Alpha phase-coding supports feature binding during working memory maintenance

Mattia F. Pagnotta[1,*], Aniol Santo-Angles[2], Ainsley Temudo[2], Joao Barbosa[3,4], Albert Compte[3], Mark D'Esposito[1,5], Kartik K. Sreenivasan[2,6]

[1] Helen Wills Neuroscience Institute, University of California Berkeley, Berkeley, United States of America

[2] Division of Science and Mathematics, New York University Abu Dhabi, Abu Dhabi, United Arab Emirates

[3] Institut d'Investigacions Biomèdiques August Pi i Sunyer (IDIBAPS), Barcelona, Spain

[4] Laboratoire de Neurosciences Cognitives et Computationnelles, INSERM U960, École Normale Supérieure - PSL Research University, 75005 Paris, France

[5] Department of Psychology, University of California Berkeley, Berkeley, United States of America

[6] Center for Brain and Health, New York University Abu Dhabi, Abu Dhabi, United Arab Emirates

[*] Correspondence and material requests should be addressed to M.F.P. (email: pagnotta@berkeley.edu).

Abstract

Working memory (WM) is the ability to retain and manipulate information in mind, which allows mnemonic representations to flexibly guide behavior. Successful WM requires that objects' individual features are bound into cohesive representations, however the mechanisms supporting feature binding remain unclear. Binding errors (or swaps) provide a window into the intrinsic limits in capacity of WM. We tested the hypothesis that binding in WM is accomplished via neural phase synchrony and swaps result from its perturbations. Using magnetoencephalography data collected from human subjects, in a task designed to induce swaps, we showed that swaps are characterized by reduced phase-locked oscillatory activity during memory retention. We found that this reduction arises from increased phase-coding variability in the alpha-band, over a distributed network of sensorimotor areas. Our findings support the notion that feature binding in WM is accomplished through phase-coding dynamics that emerge from the competition between different memories.

1. Introduction

Working memory (WM) is the ability to retain and manipulate information when it is no longer present in our environment, which introduces the possibility for abstract concepts and plans to influence our behavior (Baddeley & Hitch, 1974; D'Esposito & Postle, 2015; Goldman-Rakic, 1987; G. A. Miller et al., 1960). The well-described capacity limits of WM (Baddeley, 2003, 2012; Cowan, 2001) hence represent a key constraint on higher order cognition such as cognitive control (Badre, 2020). In real life, WM capacity is not just dependent on the maintenance of individual visual features (e.g., color, orientation, shape, and spatial location), but also of the conjunctions or bindings between them (Luck & Vogel, 1997; Ma et al., 2014; Schneegans & Bays, 2019; Vogel et al., 2001). Thus, understanding how information is bound in WM is of critical importance to understand its limits.

One proposed mechanism for feature binding in WM is temporal synchrony between populations of neurons that encode individual features (von Der Malsburg, 1994, 1995). This idea relies on the notion of temporal or phase-coding—that is, the two different neuronal populations encoding the information of distinct features of a conjunction fire at the same phase of an ongoing oscillation, enabling these features to be bound together in WM (Bush & Burgess, 2020; O'Keefe & Burgess, 2005; Panzeri et al., 2010). While the idea of phase-coding is broadly consistent with empirical evidence of synchronized oscillatory dynamics within or between different brain regions during WM (Fell & Axmacher, 2011; Jokisch & Jensen, 2007; Kornblith et al., 2016; Palva et al., 2011; Pesaran et al., 2002; Siegel et al., 2009), computational models that have attempted to implement phase-coding to explain feature binding suffer from an important limitation (Pina et al., 2018; Raffone & Wolters, 2001). Namely, since the information is encoded by the timing of individual spikes, these models require an encoder and decoder that rely on a precise spike coincidence detection to gate information in and out of WM; the biological plausibility of such a complex synchrony detection mechanism is challenged by the high variability of interspike intervals (Compte, et al., 2003; Schneegans & Bays, 2019; Shadlen & Movshon, 1999).

To overcome this limitation, a recent model proposed that, while the maintenance of conjunctions of features is accomplished through synchronization of oscillatory activity (i.e., phase-coding), both encoding and decoding are accomplished through rate coding (Barbosa et al., 2021)—that is, the encoding of information by the spike count (rate) within the temporal window, regardless of their timing within the window (Panzeri et al., 2010)—which does not require precise temporal coincidence. This hybrid model provides a plausible biologically-constrained neural architecture of the cortico-cortical circuits that implement feature binding in WM, in which phase-coding emerges from the lateral inhibition between the different neuronal populations that store information about distinct (competing) conjunctions (Buzsáki, 2006). In the present study, we sought to test a central prediction of this model: that disruptions in oscillatory phase should be associated with misbinding or 'swap' errors (Barbosa et al., 2021), where an inaccurate response to the target item is accurate relative to a non-target item (e.g., if a subject shown a red square and a blue circle mistakenly reports the color of the circle as red) (Ma et al., 2014;

Schneegans & Bays, 2019). If this prediction were validated, it would provide a key empirical demonstration that phase-coding serves as an organizational principle during WM maintenance.

We tested our hypothesis using magnetoencephalography (MEG) recordings collected from human subjects performing a task designed to induce swap errors. Our results showed a characteristic within-trial phase-locking in the alpha-band during WM maintenance. In parietal-occipital sensors contralateral to the visual stimuli, the consistency over trials of such alpha phase-locking was reduced in swap trials compared to high-performance on-target trials. Importantly, these effects did not generalize to other WM errors, suggesting that such deterioration in phase-locked oscillatory activity is a hallmark of swaps. To understand why phase relationships are less preserved in swap trials, we considered a measure of variability in the instantaneous frequency of alpha oscillations and showed that swaps are characterized by increased variability. We further localized these effects in contralateral areas in premotor, motor, parietal, and visual cortices. These results suggest that during WM maintenance feature binding is accomplished via alpha phase-coding, while swaps are produced by unstable phase-locked activity in distributed sensorimotor areas.

2. Results

2.1. Identification of swap errors

We collected MEG data from 26 human subjects who performed a delayed-response task designed to induce swaps. Subjects were briefly shown a lateralized display of 3 circles (stimulus presentation: 0.2 s) and held their colors and locations in WM (Fig. 1a). After a brief memory retention interval (delay period: 2 s), subjects reported the location of each of the circles, which were sequentially cued by their color (report period: self-paced) in a random order. We used continuous measures of response errors and employed a maximum likelihood approach to distinguish high-performance (HP) target trials (location of all circles reported accurately; Fig. 1a) from swaps (location of one of the non-cued circles mistakenly reported; Fig. 1b), and from low-performance (LP) trials (location of one or more circles reported inaccurately) based on the subjects' responses (Schneegans & Bays, 2016) (https://www.paulbays.com/toolbox/; see section 4.2).

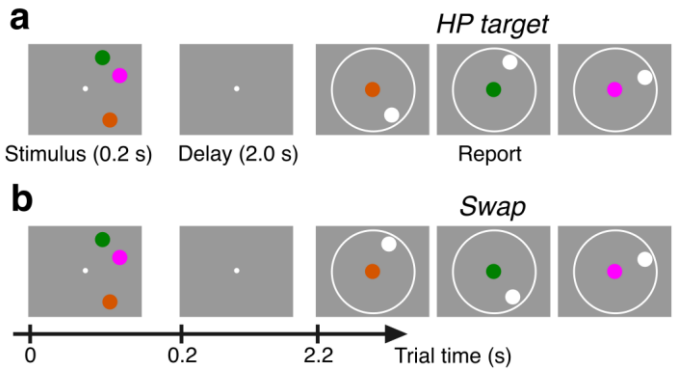


Fig. 1 Experimental design. **a–b** Schematic representation of the temporal structure of each trial, with examples of high-performance (HP) target response (a) and swap error (b).

2.2. Less stable phase-locked dynamics in swaps during WM maintenance

Our hypothesis entailed that HP trials are characterized by stable phase-locked activity during WM maintenance, while swap trials are induced by noisy fluctuations in the phase of oscillatory activity related to the maintenance of individual features (Barbosa et al., 2021). To test this prediction we used the Phase Preservation Index (PPI) (Mazaheri & Jensen, 2006), which captures for each time point the consistency over trials of the within-trial phase differences with respect to a reference time (in our case, the memory delay onset 0.2 s; Fig. 1)—that is, their level of phase clustering in polar space (M. X. Cohen, 2014a). Thus, PPI provides a measure of the consistency of frequency-specific local phase-locking over trials, as a function of time. We estimated PPI for each MEG sensor over time points in the delay period (0.2–2.2 s; Fig. 1) and frequencies 1–50 Hz separately for each

trial type, and then compared PPI between HP targets and swaps. Since PPI is sensitive to the number of trials, we performed the analysis by equating the number of HP target and swap trials (see section 4.5).

PPI was higher in HP targets than swaps in the alpha-band. A cluster was found in the observed data, extending over frequencies between 8–14 Hz and latencies around ~0.5 s after delay onset (Fig. 2a). PPI differences between trial types were most pronounced over 25 parietal-occipital sensors, contralateral to stimulus presentation (Fig. 2b), with medium effect size (Cohen's *d* in the range 0.170–0.576 across sensors in the observed cluster). The cluster-based permutation test revealed that PPI was significantly higher in HP targets compared to swaps (p_{perm}=0.0160). We note, however, that it is difficult to draw conclusions about the precise frequency and timing of effects due to the intrinsic limitations of the cluster-based permutation testing (Maris & Oostenveld, 2007; Sassenhagen & Draschkow, 2019), which only provides an approximation of the effect extent, as well as of PPI for the timing effects. PPI is estimated with respect to a reference time point, effectively inflating estimates that are close to said reference. This can be appreciated by examining the time course of PPI at different frequencies (e.g., in HP targets; Fig. 2c): estimates close to the reference are close to 1 in value and tend to decrease over time to an asymptotic value. Analogous behavior has been observed in model simulations (Barbosa et al., 2021). Our results nonetheless demonstrate that phase preservation is reduced for swaps in the alpha-band, showing a steeper decrease than HP targets. Furthermore, we calculated the z-scores of PPI over frequencies for each time point in the delay, separately for each trial type. We found positive z-scores at low frequencies in the delta/theta range (up to ~6 Hz) and negative z-scores at higher frequencies in the beta/gamma range (above ~15 Hz), while the middle frequencies in the alpha-band (~10 Hz) were characterized by positive z-scores that were higher than those at adjacent frequencies, both low and high (Fig. 2d; see also Supplementary Fig. 2a–b). These results reveal that while there is a general tendency for PPI to decrease faster at higher frequencies than lower frequencies, alpha deviates from this trend.

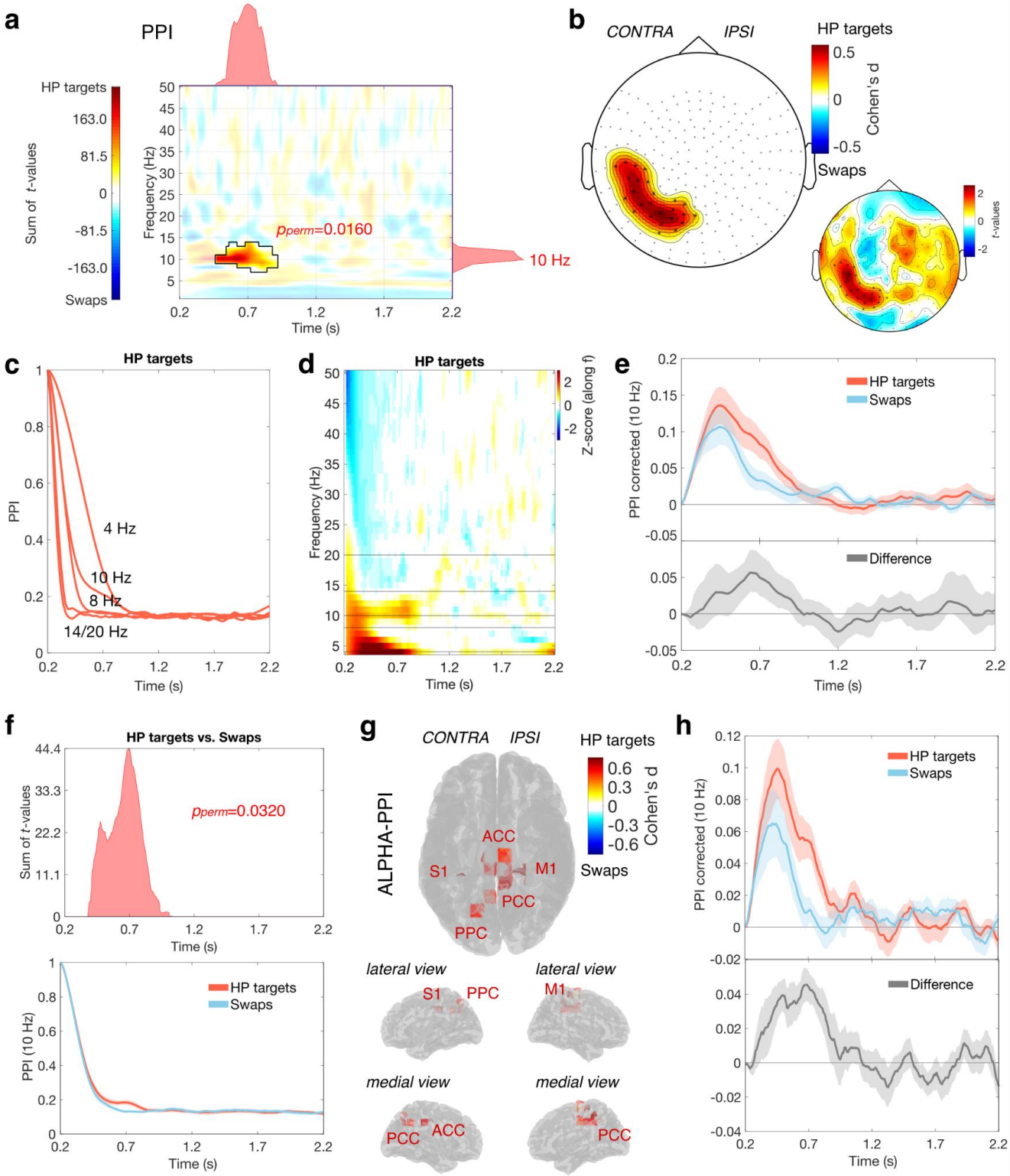


Fig. 2 PPI differences between HP targets and swaps. **a** Time-frequency distribution of the sum of PPI differences across MEG sensors (*t*-values). The black contour highlights the statistically significant cluster. The marginal plots on the right and on top represent respectively the time-collapsed frequency distribution and frequency-collapsed time distribution of the differences between trial types. **b** Topography plot with superimposed effect sizes of PPI

differences between HP targets and swaps, for each MEG sensor of the observed cluster (*t*-values are shown on the smaller topography plot on the right). **c** Time course of PPI in HP targets at different frequencies, highlighted by the black horizontal lines in the next panel. **d** Z-scores along frequencies for each time point in the delay, in HP targets. **e** Time course of PPI corrected at 10 Hz for HP targets (red) and swaps (blue), together with the difference between the two (gray). The red and blue shadings (top) represent the standard error of the mean, while the gray shading (difference; bottom) represents 95% confidence intervals (CIs). CIs were estimated using the bias-corrected and accelerated method on a bootstrap distribution of across subjects differences, obtained by resampling with replacement 10,000 times. **f** Time distribution of the sum of 10 Hz PPI differences between HP targets and swaps, across source points in source-space (*t*-values; top), and time course of the average PPI across source points of the positive cluster for each condition (bottom). The shadings represent the standard error of the mean. **g** Glass brain in MNI coordinates with superimposed effect sizes of the observed differences between HP targets and swaps. **h** Time course of PPI corrected at 10 Hz for HP targets (red) and swaps (blue) across source points of the cluster in the observed data, together with the difference between the two (gray; shadings are as in panel e).

To confirm that the observed PPI effects were due to genuine within-trial phase synchronization in the alpha-band, we performed the following control analysis. We repeated the PPI estimation (including the procedure to equate the number of trials between HP targets and swaps – see section 4.5), but this time after randomly shuffling the signals from each MEG sensor over time, which disrupted any phase relationships present in the data. We then computed 'PPI corrected' estimates as the difference between PPI estimates from the main analysis and those obtained from the control analysis. The results obtained from shuffled data showed that the speed of PPI reduction over time increases monotonically with frequency, including for the alpha-band, and they confirmed the presence of a plateau/asymptote for PPI at all frequencies (~0.13 on average across subjects; Supplementary Fig. 2c–d). The PPI corrected results showed, instead, that phase synchronization takes place specifically at alpha frequencies in our WM task (Supplementary Fig. 2e–f), and it is characterized by more stable dynamics in HP targets than swaps during the delay period (Fig. 2e). In a separate control analysis, we compared PPI estimates between HP targets and LP trials (see section 4.2) and did not find any statistically significant differences, suggesting that PPI differences in alpha are specific to swaps and are not a property of impaired behavioral performance in general.

To identify the underlying cortical sources exhibiting differences in local phase-locked activity between HP targets and swaps, we used a source reconstruction technique to localize MEG sources of activity, and repeated the PPI analysis in source-space (see section 4.7). PPI estimation was here restricted to the frequency of interest of 10 Hz, which was selected in a data-driven way as the peak of the time-collapsed frequency distribution of observed PPI differences in sensor-space (see Fig. 2a). In source-space, a positive cluster (higher PPI in HP targets than swaps) was observed in the data (Fig. 2f), which included 17 source points, localized in the cingulate cortex

and cortical areas contralateral to stimulus presentation (Fig. 2g). Effect size peaked in source points localized in the ventral anterior cingulate cortex–ACC and dorsal posterior cingulate cortex–PCC (MNI coordinates x=-17.5, y=-20, z=42.5 and x=7.5, y=-44.5, z=55; d=0.694 and d=0.702, respectively), ipsilateral primary motor cortex–M1 (MNI coordinates x=20, y=-32.5, z=67; d=0.732), and contralateral areas in primary somatosensory cortex–S1 and precuneus/posterior parietal cortex–PPC (MNI coordinates x=-30, y=-32.5, z=42.5 and x=-17.5, y=-69.5, z=42.5; d=0.766 and d=0.616, respectively). The cluster-based permutation test identified that there was a significant difference in 10 Hz PPI between HP targets and swaps (p_{perm}=0.0320). Like in the sensor-space analysis, the PPI corrected results showed that alpha phase-locked activity is characterized by more stable dynamics in HP targets than swaps during the delay period (Fig. 2h).

2.3. Increased phase-coding variability in swaps

The computational model by Barbosa and colleagues suggests that phase-locked dynamics allow maintaining color-location conjunctions in WM, but these phase-locked dynamics can be broken by abrupt noisy fluctuations, which lead to misbinding of memorized features and swaps (Barbosa et al., 2021). We confirmed the neurophysiological prediction of this model, in that trials containing swaps have a lower PPI compared to HP targets. However, PPI does not directly address the nature of the possible instabilities in the phase of alpha oscillations (Fig. 3a). For this, we calculated the frequency sliding (FS), which captures the instantaneous temporal fluctuations in oscillation peak frequency (Fig. 3b)—that is, time-varying changes in the instantaneous frequency of the oscillator (M. X. Cohen, 2014b). Moreover, unlike PPI, FS is a reference-independent measure; thus, it allows to overcome some limitations of PPI (see section 2.2) and provide a better understanding of the timing of observed effects. This is a useful feature also because previous studies have recognized that not only maintenance but also encoding and recall in WM are noisy processes (Schneegans & Bays, 2019); consequently, swaps may result from a failure in correctly binding visual features either during maintenance (Mallett et al., 2022; Pertzov et al., 2012) or during the initial encoding in WM (Golomb, 2015; Golomb et al., 2014). We used a data-driven approach to select a frequency of interest at the peak of the time-collapsed frequency distribution of PPI differences (10 Hz; see Fig. 2a), and derived FS at single-trial level at 10 Hz.

Continuous noisy fluctuations in alpha-FS during the delay period as well as noisy fluctuations that happen suddenly (Fig. 3c) would both result in a degradation of alpha phase preservation. To disambiguate these two scenarios, we estimated a measure of 'alpha-FS variability (FSV) over time' in a sliding window, during the time between the stimulus presentation and the delay period (0–2.2 s), by taking the median between trials of the standard deviation of FS over time points, separately for HP targets and swaps (see section 4.6). We also estimated the median standard deviation of FS over trials on the sliding window to derive a measure of 'alpha-FSV over trials', separately for HP targets and swaps (see section 4.6). Differences between HP targets and swaps in both measures of alpha-FSV would indicate that swaps are induced by noisy fluctuations that are more sustained over

time (scenario 1 in Fig. 3c), while if the source of variability is predominantly over trials, this would suggest that noisy fluctuations in swaps happen quickly and more suddenly (scenario 2 in Fig. 3c).

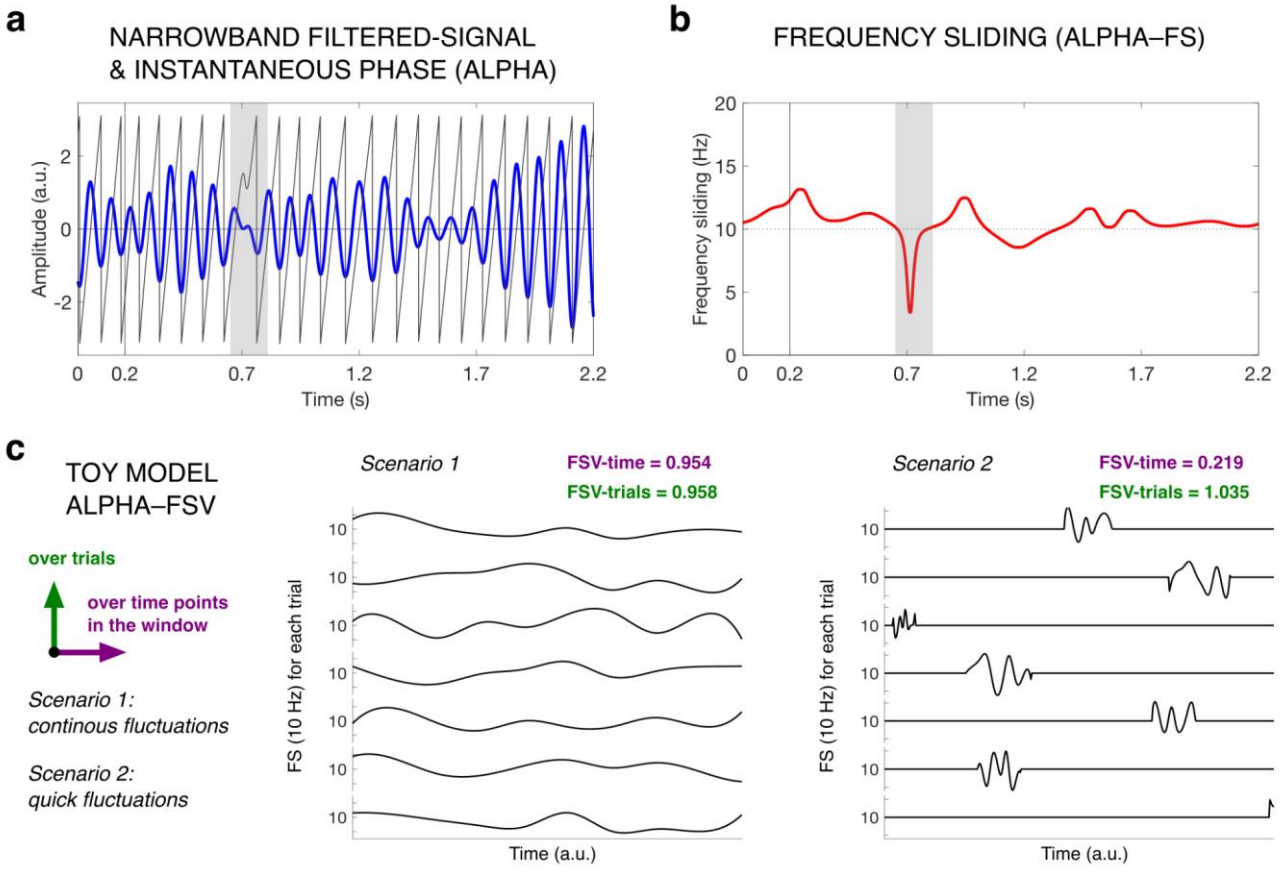


Fig. 3 Instantaneous phase fluctuations, frequency sliding (FS), and frequency sliding variability (FSV). **a** Example from a swap trial of the narrowband filtered-signal in the alpha-band (in blue) and its instantaneous phase (in black) during the delay period. The signal was obtained from the first subject, from one of the contralateral parietal sensors where we observed significant PPI difference. **b** Single-trial FS estimated from the signal represented in panel a. In panels a–b, the gray shadings highlight a noisy fluctuation in the alpha phase and how this can be captured by FS. **c** Toy model illustrating the estimation of alpha-FSV over time points or over trials, and alternative scenarios for how a reduction in alpha phase preservation in swaps is produced by noisy fluctuations in instantaneous alpha frequency with different temporal profiles.

We did not find any statistically significant differences between HP targets and swaps in the alpha-FSV over time (the lowest *p*-value from the permutation test among the identified clusters was p_{perm}=0.368). Alpha-FSV over trials, instead, was significantly higher in swaps compared to HP targets (p_{perm}=0.0180). In the observed data, a cluster was observed extending during the second half of the delay period (approximately between 1.182–2.045 s; Fig. 4a). The differences in alpha-FSV over trials were most prominent in 20 contralateral, parietal-

occipital sensors (effect size ranging between *d*=0.088 and *d*=0.321 across cluster sensors; Fig. 4b). This and the previous PPI analysis provided convergent findings, as the groups of electrodes showing the biggest effects largely overlapped between the two analyses (see also Fig. 2b).

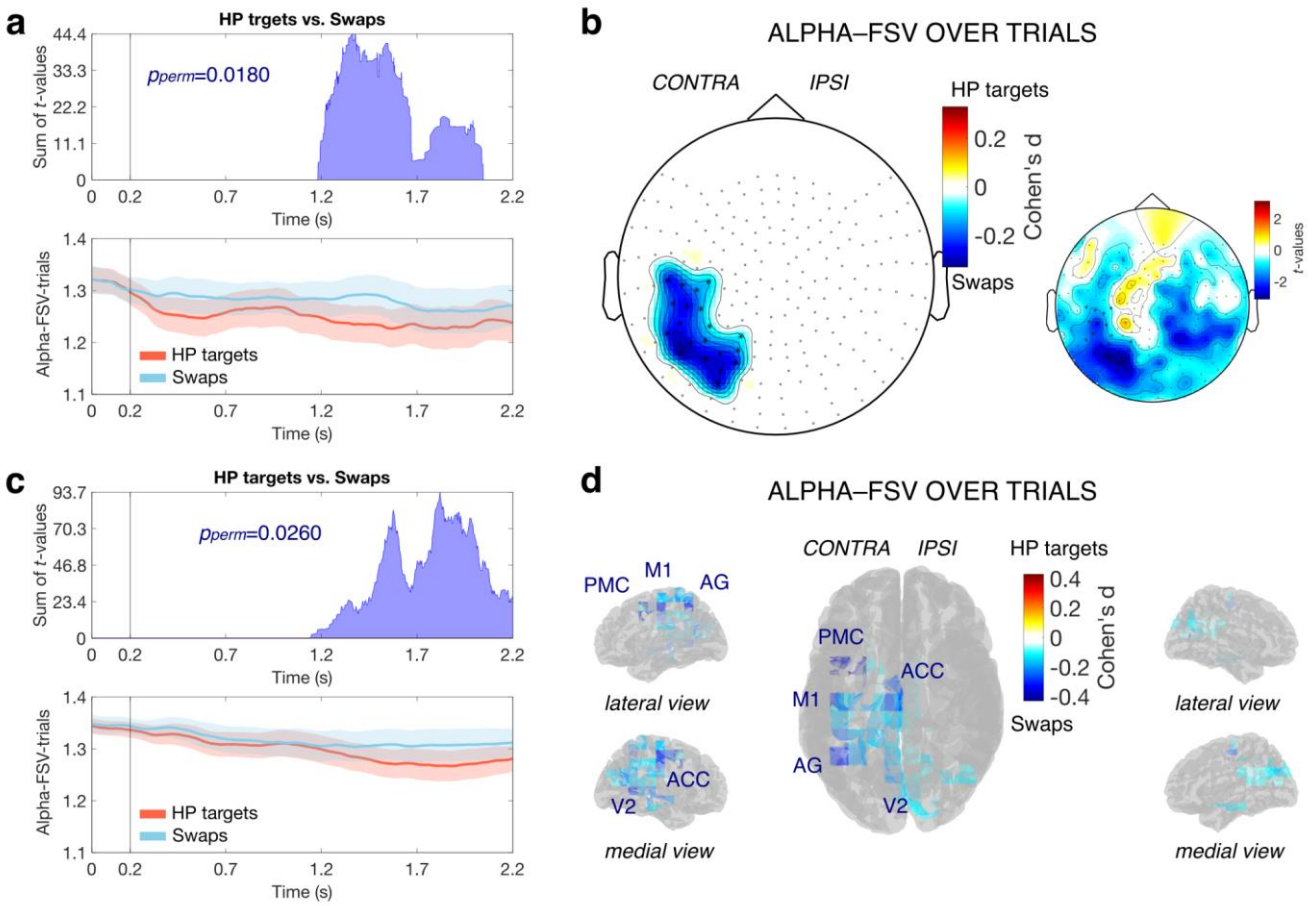


Fig. 4 Differences in alpha-FSV over trials between HP targets and swaps. **a** Time distribution of the sum of alpha-FSV differences across MEG sensors (*t*-values; top), and time course of the average alpha-FSV across sensors of the cluster found in the data for each condition (bottom). The shadings represent the standard error of the mean. **b** Topography plot with superimposed effect sizes of the differences between HP targets and swaps, for each MEG sensor of the observed cluster (*t*-values are shown on the smaller topography plot on the right). **c** Time distribution of the sum of alpha-FSV differences across source points in source-space (*t*-values; top), and time course of the average alpha-FSV across source points of the observed cluster for each condition (bottom). The shadings represent the standard error of the mean. **d** Glass brain in MNI coordinates with superimposed effect sizes of the differences between HP targets and swaps.

As for the PPI, after using a source reconstruction technique, we repeated the FSV analysis in source-space (see section 4.7). The source-space results confirmed that alpha-FSV over trials was higher in swaps than HP targets. A negative cluster was found extending approximately over latencies 1.148–2.200 s (Fig. 4c) and 92

source points, localized for the most part in contralateral cortical areas (Fig. 4d). Effect size peaked in source points localized in premotor cortex–PMC (MNI coordinates x=-30, y=5, z=55 and x=-42.5, y=5, z=42; d=0.430 and d=0.386, respectively), motor area M1 (MNI coordinates x=-42.5, y=-20, z=55; d=0.361), angular gyrus–AG (MNI coordinates x=-42.5, y=-57, z=55; d=0.305), visual area V2 (MNI coordinates x=-5, y=-57, z=5; d=0.291), and anterior cingulate–ACC (MNI coordinates x=-5, y=-20, z=42; d=0.290). The cluster-based permutation test indicated that there was a significant difference in alpha-FSV over trials, between swaps and HP targets (p_{perm}=0.0260). In line with the sensor-space results, we did not find any statistically significant differences in alpha-FSV over time points (lowest p_{perm}=0.833 among clusters). Together these results suggest that, in areas contralateral to stimulus presentation, correct feature binding in WM is supported by stable phase-coding in the alpha-band, while swap errors are the result of its perturbations. In swaps, noisy fluctuations in the alpha oscillations peak frequency happen quickly and abruptly rather than in a more sustained fashion over time.

2.4. Distinct possible sources of swaps

In general, the term ‘swap’ does not necessarily imply a symmetric exchange of features between two memorized items (e.g., spatial locations) (Bays, 2016). In our experiment, we can hence separate swap trials into different types, based on the number of non-target reports, and define ‘symmetric swaps’ (at least two non-target reports) and ‘non-symmetric swaps’ (single non-target report). These two swap types might have distinct sources. In a follow-up analysis, we investigated whether differences in phase-coding variability may help to distinguish such sources.

For the alpha-FSV over trials, a negative cluster was found in the observed data (i.e., alpha-FSV higher in non-symmetric than symmetric swaps), extending over latencies between stimulus presentation and the beginning of the delay period (~0–0.740 s; Fig. 5a). This cluster included 63 sensors, mostly in frontal and sensorimotor sites, over both hemispheres (Cohen's d in the range 0.071–0.617 across sensors; Fig. 5b). The cluster-based permutation test revealed a significant difference between swap types in alpha-FSV over trials (p_{perm}=0.0060). We did not observe significant differences at longer latencies for alpha-FSV higher in symmetric than non-symmetric swaps. Furthermore, we did not find any statistically significant differences between swaps types in the alpha-FSV over time (lowest p_{perm}=0.443 among clusters). In the source-space analysis, there was a trend towards increased alpha-FSV over trials for non-symmetric compared to symmetric swaps, in areas of the prefrontal cortex, frontal eye fields, temporal pole, and AG, in particular in the contralateral hemisphere, which is in line with the sensor-space results; however, these effects did not reach statistical significance with a rigorous statistical threshold (positive cluster extending approximately over latencies 0–1.844 s; p_{perm}=0.0919; see Supplementary Fig. 5).

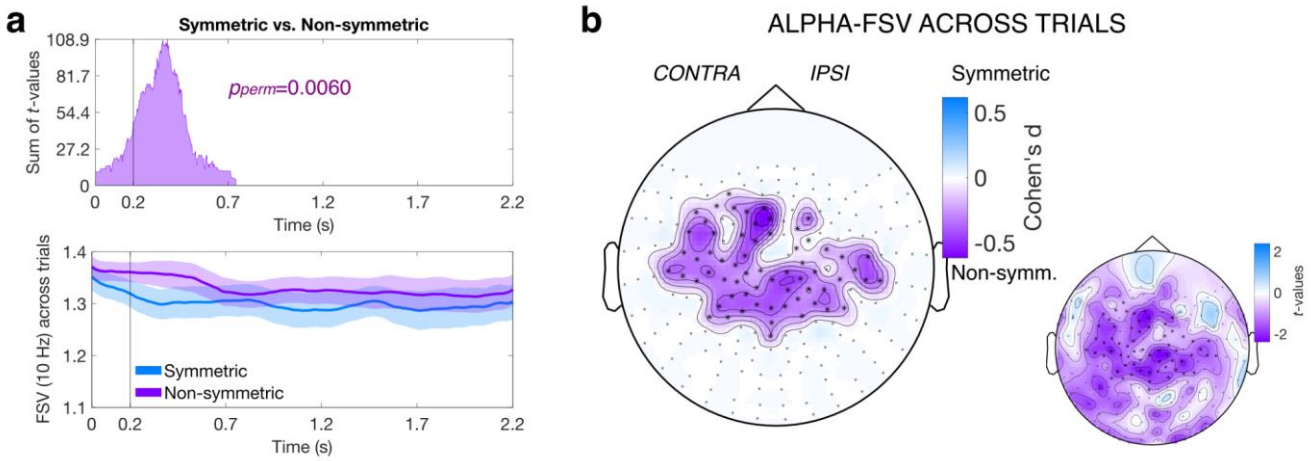


Fig. 5 Differences in alpha-FSV over trials between symmetric and non-symmetric swaps. **a** Time distribution of the sum of alpha-FSV differences across MEG sensors (*t*-values; top), and time course of the average alpha-FSV across sensors of the observed cluster for each condition (bottom). The shadings represent the standard error of the mean. **b** Topography plot with superimposed effect sizes of alpha-FSV differences between symmetric and non-symmetric swaps, for each MEG sensor of the observed cluster (*t*-values are shown on the smaller topography plot on the right).

3. Discussion

Previous models of feature binding have proposed alternative neural processes for how conjunctions of features are memorized (Ma et al., 2014; Schneegans & Bays, 2019). An influential class of theories and models proposed that feature binding in WM is accomplished via the phase synchronization of signals of neurons (phase-coding) that store the different feature values corresponding to the memorized item (Barbosa et al., 2021; Fell & Axmacher, 2011; Pina et al., 2018; Raffone & Wolters, 2001; von Der Malsburg, 1994, 1995). Here, we tested experimentally some of the neurophysiological predictions that can be derived from these models. We found that the correct feature binding is supported by stable phase-locked oscillatory activity at alpha frequencies, while swaps (binding errors in memory) are characterized by reduced phase preservation during WM maintenance. This reduction did not generalize to other types of errors, suggesting that alpha phase inconsistencies are a hallmark of swaps. Further, we found that these phase inconsistencies arose from increased phase-coding variability, in parietal and occipital MEG sensors that were contralateral to stimulus presentation. These were mostly the same sensors where the phase preservation was reduced in swaps. Thus, the two analyses independently provided convergent evidence that, in swaps, the sources of reduced neural phase synchrony and higher phase-coding variability are in contralateral areas. We localized the increase in phase-coding variability for swaps in premotor, motor, parietal, and visual cortical areas, in the hemisphere contralateral to stimulus presentation. Together our results suggest that feature binding in WM is accomplished through alpha phase-coding dynamics that emerge from the competition between different memories, which supports the theory of time-based binding for cognitive flexibility (Senoussi et al., 2022).

The competition between different memories may result from lateral inhibition between neural networks in sensory and motor areas. A recent computational model has proposed that the maintenance of conjunctions is accomplished through phase-coding between two one-dimensional attractor networks (Barbosa et al., 2021). In the model, oscillatory dynamics emerge from the competition between activity bumps corresponding to the different memory items, accomplished by gamma-aminobutyric acid (GABA)-mediated lateral inhibition between them. In a similar way, in brain sensorimotor areas the competition between mnemonic representations may be mediated by feedback inhibition via the activation of specific GABAergic inhibitory interneurons (Z. J. Huang et al., 2007; The Petilla Interneuron Nomenclature Group (PING), 2008). Brain activity is known to be dominated by an alternation in the firing of different neuronal populations (Stringer et al., 2019), and the ‘gating by inhibition hypothesis’ has proposed that alpha activity in sensory regions implements a mechanism of pulsed inhibition that silences neuronal firing (Jensen et al., 2014; Jensen & Mazaheri, 2010), mediating the observed alternation in neuronal firing. Furthermore, a recent study has shown that inhibitory interneurons types can be arranged along an axis given by their transcriptomic expression, which correlates with whether the neurons are more active during alert or oscillatory states, revealing that neurons more active during oscillatory states tend to have more inhibitory acetylcholine receptors (Bugeon et al., 2022). This specific type of inhibitory interneurons may mediate lateral

inhibition between pairs (in our case three) of sensorimotor neuronal populations that store the defining features of each item (color-location), producing an alternation in the neuronal firing among them, which gives rise at large-scale to the observed oscillatory dynamics in the alpha-band. Our results suggest that such alternating, phase-coding mechanism supports feature binding during WM maintenance: stable target dynamics result in high-performance reports of cued items, while noisy fluctuations produce abrupt alpha phase instabilities, inducing shifts to non-target dynamics that ultimately result in swaps.

Similar dynamics, mediated by oscillatory activity at different frequencies, have been observed using invasive recordings in the monkeys' prefrontal cortex (PFC) (Lundqvist et al., 2023). Nonhuman primate WM studies have found an inverse relationship between gamma (~40–120 Hz) power on one side, and beta (~20–35 Hz) power on the other, mediated by oscillatory bursts (Brincat et al., 2021; Lundqvist et al., 2016, 2018). During WM maintenance anti-correlated beta and gamma dynamics have been found to underlie a reduction in spiking variability (Lundqvist et al., 2022), which supports the notion that mnemonic representations are maintained through mechanisms of phase-coding (Fell & Axmacher, 2011; Siegel et al., 2009). While several other nonhuman primate studies have shown that areas in PFC store mnemonic representations during maintenance, human studies have provided evidence that PFC controls WM content stored in sensorimotor areas (D'Esposito & Postle, 2015; Serences, 2016; Sreenivasan et al., 2014). By using longitudinal WM training in human subjects, a recent fMRI study has shown that item-selective mnemonic representations become detectable in PFC over long-term learning, providing a plausible account for the discrepancies observed between human and nonhuman primate studies (J. A. Miller et al., 2022). Hence, the observed phase-coding mechanisms, supporting the storage of mnemonic representations during WM maintenance, may gradually spread from posterior alpha to frontal beta over the course of training or, more in general, when task-specific categories, associations, and rules are learned. It is also worth highlighting that a recent optogenetic study in monkeys has provided evidence that, in the lateral PFC, the neurons encoding WM representations may not be needed for mere WM maintenance, showing that task performance was strongly impaired by optogenetic inactivation of LPFC only during the test period of the task (after memory delay), but it was unaffected by delay period optogenetic LPFC inactivation (Mendoza-Halliday et al., 2024). These findings suggest that PFC may store abstract rules and contexts in WM, while sensorimotor areas may store in WM visual stimulus information.

In line with this idea, 'sensorimotor recruitment' models of WM have suggested that visual information is maintained in the same stimulus-selective regions that are responsible for perceiving that information (D'Esposito & Postle, 2015; Harrison & Tong, 2009; Pasternak & Greenlee, 2005; Scimeca et al., 2018; Serences et al., 2009). Our results provide supporting evidence for this sensorimotor recruitment account. We found increased variability in alpha phase-coding for swaps over a distributed group of regions in sensory and motor cortices. This is also in line with the 'distributed system view' of WM (Christophel et al., 2017; Lorenc & Sreenivasan, 2021), which considers that WM storage is distributed across multiple brain regions, among which both visual and parietal cortices play an important role in storing visual information during WM maintenance. The model proposed by

Barbosa and colleagues relies on two ring-attractor networks, one for each feature space (color and spatial location) (Barbosa et al., 2021), thus it explicitly simulates the independent storage of individual features constituting the item, in line with the increasing evidence that different features are stored in independent brain systems (different cortical areas) (Schneegans & Bays, 2019; Wang et al., 2017). In our task, independent feature storage for color and spatial location may be accomplished by different brain systems. The colors of the items may be stored in color sensitive areas in visual cortex, while their spatial locations may be integrated into salience maps by premotor/motor and parietal areas, by storing them in systems that represent them not only as perceptual information but also as targets for future motor responses during the report period. Sudden phase instabilities in either system produce abrupt changes in the phase relationship between the two systems, changing the correct conjunctions between features and producing binding errors.

The presence of binding errors has been examined also in cued-recall tasks where two non-spatial features were used as the cue and report features (Gorgoraptis et al., 2011; Pertzov et al., 2017; Pertzov & Husain, 2014), however spatial location has been recognized as having a privileged role in feature binding and in the generation of swaps compared to non-spatial features (Pertzov & Husain, 2014; Schneegans & Bays, 2017). Previous studies have shown that the likelihood of swaps depends on the feature employed as the memory cue, and the frequency of swaps is higher when spatial location is the report feature compared to when it is the cue feature (McMaster et al., 2022; Rajsic et al., 2017; Rajsic & Wilson, 2014). In the present study, we employed color as the cue feature and spatial location as the report feature. While the number of swaps would be lower when using location as cue feature and color as report feature, we expect that our results would be the same in terms of modulations of alpha phase dynamics. Future work should explicitly evaluate this scenario to confirm this hypothesis.

The origin of swaps in WM is still a topic of current debate. Some studies have proposed that swaps derive from an informed guessing strategy, rather than being true binding errors in memory (L. Huang, 2020; Pratte, 2019). In particular, by employing trials in which the cued item was not in memory at all, Pratte has suggested that the majority of swaps are simply a guessing strategy to forgotten items (Pratte, 2019). However, a more recent study has provided evidence that is inconsistent with the guessing strategy account, suggesting that swaps are better accounted for by cue-dimension variability (McMaster et al., 2022). Furthermore, to address the lack of neuroimaging evidence for swaps, another study has combined behavioral response modeling with functional magnetic resonance imaging (fMRI) modeling (Mallett et al., 2022), and has shown that it was possible to reconstruct a mnemonic representation of the swapped location but not of the cued location, suggesting that swaps involve the active maintenance of a non-target memory item, before the time of response. Adding on this neuroimaging evidence, we leveraged the MEG high temporal resolution and showed increased variability in alpha phase-coding for swaps compared to HP targets during the delay period (WM maintenance). This supports the notion that swaps are not mere response guesses but true binding errors in memory, occurring even before response time.

In a follow-up analysis, we found higher alpha phase-coding variability in non-symmetric swaps (single non-target report) than symmetric swaps (at least two non-target reports). These effects were observed predominantly at latencies between the stimulus presentation and the beginning of the delay period, over medial, frontal, and sensorimotor sensors. These results suggest that non-symmetric swaps may also capture situations in which a wrong association between visual features is encoded during the initial perception and encoding of the items' features into WM, even before memory retention (Golomb, 2015; Golomb et al., 2014), and/or in the early moments of WM maintenance. Based on the topology of the differences, these 'encoding/early maintenance' swaps may derive from a different network of brain areas. A caveat of this analysis is, however, that the number of trials was limited for the two swap types. Future studies are required to demonstrate whether distinct sources of swaps can be characterized on the basis of modulations in phase-coding variability.

In this study, we only investigated the phase dynamics throughout the delay period, during WM maintenance. In the previous hybrid model (Barbosa et al., 2021), a phase-coding mechanism accomplishes the maintenance of color-location conjunctions in WM, while both WM encoding and decoding are accomplished through a rate code (flat pulses), which does not require precise temporal coincidence. These functions may be controlled by the PFC. What makes PFC the best candidate for WM control are its anatomical connections and position. PFC is a great integrator of information in the brain (Menon & D'Esposito, 2022; E. K. Miller & Cohen, 2001), which receives feedforward inputs from almost everywhere, and it is also connected to the basal ganglia through corticostriatal loops (Haber, 2003; Sherman & Guillery, 1996). These connections from PFC to striatum, pallidum, thalamus, and then back to the same portion of cortex, may allow for recursive processing, feeding back top-down signals to sensorimotor areas, on the basis of the currently relevant high-level representations, task rules, abstract principles, and contexts, to guide behavior and action selection (Badre & Nee, 2018). Some studies suggest that a transient delta-to-theta (~2–8 Hz) power increase over frontal regions may underlie such top-down control over WM (de Vries et al., 2018; Wallis et al., 2015). These low-frequency oscillatory dynamics should be further investigated in the future, in the attempt to characterize how they implement inter-areal communication and information transfer to control the storage of the relevant memoranda representations in feature-sensory neuronal populations.

4. Methods

4.1. Subjects

This study was performed in compliance with the Declaration of Helsinki on "Medical Research Involving Human Subjects", and was approved by the IRB ethics committee at the New York University (NYU) Abu Dhabi. Healthy human subjects from the community of NYU Abu Dhabi were recruited to participate in the study. Subjects participated as paid volunteers (50 AED/h) and gave written informed consent prior to the experimental sessions. All subjects had normal or corrected-to-normal vision. In a preliminary behavioral session, subjects performed the WM task presented in Fig. 1. Only subjects who made 'symmetric swaps' (see section 4.2) in at least 5% of the trials were invited for a following MEG session. Twenty-six subjects participated in the MEG session (age range 19–37 years, *M*=24.33, *SD*=4.57; 4 female; all right-handed). Previous papers reporting PPI have ranged from 6 to 12 subjects and typically do not report effect size (Bangel et al., 2023; Desideri et al., 2019; Kawasaki et al., 2014; Yamanaka & Yamamoto, 2010); we based our sample size on an estimated effect size of 0.5 and power of 0.8. Both behavioral and MEG sessions were characterized by the same experimental design and procedures (see below).

4.2. Stimuli, design, and procedures

Visual stimuli consisted of 3 different items (colored circles), each of which measured 0.55 degrees of visual angle (DVA) in diameter, presented at an eccentricity of 4.5 DVA (Fig. 1). Stimulus items were presented on a gray background for 0.2 s, followed by a retention interval (delay period) during which the screen was blank (2 s). A white fixation point (0.2 DVA) was presented at the center of the screen during both stimulus presentation and delay period. In a pre-fixation period (1 s before the beginning of each trial), the fixation point changed color to red when the subjects broke fixation. Trials in which the subjects broke fixation were removed from successive analyses.

Subjects were asked to remember the 3 stimulus items, and to hold in mind their colors and locations during the delay period. After memory delay, subjects were asked to report the location of each of the items (report period), which were sequentially cued by presenting their color at the center of the screen. Subjects had no control over the order in which the items were cued during the report period. During the report, a white circle (same size and eccentricity of the items) was also presented on the screen. Subjects were asked to respond by adjusting the position of this white circle by using a rotating dial, and pressing a button on a response box to confirm their response. The report period was self-paced and was followed by a feedback period (1 s), showing the cued items ringed by white circles and the subjects' responses ringed by black circles. After the feedback, a self-paced inter-trial-interval (1 s) was used to allow the subjects to control their blinking. Subjects had to press a button to move onto the next trial.

Each subject performed 500 trials of the WM task in one session, divided into 10 blocks (50 trials each). During stimulus presentation, the items were always presented on one single hemifield (5 blocks on the left, 5 blocks on the right; randomly ordered), never across hemifields. In each trial, the spatial location of each stimulus item was randomly sampled from a space that depended on the hemifield used for that block, obtained by excluding locations that were 10 DVA away from the vertical meridian, and using a minimum gap of 15 DVA. This was done to prevent physical overlap between items and to avoid swaps purely induced by misperception of spatial location. The color of each item was selected randomly from a color wheel of 180 color segments, using a minimum gap of 15 color segments between items to avoid swaps that were purely induced by colors misperception.

A maximum likelihood approach was used to assign to each response a likelihood of being a binding error (non-target) or a target response. The approach was based on a mixture model that comprised the probability of correctly reporting the target item, the probability of incorrectly reporting a non-target item, and the probability of responding randomly (guess) (Schneegans & Bays, 2016) (https://www.paulbays.com/toolbox/). This procedure allowed categorizing each trial on the basis of the subjects' responses: trials with a probability of correctly reporting the target item greater than 0.95 in all three responses were considered high-performance (HP) targets, while those with a probability of incorrectly reporting a non-target item greater than 0.70 in at least one response were considered swaps. The remaining trials were considered low-performance (LP) reports. After MEG data preprocessing (see section 4.3), the number of trials across subjects was in the range 16–439 for HP targets (*M*=92.81, *SD*=88.06), 22–196 for swaps (*M*=116.38, *SD*=47.79), and 44–328 for LP trials (*M*=198.73, *SD*=64.03). For a specific analysis on swaps (see section 2.4), these were further divided into symmetric (at least two non-target reports) and non-symmetric swaps (single non-target report). The number of trials for the two swap types was in the range 7–81 (*M*=46.50, *SD*=18.51) and 15–123 (*M*=69.88, *SD*=31.70) across subjects, respectively.

4.3. MEG data acquisition and preprocessing

The locations of the marker coils on each subject, as well as fiducial locations (including nasion–NAS, left and right Pre-Auricular points–LPA/RPA, left and right PreFrontal–LPF/RPF), were recorded before MEG data acquisition. The shape of the subject's head was also recorded using a Polhemus dual source handheld FastSCAN-II system (Colchester, VT, United States of America), collecting between 2,500 and 4,500 points per subject. MEG data were acquired continuously using a 208-channel axial gradiometer system (Kanazawa Institute of Technology, Kanazawa, Japan), with sampling frequency of 1,000 Hz, and applying an online low-pass filter (cutoff frequency 200 Hz). During recording, the subject was lying in the supine position, while performing the WM task. The visual stimuli were generated using MATLAB (The MathWorks Inc., Natick, USA) and projected onto a screen positioned at 85 cm from the face of the subject.

A noise reduction procedure was applied to the continuous MEG data using the rotated spectral Principal Component Analysis (rsPCA), as implemented in the MEG laboratory software MEG160 (Yokogawa Electric

Corporation, Tokyo, Japan), with block width of 5000 and 30 shifts, by using eight magnetometer reference channels that were located away from the subject's head. The FastICA algorithm implemented in MNE-Python (Hyvarinen, 1999) was used to decompose the MEG data by independent component analysis (ICA), and components identified as eye blinks or heartbeat artifacts were removed from the data. Trials were defined using the time epoch from 1 s before stimulus onset (pre-fixation period) to 2.7 s after stimulus onset. Temporal demean and low-pass filtering (cutoff frequency 140 Hz) procedures were then applied to the data using FieldTrip (Oostenveld et al., 2011) (http://www.ru.nl/neuroimaging/fieldtrip/). Noisy trials and channels were automatically identified (high variance) and rejected by visual inspection. Trials showing residual ocular or muscular artifacts were rejected by visual inspection. Rejected channels were interpolated using the average of all their neighbors. As a final step, the MEG channels were switched between right and left side in all trials in which the stimulus was presented on the left hemifield. This procedure was performed to compound together the trials from both left and right stimulus presentations, as if the stimulus presentation occurred always on the right hemifield.

4.4. Statistics

All statistical comparisons were performed with a cluster-based permutation approach (Maris & Oostenveld, 2007), using a two-tailed dependent *t*-test (df=25, p<0.05, alpha level distributed over both tails by multiplying the p-values with a factor of 2 prior to thresholding), 1,000 permutations, and p_{perm}<0.05 for the permutation test. The same settings were used in every comparison, unless specified otherwise, while the space for clustering in each analysis is specified in the sections below.

In each comparison between trial types, effect sizes of the differences were estimated using Cohen's *d* (J. Cohen, 1992), after averaging the estimates separately per trial type, either over the time points and frequencies identified by the significant cluster (Phase Preservation Index–PPI analysis; see section 4.5), or simply over time points (frequency sliding variability–FSV analyses; see section 4.6–4.7).

4.5. Phase Preservation Index

The Phase Preservation Index–PPI (Mazaheri & Jensen, 2006) was used to estimate the consistency over trials of time-frequency phase differences with respect to a reference phase, separately for each trial type (HP, LP, and swaps). Specifically, PPI measures the intertrial consistency in phase differences as a function of time (t), with respect to a reference time point (t_{ref}), at a specific frequency of interest f_0:

$$PPI(f_0, t) = \frac{1}{N}\left|\sum_{k=1}^{N} e^{i\phi^k(f_0,t_{ref})-i\phi^k(f_0,t)}\right| \quad (1)$$

For each trial ($k=1,\ldots,N$) the instantaneous phase $\Phi^k(f_0,t)$ was derived using a Morlet wavelet transform (central frequency parameter ω_0=6; zero-padding to solve edge effects problem) (Pagnotta et al., 2018a, 2018b; Torrence & Compo, 1998), in each MEG sensor. The frequency of interest f_0 varied in the range 1–50 Hz (1 Hz steps). The

delay onset was used as the reference time point (t_{ref}=0.2 s) for PPI estimation. Since PPI is sensitive to the number of trials, this was balanced between trial types (HP targets and swaps) at the level of each subject. This procedure was performed by i) considering 100 combinations of trials subsets in the type with more trials, to match the number in the type with less, ii) estimating PPI for each subset, and finally iii) computing the median of PPI estimates across subsets. PPI was estimated only once in the type with less trials. PPI was then compared between HP targets and swaps using a cluster-based permutation approach (see section 4.4), over all time points in the delay period (0.2–2.2 s), frequencies (1–50 Hz), and MEG sensors. Neighboring sensors were here defined based on their 3D geometrical distance (max distance of 25 mm for neighbors definition).

4.6. Frequency sliding (instantaneous frequency)

The frequency sliding–FS (M. X. Cohen, 2014b) was employed to capture the single-trial temporal fluctuations in oscillation peak frequency. FS estimation relies on five core steps: (i) extract the phase angle time series at a certain frequency (Fig. 6a); (ii) unwrap the phase angle time series (Fig. 6a); (iii) compute the first order derivative over time (measure in rad); (iv) scale the measure by multiplying by the sampling frequency and dividing by 2π (measure converted into Hz; Fig. 6b); (v) apply a median filter to attenuate neurophysiological noise spikes (Fig. 6b). A Morlet wavelet transform was employed to extract the phase angle time series (step i), using ω_0=6 and zero-padding (Pagnotta et al., 2018a, 2018b; Torrence & Compo, 1998). An order of 10 and a maximum window size of 0.1 s were used for the median filter settings (step v), which reassigned each time point to be the median of a distribution made from surrounding points (M. X. Cohen, 2014b).

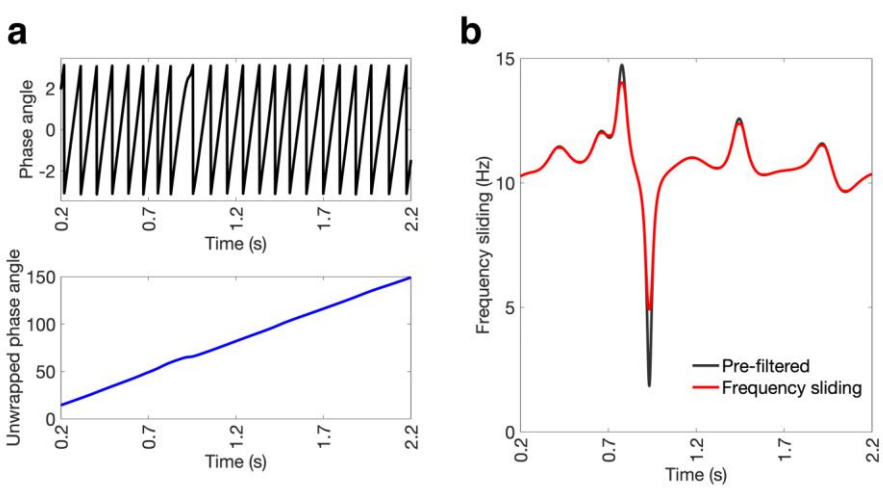


Fig. 6 Estimation of frequency sliding–FS (instantaneous frequency). **a** Example of phase angle time series at 10 Hz, before (top, in black) and after (bottom, in blue) the unwrapping procedure. **b** Frequency sliding pre-filtered (black) and after median filtering (red) for the phase angle time series shown in panel a.

To derive measures of FS variability (FSV) over time and over trials, single-trial FS estimates were first derived at 10 Hz, separately for each trial type; then, the median between trials of the standard deviation of FS over time points (alpha-FSV over time) and the median between time points of the standard deviation of FS over trials (alpha-FSV over trials) were estimated on a sliding window of 0.4 s (i.e., four cycles of the 10 Hz oscillation), with maximum overlap in time. The two measures of alpha-FSV were successively compared between trial types using a cluster-based permutation approach (see section 4.4), over time points during the stimulus presentation and in the delay period (0–2.2 s) and sensors, where neighboring sensors were defined based on their 3D geometrical distance (max distance of 25 mm for neighbors definition).

4.7. Source reconstruction

The solution of the MEG forward problem requires models of the head volume conduction, the sensors, and the cortical sources. To derive these models, a transformation matrix (rotation and translation) was estimated through an automatic procedure that aligned the five fiducials (NAS, LPA, RPA, LPF, and RPF) between subject-specific Polhemus head shape and markers. The transformation matrix was then used to align the sensors' locations to the head shape. The boundary element method (BEM) template provided in FieldTrip (Oostenveld et al., 2011) was used for the head volume conduction. The template head model was coregistered with the head shape, and the transformation was successively refined using a method that relies on fitting spheres to the head shapes (which allows deriving translation and global scaling). The head volume conduction was computed based on the refined template head models, by creating a single-shell model on the basis of the brain compartment, and taking its inner shell. As a final step, the head model and head shape were manually realigned, and the registration was also checked manually. The source model was constructed using template brain coordinates (12.5 mm resolution; 10 mm outward shift; 1,499 equivalent current dipoles), and then transformed into head coordinates. The subject-specific lead field matrix was computed considering a forward operator with unconstrained orientation, where each source point was modeled as three orthogonal equivalent current dipoles.

Once the lead field matrix was computed, the MEG inverse problem was solved using the linearly constrained minimum variance (LCMV) beamformer (Van Veen et al., 1997). An estimate of sensor-space covariance matrix was obtained from the 1 s pre-fixation period. Source-reconstructed time series were extracted from all source points using a method based on singular value decomposition (Rubega et al., 2019), which allows to derive single-trial time series for each source point that best explain the variability in dipole orientations and signal strengths across trials for that point. After source reconstruction, measures of PPI, single-trial FS and FSV at 10 Hz were derived for each source point, following the same procedures used in the sensor-space analyses (see sections 4.5–4.6). PPI and the two measures of FSV were then compared between HP targets and swaps using a cluster-based permutation approach (see section 4.4), over time points either in the delay period only (0.2–2.2 s; for PPI) or in both stimulus presentation and delay period (0–2.2 s; for alpha-FSV), and source points in source-

space. Neighboring points were defined based on their distance in MNI space (max distance of 15 mm for neighbors definition).

References

Baddeley, A. D. (2003). Working memory: Looking back and looking forward. *Nature Reviews Neuroscience*, *4*(10), 829–839. https://doi.org/10.1038/nrn1201

Baddeley, A. D. (2012). Working Memory: Theories, Models, and Controversies. *Annual Review of Psychology*, *63*(1), 1–29. https://doi.org/10.1146/annurev-psych-120710-100422

Baddeley, A. D., & Hitch, G. (1974). Working Memory. In *Psychology of Learning and Motivation* (Vol. 8, pp. 47–89). Elsevier. https://doi.org/10.1016/S0079-7421(08)60452-1

Badre, D. (2020). *On Task: How Our Brain Gets Things Done*. Princeton University Press. http://www.jstor.org/stable/10.2307/j.ctv11hprj1

Badre, D., & Nee, D. E. (2018). Frontal Cortex and the Hierarchical Control of Behavior. *Trends in Cognitive Sciences*, *22*(2), 170–188. https://doi.org/10.1016/j.tics.2017.11.005

Bangel, K. A., Bais, M., Eijsker, N., Schuurman, P. R., Van Den Munckhof, P., Figee, M., Smit, D. J. A., & Denys, D. (2023). Acute effects of deep brain stimulation on brain function in obsessive–compulsive disorder. *Clinical Neurophysiology*, *148*, 109–117. https://doi.org/10.1016/j.clinph.2022.12.012

Barbosa, J., Babushkin, V., Temudo, A., Sreenivasan, K. K., & Compte, A. (2021). Across-Area Synchronization Supports Feature Integration in a Biophysical Network Model of Working Memory. *Frontiers in Neural Circuits*, *15*, 716965. https://doi.org/10.3389/fncir.2021.716965

Bays, P. M. (2016). Evaluating and excluding swap errors in analogue tests of working memory. *Scientific Reports*, *6*(1), 19203. https://doi.org/10.1038/srep19203

Brincat, S. L., Donoghue, J. A., Mahnke, M. K., Kornblith, S., Lundqvist, M., & Miller, E. K. (2021). Interhemispheric transfer of working memories. *Neuron*, *109*(6), 1055-1066.e4. https://doi.org/10.1016/j.neuron.2021.01.016

Bugeon, S., Duffield, J., Dipoppa, M., Ritoux, A., Prankerd, I., Nicoloutsopoulos, D., Orme, D., Shinn, M., Peng, H., Forrest, H., Viduolyte, A., Reddy, C. B., Isogai, Y., Carandini, M., & Harris, K. D. (2022). A transcriptomic axis predicts state modulation of cortical interneurons. *Nature*, *607*(7918), 330–338. https://doi.org/10.1038/s41586-022-04915-7

Bush, D., & Burgess, N. (2020). Advantages and detection of phase coding in the absence of rhythmicity. *Hippocampus*, *30*(7), 745–762. https://doi.org/10.1002/hipo.23199

Buzsáki, G. (2006). *Rhythms of the Brain*. Oxford University Press. https://doi.org/10.1093/acprof:oso/9780195301069.001.0001

Christophel, T. B., Klink, P. C., Spitzer, B., Roelfsema, P. R., & Haynes, J.-D. (2017). The Distributed Nature of Working Memory. *Trends in Cognitive Sciences*, *21*(2), 111–124. https://doi.org/10.1016/j.tics.2016.12.007

Cohen, J. (1992). A power primer. *Psychological Bulletin*, *112*(1), 155–159. https://doi.org/10.1037/0033-2909.112.1.155

Cohen, M. X. (2014a). *Analyzing Neural Time Series Data: Theory and Practice*. MIT Press.

Cohen, M. X. (2014b). Fluctuations in Oscillation Frequency Control Spike Timing and Coordinate Neural Networks. *Journal of Neuroscience*, *34*(27), 8988–8998. https://doi.org/10.1523/JNEUROSCI.0261-14.2014

Compte, A., Constantinidis, C., Tegnér, J., Raghavachari, S., Chafee, M. V., Goldman-Rakic, P. S., & Wang, X.-J. (2003). Temporally Irregular Mnemonic Persistent Activity in Prefrontal Neurons of Monkeys During a Delayed Response Task. *Journal of Neurophysiology*, *90*(5), 3441–3454. https://doi.org/10.1152/jn.00949.2002

Cowan, N. (2001). The magical number 4 in short-term memory: A reconsideration of mental storage capacity. *Behavioral and Brain Sciences*, *24*(1), 87–114. https://doi.org/10.1017/S0140525X01003922

de Vries, I. E. J., van Driel, J., Karacaoglu, M., & Olivers, C. N. L. (2018). Priority Switches in Visual Working Memory are Supported by Frontal Delta and Posterior Alpha Interactions. *Cerebral Cortex*, *28*(11), 4090–4104. https://doi.org/10.1093/cercor/bhy223

Desideri, D., Zrenner, C., Ziemann, U., & Belardinelli, P. (2019). Phase of sensorimotor μ-oscillation modulates cortical responses to transcranial magnetic stimulation of the human motor cortex. *The Journal of Physiology*, *597*(23), 5671–5686. https://doi.org/10.1113/JP278638

D'Esposito, M., & Postle, B. R. (2015). The Cognitive Neuroscience of Working Memory. *Annual Review of*

Psychology, *66*(1), 115–142. https://doi.org/10.1146/annurev-psych-010814-015031

Fell, J., & Axmacher, N. (2011). The role of phase synchronization in memory processes. *Nature Reviews Neuroscience*, *12*(2), Article 2. https://doi.org/10.1038/nrn2979

Goldman-Rakic, P. S. (1987). Circuitry of Primate Prefrontal Cortex and Regulation of Behavior by Representational Memory. In *Comprehensive Physiology* (1st ed., pp. 373–417). Wiley. https://onlinelibrary.wiley.com/doi/10.1002/cphy.cp010509

Golomb, J. D. (2015). Divided spatial attention and feature-mixing errors. *Attention, Perception, & Psychophysics*, *77*(8), 2562–2569. https://doi.org/10.3758/s13414-015-0951-0

Golomb, J. D., L'Heureux, Z. E., & Kanwisher, N. (2014). Feature-Binding Errors After Eye Movements and Shifts of Attention. *Psychological Science*, *25*(5), 1067–1078. https://doi.org/10.1177/0956797614522068

Gorgoraptis, N., Catalao, R. F. G., Bays, P. M., & Husain, M. (2011). Dynamic Updating of Working Memory Resources for Visual Objects. *Journal of Neuroscience*, *31*(23), 8502–8511. https://doi.org/10.1523/JNEUROSCI.0208-11.2011

Haber, S. N. (2003). The primate basal ganglia: Parallel and integrative networks. *Journal of Chemical Neuroanatomy*, *26*(4), 317–330. https://doi.org/10.1016/j.jchemneu.2003.10.003

Harrison, S. A., & Tong, F. (2009). Decoding reveals the contents of visual working memory in early visual areas. *Nature*, *458*(7238), 632–635. https://doi.org/10.1038/nature07832

Huang, L. (2020). Distinguishing target biases and strategic guesses in visual working memory. *Attention, Perception, & Psychophysics*, *82*(3), 1258–1270. https://doi.org/10.3758/s13414-019-01913-2

Huang, Z. J., Di Cristo, G., & Ango, F. (2007). Development of GABA innervation in the cerebral and cerebellar cortices. *Nature Reviews Neuroscience*, *8*(9), 673–686. https://doi.org/10.1038/nrn2188

Hyvarinen, A. (1999). Fast and robust fixed-point algorithms for independent component analysis. *IEEE Transactions on Neural Networks*, *10*(3), 626–634. https://doi.org/10.1109/72.761722

Jensen, O., Gips, B., Bergmann, T. O., & Bonnefond, M. (2014). Temporal coding organized by coupled alpha and gamma oscillations prioritize visual processing. *Trends in Neurosciences*, *37*(7), 357–369.

https://doi.org/10.1016/j.tins.2014.04.001

Jensen, O., & Mazaheri, A. (2010). Shaping Functional Architecture by Oscillatory Alpha Activity: Gating by Inhibition. *Frontiers in Human Neuroscience*, *4*. https://doi.org/10.3389/fnhum.2010.00186

Jokisch, D., & Jensen, O. (2007). Modulation of Gamma and Alpha Activity during a Working Memory Task Engaging the Dorsal or Ventral Stream. *Journal of Neuroscience*, *27*(12), 3244–3251. https://doi.org/10.1523/JNEUROSCI.5399-06.2007

Kawasaki, M., Uno, Y., Mori, J., Kobata, K., & Kitajo, K. (2014). Transcranial magnetic stimulation-induced global propagation of transient phase resetting associated with directional information flow. *Frontiers in Human Neuroscience*, *8*. https://doi.org/10.3389/fnhum.2014.00173

Kornblith, S., Buschman, T. J., & Miller, E. K. (2016). Stimulus Load and Oscillatory Activity in Higher Cortex. *Cerebral Cortex*, *26*(9), 3772–3784. https://doi.org/10.1093/cercor/bhv182

Lorenc, E. S., & Sreenivasan, K. K. (2021). Reframing the debate: The distributed systems view of working memory. *Visual Cognition*, *29*(7), 416–424. https://doi.org/10.1080/13506285.2021.1899091

Luck, S. J., & Vogel, E. K. (1997). The capacity of visual working memory for features and conjunctions. *Nature*, *390*(6657), 279–281. https://doi.org/10.1038/36846

Lundqvist, M., Brincat, S. L., Rose, J., Warden, M. R., Buschman, T. J., Miller, E. K., & Herman, P. (2023). Working memory control dynamics follow principles of spatial computing. *Nature Communications*, *14*(1), Article 1. https://doi.org/10.1038/s41467-023-36555-4

Lundqvist, M., Herman, P., Warden, M. R., Brincat, S. L., & Miller, E. K. (2018). Gamma and beta bursts during working memory readout suggest roles in its volitional control. *Nature Communications*, *9*(1), 394. https://doi.org/10.1038/s41467-017-02791-8

Lundqvist, M., Rose, J., Brincat, S. L., Warden, M. R., Buschman, T. J., Herman, P., & Miller, E. K. (2022). Reduced variability of bursting activity during working memory. *Scientific Reports*, *12*(1), 15050. https://doi.org/10.1038/s41598-022-18577-y

Lundqvist, M., Rose, J., Herman, P., Brincat, S. L., Buschman, T. J., & Miller, E. K. (2016). Gamma and Beta Bursts Underlie Working Memory. *Neuron*, *90*(1), 152–164.

https://doi.org/10.1016/j.neuron.2016.02.028

Ma, W. J., Husain, M., & Bays, P. M. (2014). Changing concepts of working memory. *Nature Neuroscience*, *17*(3), 347–356. https://doi.org/10.1038/nn.3655

Mallett, R., Lorenc, E. S., & Lewis-Peacock, J. A. (2022). Working Memory Swap Errors Have Identifiable Neural Representations. *Journal of Cognitive Neuroscience*, *34*(5), 776–786. https://doi.org/10.1162/jocn_a_01831

Maris, E., & Oostenveld, R. (2007). Nonparametric statistical testing of EEG- and MEG-data. *Journal of Neuroscience Methods*, *164*(1), 177–190. https://doi.org/10.1016/j.jneumeth.2007.03.024

Mazaheri, A., & Jensen, O. (2006). Posterior α activity is not phase-reset by visual stimuli. *Proceedings of the National Academy of Sciences*, *103*(8), 2948–2952. https://doi.org/10.1073/pnas.0505785103

McMaster, J. M. V., Tomić, I., Schneegans, S., & Bays, P. M. (2022). Swap errors in visual working memory are fully explained by cue-feature variability. *Cognitive Psychology*, *137*, 101493. https://doi.org/10.1016/j.cogpsych.2022.101493

Mendoza-Halliday, D., Xu, H., Azevedo, F. A. C., & Desimone, R. (2024). Dissociable neuronal substrates of visual feature attention and working memory. *Neuron*. https://doi.org/10.1016/j.neuron.2023.12.007

Menon, V., & D'Esposito, M. (2022). The role of PFC networks in cognitive control and executive function. *Neuropsychopharmacology*, *47*(1), 90–103. https://doi.org/10.1038/s41386-021-01152-w

Miller, E. K., & Cohen, J. D. (2001). An Integrative Theory of Prefrontal Cortex Function. *Annual Review of Neuroscience*, *24*(1), 167–202. https://doi.org/10.1146/annurev.neuro.24.1.167

Miller, G. A., Galanter, E., & Pribram, K. H. (1960). *Plans and the structure of behavior* (pp. x, 222). Henry Holt and Co. https://doi.org/10.1037/10039-000

Miller, J. A., Tambini, A., Kiyonaga, A., & D'Esposito, M. (2022). Long-term learning transforms prefrontal cortex representations during working memory. *Neuron*, S0896627322008583. https://doi.org/10.1016/j.neuron.2022.09.019

O'Keefe, J., & Burgess, N. (2005). Dual phase and rate coding in hippocampal place cells: Theoretical significance and relationship to entorhinal grid cells. *Hippocampus*, *15*(7), 853–866.

https://doi.org/10.1002/hipo.20115

Oostenveld, R., Fries, P., Maris, E., & Schoffelen, J.-M. (2011). FieldTrip: Open Source Software for Advanced Analysis of MEG, EEG, and Invasive Electrophysiological Data. *Computational Intelligence and Neuroscience*, *2011*, 1–9. https://doi.org/10.1155/2011/156869

Pagnotta, M. F., Dhamala, M., & Plomp, G. (2018a). Assessing the performance of Granger–Geweke causality: Benchmark dataset and simulation framework. *Data in Brief*, *21*, 833–851. https://doi.org/10.1016/j.dib.2018.10.034

Pagnotta, M. F., Dhamala, M., & Plomp, G. (2018b). Benchmarking nonparametric Granger causality: Robustness against downsampling and influence of spectral decomposition parameters. *NeuroImage*, *183*, 478–494. https://doi.org/10.1016/j.neuroimage.2018.07.046

Palva, S., Kulashekhar, S., Hamalainen, M., & Palva, J. M. (2011). Localization of Cortical Phase and Amplitude Dynamics during Visual Working Memory Encoding and Retention. *Journal of Neuroscience*, *31*(13), 5013–5025. https://doi.org/10.1523/JNEUROSCI.5592-10.2011

Panzeri, S., Brunel, N., Logothetis, N. K., & Kayser, C. (2010). Sensory neural codes using multiplexed temporal scales. *Trends in Neurosciences*, *33*(3), 111–120. https://doi.org/10.1016/j.tins.2009.12.001

Pasternak, T., & Greenlee, M. W. (2005). Working memory in primate sensory systems. *Nature Reviews Neuroscience*, *6*(2), 97–107. https://doi.org/10.1038/nrn1603

Pertzov, Y., Dong, M. Y., Peich, M.-C., & Husain, M. (2012). Forgetting What Was Where: The Fragility of Object-Location Binding. *PLoS ONE*, *7*(10), e48214. https://doi.org/10.1371/journal.pone.0048214

Pertzov, Y., & Husain, M. (2014). The privileged role of location in visual working memory. *Attention, Perception, & Psychophysics*, *76*(7), 1914–1924. https://doi.org/10.3758/s13414-013-0541-y

Pertzov, Y., Manohar, S., & Husain, M. (2017). Rapid forgetting results from competition over time between items in visual working memory. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *43*(4), 528–536. https://doi.org/10.1037/xlm0000328

Pesaran, B., Pezaris, J. S., Sahani, M., Mitra, P. P., & Andersen, R. A. (2002). Temporal structure in neuronal activity during working memory in macaque parietal cortex. *Nature Neuroscience*, *5*(8), 805–811.

https://doi.org/10.1038/nn890

Pina, J. E., Bodner, M., & Ermentrout, B. (2018). Oscillations in working memory and neural binding: A mechanism for multiple memories and their interactions. *PLOS Computational Biology*, *14*(11), e1006517. https://doi.org/10.1371/journal.pcbi.1006517

Pratte, M. S. (2019). Swap errors in spatial working memory are guesses. *Psychonomic Bulletin & Review*, *26*(3), 958–966. https://doi.org/10.3758/s13423-018-1524-8

Raffone, A., & Wolters, G. (2001). A Cortical Mechanism for Binding in Visual Working Memory. *Journal of Cognitive Neuroscience*, *13*(6), 766–785. https://doi.org/10.1162/08989290152541430

Rajsic, J., Swan, G., Wilson, D. E., & Pratt, J. (2017). Accessibility limits recall from visual working memory. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *43*, 1415–1431. https://doi.org/10.1037/xlm0000387

Rajsic, J., & Wilson, D. E. (2014). Asymmetrical access to color and location in visual working memory. *Attention, Perception, & Psychophysics*, *76*(7), 1902–1913. https://doi.org/10.3758/s13414-014-0723-2

Rubega, M., Carboni, M., Seeber, M., Pascucci, D., Tourbier, S., Toscano, G., Van Mierlo, P., Hagmann, P., Plomp, G., Vulliemoz, S., & Michel, C. M. (2019). Estimating EEG Source Dipole Orientation Based on Singular-value Decomposition for Connectivity Analysis. *Brain Topography*. https://doi.org/10.1007/s10548-018-0691-2

Sassenhagen, J., & Draschkow, D. (2019). Cluster-based permutation tests of MEG/EEG data do not establish significance of effect latency or location. *Psychophysiology*, *56*(6), e13335. https://doi.org/10.1111/psyp.13335

Schneegans, S., & Bays, P. M. (2016). No fixed item limit in visuospatial working memory. *Cortex*, *83*, 181–193. https://doi.org/10.1016/j.cortex.2016.07.021

Schneegans, S., & Bays, P. M. (2017). Neural Architecture for Feature Binding in Visual Working Memory. *The Journal of Neuroscience*, *37*(14), 3913–3925. https://doi.org/10.1523/JNEUROSCI.3493-16.2017

Schneegans, S., & Bays, P. M. (2019). New perspectives on binding in visual working memory. *British Journal of Psychology*, *110*(2), 207–244. https://doi.org/10.1111/bjop.12345

Scimeca, J. M., Kiyonaga, A., & D'Esposito, M. (2018). Reaffirming the Sensory Recruitment Account of Working Memory. *Trends in Cognitive Sciences*, *22*(3), 190–192. https://doi.org/10.1016/j.tics.2017.12.007

Senoussi, M., Verbeke, P., & Verguts, T. (2022). Time-Based Binding as a Solution to and a Limitation for Flexible Cognition. *Frontiers in Psychology*, *12*, 798061. https://doi.org/10.3389/fpsyg.2021.798061

Serences, J. T. (2016). Neural mechanisms of information storage in visual short-term memory. *Vision Research*, *128*, 53–67. https://doi.org/10.1016/j.visres.2016.09.010

Serences, J. T., Ester, E. F., Vogel, E. K., & Awh, E. (2009). Stimulus-Specific Delay Activity in Human Primary Visual Cortex. *Psychological Science*, *20*(2), 207–214. https://doi.org/10.1111/j.1467-9280.2009.02276.x

Shadlen, M. N., & Movshon, J. A. (1999). Synchrony Unbound: A Critical Evaluation of the Temporal Binding Hypothesis. *Neuron*, *24*(1), 67–77. https://doi.org/10.1016/S0896-6273(00)80822-3

Sherman, S. M., & Guillery, R. W. (1996). Functional organization of thalamocortical relays. *Journal of Neurophysiology*, *76*(3), 1367–1395. https://doi.org/10.1152/jn.1996.76.3.1367

Siegel, M., Warden, M. R., & Miller, E. K. (2009). Phase-dependent neuronal coding of objects in short-term memory. *Proceedings of the National Academy of Sciences*, *106*(50), 21341–21346. https://doi.org/10.1073/pnas.0908193106

Sreenivasan, K. K., Curtis, C. E., & D'Esposito, M. (2014). Revisiting the role of persistent neural activity during working memory. *Trends in Cognitive Sciences*, *18*(2), 82–89. https://doi.org/10.1016/j.tics.2013.12.001

Stringer, C., Pachitariu, M., Steinmetz, N., Reddy, C. B., Carandini, M., & Harris, K. D. (2019). Spontaneous behaviors drive multidimensional, brainwide activity. *Science*, *364*(6437), eaav7893. https://doi.org/10.1126/science.aav7893

The Petilla Interneuron Nomenclature Group (PING). (2008). Petilla terminology: Nomenclature of features of GABAergic interneurons of the cerebral cortex. *Nature Reviews Neuroscience*, *9*(7), 557–568. https://doi.org/10.1038/nrn2402

Torrence, C., & Compo, G. P. (1998). A Practical Guide to Wavelet Analysis. *Bulletin of the American Meteorological Society*, *79*(1), 61–78. https://doi.org/10.1175/1520-0477(1998)079<0061:APGTWA>2.0.CO;2

Van Veen, B. D., Van Drongelen, W., Yuchtman, M., & Suzuki, A. (1997). Localization of brain electrical activity via linearly constrained minimum variance spatial filtering. *IEEE Transactions on Biomedical Engineering*, *44*(9), 867–880. https://doi.org/10.1109/10.623056

Vogel, E. K., Woodman, G. F., & Luck, S. J. (2001). Storage of features, conjunctions, and objects in visual working memory. *Journal of Experimental Psychology: Human Perception and Performance*, *27*(1), 92–114. https://doi.org/10.1037/0096-1523.27.1.92

von Der Malsburg, C. (1994). The Correlation Theory of Brain Function. In E. Domany, J. L. Van Hemmen, & K. Schulten (Eds.), *Models of Neural Networks* (pp. 95–119). Springer New York. https://doi.org/10.1007/978-1-4612-4320-5_2

von Der Malsburg, C. (1995). Binding in models of perception and brain function. *Current Opinion in Neurobiology*, *5*(4), 520–526. https://doi.org/10.1016/0959-4388(95)80014-X

Wallis, G., Stokes, M., Cousijn, H., Woolrich, M., & Nobre, A. C. (2015). Frontoparietal and Cingulo-opercular Networks Play Dissociable Roles in Control of Working Memory. *Journal of Cognitive Neuroscience*, *27*(10), 2019–2034. https://doi.org/10.1162/jocn_a_00838

Wang, B., Cao, X., Theeuwes, J., Olivers, C. N. L., & Wang, Z. (2017). Separate capacities for storing different features in visual working memory. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *43*(2), 226–236. https://doi.org/10.1037/xlm0000295

Yamanaka, K., & Yamamoto, Y. (2010). Single-trial EEG Power and Phase Dynamics Associated with Voluntary Response Inhibition. *Journal of Cognitive Neuroscience*, *22*(4), 714–727. https://doi.org/10.1162/jocn.2009.21258

Acknowledgements

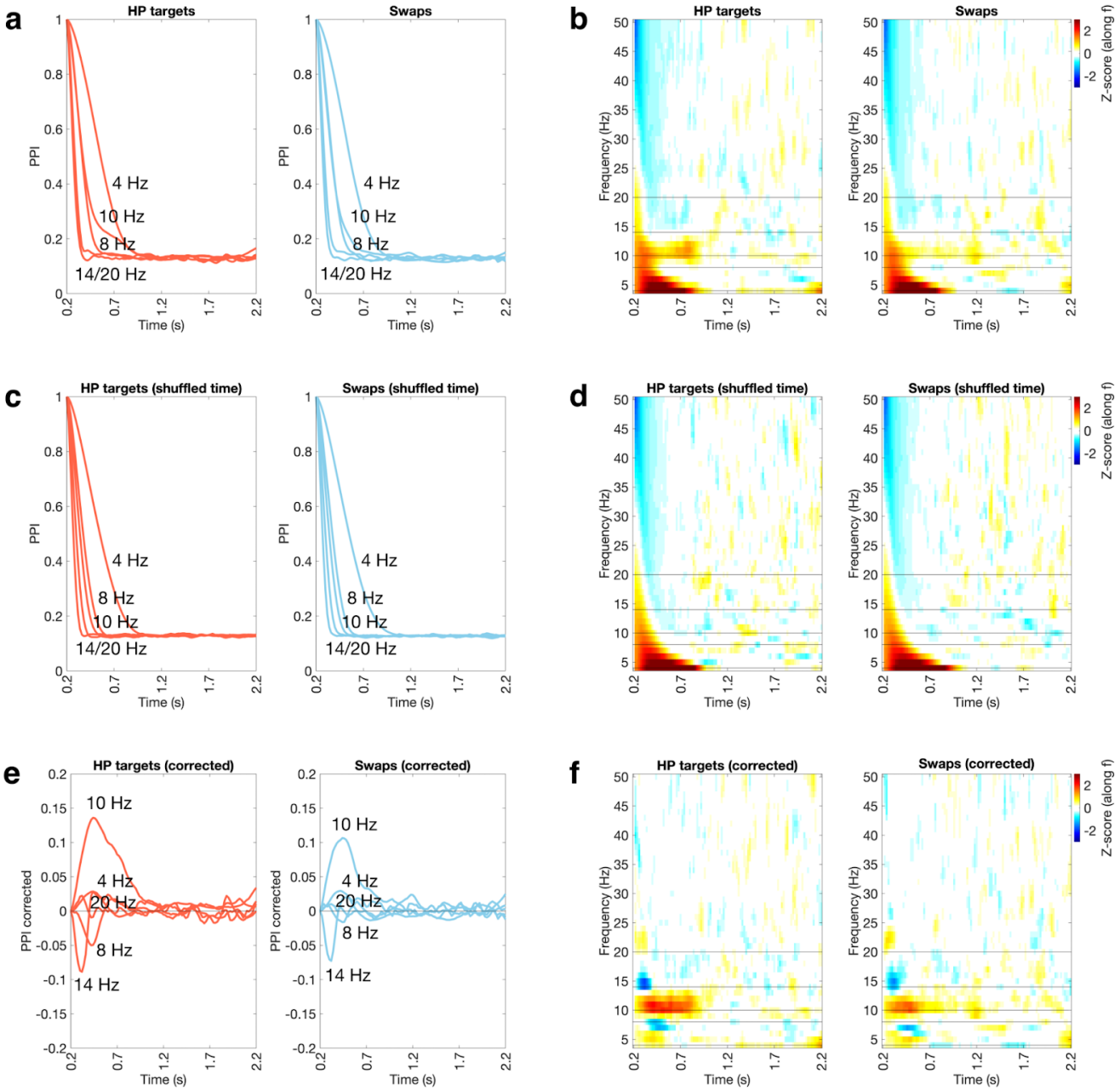
This work was supported by the Swiss National Science Foundation (Grants P2FRP3_195083 and P500PB_214404 to M.F.P.), the National Institutes of Health (Grant R01MH063901 to M.D.), and the NYU Research Enhancement Fund and the NYUAD Center for Brain and Health, funded by Tamkeen under NYUAD Research Institute (grant CG012 to K.K.S.). This research was conducted using resources provided by the High Performance Computing Cluster and the Core Technology Platform at NYUAD. A.C. and J.B. acknowledge support from the Spanish Ministry of Science, Competitiveness and Universities co-funded by the European Regional Development Fund (Refs: BFU 2015-65318-R, RTI2018-094190-B-I00). J.B. was supported by the Bial Foundation (ref: 356/19).

Competing interests

The authors declare that the study has been conducted in the absence of any conflict of interest.

Supplementary material

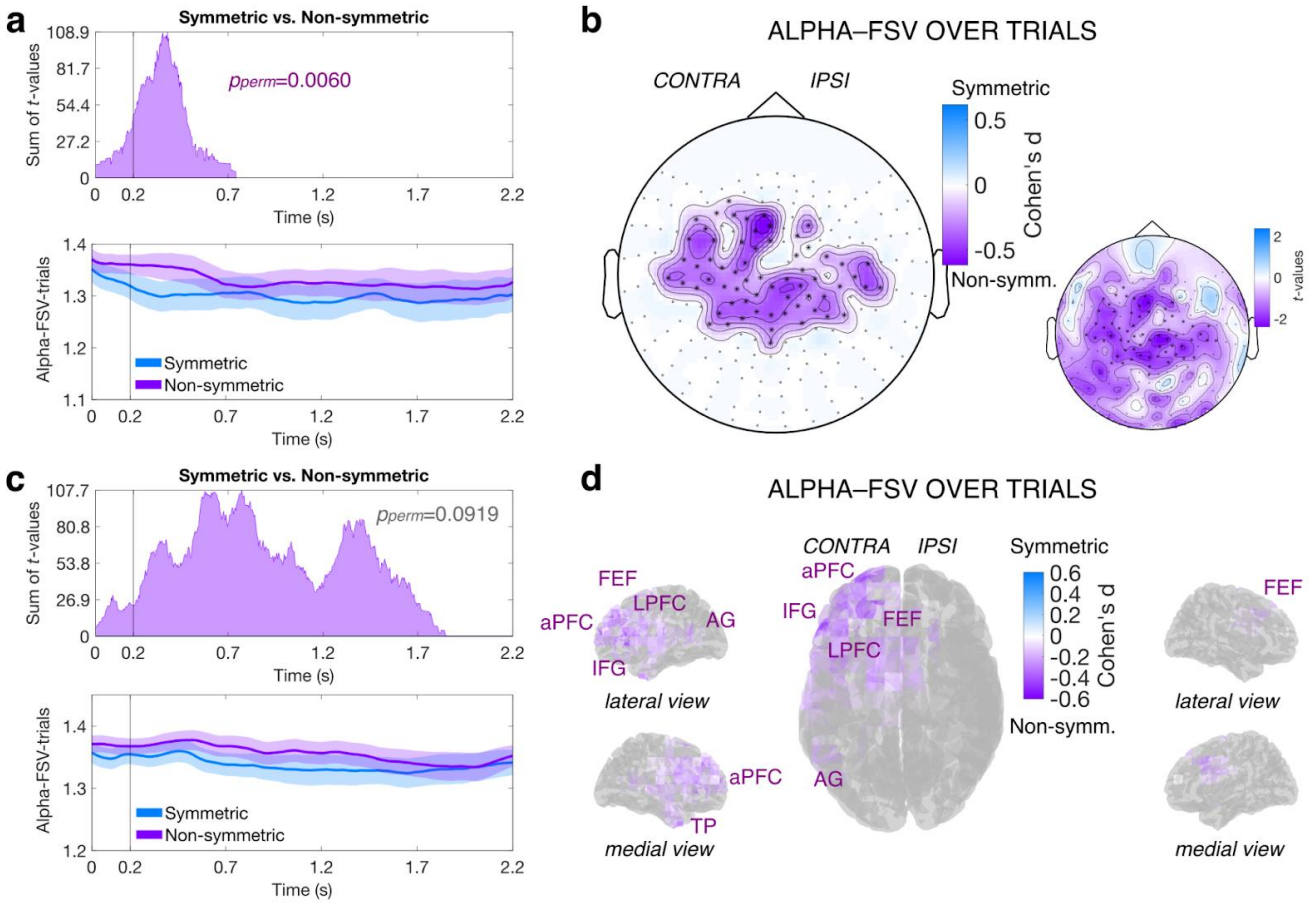
SM–2.2. Reduced phase preservation in swaps during WM maintenance



Supplementary Fig. 2 PPI estimates between HP targets and swaps. **a** Time course of PPI in HP targets (left) and swaps (right) at different frequencies, highlighted by the black horizontal lines in the next panel. **b** PPI: z-scores along frequencies for each time point in the delay, in HP targets (left) and swaps (right). **c** Time course of PPI obtained from shuffling the signals over time in HP targets (left) and swaps (right) at different frequencies, highlighted by the black horizontal lines in the next panel. **d** PPI (shuffled signals over time): z-scores along frequencies for each time point in the delay, in HP targets (left) and swaps (right). **e** Time course of PPI corrected

in HP targets (left) and swaps (right) at different frequencies, highlighted by the black horizontal lines in the next panel. **f** PPI corrected: z-scores along frequencies for each time point in the delay, in HP targets (left) and swaps (right).

SM–2.4. Distinct possible sources of swaps



Supplementary Fig. 5 Differences in alpha-FSV over trials between symmetric and non-symmetric swaps. **a** Time distribution of the sum of alpha-FSV differences across MEG sensors (*t*-values; top), and time course of the average alpha-FSV across sensors of the observed cluster for each condition (bottom). The shadings represent the standard error of the mean. **b** Topography plot with superimposed effect sizes of alpha-FSV differences between symmetric and non-symmetric swaps, for each MEG sensor of the cluster (t-values are shown on the smaller topography plot on the right). **c** Time distribution of the sum of alpha-FSV differences across source points in source-space (*t*-values; top), and time course of the average alpha-FSV across source points of the cluster with the lowest *p*-value from the permutation test (p_{perm}=0.0919) for each condition (bottom). The shadings represent the standard error of the mean. **d** Glass brain in MNI coordinates with superimposed effect sizes of differences between symmetric and non-symmetric swaps.