska-Klimowska, et al. An adversarial collaboration to critically evaluate theories of consciousness. *BioRxiv* pp. 2023–06 (2023).

12. D. Wang and Y. Sun. The pairwise phase consistency in cortical network and its relationship with neuronal activation. In *BIO Web of Conferences*, volume 8, p. 02006. EDP Sciences (2017).

13. M. Vinck, M. van Wingerden, T. Womelsdorf, P. Fries, and C. M. A. Pennartz. The pairwise phase consistency: A bias-free measure of rhythmic neuronal synchronization. *NeuroImage* **51**, 112 (2010).

14. O. M. Cliff, A. G. Bryant, J. T. Lizier, N. Tsuchiya, and B. D. Fulcher. Unifying pairwise interactions in complex dynamics. *Nature Computational Science* **3**, 883 (2023).

15. A. G. Bryant, K. Aquino, L. Parkes, A. Fornito, and B. D. Fulcher. Extracting interpretable signatures of whole-brain dynamics through systematic comparison. *PLOS Computational Biology* **20**, e1012692 (2024).

16. J. F. Hipp, D. J. Hawellek, M. Corbetta, M. Siegel, and A. K. Engel. Large-scale cortical correlation structure of spontaneous oscillatory activity. *Nature Neuroscience* **15**, 884 (2012). Publisher: Nature Publishing Group.

17. S. Khan, J. A. Hashmi, F. Mamashli, et al. Maturation trajectories of cortical resting-state networks depend on the mediating frequency band. *NeuroImage* **174**, 57 (2018).

18. M. Oizumi, S.-i. Amari, T. Yanagawa, N. Fujii, and N. Tsuchiya. Measuring integrated information from the decoding perspective. *PLoS computational biology* **12**, e1004654 (2016).

19. M. Breakspear. Dynamic models of large-scale brain activity. *Nature neuroscience* **20**, 340 (2017).

20. L. Liu, O. Ferrante, T. Ghafari, et al. Cogitate M-EEG data [dataset] (2024).

21. A. Gramfort, M. Luessi, E. Larson, et al. MEG and EEG data analysis with MNE-Python. *Frontiers in Neuroscience* **7**, 1 (2013).

22. A. Schaefer, R. Kong, E. M. Gordon, et al. Local-global parcellation of the human cerebral cortex from intrinsic functional connectivity MRI. *Cerebral cortex* **28**, 3095 (2018).

23. B. T. Yeo, F. M. Krienen, J. Sepulcre, et al. The organization of the human cerebral cortex estimated by intrinsic functional connectivity. *Journal of neurophysiology* (2011).

24. O. Ferrante, L. Liu, T. Minarik, et al. Flux: A pipeline for meg analysis. *NeuroImage* **253**, 119047 (2022).

25. S. Taulu, M. Kajola, and J. Simola. Suppression of interference and artifacts by the signal space separation method. *Brain Topography* **16**, 269 (2004).

26. A. Hyvärinen and E. Oja. Independent component analysis: Algorithms and applications. *Neural Networks* **13**, 411 (2000).

27. A. M. Dale, A. K. Liu, B. R. Fischl, et al. Dynamic statistical parametric mapping: Combining fMRI and MEG for high-resolution imaging of cortical activity. *Neuron* **26**, 55 (2000).

28. T. Grootswagers, S. G. Wardle, and T. A. Carlson. Decoding dynamic brain patterns from evoked responses: A tutorial on multivariate pattern analysis applied to time series neuroimaging data. *Journal of cognitive neuroscience* **29**, 677 (2017).

29. P. Virtanen, R. Gommers, T. E. Oliphant, et al. SciPy 1.0: Fundamental algorithms for scientific computing in python. *Nature Methods* **17**, 261 (2020).

30. F. Pedregosa, G. Varoquaux, A. Gramfort, et al. Scikit-learn: Machine learning in Python. *the Journal of machine learning research* **12**, 2825 (2011). Publisher: JMLR. org.

31. W. Gerstner, W. M. Kistler, R. Naud, and L. Paninski. *Neuronal dynamics: From single neurons to networks and models of cognition*. Cambridge University Press (2014).

32. P. Dayan and L. F. Abbott. *Theoretical neuroscience: computational and mathematical modeling of neural systems*. MIT press (2005).

33. H. R. Wilson. Spikes, decisions, and actions: the dynamical foundations of neuroscience. *(Oxford University Press)* (1999).

34. S. Kullback and R. A. Leibler. On information and sufficiency. *The annals of mathematical statistics* **22**, 79 (1951).

35. L. N. Vaserstein. Markov processes over denumerable products of spaces, describing large systems of automata. *Problemy Peredachi Informatsii* **5**, 64 (1969).

36. A. G. Bryant and C. Whyte. anniegbryant/meg_functional_connectivity (2025).

37. O. Ratcliffe, K. Shapiro, and B. P. Staresina. Fronto-medial theta coordinates posterior maintenance of working memory content. *Current Biology* **32**, 2121 (2022).

38. C. L. Scrivener, J. A. Teed, and E. H. Silson. Visual imagery of familiar people and places in category selective cortex. *bioRxiv* pp. 2024–09 (2024).

39. V. Brumberg. *Essential Relativistic Celestial Mechanics*. CRC Press (1991).

40. F. Petitjean, A. Ketterlin, and P. Gançarski. A global averaging method for dynamic time warping, with applications to clustering. *Pattern recognition* **44**, 678 (2011).

41. D. Schultz and B. Jain. Nonsmooth analysis and subgradient methods for averaging in dynamic time warping spaces. *Pattern Recognition* **74**, 340 (2018).

42. I. Deffke, T. Sander, J. Heidenreich, et al. MEG/EEG sources of the 170-ms response to faces are co-localized in the fusiform gyrus. *Neuroimage* **35**, 1495 (2007).

43. N. Hadjikhani, K. Kveraga, P. Naik, and S. P. Ahlfors. Early (m170) activation of face-specific cortex by face-like objects. *Neuroreport* **20**, 403 (2009).

44. M. Cuturi and M. Blondel. Soft-dtw: A differentiable loss function for time-series. In *International conference on machine learning*, pp. 894–903. PMLR (2017).

45. N. Negro. (Dis) confirming theories of consciousness and their predictions: towards a Lakatosian consciousness science. *Neuroscience of Consciousness* **2024**, niae012 (2024).

46. A. W. Corcoran, J. Hohwy, and K. J. Friston. Accelerating scientific progress through Bayesian adversarial collaboration. *Neuron* **111**, 3505 (2023).

47. P. Godfrey-Smith. Theory and reality: An introduction to the philosophy of science. In *Theory and reality*. University of Chicago Press (2009).

48. D. H. Wolpert. The lack of a priori distinctions between learning algorithms. *Neural Computation* **8**, 1341 (1996).

49. D. Wolpert and W. Macready. No free lunch theorems for optimization. *IEEE Transactions on Evolutionary Computation* **1**, 67 (1997). Conference Name: IEEE Transactions on Evolutionary Computation.

50. A. Gramfort. MEG and EEG data analysis with MNE-Python. *Frontiers in Neuroscience* **7** (2013).

51. S. Silva Pereira, R. Hindriks, S. Mühlberg, et al. Effect of field spread on resting-state magneto-encephalography functional network analysis: A computational modeling study. *Brain Connectivity* **7**, 541 (2017).

52. A. Arnatkeviciūtė, B. D. Fulcher, R. Pocock, and A. Fornito. Hub connectivity, neuronal diversity, and gene expression in the Caenorhabditis elegans connectome. *PLOS Computational Biology* **14**, e1005989 (2018). Publisher: Public Library of Science.

53. J. F. Mejias, J. D. Murray, H. Kennedy, and X.-J. Wang. Feedforward and feedback frequency-dependent interactions in a large-scale laminar network of the primate cortex. *Science advances* **2**, e1601335 (2016).

54. L. Novelli, K. Friston, and A. Razi. Spectral dynamic causal modeling: A didactic introduction and its relationship with functional connectivity. *Network Neuroscience* **8**, 178 (2024).

55. P. Zeidman, K. Friston, and T. Parr. A primer on variational laplace (VL). *NeuroImage* **279**, 120310 (2023).

56. B. D. Fulcher, M. A. Little, and N. S. Jones. Highly comparative time-series analysis: The empirical structure of time series and their methods. *Journal of the Royal Society Interface* **10**, 20130048 (2013).

57. B. D. Fulcher and N. S. Jones. hctsa: A computational framework for automated time-series phenotyping using massive feature extraction. *Cell systems* **5**, 527 (2017).

58. G. Shafiei, B. D. Fulcher, B. Voytek, et al. Neurophysiological signatures of cortical micro-architecture. *Nature communications* **14** (2023).

59. A. Santoro, F. Battiston, M. Lucas, G. Petri, and E. Amico. Higher-order connectomics of human brain function reveals local topological signatures of task decoding, individual identification, and behavior. *Nature Communications* **15**, 10244 (2024).

60. B. Camino-Pontes, A. Jimenez-Marin, I. Tellaetxe-Elorriaga, et al. Brain structural modules associated to functional high-order interactions in the human brain. *bioRxiv* pp. 2025–03 (2025).

61. C. J. Whyte, A. W. Corcoran, J. Robinson, et al. On the minimal theory of consciousness implicit in active inference. *arXiv preprint arXiv:2410.06633* (2024).

62. C. J. Whyte, B. R. Munn, J. Aru, et al. A burst-dependent thalamocortical substrate for visual rivalry. *bioRxiv* pp. 2023–07 (2023).

63. U. Klatzmann, S. Froudist-Walsh, D. P. Bliss, et al. A dynamic bifurcation mechanism explains cortex-wide neural correlates of conscious access. *Cell Reports* **44** (2025).

64. S. Dehaene, C. Sergent, and J.-P. Changeux. A neuronal network model linking subjective reports and objective physiological data during conscious perception. *Proceedings of the National Academy of Sciences* **100**, 8520 (2003).

65. S. Dehaene, M. Kerszberg, and J.-P. Changeux. A neuronal model of a global workspace in effortful cognitive tasks. *Proceedings of the national Academy of Sciences* **95**, 14529 (1998).

66. S. M. Fleming. Awareness as inference in a higher-order state space. *Neuroscience of consciousness* (2020).

Supporting Information Appendix for 'A data-driven approach to identifying and evaluating connectivity-based neural correlates of consciousness'

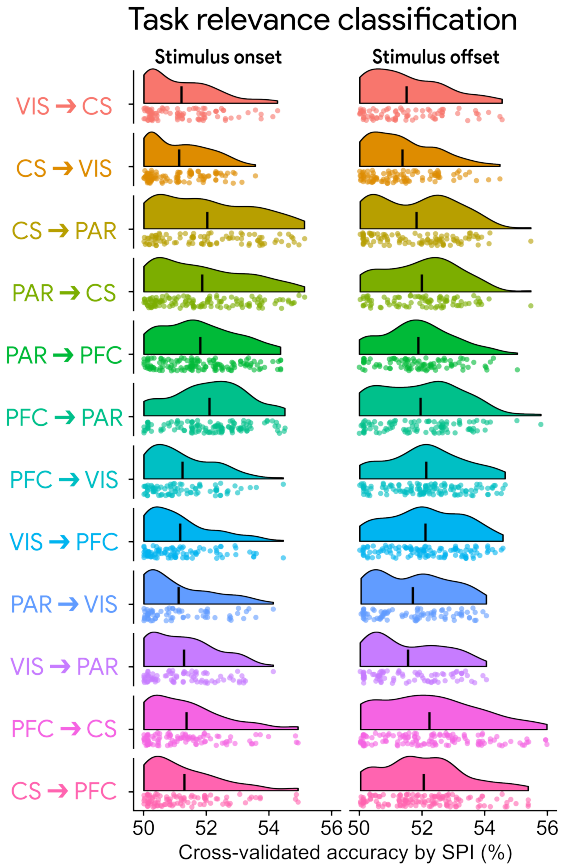

Figure S1. Domain-independent task relevance is not well classified across FC measures and region–region pairs. Each raincloud plot shows the distribution of classification accuracy values for each FC measure in the corresponding region–region pair, partitioned by stimulus onset or stimulus offset period. For visualization purposes, the x-axis was truncated at 50% accuracy for both stimulus presentation periods, such that SPIs yielding lower than 50% accuracy are not shown here.

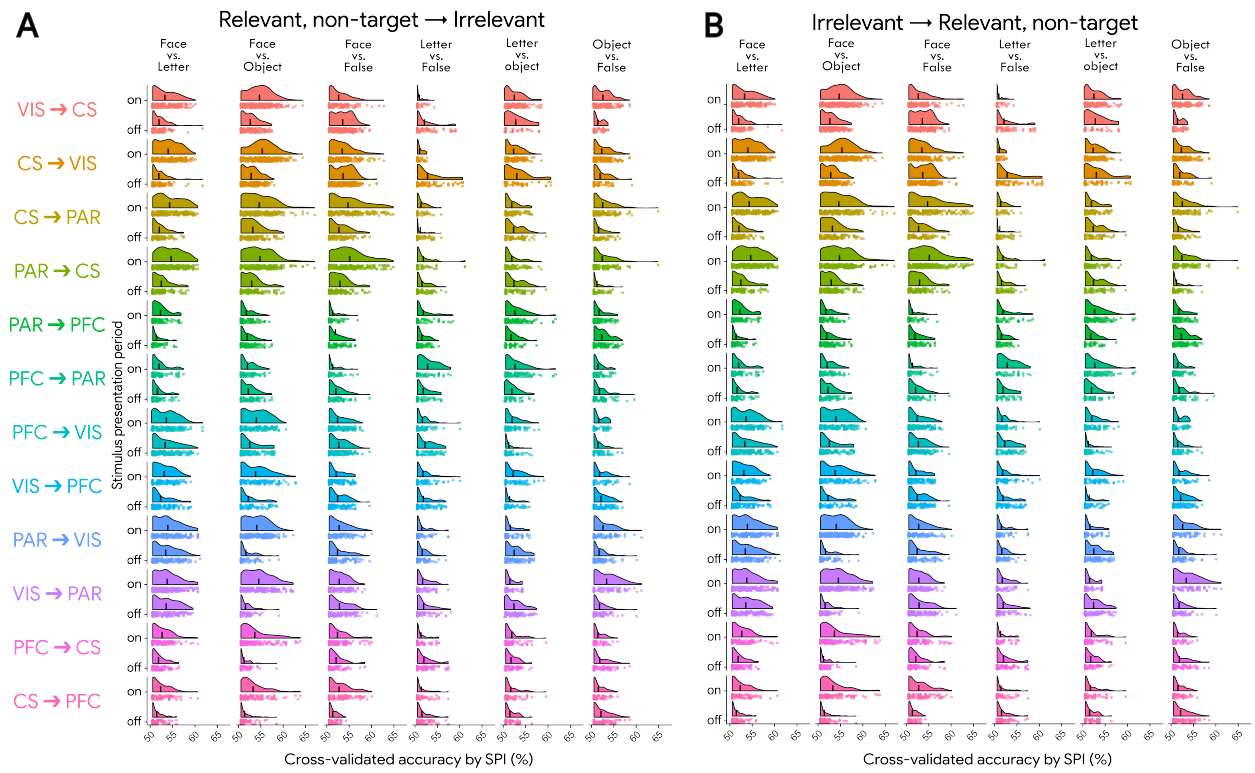

Figure S2. Cross-task classification between stimulus pairs shows comparable performance to that seen in the main stimulus type classification. Each raincloud plot shows the distribution of classification accuracy values for each FC measure in the corresponding region–region pair, resulting from classifiers trained on stimulus pairs in one relevance setting (e.g., 'Relevant non-target') and tested on the same stimulus pairs in the other relevance setting (e.g., 'Irrelevant'). For visualization purposes, the x-axis was truncated at 50% accuracy for both stimulus presentation periods, such that SPIs yielding lower than 50% accuracy are not shown here.

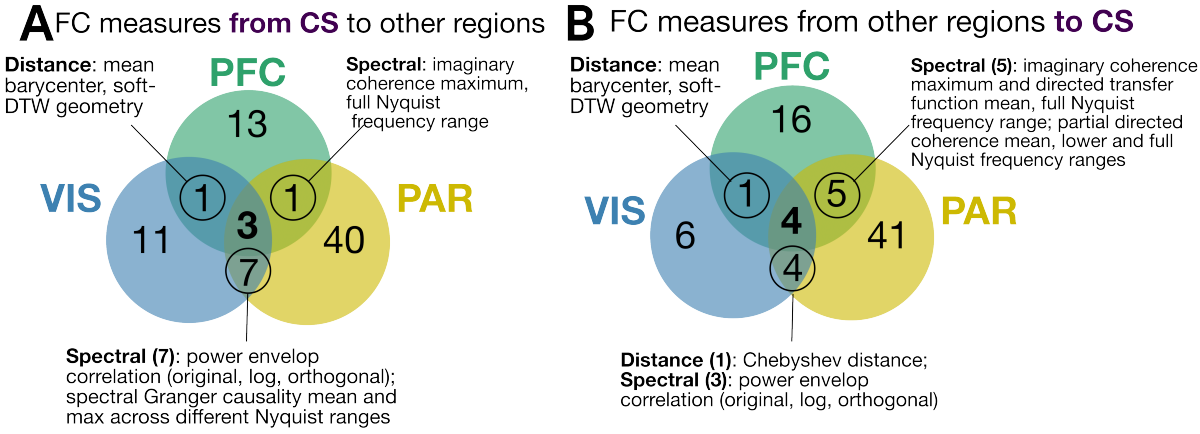

Figure S3. Detailed breakdown of the top-performing FC measures shared between two functional pathways. A The same Venn diagram from Fig. 4A is reproduced, with detailed annotation of the FC measures comprising the intersection between two functional pathways projecting **from** CS to VIS, PFC, or PAR. **A** The same Venn diagram from Fig. 4B is reproduced, with detailed annotation of the FC measures comprising the intersection between two functional pathways projecting **to** CS from VIS, PFC, or PAR.

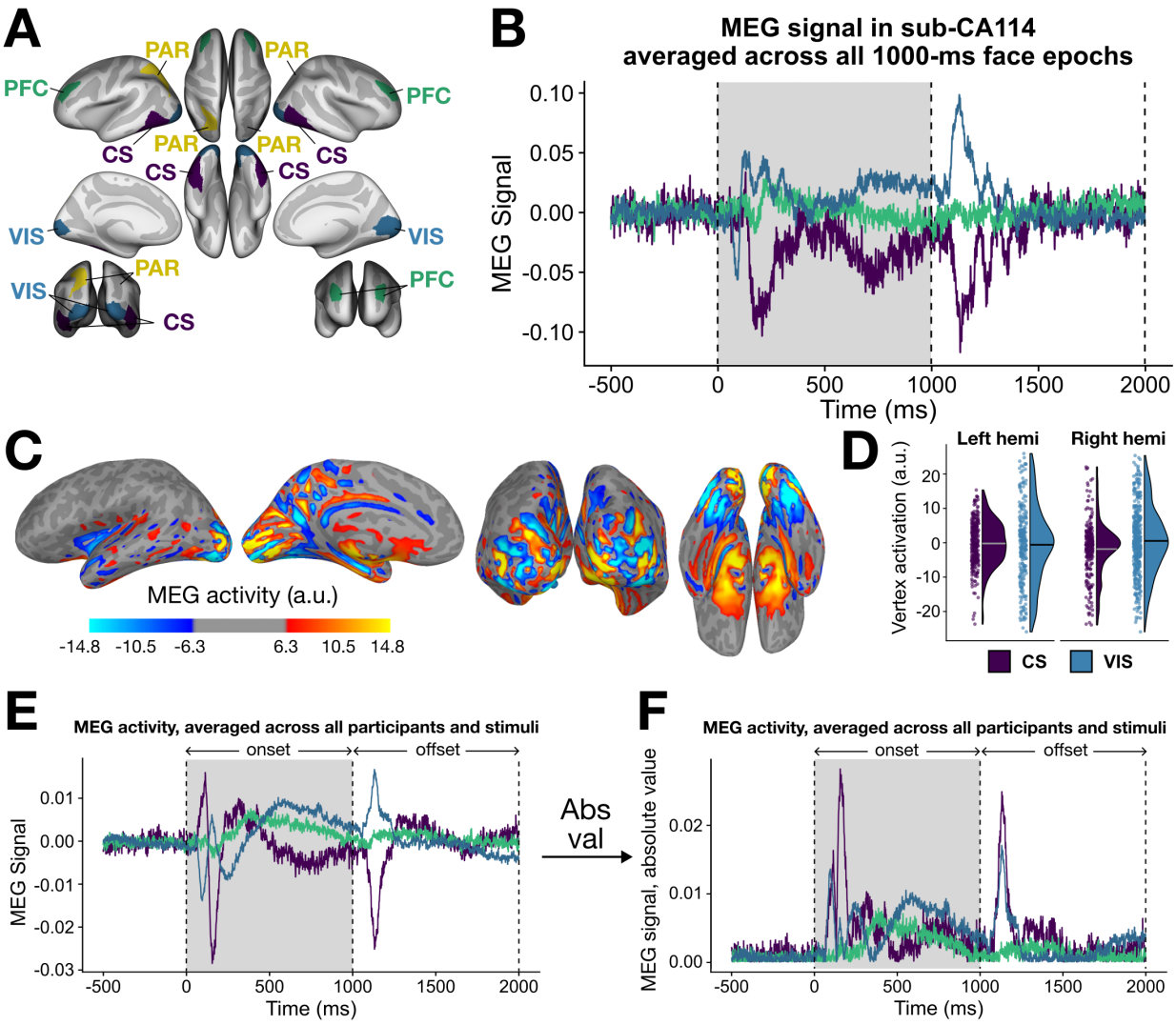

Figure S4. Analyzing regional polarity in MEG signal supports computing the absolute value of MEG amplitude. A. The four regions of interest examined in this study are shown on the cortical surface in `fsaverage` space. **B.** The average MEG signal is shown for one participant (sub-CA114), with activity averaged across all epochs per region. The regions are colored as in **A**. Stimulus onset is denoted with a shaded gray rectangle (occurring between 0 to 1000 ms), followed by the 1000-ms offset period. **C.** The raw MEG signal is plotted at the point of peak brain-wide intensity, corresponding to 153 ms post-onset. Larger activity magnitudes indicate greater electrical current, arising primarily from postsynaptic activity in pyramidal neurons in the cortex, with the sign indicating the orientation relative to the MEG sensors. **D.** Vertex-wise activation values are summarized in raincloud plots for the CS and VIS regions in the left and right hemispheres, respectively. Horizontal bars in each half-violin represent the mean activation value in artificial units (a.u.). **E.** The average MEG signal is shown, computed across all participants (N=94) and all epoch types, per brain region. The regions are colored as in **A**. Stimulus onset is denoted with a shaded gray rectangle (occurring between 0 to 1000ms), followed by the 1000-ms offset period. **F.** The absolute value is obtained from the average MEG time series shown in **E**.

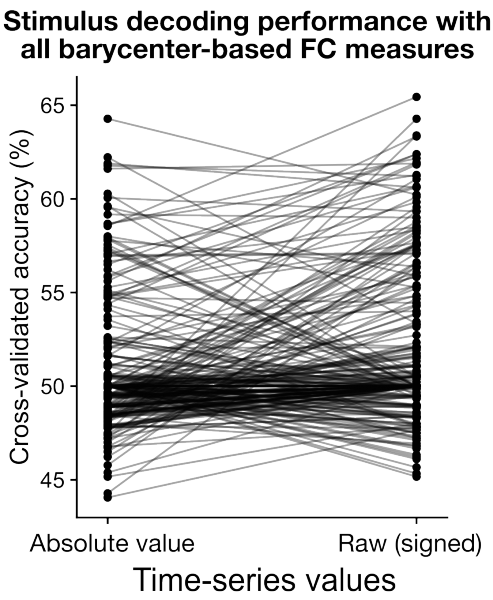

Figure S5. Stimulus decoding performance with barycenter-based FC measures does not systematically vary between absolute value versus raw (signed) empirical MEG time series.

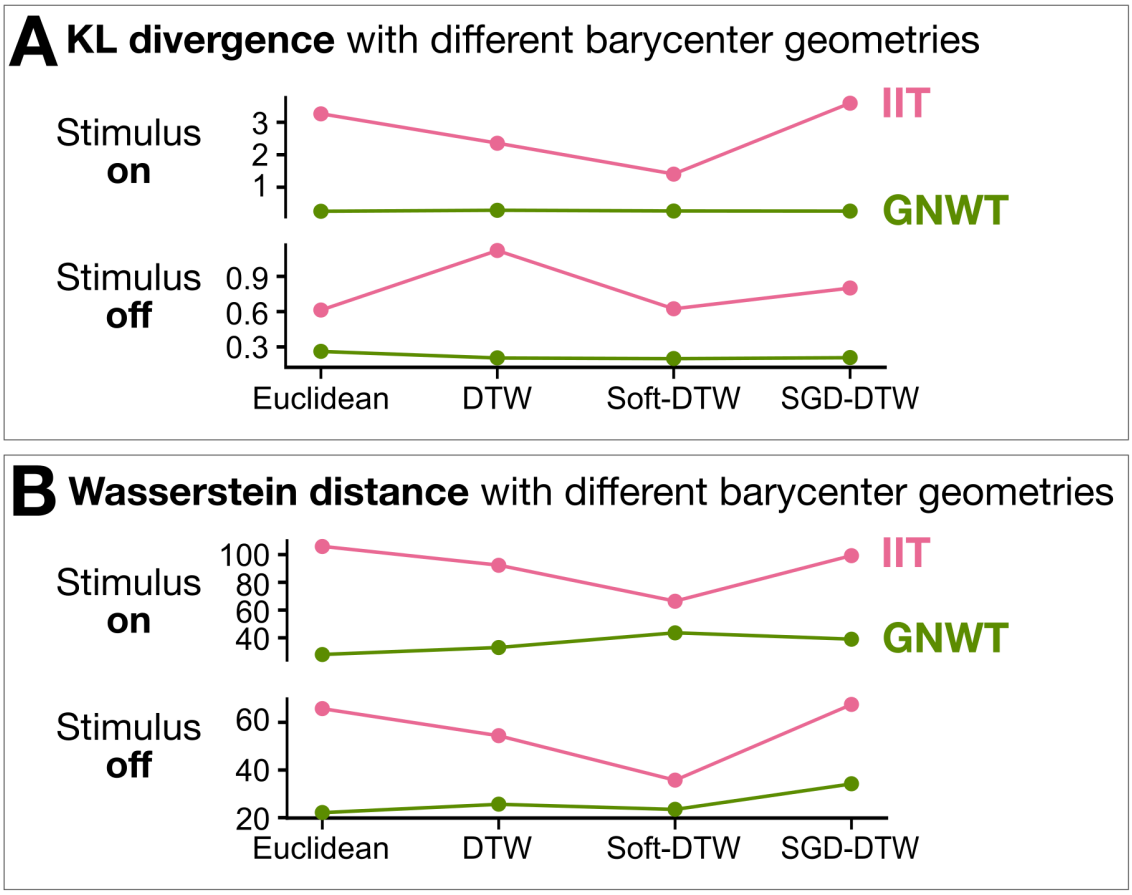

Figure S6. K-L divergence and Wasserstein distance between empirical vs. model-based maximum squared barycenter timepoints show consistent trends across different Euclidean and non-Euclidean barycenter geometries.

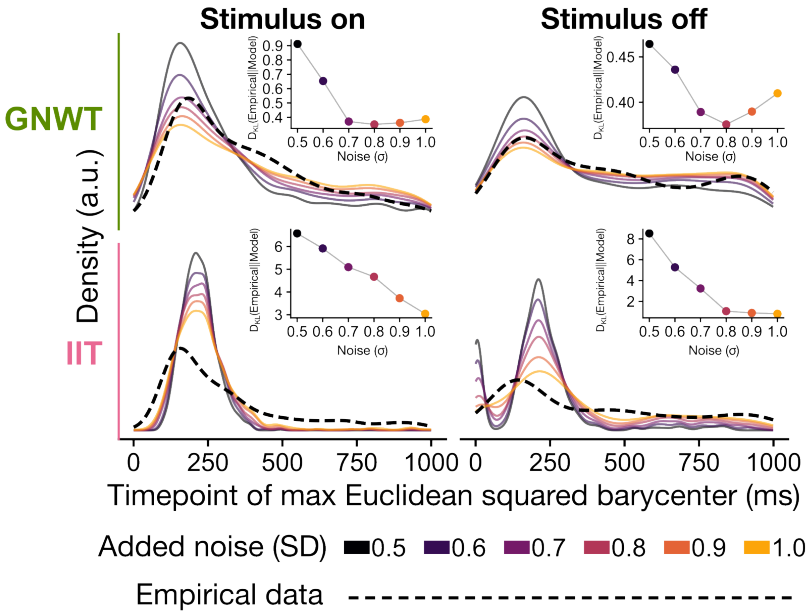

Figure S7. K-L divergence patterns are robust across added noise levels for simulated time series between GNWT and IIT. In each stimulus presentation period ('on' vs. 'off') and model context ('GNWT' vs. 'IIT'), the distribution of timepoints for the maximum squared Euclidean barycenter is plotted for the empirical MEG data (thicker dashed black line) as well as for model-simulated data with noise thresholds ranging from $0.5 \leq \sigma \leq 1.0$. The inset in each graph depicts the K-L divergence between the empirical versus model-simulated time series at the corresponding noise level.

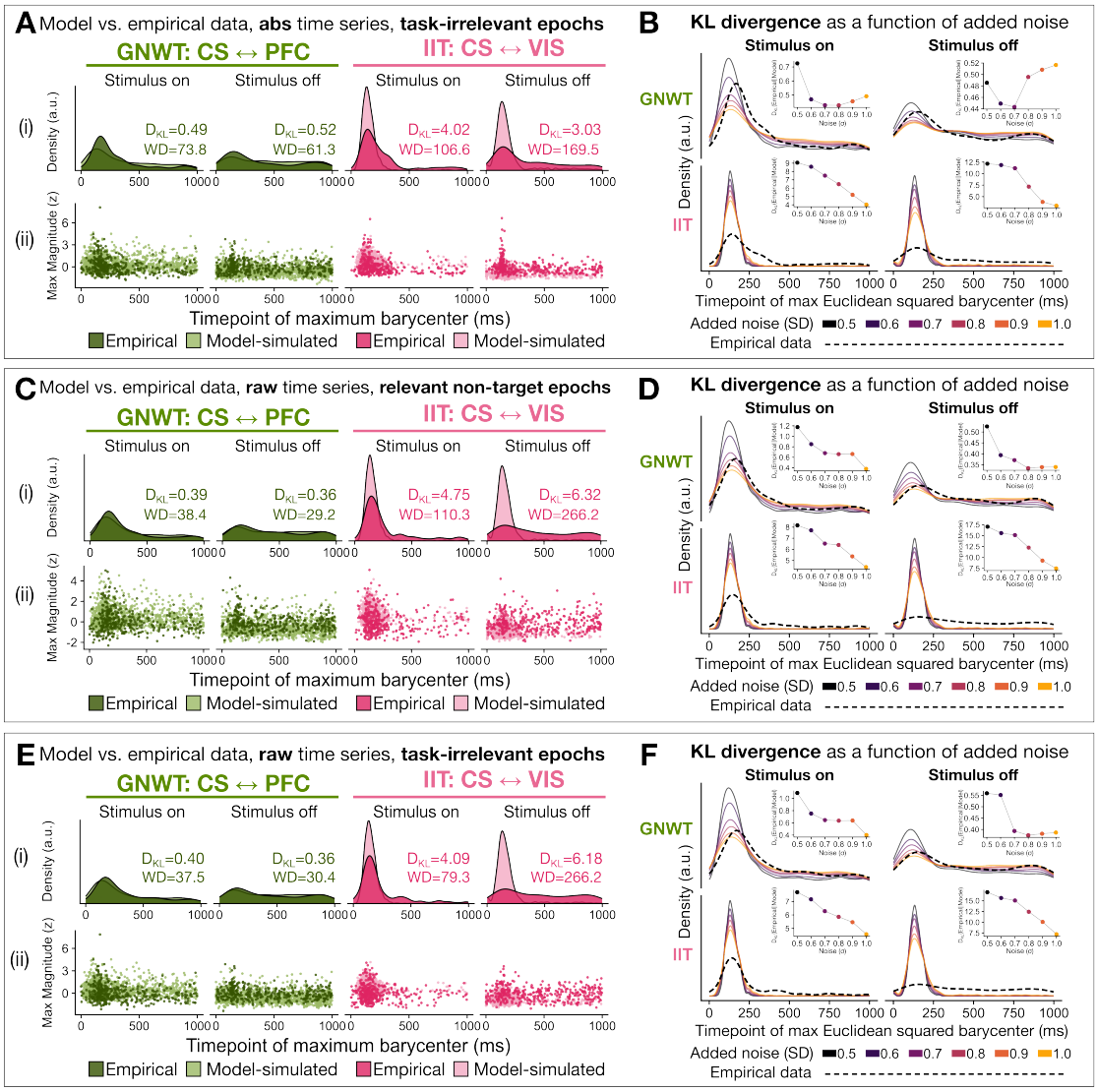

Figure S8. Timing of the maximum squared Euclidean center is better approximated by the GNWT-based than IIT-based model across stimulus relevance settings, with either raw or absolute value time series. A, C, E. In each setting, we compare the distribution of timepoints corresponding to the maximum squared Euclidean barycenter for the empirical data (darker shading) versus model-simulated time series (lighter shading) for GNWT in green, corresponding to CS ↔ PFC, and for IIT in pink, corresponding to CS ↔ VIS in pink. The kernel-estimated densities are shown in each panel in (i), with the K-L divergence (D_{KL}) and Wasserstein distance (WD) given for each empirical vs. model-simulated comparison. The timepoints and corresponding maximum barycenter magnitudes are shown in each (ii) panel. Here, we compare (**A**) absolute time series in task-irrelevant epochs; (**C**) raw time series in relevant non-target epochs; and (**E**) raw time series in task-irrelevent epochs. **B, D, F.** In each stimulus presentation period ('on' vs. 'off') and model context ('GNWT' vs. 'IIT'), the distribution of timepoints for the maximum squared Euclidean barycenter is plotted for the empirical MEG data (thicker dashed black line) as well as for model-simulated data with noise thresholds ranging from $0.5 \leq \sigma \leq 1.0$. The inset in each graph depicts the K-L divergence between the empirical versus model-simulated time series at the corresponding noise level.