Prefrontal neuronal dynamics in the absence of task execution

Shusen Pu[1,2+], Wenhao Dang[1+], Xue-Lian Qi[3], Christos Constantinidis[1,4,5]

1. Department of Biomedical Engineering, Vanderbilt University, Nashville TN 37235, USA.
2. Department of Mathematics and Statistics, University of West Florida, Pensacola, FL 32514, USA.
3. Department of Neurobiology & Anatomy, Wake Forest School of Medicine, Winston-Salem, NC 27157, USA.
4. Neuroscience Program, Vanderbilt University, Nashville TN 37235, USA.
5. Department of Ophthalmology and Visual Sciences, Vanderbilt University Medical Center, Nashville TN 37232, USA.

+These authors contributed equally to this work

Keywords: monkey; population coding; prefrontal cortex; working memory

Lead Contact / Corresponding Author:
Christos.Constantinidis.1@vanderbilt.edu

SUMMARY

Activity across prefrontal cortical populations represents stimuli in a low-dimensional manifold that often undergoes transformations during the execution of cognitive tasks. These dynamics are commonly thought to be determined by whatever cognitive operation is being performed at the time, for example, rotation of stimulus representation in population activity during working memory is thought to reduce interference by subsequent distractors. The assumption that cognitive operations dictate stimulus transformations has not yet been tested experimentally. Here we show that the activity space of prefrontal neurons in naïve monkeys exhibits significant rotations, some aspects of which remain remarkably unchanged after training to perform working memory tasks. Training does induce changes in population dynamics, which further distinguish correct and error trials during task execution. Our results reveal automatic functions of prefrontal neural circuits that form the substrate of cognitive flexibility.

INTRODUCTION

Prefrontal neurons represent sensory stimuli in a dynamic, task-specific manner [1, 2, 3]. Analysis of population activity with methods of dimensionality reduction reveals that stimuli can be represented in a low dimensional space, or manifold, with the majority of the firing rate variance captured by a few dimensions [4, 5]. Furthermore, the representation of stimuli in the reduced space is not static but changes dynamically during the course of a trial, as the subjects perform a cognitive operation in the context of a task [6, 7, 8]. The stimulus manifold may therefore be projected to a different subspace, rotated, or otherwise geometrically transformed [6, 7, 8].

Transformations of stimuli in state space have been thought to be specific for the cognitive operations being performed. Orthogonal rotation of a stimulus representation has been proposed as a possible mechanism for reducing interference between sensory and memory representations, protecting the memory of an initial stimulus from the interference of a subsequent stimulus presentation [9]. The same stimuli are represented in different subspaces when used in the context of different tasks [10] and errors are characterized by changes in stimulus representation geometry [11].

This evidence appears to link the geometric transformation of stimulus information in neuronal populations to cognitive operations [12], however, a critical assumption that task execution is causal to such transformations has not been tested experimentally. It is thus unknown how learning to perform a task in the first place changes the representation of stimuli. Here we test this idea by analyzing population of prefrontal neurons in subjects before and after they were trained to perform cognitive tasks involving identical stimuli. We checked multiple aspects of the activity space,

including subspace alignment, geometrical similarity, and dynamics, to determine the relative contribution of training and regional specificity in the formation of the observed low dimensional geometry.

RESULTS

Neurophysiological recordings were collected from the lateral PFC of six monkeys before and after they were trained to perform working memory tasks [13, 14]. Once fully trained, the monkeys viewed two stimuli appearing in sequence with intervening delay periods between them and reported whether the second stimulus was the same as the first (Fig. 1A-B). The stimulus sets used in these experiments varied in terms of spatial location or shape (Fig. 1A-B). Recordings were also obtained from these animals, viewing the same stimuli presented with the same timing prior to any training in the task. A total of 1164 neurons from six monkeys in five prefrontal subdivisions (Fig. 1C) were recorded during the passive, pre-training viewing of spatial stimuli; 847 neurons were recorded during the passive, pre-training viewing of feature stimuli. Additionally, 1031 neurons from three monkeys were obtained after training, when the animals performed actively the spatial working memory task; 793 neurons were obtained during the active feature working memory task.

Spatial stimuli are geometrically arranged in a low dimensional subspace

We applied Principal Component Analysis (PCA) on the covariance matrix of the zero-centered data, and selected eigenvectors sorted in decreasing order in terms of explained variance. In agreement with previous studies that have suggested that prefrontal neurons

represent stimuli in a low-dimensional manifold [4, 5], we found that the first three principal components generally explain more than 50% of the variance (Fig. S1) across brain regions and training phases, validating the approach of visualizing and measuring geometry structure in low dimensional space.

Previous studies also revealed that location information is arranged in a well-structured geometry resembling the physical appearance of visual stimuli in working memory [6, 15], however, it is not clear if this representation is learned through task training. Visualizing in the subspace defined by the first two principal components confirmed that spatial stimuli roughly fell on a ring with the same relative positioning in the low-dimensional subspace (Fig. 2A), and this geometry did not depend on training to perform a task. We calculated the unexplained variance between the empirical geometry and the 'standard' geometry that maximizes the distance between diametric locations, and found that spatial representations were well fitted both before and after training using this approach (Fig. 2B). Across prefrontal subdivisions, the mid-dorsal region had the best match for the hypothetical arrangement, indicated by the overall lowest unexplained variance.

Some subspace transformations are task independent

Low-dimensional stimulus representations in neuronal population activity are transformed during cognitive operations [7, 8, 9] and we therefore sought to test how such transformations differ systematically before and after training to perform a cognitive task. We quantified transformations based on the rotation of subspaces in different task epochs or task conditions.

Orthogonality of stimulus representations at different task intervals has been described previously in the context of behavioral tasks [9], and we sought to compare how training affects such transformations in prefrontal areas. Strikingly, representations of the same stimuli during the cue and match intervals already exhibited significant angles of rotation in multiple brain areas for both the spatial and feature sets before any task training (Fig. 3A-B). We also found that the pattern of transformation between the cue and match stimuli was highly area specific, that is, similar angles of plane rotation were seen across different prefrontal areas before and after training (Fig. 3B). The Pearson correlation coefficient between pre- and post-training rotation angles was $r=0.78$ ($p=0.023$, Fig. S2 left). To further confirm the subspace orthogonality observed in the angle measure, we calculated the variance accounted for (VAF) ratio between the cue and the match period. The results qualitatively agree with the angle measure (Fig. 3C), in that high angle measurements always correspond to low VAF. Once again, we found that the relative order of values across subdivisions matched between the pre- and post-training phases. This result indicates that populations of neurons in different subregions of the prefrontal cortex transform matching stimuli in a stereotypical fashion, independent of training and task execution.

Some representation transformations were only observed after training, for example involving the representation of stimuli in match and nonmatch conditions. In the spatial task, and specifically in the most selective, mid-dorsal region of PFC[16], the angle of rotation between the match and nonmatch representations (both during the second stimulus presentation period) changed considerably after training (Fig. 4A). Across areas, there was little correlation between the plane angles observed before and after training

across areas, and the correlation coefficient was r=-0.29 (p=0.491, Fig. S2 middle). The VAF ratio measurement agreed with angle measurement in this case too, evidenced by a major left shift of the distribution for the mid-dorsal region in the spatial task. A similar effect of training was observed for the subspace angle between the cue and the first delay period, with anterior subregions of the PFC exhibiting increased rotation for the spatial task, and posterior subdivisions exhibiting increased rotation for the feature task (Fig. S3).

Task dependent and independent representation geometry

In addition to subspace transformations, we also examined the within-subspace geometry. Two indexes were calculated to represent relative representational amplitude and structure similarity between task epochs or conditions: the unexplained variance, which quantified geometrical structure similarity after controlling the scale and orientation differences of representation in respective subspaces, and the scale factor, which was the ratio of within-subspace to total variance. We additionally used a within epoch measure as control (i.e. comparing the geometry of data from the same dataset to itself, see Methods). For the unexplained variance, the more similar the two conditions are compared to each other, the closer the measurement should be to 0; for the scale factor, more similar measurements correspond to values close to 1.

In the geometry similarity analysis, we found that the effect of training for the most selective subdivision, the mid-dorsal PFC, was to decrease the similarity between epochs. This was true for the comparison of the two stimulus presentation epochs (cue vs match): the spatial unexplained variance increased from a pre-training value of

0.097±0.012, to a post-training value of 0.25±0.01; the feature pre-training unexplained variance similarly increased from 0.11±0.003, to a post-training value of 0.24±0.014 (Fig. 3C). The same relationship held between the cue stimulus and first delay period epochs: the spatial pre-training unexplained variance increased from 0.1±0.003 to a post-training value of 0.22±0.007 (Fig. S3C).

The representation scale measured across the cue and sample epochs aggregated around 1, both before and after training (Fig. 3D and Fig. 4C). This means that the stimulus representation generally spanned roughly the same amount of area in the respective subspace, either when compared between the cue and sample (first and second stimulus presentation) period, or between match and nonmatch conditions within the same period, thus indicating a stable magnitude of population responses. For the cue vs. delay comparison, the scale factor measure fell below 1 (Fig. S3C) due in part to the fact that, even with persistent delay activity, the overall average firing rate in the delay is lower compared to the stimulus presentation epochs. The stability of the scale factor during the stimulus presentation periods suggests that on the population level, changes in the correlation structure induced by training did not result in an overall increase of stimulus-driven firing rate changes; whereas the amplitude of stimulus representation was held fairly constant.

Subspace rotation and dynamics correlates with behavior performance

We also compared the dynamic trajectory of population activity before and after training. As expected, pre-training activity was largely confined to a narrow subspace during the

delay period, while post-training activity became increasingly dynamic in an expanded space (Fig. S4A, B). To further quantify this change, we randomly sampled half of the trials from all periods and calculated its PCA space as the base to project the remaining half trials (see Methods). The mean ratios of the decoding space area between fixation, first delay and second delay period (using the fixation interval as reference) were 1 : 1.07 : 1.26 for the pre-training data and 1 : 8.62 : 1.83 for the post-training data (Fig. S4C-E).

We also sought to test whether the post-training population rotation in the match relative to the nonmatch condition was linked to improved task performance. We thus compared the subspace rotation angle in correct and error trials in a subset of neurons with sufficient error trials across conditions (n=295 in the spatial working memory task, and n=201 in the feature working memory task), pooling data across all areas (see Methods). In correct trials (Fig. 5A, C), match and nonmatch trials were nearly orthogonal to each other (79.9±5.4 and 78.0±14.5 degrees for the spatial and feature stimuli, respectively), consistent with the post-training findings of the full population (Fig. 3). In error trials, this rotation was significantly reduced (Fig. 5B, D,19.7±9.0 and 18.8±14.5 respectively). To determine how population responses evolved across the length of the trial, we investigated the dynamics of representations for match and nonmatch trials in the reduced PCA space. Example trajectories for a single stimulus condition are plotted for the mid-dorsal area in Fig. 5E. These, too, resembled the trajectories in state space described previously for neuronal activity in animals trained to perform behavioral tasks involving perceptual decisions [8]. In correct trials, as expected, match and nonmatch trajectories stayed close to each other until the sample period and then diverged during the sample presentation and the delay period that followed it. In

error trials, an abnormal rotation was already present in the cue period and match and nonmatch trajectories were less-distinguishable (Fig. S5). The result suggests that incomplete rotation of the stimulus subspace is more likely to result in errors. However, this failure of population activity to follow the desired trajectory is present early in the trial and does not depend specifically on orthogonalization of the nonmatch stimulus, which, as we showed here, was present even in naïve animals.

DISCUSSION

Computation in neural activity is characterized by the transformation of representations across stages of processing [12]. Our results demonstrate that neuronal populations perform stimulus transformations and exhibit dynamics that are currently thought to represent cognitive operations [5, 8, 9, 10], even in the absence of task execution, in animals naïve to any such training. Our results in the same dataset from trained animals corroborated previous findings: sequential cue and match stimuli from the same trials were represented in orthogonal planes; match and nonmatch stimuli exhibited large state space separations; and error trials exhibited smaller plane rotations than correct ones. These positive results only accentuate the finding that naïve animals, viewing passively the same stimuli, produced similar rotations as trained animals for the cue vs. match comparison. The result indicates that some dynamic transformations of population-based stimulus representations are not caused by task operations requiring explicit training.

Neural basis of cognitive flexibility

Natural plasticity allows for significant training-induced improvements of working memory, particularly through changes in PFC activity [17, 18, 19, 20, 21]. Representation of stimuli in neural activity has also been shown to be dynamic at much faster timescales, during the time course of a single behavioral session [22] or, depending on trial events, on a moment-by-moment basis [1]. The transformation of stimulus representations has previously been considered indicative of flexible task execution, e.g. rotating representations may minimize interference between multiple stimuli, thus allowing additional stimuli to be encoded with greater efficiency [9]. Our current results suggest that

the prefrontal cortex may perform these operations automatically, even in the absence of task training or execution. Furthermore, the scale of stimulus representation was largely unchanged after training. The idea of stable population information under shifting neural responses within the range of days to weeks is an increasingly appreciated phenomenon in studies involving post-learning plasiticity[23, 24]. Our results here showed a similar phenomenon after learning a new task.

Regardless of transformations, data across prefrontal areas and task conditions were generally well fit by low dimensional representations. Low dimensionality in working memory representations is considered a hallmark of generalizability, while high dimensionality is correlated with better discriminability [25]. In our experiment, stimuli were highly discriminable from each other from the outset, and task training involved incorporating them into the context of a new task rather than learning fine features of the stimuli [26], suggesting that an increase in dimensionality was not required to perform the task after training.

Regional Specialization

Prior studies have debated whether different types of information are represented across the dorsal-ventral and anterior-posterior axes of the prefrontal cortex has been contentious [27, 28, 29, 30]. Converging evidence, however, suggests that anterior subdivisions of the prefrontal cortex generally represent more abstract information, and their activation depends to a greater extent on the task subjects have been trained to perform [31]. Consistent with this idea, we observed that anterior areas consistently exhibited high rotation values for match and nonmatch stimuli relative to the cue (and relative to each

other), both before and after training. Our results raise the possibility that this innate stimulus transformation endows anterior areas of the prefrontal cortex with greater capacity for plasticity.

METHODS

Data obtained from six male rhesus monkeys (*Macaca mulatta*), ages 5–9 years old, weighing 5–12 kg, as previously documented [14], were analyzed in this study. None of these animals had any prior experimentation experience at the onset of our study. Monkeys were either single-housed or pair-housed in communal rooms with sensory interactions with other monkeys. All experimental procedures followed guidelines set by the U.S. Public Health Service Policy on Humane Care and Use of Laboratory Animals and the National Research Council's Guide for the Care and Use of Laboratory Animals and were reviewed and approved by the Wake Forest University Institutional Animal Care and Use Committee.

Monkeys sat with their heads fixed in a primate chair while viewing a monitor positioned 68 cm away from their eyes with dim ambient illumination and were required to fixate on a 0.2° white square appearing in the center of the screen. During each trial, the animals were required to maintain fixation on a 0.2° white square appearing in the center of the screen while visual stimuli were presented either at a peripheral location or over the fovea, in order to receive a liquid reward (typically fruit juice). Any break of fixation immediately terminated the trial and no reward was given. Eye position was monitored throughout the trial using a non-invasive, infrared eye position scanning system (model RK-716; ISCAN, Burlington, MA). The system achieved a < 0.3° resolution around the center of vision. Eye position was sampled at 240 Hz, digitized and recorded. The visual stimulus display, monitoring of eye position, and synchronization of stimuli with neurophysiological data was performed with in-house software implemented

in the MATLAB environment (Mathworks, Natick, MA), utilizing the Psychophysics Toolbox [32].

Pre-training task

Following a brief period of fixation training and acclimation to the stimuli, monkeys were required to fixate on a center position while stimuli were displayed on the screen. The stimuli shown in the pre-training, passive, spatial task consisted of white 2° squares, presented in one of nine possible locations arranged in a 3 × 3 grid with 10° of distance between adjacent stimuli. Only the eight peripheral locations are analyzed here, as the center location never appeared as a nonmatch. The stimuli shown in the pre-training passive feature task consisted of white 2° geometric shapes drawn from a set comprising a circle, a diamond, the letter H, the hashtag symbol, the plus sign, a square, a triangle, and an inverted Y-letter. These stimuli could be presented in one of the possible locations of the spatial grid.

The presentation began with a fixation interval of 1 s where only the fixation point was displayed, followed by 500 ms of stimulus presentation (referred to hereafter as cue), followed by a 1.5 s “delay” interval where, again, only the fixation point was displayed. A second stimulus (referred to hereafter as sample) was subsequently shown for 500 ms. In the spatial task, this second stimulus would be either identical in location to the initial stimulus, or diametrically opposite to the first stimulus. In the feature task, this second stimulus would appear in the same location as the initial stimulus and would either be an identical shape or the corresponding non-match shape (each shape was paired with one non-match shape). Only one nonmatch stimulus was paired with each cue, so

that the number of match and nonmatch trials were balanced in each set. In both the spatial and feature task, this second stimulus display was followed by another “delay” period of 1.5 s where only the fixation point was displayed. The location and identity of stimuli were of no behavioral relevance to the monkeys during the “pre-training” phase, as fixation was the only necessary action for obtaining rewards.

Post-training task

Three monkeys, data of which are analyzed here, were trained to perform active working memory tasks that involved the presentation of identical stimuli as the spatial and feature tasks during the “pre-training” phase. Monkeys were required to remember the spatial location and/or shape of the first presented stimulus, and report whether the second stimulus was identical to the first or not, via saccading to one of two target stimuli (green for matching stimuli, blue for non-matching). Each target stimulus could appear at one of two locations orthogonal to the cue/sample stimuli, pseudo-randomized in each trial. In the spatial task, the second stimulus could be either identical in location to the initial stimulus, or diametrically opposite to it. In the feature task, this second stimulus would appear in the same location as the initial stimulus and would either be an identical shape or the corresponding non-match shape.

Surgery and neurophysiology

A 20 mm diameter craniotomy was performed over the PFC and a recording cylinder was implanted over the site. The location of the cylinder was visualized through anatomical magnetic resonance imaging (MRI) and stereotaxic coordinates post-surgery. For two of

the four monkeys that were trained to complete active spatial and feature working memory tasks, the recording cylinder was moved after an initial round of recordings in the post-training phase to sample an additional surface of the PFC.

Anatomical localization

Each monkey underwent an MRI scan prior to neurophysiological recordings. Electrode penetrations were mapped onto the cortical surface. We identified six lateral PFC regions: a posterior-dorsal region that included area 8A, a mid-dorsal region that included areas 8B and 9/ 46, an anterior-dorsal region that included area 9 and area 46, a posterior-ventral region that included area 45, an anterior-ventral region that included area 47/12, and a frontopolar region that included area 10. However, the frontopolar region was not sampled sufficiently to be included in the present analyses.

Neuronal recordings

Neural recordings were carried out in the aforementioned areas of the PFC both before and after training in each WM task. Extracellular recordings were performed with multiple microelectrodes that were either glass- or epoxylite-coated tungsten, with a 250 μm diameter and 1–4 MΩ impedance at 1 kHz (Alpha-Omega Engineering, Nazareth, Israel). A Microdrive system (EPS drive, Alpha- Omega Engineering) advanced arrays of up to 8 microelectrodes, spaced 0.2–1.5 mm apart, through the dura and into the PFC. The signal from each electrode was amplified and band-pass filtered between 500 Hz and 8 kHz while being recorded with a modular data acquisition system (APM system, FHC, Bowdoin, ME). Waveforms that exceeded a user-defined threshold were sampled at 25 μs

resolution, digitized, and stored for offline analysis. Neurons were sampled in an unbiased fashion, collecting data from all units isolated from our electrodes, with no regard to the response properties of the isolated neurons. A semi-automated cluster analysis relied on the KlustaKwik algorithm, which applied principal component analysis of the waveforms to sort recorded spike waveforms into separate units. To ensure a stable firing rate in the analyzed recordings, we identified recordings in which a significant effect of trial sequence was evident at the baseline firing rate (ANOVA, $p < 0.05$), e.g., due to a neuron disappearing or appearing during a run, as we were collecting data from multiple electrodes. Data from these sessions were truncated so that analysis was only performed on a range of trials with stable firing rates. Less than 10% of neurons were corrected in this way. Identical data collection procedures, recording equipment, and spike sorting algorithms were used before and after training in order to prevent any analytical confounds.

Dimensionality reduction and rotation of subspaces

We applied principal components analysis (PCA) to visualize the neural population activity manifolds during the spatial and feature tasks. PCA was performed on the mean firing rate of neurons across different prefrontal regions, for data collected both before and after training. We examined the submanifolds of neural activity as a function of the spatial and feature stimulus sets.

In the spatial task, we collected data from five regions during five epochs (fixation, cue, delay1, sample, delay2) before and after the monkeys were trained to perform the task. In each region, trials were collected when the cue stimuli appeared at

$L = 8$ locations. We then selected the recordings with more than 16 trials at each location. For each neuron, the average firing rate of the match trials (and nonmatch trials) during a given period at the eight locations formed a column entry (8x1 vector) for the population activity matrix A_1 (A_2 for nonmatch trials). During each phase of the experiment (fixation, cue, delay1, sample, delay2) we defined the corresponding activity matrix A_1 (A_2) for each prefrontal region as an $8 \times N$ matrix, where 8 is the number of locations where the cue occurred and N is the number of neurons. To find the rotation between two periods, we aligned the activity matrix into one single matrix B, which is a $16 \times N$ matrix with the first eight rows containing the activities in one period and the rest containing for another period. Then we normalized B by subtracting the mean across each column to guarantee the matrix was zero-centered. The PCA was applied on the centered data using singular value decomposition (SVD). We selected the first three eigenvectors of the covariance matrix of the zero-centered data and sorted them by decreasing order in terms of explaining the variance. The first three principal components explained an average of 73% of the response matrix variance across all examined periods and locations. Similar procedures were applied to visualize the responses in the feature task, where we grouped the data according to the shape of cue stimulus. Specifically, there were eight different shapes in the feature task. Therefore, the corresponding activity matrix had a dimension of $8 \times N$, in which N is the number of neurons and each row stands for a shape in the feature task (see Figure1 B). For each pair of comparison for the rotations, we aligned the data in the same way as described above.

To visualize the population representations, we projected the centered activity matrix into a three-dimensional (3D) PCA space. After projection, the population

representations of the eight spatial locations (and eight shapes in the feature task) formed eight points in the reduced 3D space. The eight projected points seemed to lie on the same plane in both the spatial and feature tasks. Following a similar procedure, we further applied PCA on the eight points in the reduced plane, with a dimension of 8×3, to find its best-fit plane. The high proportion of variance explained by the fitted plane validated our hypothesis that the reduced space could be presented in 2D.

We defined the angles between different conditions based on the best-fit planes that we constructed above. Given the vectors spanning the best-fit plane (P_1) for one period are $\vec{v}_1$ and $\vec{v}_2$ (obtained from the PCs of the reduced eight points), and the best-fit plane (P_2) for another period in the same region are $\vec{v}_3$ and $\vec{v}_4$, the angle of between P_1 and P_2 was calculated as:

$$\langle P_1, P_2 \rangle = \cos^{-1}\left(\frac{|(\vec{v}_1 \times \vec{v}_2) \cdot (\vec{v}_3 \times \vec{v}_4)|}{|\vec{v}_1 \times \vec{v}_2| \cdot |\vec{v}_3 \times \vec{v}_4|}\right),$$

where "$\langle P_1, P_2 \rangle$" denotes the angle between P_1 and P_2, $\cos^{-1}$ is the inverse cosine function, and "$\vec{v}_1 \times \vec{v}_2$" is the cross product that finds the vector that is perpendicular to the plane spanned by $\vec{v}_1$ and $\vec{v}_2$, the "·" sign stands for the dot product, and "|x|" returns the absolute value (for a scaler) or length (for a vector) of x.

To distinguish between a true large rotation of stimulus representation and a spurious rotation coming about in a population of neurons exhibiting low selectivity and high variability, we created a control condition for every rotation-angle measurement. Specifically, we first individually fitted a Poisson distribution for each cell and each stimulus class (e.g. the circle stimulus in the shape working memory task), using the same number of trials as in the empirical data angle calculation. Then we randomly sampled the same number of trials from the simulated distribution. This controlled

dataset has the same level of variability and selectivity statistics, and thus provides a baseline rotation within each epoch. Only cells with at least 6 match and 6 nonmatch trials for each stimulus category were included in the rotation angle analysis. Variability was estimated by randomly drawing 80% of the cells from each prefrontal subdivision over 100 iterations.

Variance accounted for ratio (VAF)

To cross-check the rotation angle between the 2D subspace of different task epochs, we measured the Variance Accounted For (VAF) ratio for each angle measurement[15]. The VAF ratio for epoch subspace pair (1,2) was defined as

$$VAF_{1,2} = \frac{Var(v_2 v_2^T v_1 L_1^T)}{Var(v_1 L_1^T)}$$

Where v_i $(i = 1,2)$ is a 3×2 matrix representing two subspace axes, and L_1 represents the stimulus projection in the first subspace. VAF ranges from 0 to 1; large values indicate better alignment of subspaces, while values close to 0 suggest orthogonality. Only cells with at least 6 match and 6 nonmatch trials for each stimulus category were included in the VAF analysis.

Geometry of low-dimensional representation

A matrix L of size 8×2 was used to represent the projection of 8 locations or shapes on the 2D plane. To quantify the amplitude of stimuli representations in two conditions, we used a scale factor, which is simply the variance ratio between two geometrical arrangement calculated as follows

$$Scale\ factor = \frac{\sqrt{\sum_p \sum_q C_{second}^2(p,q)}}{\sqrt{\sum_p \sum_q C_{first}^2(p,q)}}$$

$$C_{(p,q)} = L_{(p,q)} - \bar{L}$$

Where C_{first} and C_{second} represent mean-subtracted projections on a 2D plane of the first and second condition in a comparison, respectively, and where p=1, … ,8 and q=1,2, representing two coordinates for eight shapes in a 2D subspace. To quantify the arrangement consistency between experimental conditions, and between experiment data and hypothetical predictions, we first used the Kabsch algorithm [33] to find the optimal rotation that aligns the low-dimensional representation in different subspaces. Then we calculated the covariance matrix $H = (Scale\ factor * C_1)^T C_2$, and determined the SVD of H as $H = UDV^T$, where D is a rectangular diagonal matrix with non-negative real values. Thereafter, the optimal rotation matrix R was calculated

$$R = V \begin{bmatrix} 1 & 0 & 0 \\ 0 & 1 & 0 \\ 0 & 0 & d \end{bmatrix} U^T$$

$$d = sign(\det(VU^T))$$

Lastly, the difference between two low-dimensional manifolds after controlling for scale was quantified as

$$Unexplained\ variance = \frac{2 \times \sqrt{\sum\sum(F_{first} - C_{second})^2}}{\sqrt{\sum_p \sum_q F_{first}^2(p,q)} + \sqrt{\sum_p \sum_q C_{second}^2(p,q)}}$$

Where F_{first} = R(Scale factor*C_{first}) represents the low-dimensional geometry of the first condition after scaling, translation and rotation.

When comparing the empirical geometry with a hypothetical geometry (Fig. 2), we assumed that the projection of eight stimuli should fall approximately on a ring, and that the distance between each cue and its nonmatch pair should be maximized. For the spatial task, the match-nonmatch structure is symmetric (e.g. when A is cue, B is nonmatch; and when B is cue A is nonmatch). These two assumptions should lead to a predicted arrangement similar to the physical appearance of the spatial stimuli, where the nonmatch pair always perfectly falls on two ends of a diagonal line.

Only cells with at least 6 match and 6 nonmatch trials for each stimulus category were included in the geometrical similarity analysis. Standard error was measured by random drawing 80% of cells in each subdivision in 100 iterations.

Time course of subspace rotation

In addition to visualizing the rotation between different phases of the WM task, we investigated the dynamics of presentations in the reduced PCA space. To construct the trajectory, we discretized the whole task with a timestep of $\Delta_t = 50\ ms$ and an interval of $l_t = 250\ ms$. For each timestep k, we recorded the mean firing rates for the match and nonmatch trials in the interval $[t_0 + (k-1)\Delta_t - 250, t_0 + (k-1)\Delta_t]$ as the entries for population activity matrix A_1 and A_2 (defined in the above section). Similar as the procedure of constructing the PCA space, we projected the neural responses for each condition into their first three principal components. To plot the trajectories, we started from 250 milliseconds after the start of fixation, i.e., set the initial time $t_0 = 250\ ms$. Moreover, we use the first three PCAs in the match trials during the cue period as the common basis to calculate the coordinates. In other words, we projected the activity

matrix into the same common subspace spanned by the PCA space of the match trials. Note that we can pick up any common subspaces to project the data, and the similar geometry relations in Figure 3 should hold for all bases. As illustrated in Figure 3, we explored the dynamics of representation for each location (shape) in the spatial (feature) task, where we plot every six points from the dynamics, i.e., we set an increment of $dt = 300\ ms$ for the purpose of visualization.

To further investigate the changes in decoding spaces, we randomly sampled half of the trials from all periods (fixation, cue, delay1, sample and delay2) and calculated its PCA space as the base to project the remaining half trials. The area of the projection for a given period indicates how much information the state space contains from all eight locations, which we defined as the decoding subspace ($\mathrm{S_t}$). Take the spatial task for example, in the reduced space formed by PC1 and PC2, the eight locations were presented by eight points. To quantify the size of decoding subspace, we first found the boundaries of the eight points, which was a set of points representing a single conforming 2-D boundary around the eight points (see Figure S4C and D). Then we calculated the areas of the polygons defined by the boundary points, which we defined as the areas of the decoding space ($\mathrm{A_t}$). We calculated the ratios of the areas of decoding spaces based on the fixation period (A_{fix}). For example (as in Figure S4C), the ratios (R) between fixation, delay1 and delay2 were defined as

$$R = \frac{\mathrm{A}_{fix}}{\mathrm{A}_{fix}} : \frac{\mathrm{A}_{delay1}}{\mathrm{A}_{fix}} : \frac{\mathrm{A}_{delay2}}{\mathrm{A}_{fix}}\ .$$

Note that we divided the area at fixation period to normalize the data so that we could compare the ratios from different samples. We repeated the random sampling across 10,000 iterations, and the statistics of the ratios were reported in the results section.

Significance testing of rotations with bootstrap

Our previous analysis was based on all data recorded from neurons that had more than 16 trials for all cue conditions and in each region in Figure 1C. To test the significance and robustness of the rotations in the representations, we applied a bootstrap method to sample 100 iterations from the original data. On each iteration, 80% of the cells were drawn with replacement. The angles were calculated following the same procedure as above. Mean angles from 100 iterations obtained from bootstrap plus/minus standard errors of the mean were superimposed as error bars in figures.

Analysis of error trials

In the post-training spatial task, most cells did not have an error trial. Therefore, the corresponding activity matrix (as defined above) for the error trials was very sparse. In the spatial task, there were 295 neurons where error trials presented at more than half of the eight locations, i.e., 295 rows of the population activity matrix had more than four non-empty entries. We selected those 295 neurons for the error trial analysis in the spatial task. Following the same criterion, we selected 201 neurons containing error trials in the feature task. We first aligned the activity matrix for both cases in the same order, i.e., the same neuron occupied the same row in both matrices. In most cases, for a given stimulus, there were more correct trials than error trials, which led to a larger number of samples in most entries of the activity matrix. Each entry of the activity matrix was calculated from the mean activity across all trials. We thus matched the number of trials in each entry to allow comparison between correct and error trials. Specifically, for each empty entry in

the error matrix, we removed the corresponding entry in the correct matrix. Additionally, we selected the same number (minimum of the corresponding entries) of trials in both cases to ensure the same number of samples in both the correct and error trials.

Dimensionality reduction was performed on the error trials using PCA. Note that there are several possible versions of PCA for a sparse matrix; we relied on a probabilistic principal component (PPCA) with variational Bayesian learning [34] as this method worked best for our neural data.

Acknowledgments

Research reported in this paper was supported by the National Eye Institute of the National Institutes of Health under award number R01 EY017077. We wish to thank Rye Jaffe, Junda Zhu, and Zhengyang Wang for helpful comments on the manuscript.

Author Contributions

C.C. conceived and designed the experiments. C.C. and X. L. Q. performed experiments. S.P, W.D., and C.C. performed data analysis. C.C., S. P. and W.D. wrote the manuscript with input from all authors.

Declaration of Interests.

The authors declare no competing interest

REFERENCES

1. Mante V, Sussillo D, Shenoy KV, Newsome WT. Context-dependent computation by recurrent dynamics in prefrontal cortex. *Nature* **503**, 78-84 (2013).

2. Rainer G, Asaad WF, Miller EK. Selective representation of relevant information by neurons in the primate prefrontal cortex. *Nature* **393**, 577-579 (1998).

3. Rigotti M*, et al.* The importance of mixed selectivity in complex cognitive tasks. *Nature* **497**, 585-590 (2013).

4. Ebitz RB, Hayden BY. The population doctrine in cognitive neuroscience. *Neuron*, (2021).

5. Cueva CJ*, et al.* Low-dimensional dynamics for working memory and time encoding. *Proc Natl Acad Sci U S A* **117**, 23021-23032 (2020).

6. Murray JD, Bernacchia A, Roy NA, Constantinidis C, Romo R, Wang XJ. Stable population coding for working memory coexists with heterogeneous neural dynamics in prefrontal cortex. *Proc Natl Acad Sci U S A* **114**, 394-399 (2017).

7. Tang C, Herikstad R, Parthasarathy A, Libedinsky C, Yen SC. Minimally dependent activity subspaces for working memory and motor preparation in the lateral prefrontal cortex. *Elife* **9**, (2020).

8. Okazawa G, Hatch CE, Mancoo A, Machens CK, Kiani R. Representational geometry of perceptual decisions in the monkey parietal cortex. *Cell* **184**, 3748-3761 e3718 (2021).

9. Libby A, Buschman TJ. Rotational dynamics reduce interference between sensory and memory representations. *Nat Neurosci* **24**, 715-726 (2021).

10. Minxha J, Adolphs R, Fusi S, Mamelak AN, Rutishauser U. Flexible recruitment of memory-based choice representations by the human medial frontal cortex. *Science* **368**, (2020).

11. Panichello MF, Buschman TJ. Shared mechanisms underlie the control of working memory and attention. *Nature* **592**, 601-605 (2021).

12. Kriegeskorte N, Wei XX. Neural tuning and representational geometry. *Nat Rev Neurosci* **22**, 703-718 (2021).

13. Meyer T, Qi XL, Stanford TR, Constantinidis C. Stimulus selectivity in dorsal and ventral prefrontal cortex after training in working memory tasks. *J Neurosci* **31**, 6266-6276 (2011).

14. Riley MR, Qi XL, Zhou X, Constantinidis C. Anterior-posterior gradient of plasticity in primate prefrontal cortex. *Nat Commun* **9**, 3790 (2018).

15. Xie Y*, et al.* Geometry of sequence working memory in macaque prefrontal cortex. *Science* **375**, 632-639 (2022).

16. Dang W, Jaffe RJ, Qi XL, Constantinidis C. Emergence of Nonlinear Mixed Selectivity in Prefrontal Cortex after Training. *J Neurosci* **41**, 7420-7434 (2021).

17. Dux PE, Tombu MN, Harrison S, Rogers BP, Tong F, Marois R. Training improves multitasking performance by increasing the speed of information processing in human prefrontal cortex. *Neuron* **63**, 127-138 (2009).

18. Klingberg T*, et al.* Computerized Training of Working Memory in Children with ADHD - a Randomized, Controlled Trial. *J Am Acad Child Adolesc Psychiatry* **44**, 177-186 (2005).

19. Jaeggi SM, Buschkuehl M, Jonides J, Perrig WJ. Improving fluid intelligence with training on working memory. *ProcNatlAcadSciUSA* **105**, 6829-6833 (2008).

20. Constantinidis C, Klingberg T. The neuroscience of working memory capacity and training. *Nat Rev Neurosci* **17**, 438-449 (2016).

21. Qi XL, Constantinidis C. Neural changes after training to perform cognitive tasks. *Behav Brain Res* **241**, 235-243 (2013).

22. Asaad WF, Rainer G, Miller EK. Neural activity in the primate prefrontal cortex during associative learning. *Neuron* **21**, 1399-1407 (1998).

23. Rule ME, Loback AR, Raman DV, Driscoll LN, Harvey CD, O'Leary T. Stable task information from an unstable neural population. *Elife* **9**, (2020).

24. Rule ME, O'Leary T. Self-healing codes: How stable neural populations can track continually reconfiguring neural representations. *Proc Natl Acad Sci U S A* **119**, (2022).

25. Bernardi S, Benna MK, Rigotti M, Munuera J, Fusi S, Salzman CD. The Geometry of Abstraction in the Hippocampus and Prefrontal Cortex. *Cell* **183**, 954-967 e921 (2020).

26. Meyers EM, Qi XL, Constantinidis C. Incorporation of new information into prefrontal cortical activity after learning working memory tasks. *Proc Natl Acad Sci U S A* **109**, 4651-4656 (2012).

27. Wilson FA, Scalaidhe SP, Goldman-Rakic PS. Dissociation of object and spatial processing domains in primate prefrontal cortex. *Science* **260**, 1955-1958 (1993).

28. Badre D, D'Esposito M. Is the rostro-caudal axis of the frontal lobe hierarchical? *Nat Rev Neurosci* **10**, 659-669 (2009).

29. Rao SC, Rainer G, Miller EK. Integration of what and where in the primate prefrontal cortex. *Science* **276**, 821-824 (1997).

30. Owen AM, Stern CE, Look RB, Tracey I, Rosen BR, Petrides M. Functional organization of spatial and nonspatial working memory processing within the human lateral frontal cortex. *Proc Natl Acad Sci U S A* **95**, 7721-7726 (1998).

31. Constantinidis C, Qi XL. Representation of Spatial and Feature Information in the Monkey Dorsal and Ventral Prefrontal Cortex. *Front Integr Neurosci* **12**, 31 (2018).

32. Meyer T, Constantinidis C. A software solution for the control of visual behavioral experimentation. *J Neurosci Methods* **142**, 27-34. (2005).

33. Kabsch W. A solution for the best rotation to relate two sets of vectors. *Acta Crystallogr* **A32**, 922 (1976).

34. Ilin A, Raiko T. Practical Approaches to Principal Component Analysis in the Presence of Missing Values. *Journal of Machine Learning Research*, 1957-2000 (2010).

FIGURES

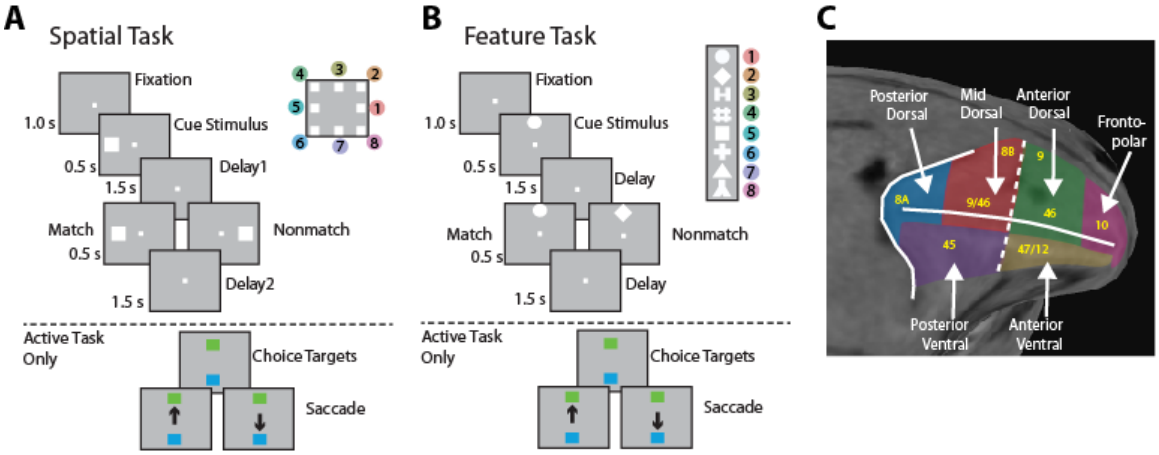


Figure 1. **Task structure and recording areas**. The animals were required to maintain center fixation throughout both active and passive task trials. At the end of active tasks trials, monkeys were required to make a saccade to a green target if the stimuli matched or to a blue target if the stimuli did not match. (A) Spatial match-to-sample task; the eight cue locations analyzed are shown in the inset. (B) Shape feature match-to-sample task; eight possible shapes in a session shown are shown in the inset. (C). Neurons were recorded from 5 subdivisions of lateral PFC, which we refer to as Anterior Dorsal (AD), Anterior Ventral (AV), Mid-Dorsal (MD), Posterior Dorsal (PD), and Posterior Ventral (PV).

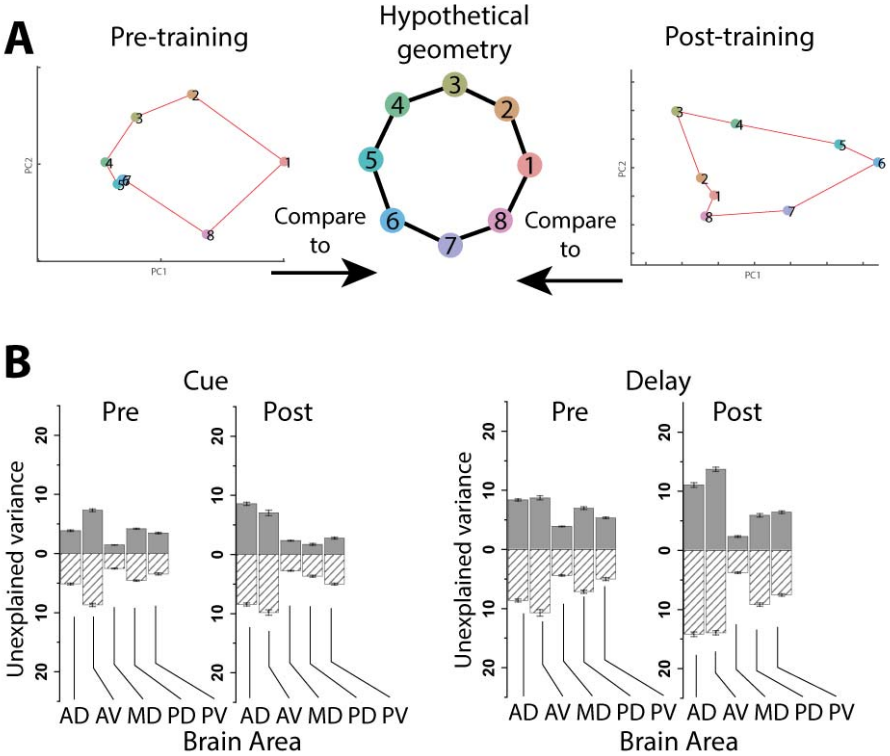


Figure 2. **Comparison of within-subspace geometry for both tasks.** (A) A graphical summary of how representational geometry was compared. Geometries were calculated from cue period activity of the spatial task, pre- and post-training. The geometries were compared to a hypothetical geometry that maximizes the distance between match and nonmatch stimulus pairs (see Methods). (B) Geometrical similarity to the hypothetical configuration for both tasks in the cue and the first delay period. Filled bars represent empirical data, and dashed bars represent similarity measure after shuffling shape labels as a control. The bar height represents deviation from both the circle configuration, and from the 'standard' order of stimuli on the circle. A smaller difference between the filled and dashed bars shows a higher similarity between the representational geometry and the 'standard' hypothetical geometry. The y-axis units are arbitrary.

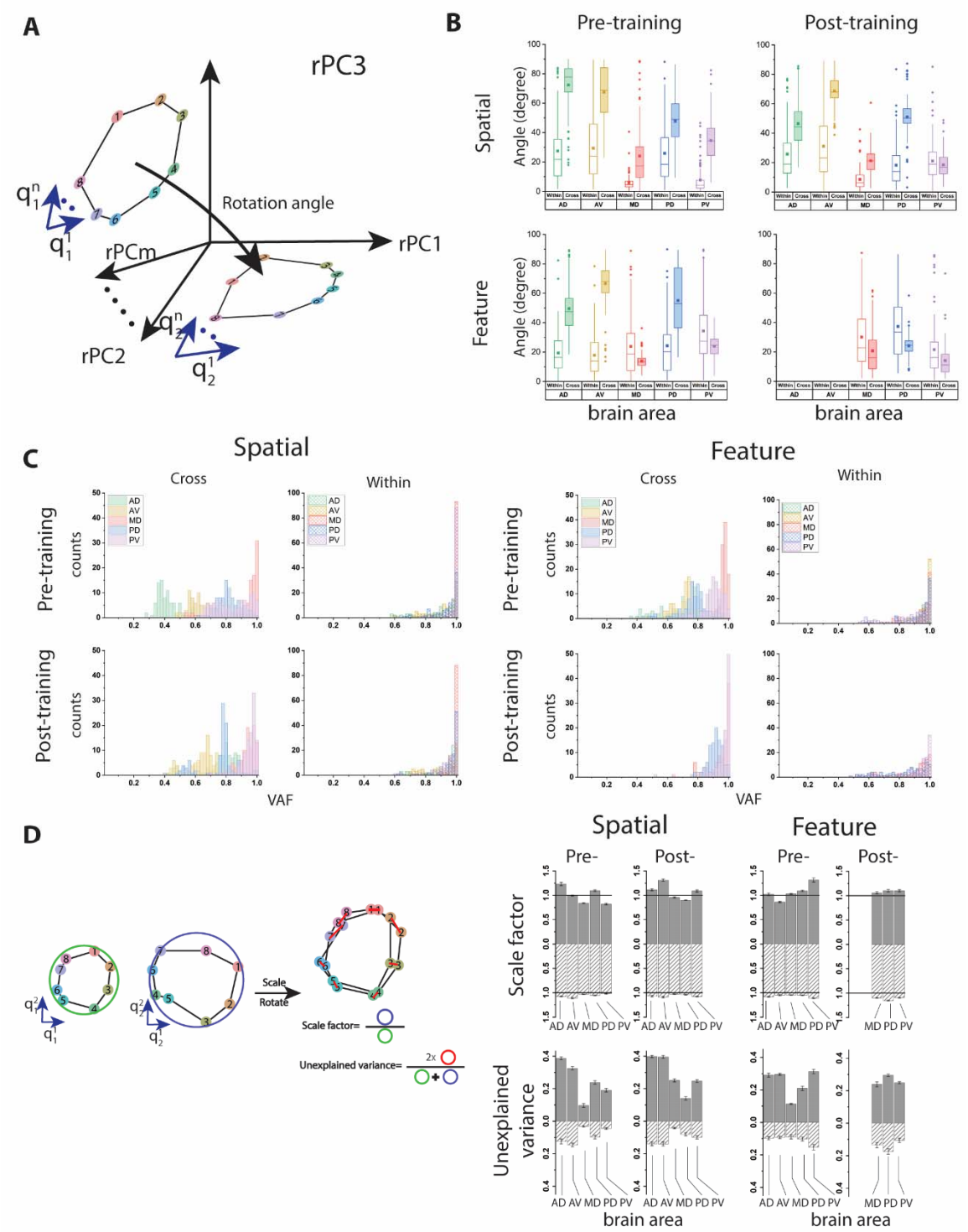


Figure 3. Subspace rotation between the cue period and the sample period in match trials. (A) Illustration of subspace rotations. Average neural response of different conditions were projected onto a 3D PCA-space, where the rotation and geometrical similarity were calculated (see Methods for details). (B) Filled boxes: angles between the cue and the sample-match condition in different PFC subdivisions for the spatial and the feature task. Unfilled boxes: angles from the control experiments obtained by randomly

drawing data from a distribution fitted to the empirical data for each task epoch. The dot and horizontal line in the box represent the mean and median, respectively. Whiskers are five percentile minimum and maximum values. (C) Variance accounted for (VAF) between the cue and the sample-match condition in difference PFC subdivisions for the spatial and the feature task**.** (D) Left: a graphic summary of subspace geometrical similarity. The scale factor was the ratio between within-subspace total variance. The unexplained variance quantifies geometrical structure similarity after controlling the scale and orientation differences of representation in respective subspaces. Right: Geometrical similarity between cue and sample-match. Gray boxes represent results from the cue-match dataset, boxes with oblique lines are results from the within cue period control group.

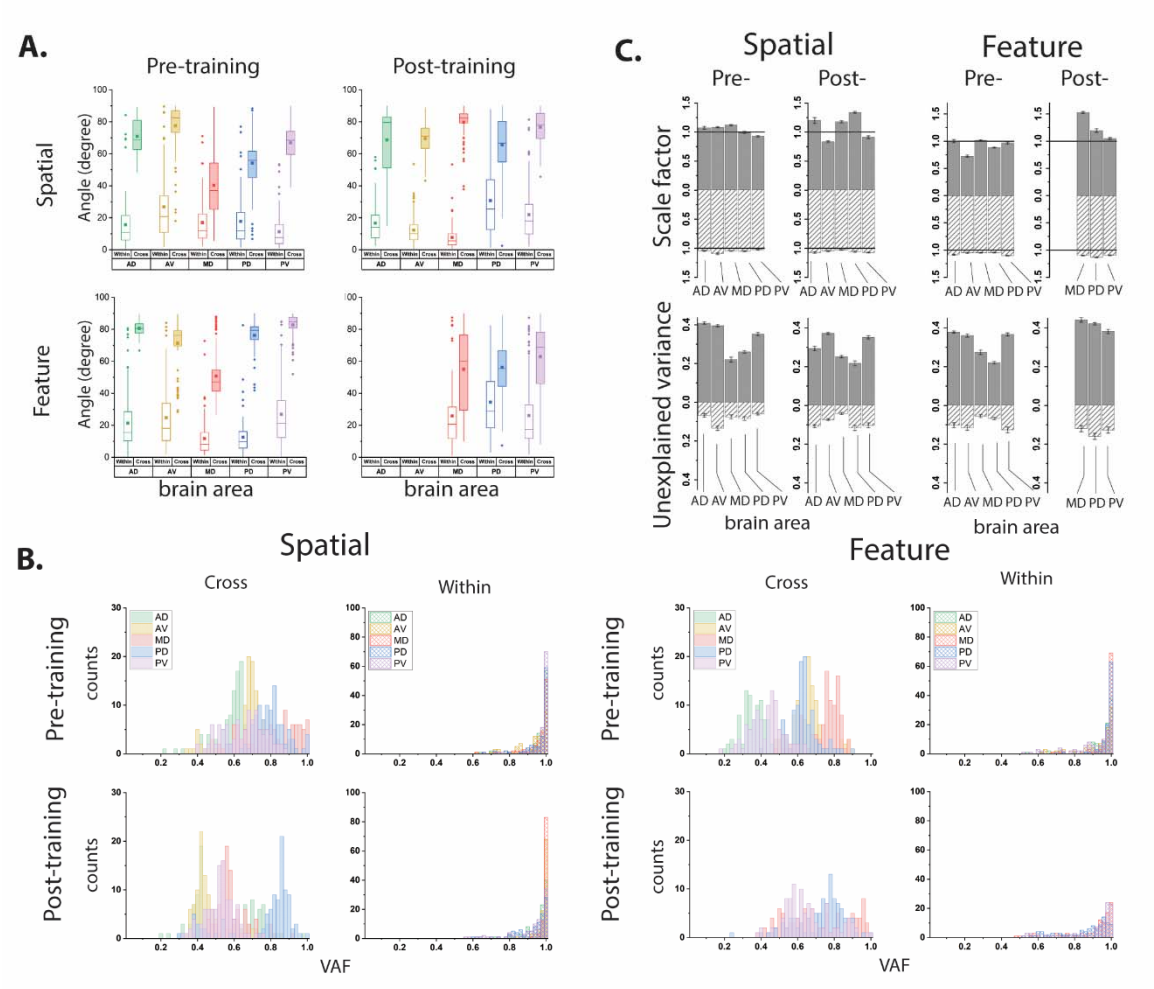


Figure 4. Subspace rotation between the match and nonmatch conditions in the sample period. (A) Primary angles between the match and non-match in difference PFC subdivisions for the spatial and the feature task. Primary angles between the cue and the sample-match condition in different PFC subdivisions for the spatial and the feature task. As a control we randomly draw data from a distribution fitting to empirical data from a particular task epoch (unfilled boxes). In this figure control is within the match condition of the sample period. (B) Variance accounted for (VAF) between the match and non-match conditions in difference PFC subdivisions for the spatial and the feature task**.** (C) Geometrical similarity between match and nonmatch in the sample period. Gray boxes represent results from match-nonmatch dataset, while boxes with oblique lines are results from the within match period control group.

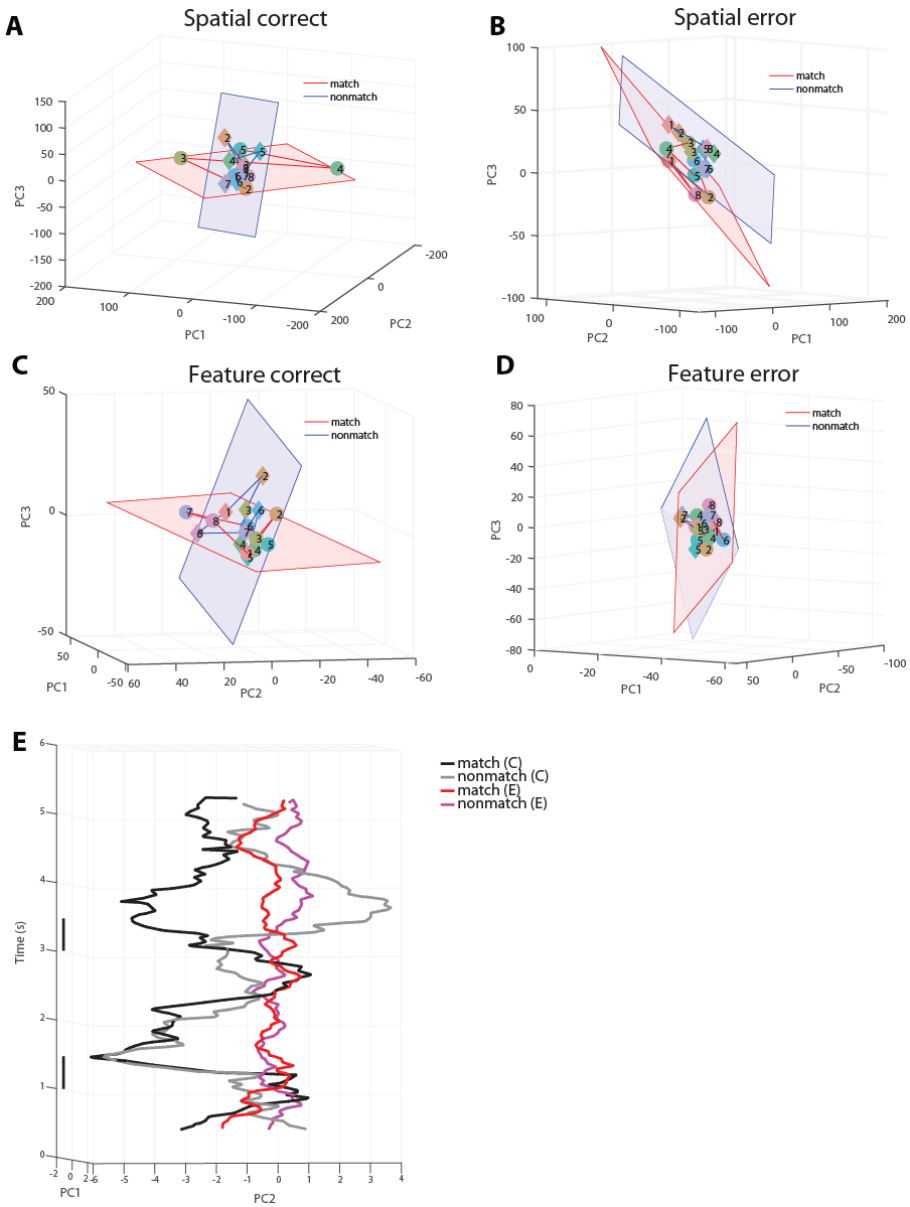
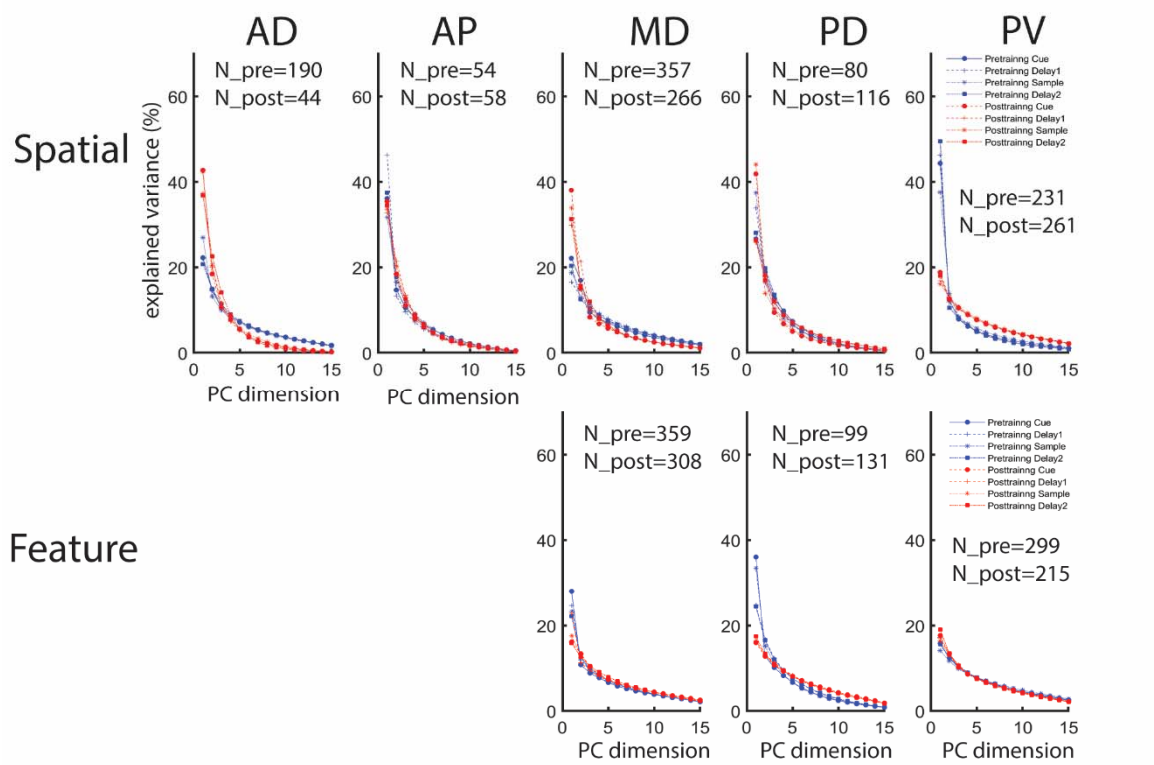
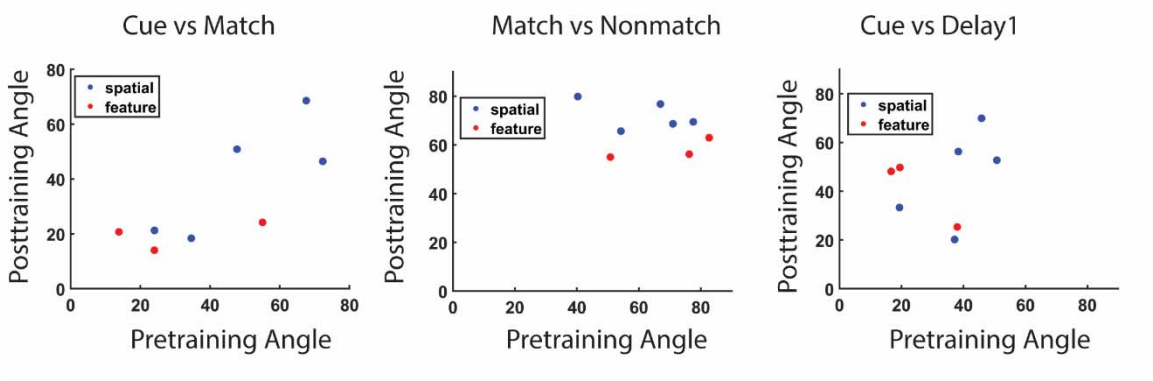


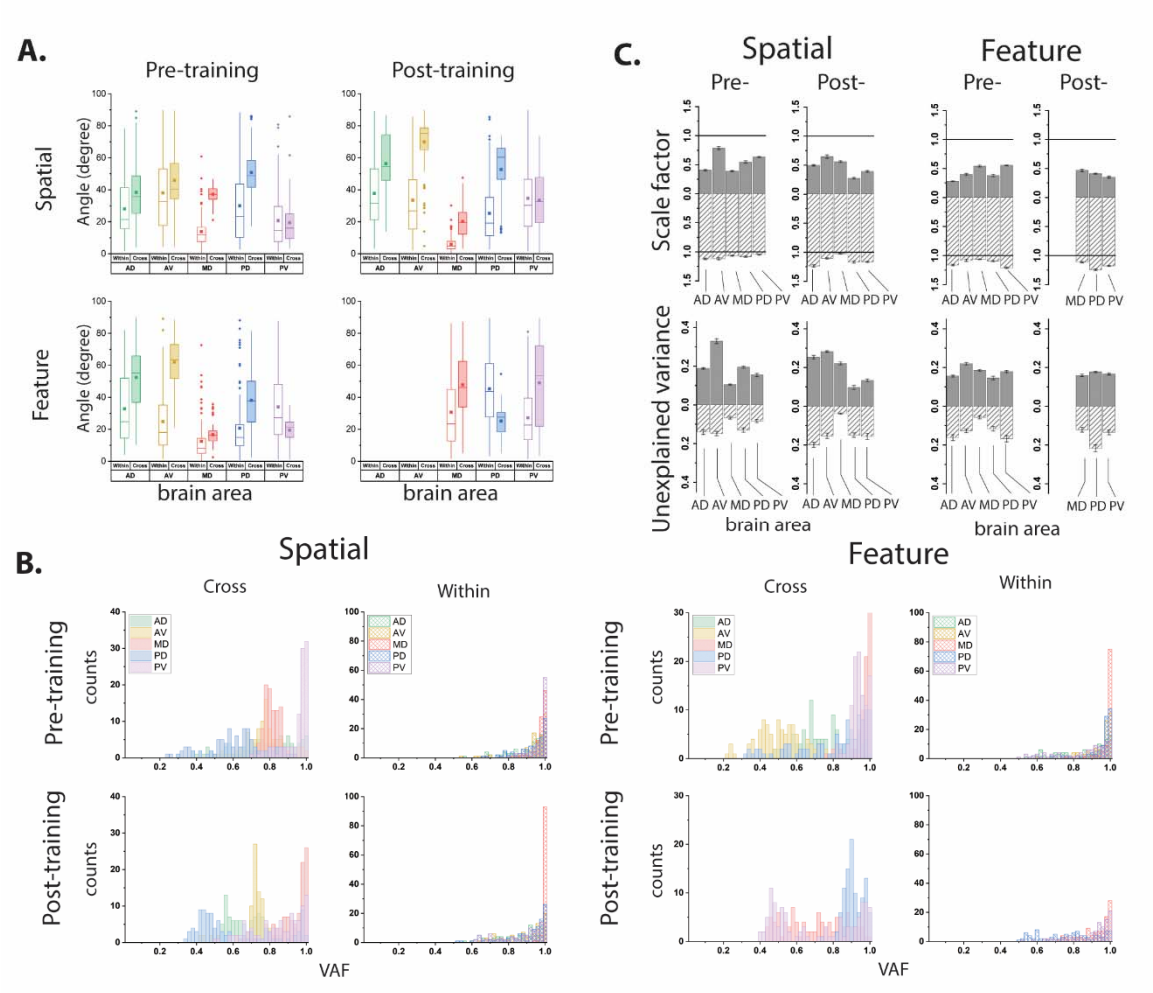
Figure 5. Population representations in correct and error trials. (A) Representation of spatial stimuli during the match and nonmatch period in correct trials (n=295 neurons). Rotation φ=76°. (B) Representation of spatial stimuli during the match and nonmatch period in the error trials (same population as in panel A). Rotation φ=15°. (C) Representation of feature stimuli during the match and nonmatch period in correct trials (n=201 neurons). Rotation φ=84°. (D) Representation of feature stimuli during the match and nonmatch period in error trials (same population as in panel C). Rotation φ=7°. (E) Dynamics of stimulus representations in correct and error trials, for one example spatial location. Solid black: match in correct trials, solid gray: nonmatch in correct trials, dashed red: match in error trials, dashed magenta: nonmatch in error trials. Always same cells are plotted (n=295 neurons).



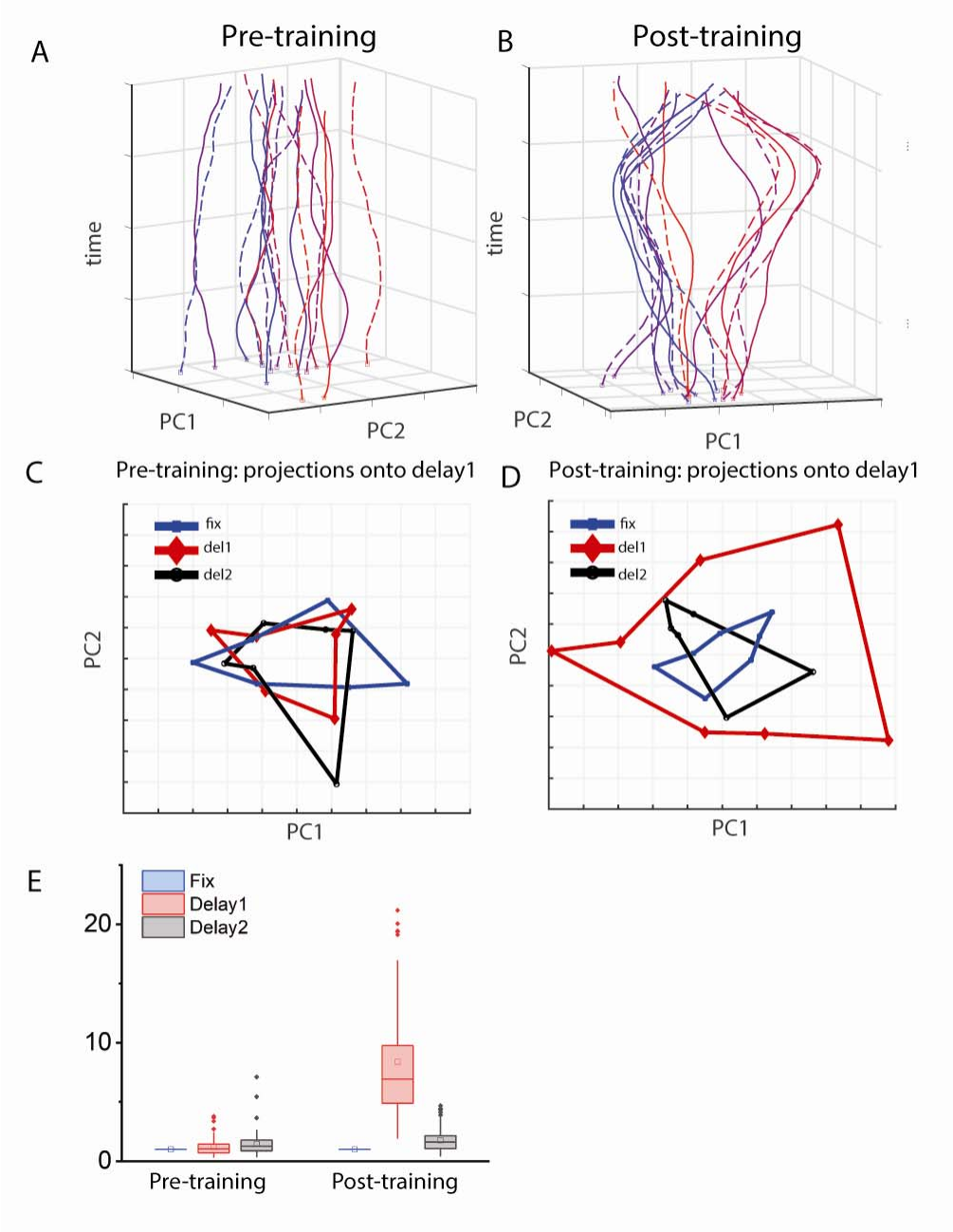
Supplemental Figure S1. Explained variance distribution for both tasks pre- and post-training in the PCA analysis. In general, fewer dimensions were needed for characterizing population activity in the spatial task, but under most conditions, three dimensions were enough to explain more than 50% of the variance. The training process has a heterogenous effect on the distribution of variance in different subdivisions of LFPC.



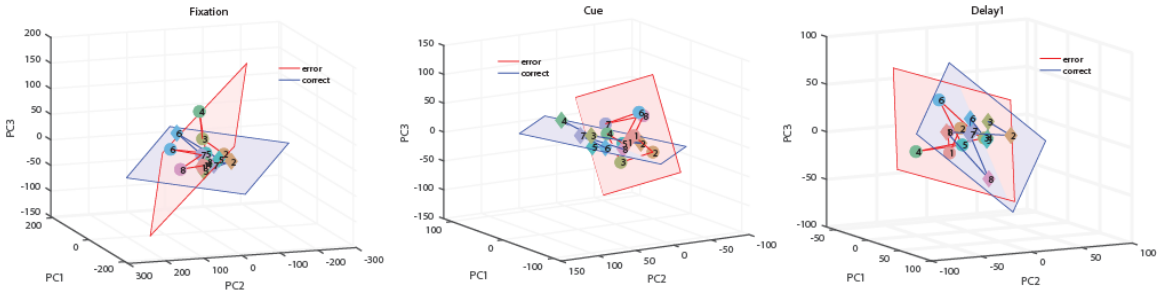
Supplemental Figure S2. Scatter plot of subspace rotation angle for the spatial and the feature task. Each dot represents the mean angle measure for one PFC subdivision, blue color dots indicate spatial task, and red dots indicate feature task.



Supplemental Figure S3. Subspace rotation between the cue and the first delay. (A) Filled boxes: primary angles between the cue and the first delay in difference PFC subdivisions for the spatial and the feature task. Unfilled boxes: control group obtained by randomly drawing data from a distribution fitting to empirical data from a particular task epoch. In this figure control is within the first delay period. (B) Variance accounted for (VAF) between the cue and the first delay in difference PFC subdivisions for the spatial and the feature task**.** (C) Geometrical similarity between the cue and the firs delay period. Gray boxes: results from cue-delay dataset; boxes with oblique lines: results from the within first delay control group.



Supplemental Figure S4. Complete population activity trajectories before and after training. (A) Trajectory of population representation in state space prior to training, for the spatial stimuli in the mid dorsal area (n=453). Solid lines: match trials, dashed lines: nonmatch trials. **(B)** As in A, for neuronal activity in the mid-dorsal area after training (n=374). **(C)** Area covered by projection of spatial stimuli prior to training, in three task epochs: fixation, delay1, and delay2. **(D)** As in C, for data obtained after training. **(E)**. Boxplot of surface areas: summarizing statistics of panels C and D with bootstrapping sampling, where each sample contained 80% of cells. Area values have been normalized by that of the fixation period.



Supplemental Figure S5. Population representations in correct and error trials. (A) Representation of spatial stimuli during the fixation period in correct and error trials (n=295 neurons). Rotation φ=45°. (B). As in panel A, for the cue period. Rotation φ=81°. (C) As in panel A for the first delay period. Rotation φ=82°.