

Bump attractor dynamics in prefrontal cortex explains behavioral precision in spatial working memory

Klaus Wimmer¹, Duane Q. Nykamp^{1,2}, Christos Constantinidis³ & Albert Compte¹

¹Institut d'Investigacions Biomèdiques August Pi i Sunyer (IDIBAPS), Barcelona, Spain

²School of Mathematics, University of Minnesota, Minneapolis MN

³Wake Forest School of Medicine, Winston-Salem NC

Keywords: Persistent activity, saccade, neural code, correlations, short-term memory

Corresponding author:

Albert Compte
Institut d'Investigacions Biomèdiques August Pi i Sunyer (IDIBAPS)
C/ Rosselló 149
08036 Barcelona
Spain
Fax: +34 93 3129408
Tel: +34 93 2275400 ext. 4151
E-mail: acompte@clinic.ub.es

Abstract

Prefrontal persistent activity during the delay of spatial working memory tasks is thought to maintain spatial location in memory. A “bump attractor” computational model can account for this physiology and its relation to behavior. However, direct experimental evidence linking parameters of prefrontal firing to the memory report in individual trials is lacking, and to date no demonstration exists that bump attractor dynamics underlies spatial working memory. Here, we demonstrate in the monkey model-derived predictive relationships between the variability of prefrontal activity in the delay and the fine details of recalled spatial location, evident in trial-to-trial imprecise oculomotor responses. The results support a diffusing bump representation for spatial working memory instantiated in persistent prefrontal activity. The findings reinforce persistent activity as a basis for spatial working memory, provide evidence for a continuous prefrontal representation of memorized space, and offer experimental support for bump attractor dynamics mediating cognitive tasks in the cortex.

The neural basis of spatial working memory has been studied extensively with oculomotor delayed response tasks in awake behaving monkeys¹⁻³. Experiments report persistent neural activity in prefrontal cortex (PFC) during the delay period after a spatial stimulus and before the memory-guided saccadic response. Notably, this neural activity is selective to the location of the visual cues, so that persistent activity can be described as a function of stimulus position using a bell-shaped tuning curve². These observations suggest that a continuous spatial representation in PFC, persisting during the delay period, could underlie spatial working memory. However, assumptions necessary to link this neuronal code with behavior have yet to be validated by experimental data.

Continuous, persistent population codes emerge naturally in computational models of the cortical microcircuit, typically by combining local recurrent excitation and broader feedback inhibition⁴⁻⁸. Such network models can display a bell-shaped pattern of activity (*bump*) even in the absence of tuned external input, called bump attractor because the network structure causes neural activity to be naturally attracted toward the bump state. The bump attractor has been proposed to be the network substrate underlying spatial working memory because of two main features. First, the self-sustained attractor in the absence of tuned external input is precisely the condition required for tuned persistent activity during the stimulus-devoid delay period. Second, the center of the bump can be located at any continuously varying location across the network; hence, the bump location provides a substrate for encoding a continuous variable such as spatial location.

In the absence of direct experimental evidence for this model, it has been a matter of debate whether activity in PFC encodes location in a continuous or in a categorical fashion⁹⁻¹¹, or even if working memory depends on a neural code based on persistent activity at all, or alternatively it is better described by some slow transient dynamics¹²⁻¹⁶. Consistent with a continuous code, behavioral data shows a progressive spatial spread of inaccurate saccadic reports as the delay period is extended in spatial working memory tasks^{2,17,18}. However, the possibility remains that PFC develops a discrete categorical representation after repeated training with a small number of visual cues, and a different mechanism related to elapsed time, not stimulus identity, is responsible for delay-dependent behavioral inaccuracies.

A unique feature of the bump attractor model for spatial working memory is that it predicts a particular relationship between neural activity and behavioral inaccuracies^{2,17,18}. Since the location of the bump of

activity can vary continuously along the network, the bump is not constrained to remain centered at the same location in the absence of tuned external input during the delay period. Hence, the lack of sensory guidance during the delay makes the encoding vulnerable to random activity fluctuations. Even though the shape of the bump is preserved by the network structure, the location of the center of the bump diffuses along the network^{8,9,19-21}. This feature of the model has two important consequences: drifts of bump activity in PFC should be predictive of memory-dependent inaccuracies in the behavioral report of the remembered location, and this PFC neural dynamics should be reflected in a specific pattern of pairwise neuronal correlations^{6,7,20}.

Here, we test these predictions to examine the hypothesis that the bump attractor model is the neural substrate in PFC for spatial working memory. We scrutinized an experimental dataset that includes single-neuron recordings, simultaneous neuron pair recordings, and detailed behavioral data while monkeys performed a visuo-spatial delayed response task³. We tested four specific predictions of the bump attractor model. First, tuning curves computed based on subsets of trials with disparate saccadic end points should show a bias at the end of the delay period, before the monkey makes its response. Secondly, firing rate deviations should correlate positively with behavioral responses towards the neuron's preferred location, and this correlation should develop progressively through the delay. Thirdly, neuronal variability right before saccade initiation should be higher for inaccurate than for accurate saccades. And finally, neuron pairs should be negatively correlated at the end of the delay, specifically for stimuli appearing between the two preferred locations^{6,7,20}. The results in each case support the hypothesis that a bump attractor representation in the PFC maintains spatial information during working memory.

Results

We trained two rhesus monkeys in an oculomotor spatial working memory task requiring them to remember the spatial location of a brief (0.5 s) visual stimulus and execute a saccade toward the remembered location after a delay period of 3 seconds³. By tracking eye position, we collected a behavioral dataset consisting of the coordinates of the saccadic end point for each successful trial in which the saccade correctly reported the approximate location of the memorized cue (**Fig. 1a**, Online Methods). For each monkey, we computed the mean saccadic end point for each of eight cues presented. We computed a behavioral measure of accuracy as the angular distance on the screen from each trial's saccadic end point to the mean saccadic end point for the given cue. This measure classifies trials into clockwise (CW) and counterclockwise (CCW) trials (**Fig. 1a,b**).

While the monkeys performed the task, we collected single unit responses from the dorsolateral PFC. A significant fraction of neurons in this area show tuned persistent delay activity during the mnemonic phases of the task^{2,3}, and we selected these neurons for our further analyses ($n = 204$, **Fig. 1c**). We calculated memory tuning curves by averaging the delay period activity of each neuron across trials as a function of the 8 equally spaced cues, and determined a preferred cue angle (Online Methods). By aligning tuning curves by their preferred cue and averaging, we formed the population tuning curve (**Fig. 1d**).

Despite these general attributes of prefrontal delay activity during this task, neural responses were highly heterogeneous²². Mean firing rates during the delay period varied broadly across the population (**Fig. 1e**). Moreover, the degree of delay tuning to the cue varied widely among neurons. To quantify this, we calculated the tuning strength T for each neural tuning curve ($T = 0$ for non-tuned responses, $T = 1$ indicates response to one single cue, Online Methods). We found T to be broadly distributed (**Fig. 1e**). Mean firing rate and tuning strength were correlated across neurons (tuning curves for high T and low T neurons in **Fig. 1d**), with no clustering suggesting separate functional populations (**Fig. 1e**). Some neurons showed dynamics in their delay firing rate (**Fig. 1f**) with 30/204 (39/204) cells having significantly higher (lower) activity for preferred cues in the last compared to the first second of the delay (Wilcoxon test, $P < 0.05$). However, a majority of neurons (135/204) did not show a significant activity change, and this was also the result for the population (Wilcoxon test, $P > 0.5$, **Fig. 1f**). Fano factors decreased slightly from the first to the last half of the delay period, from 1.29 ± 0.02 to

1.24 ± 0.03 (Wilcoxon test, $z = 3.6$, $P < 0.001$, $n = 196$). However, the tuning of individual neurons was remarkably consistent through the delay period (**Fig. 1g**): preferred cue angles and tuning strength T did not change significantly between the first and last seconds of the delay (preferred angle: Harrison-Kanji test, $\chi^2 = 0.38$, $P = 0.8$; tuning T : Wilcoxon test, $z = -0.83$, $P = 0.4$, $n = 204$, **Fig. 1g**). The steadiness of the coding properties through the delay suggests that PFC activity can be described by a bump code, which could be responsible for behavioral performance during the task.

With this data, we tested the *bump attractor hypothesis*, the idea that trial-averaged memory tuning curves reflect a hill of population activity of invariant shape (*bump attractor*) that encodes during the delay period the information that will determine the upcoming saccade (**Fig. 2a**). The center of mass of the bump attractor can be at any position along a continuum and encodes a continuous representation of the visual cue's position. Random fluctuations in bump attractor position^{6-8,21} lead to deviations in the read-out at the end of the delay, which result in inaccurate behavioral responses (**Fig. 2b,c**). This mechanism produces specific predictions regarding the trial-to-trial relationship between variability in neural activity and variability in behavioral responses that we tested in the data.

Tuning-curve bias in the delay predicts behavioral biases

According to our hypothesis, population activity displacements at the end of the delay underlie behavioral response deviations. These displacements of population activity should be reflected in a systematic bias of delay tuning curves derived from the sets of trials that led to CW and CCW deviations (**Supplementary Fig. 1**). For each neuron, we separated CW and CCW trials for each cue condition (**Fig. 3a**) and computed the corresponding CW and CCW tuning curves (**Fig. 3b**) as the corresponding trial-averaged firing rate versus the 8 angles of the cue location. The *tuning bias* was defined as the signed angular distance from the CCW to the CW tuning curve centers (Online Methods). With this definition, our hypothesis predicts that the tuning bias should become positive during the delay (**Supplementary Fig. 1**). We computed tuning biases for all neurons in different time windows along the trial, and combined them to obtain the time evolution of the population tuning bias. Consistent with the bump attractor hypothesis, the population tuning bias became significantly positive at the end of the delay (tuning bias $4.4^\circ \pm 2.9^\circ$ in the last second of delay, one-sided permutation test $P = 0.024$, $n = 204$), right before the behavioral response (**Fig. 3c**). To test for a possible motor origin of this signal we repeated the analysis excluding neurons with increasing rates in the delay period (positive

modulation index, **Fig. 1f**), which have been shown to represent saccade preparation²³. We still found significantly positive tuning bias in the last second of the delay (tuning bias $9.9^\circ \pm 6.6^\circ$, one-sided permutation test $P = 0.014$, $n = 101$), thus excluding a driving role for saccade preparation neurons in generating the tuning bias during the delay.

In addition, we found a quantitative agreement between the mean tuning bias computed from our 204 neurons and the mean behavioral deviation computed as the difference between the average saccade end points of the corresponding CCW and CW trials (mean tuning bias $4.4^\circ \pm 2.9^\circ$, mean behavioral deviation $7^\circ \pm 0.2^\circ$, Welch's test $t = 0.9$, $P = 0.36$, $n = 201$). This order-of-magnitude match shows that the bump attractor hypothesis in PFC can account for the magnitude of behavioral inaccuracies observed experimentally.

Correlation between delay activity and behavioral deviations

According to the bump attractor model, not only the average tuning should be related to dichotomized behavior (CW–CCW), but firing rates on a trial-by-trial basis should correlate with parametric deviations in behavioral response. In particular, a neuron increases its activity as the activity bump moves closer to its preferred location. As a result, trials for which a given neuron had stronger delay responses should result in behavioral deviations towards that neuron's preferred location. Thus, we expected a positive correlation between firing rate and behavior attraction *to the neuron's preferred location*. This effect should be especially strong for neurons with strong tuning and for cues at the tuning curve flanks (i.e., cues 1-2 positions from preferred), where responses are most sensitive to small variations in bump location.

We selected for each neuron the trials with stimuli in its tuning curve flanks (**Fig. 4a**) and matched the neuron's responses with the corresponding behavioral deviation (**Fig. 1b**). Defining behavioral deviations to be positive (negative) for saccades closer to (further from) the neuron's preferred location (color coded in **Fig. 4a**), we calculated the correlation coefficient between response deviation and behavioral deviation for each cell (**Fig. 4b**). We found that the population average of these correlations became significantly positive during the delay period (**Fig. 4c**), especially for neurons with stronger tuning T (**Fig. 4d**). This effect persisted when removing neurons with ramping-up delay activity from the analysis (correlation in last second of delay 0.029 ± 0.016 , $P = 0.041$, $n = 101$, one-sided

permutation test), thus excluding saccade preparation as the cause of this signal.

This positive correlation accrued with time into delay (**Fig. 4c**), suggesting that behavioral deviations result from accumulated errors in prefrontal activity during the delay, as in our hypothesis (**Fig. 2**). We confirmed this by looking separately at cues at different distances from the preferred location (**Fig. 5**). If the bump diffuses during the delay (**Fig. 2**), the correlation between firing rate and behavior should appear earlier in trials in which the cue was presented closer to the cell's preferred location. This occurs because it takes more time for the bump to diffuse and modulate neurons with preferred locations more distant from the cue. Indeed, periods of significant correlation between neuronal and behavioral variability appeared later in the delay as we took flank cues more distant from preferred location (**Fig. 5**).

The trial-to-trial relationship between firing rate and behavior should be restricted to neurons participating in the bump. To test this, we investigated neurons without significant delay tuning (*non-tuned neurons*, $n = 523$, Online Methods and **Supplementary Fig. 2**). We found no significant correlation between responses to flank stimuli and behavioral deviations for this dataset (**Fig. 4c,d**). However, this analysis required computing the preferred cue of each neuron, which probably suffered large estimation errors for such weakly tuned neurons. In addition, alignment to a preferred location could mask a rate-behavior relationship not related to the bump attractor hypothesis in these neurons. We therefore performed an additional analysis that did not assume a specific relationship between a neuron's tuning curve and behavioral deviations. For each neuron and each cue, we calculated the R^2 value of the linear regression between rate at the end of the delay and behavioral deviation (**Fig. 4e**). Consistent with **Fig. 4c**, the average of R^2 across tuned neurons was significant for two flank cues (**Fig. 4f**). Crucially, the mean R^2 over all cues, averaged across all cells, was significant for tuned but not for non-tuned neurons (**Fig. 4g**). Thus, non-tuned neurons did not show a detectable rate-behavior relationship in our dataset.

Late-delay behavioral modulations of Fano factors

We then tested the contribution of bump attractor diffusion to neuronal variability, as captured by the Fano factor. Following our hypothesis, bump displacements in different trials induce behavioral inaccuracies and lead to the largest variation in neuronal response for cues in the flanks of a neuron's tuning curve (**Fig. 6a,b**). For these cues, random diffusion of the bump causes maximal neuronal

activation in trials when the bump drifts towards the neuron's preferred location, and minimal activation when it drifts away (neuron at 135° in **Fig. 6b**, left). Our hypothesis predicts that the variance of neural responses to flank stimuli should be larger for trials with inaccurate compared to more accurate behavior. Moreover, this difference should be specific of flank stimuli and absent, or even inverted, for preferred or tail stimuli (**Fig. 6b**, right). Note that the variance of neural responses is also affected by independent spiking noise, typically in proportion to the mean response^{24,25}. This contribution will not interfere qualitatively with our prediction, assuming its invariance across task conditions²⁵.

We separated inaccurate trials, where the monkey made saccadic responses beyond the median absolute angular displacement (distribution tails in **Fig. 1b**), from accurate trials, which contained the same number of trials for each neuron and cue combination as the inaccurate group, only it contained the trials with the smallest angular displacement. In flank-cue trials, the Fano factor for accurate and inaccurate trials differed significantly at the end of the delay period, as predicted (last 0.5 s of delay, one-sided paired t -test $t = 1.56$, $P = 0.05$, $n = 181$, **Fig. 6c**). This difference increased parametrically as we restricted inaccurate trials to the most extreme saccadic deviations (**Supplementary Fig. 3**). The effect was specific for flank stimuli (**Fig. 6d**): modeling Fano factor with a mixed-effects ANOVA with factors cue, accuracy, monkey and neuron identity as random factor yielded a significant interaction effect of cue \times accuracy ($F[2,897] = 6.69$, $P = 0.0013$, cell size 152 or 29 depending on monkey, **Supplementary Fig. 4**). Reduced ANOVA models revealed a main effect of cue for inaccurate trials ($F[2,360] = 16.01$, $P < 0.001$), and no significant cue effect for accurate trials ($F[2,360] = 1.32$, $P = 0.27$).

Late-delay selective negative pairwise correlations

We finally tested a long-standing prediction on how spike count correlations between neurons depend on the neurons' tuning preferences and the cue in a bump attractor representation^{6,7,20}. We expected negative trial-to-trial correlations in the delay activity of two neurons responding to a cue presented right between their two preferred locations, due to random bump displacements in different trials (**Fig. 7a,b**). Only for this condition, when the cue engages the neurons in parts of their tuning curve with slopes of opposite sign, do we expect a negative correlation. Other cue conditions, or correlations for neurons with the same selectivity should show positive or vanishing correlations^{6,7}.

We selected neuron pairs in two conditions: *same-tuning pairs*, with neurons sharing preferred cue

($n = 15$, **Fig. 7c**), and *different-tuning pairs*, where neurons differed in preferred location with one intervening cue in between ($n = 10$, **Fig. 7d**). We computed correlations between responses in the pair (insets in **Fig. 7c,d**) for various cue conditions (Peak, Flank, and Tail in **Fig. 7c**, and In-flank, Peak and Out-flank in **Fig. 7d**) both in early and late delay (200 ms windows). Same-tuning pairs had stronger correlations than different-tuning pairs (3-way ANOVA with factors time, cue, and neuron selectivity difference: main effect of selectivity difference, $F[1,272] = 25.6$, $P < 0.001$; no interaction effects $P > 0.2$; **Supplementary Fig. 4e-h**). Within same-tuning pairs, no other interaction or main effect was significant (factorial ANOVA, factors time and cue, $P > 0.5$), indicating consistently positive correlations through the delay, independently of the cue (**Fig. 7e**). Instead, a significant interaction of time and cue emerged for different-tuning pairs ($F[2,114] = 3$, $P = 0.05$), which reflected pairwise correlations becoming significantly negative at the end of the delay for in-flank stimuli (**Fig. 7f**). Thus, spike count correlations changed with cue condition as predicted by the bump attractor hypothesis.

The data can distinguish alternative scenarios

To test the extent to which our experimental data could distinguish the bump attractor model from other alternative encoding hypotheses for memory maintenance in PFC, we formulated two alternative models, the discrete attractor and the decaying bump network models (**Supplementary Videos**).

The continuous bump attractor network featured strong topographic connectivity between excitatory neurons, so that a rigid bump attractor stabilized after brief network activation and it diffused in the network during the delay period due to external noisy inputs (**Fig. 8a, left**). The discrete attractor network included 8 populations, each one encoding one of the cue locations, with stronger connections within and weaker connections across populations. An external stimulus brought the system to an attractor that maintained 3 adjacent populations persistently active and subject to strong external noise (**Fig. 8a, middle**). Finally, in the decaying bump network, mnemonic information was encoded by individual neurons through an intrinsic depolarizing current that slowly decayed away after initial activation by the stimulus. The bump of activity thus slowly decayed away during the delay period (**Fig. 8a, right**).

To test whether a neural dataset with the characteristics of our experimental data can distinguish between these models, we performed simulations with the three firing rate models (Online Methods and **Supplementary Code 1-3**), and we generated three surrogate datasets matching the sample sizes in our

experiment. We picked parameters for our models that produced similar neural and behavioral data (**Fig. 8a,b**), in good qualitative agreement with experimental data (**Fig. 1**). To get a sense of the quantitative effects expected, we also tested non-mechanistic coding models that could match quantitatively the heterogeneity of experimental neural data (Online Methods and **Supplementary Figs. 5-6**).

We analyzed these surrogate datasets exactly as we did before for the experimental data. Data produced by the bump attractor model replicated our experimental findings, but none of the effects were replicated by the discrete attractor or the decaying bump models (tuning curve bias **Fig. 8c**; rate-behavior correlation **Fig. 8d**; Fano-Factor and pairwise correlations **Supplementary Fig. 6**). For these models, the lack of effects occurred because behavioral variability did not result primarily from collective neuronal variability: both in the discrete attractor and in the decaying bump models behavioral variability emerged largely from independent random fluctuations at the cellular level, and not from correlated population dynamics as in the bump attractor model. Such dynamics can occur in discrete attractor models, in the form of noise-induced transitions to adjacent attractors. However, this leads to large abrupt shifts in behavioral read-out, which result in multimodal behavioral distributions, not supported experimentally (**Fig. 1b**) unless very fine discretization approaching a continuum is assumed. We conclude that our experimental findings can discriminate between distinct mechanisms of working memory maintenance, supporting a neural representation in PFC compatible with the bump attractor hypothesis for spatial working memory.

Discussion

We have analyzed behavioral and electrophysiological data from monkeys performing a spatial working memory task and tested predictions from an attractor bump hypothesis for spatial memory maintenance. Our analyses confirmed the model predictions, supporting the hypothesis that PFC activity represents spatial memories in a fixed-shape bump of activity that is used for guiding behavior but is liable to cumulative encoding errors due to random fluctuations. Our results impact our conceptual understanding of PFC activity in spatial working memory: they (1) validate the concept of prefrontal persistent activity as the basis of spatial working memory, (2) support the continuous (or finely discretized) nature of spatial memory encoding in PFC, and (3) are consistent with bump attractor dynamics mediating cognitive function in the cortex.

The concept of persistent activity has been highly influential in formulating concrete mechanistic models for working memory^{10,19,26,27}, but it cannot account naturally for the great heterogeneity, irregularity, and dynamics of prefrontal activity during delayed response tasks^{9,14–16,28–31}. Recent studies have shown that biophysical neural network mechanisms for persistent activity can accommodate heterogeneity in firing rates and tuning properties^{32,33}, as well as irregular spiking activity^{33,34}. In addition, variable neural activity in the delay period after identical stimulus presentation can reflect insufficient conditioning to details of sensory stimuli or behavior. According to this view, variability in the delay period should be predictive of behavioral responses. Our data confirms this for PFC neural variability in spatial working memory tasks, consistent with choice probability measurements in other cognitive tasks^{28,35}. Finally, heterogeneity might reflect the presence of distinct functional populations³⁶. Consistent with this, we have found a significant relationship between trial-to-trial variability of neuronal responses and the fine details of behavioral reports only for neurons with sustained and tuned delay activity, but not for non-tuned neurons (**Fig. 4**).

Non-stationary responses are more difficult to account for within attractor network models for working memory⁹. Alternative neural mechanisms have been proposed, which rely on sequential activation of neuronal subpopulations¹³ or storage in short-term synaptic states¹², not on sustained neural spiking. It is unclear how these alternative models could accommodate our experimental findings. Here, we discarded a non-attractor model based on the slow decay of an initial bump of activity, although the data is still consistent with a diffusing bump representation that also undergoes slow decay. However,

we cannot rule out that other non-attractor models may explain some of our findings. Testing this would require building spatial working memory networks based on these other theoretical frameworks.

Also, several independent dynamical components may converge in shaping neural activity in PFC³⁷. In our task, a timing component related to the predictability of task events could partly explain some temporal dynamics observed in Fano factors or firing rates, while a stimulus encoding component would determine neural tuning. The assumption that these components are independent³⁷ is implicit in our Fano factor analysis (**Fig. 6**), where we do not pay attention to temporal dynamics and just focus on neural tuning and behavior. A related concern is that a tuned ramping-up component related to motor parameters, known to emerge in late delay²³, may dominate our results, thus reflecting saccade preparation and not working memory in PFC. We have ruled out this possibility by showing that our results persist when excluding neurons with ramping-up delay activity from the analyses.

Our results support a continuous or finely discretized spatial encoding of memorized locations in the PFC. This is remarkable because our two monkeys were tested over one year in a task that involved systematically the same 8 identical cues, at specific locations in their visual space. This could have generated a neural code that emphasized these 8 discrete locations without much selectivity for intermediate positions³⁸ (discrete attractor model, **Fig. 8**). Such task-tailored representation during working memory delays has been seen in the PFC for other type of task requirements³⁹. Our findings set apart spatial working memory in this respect, suggesting that a continuous representation for memorized space is permanently present in PFC, possibly reinforced by everyday navigation in a spatially continuous environment. This is consistent with evidence showing that prefrontal neurons have mnemonic spatial fields even in untrained animals⁴⁰.

Noise correlations are known to depend on tuning curve similarity (i.e. on signal correlations)^{41–44}, but to what extent they depend on the stimulus remains unclear^{41,42}. Here, we have found that noise correlations between PFC neurons do depend on the location of the memorized stimulus. This is further evidence that noise correlations do not just reflect fixed hard-wired connectivity but also network dynamics and ongoing computations, such as attentional^{44,45} and context-depending processing⁴³. Moreover, the data revealed negative noise correlations in very specific stimulus conditions (**Fig. 7**), thus confirming a long-standing prediction of bump attractor models^{6,7}. This finding is in contrast with another study⁴¹, which did not find evidence for negative stimulus-induced correlations in orientation-

tuned neurons in primary visual cortex (V1) of anesthetized monkeys. Different network dynamics in PFC and V1, and/or different brain state in awake vs. anesthetized animals could be responsible for this difference.

Experimental studies searching for evidence of attractor dynamics in neural activity during behavior have focused on testing whether neuronal spiking converges to discrete singular patterns upon parametric changes in sensory stimuli⁴⁶⁻⁴⁸, upon changes in reward rules⁴⁹ or during distinct phases of a multi-component cognitive task⁵⁰. These approaches can identify dynamics governed by discrete attractors, where small changes in task conditions shift the neural representation abruptly to a new attractive state. However, they cannot identify continuous attractors, such as the bump attractor, since such attractors have one continuous dimension and therefore one does not expect abrupt changes to small modifications of external conditions. Our findings demonstrate that identifying the behavioral parameters associated with the direction of marginal stability in the model, here the small angular inaccuracies in behavioral report, can provide model-dependent tests to apply to the data and validate a continuous attractor explanation for the underlying neural dynamics. Our results implicate a bump attractor code in the PFC mediating the precision of response in spatial working memory.

Acknowledgments

This work was funded by the Ministry of Economy and Competitiveness of Spain, and the European Regional Development Fund (Ref: BFU2009-09537) and by the National Science Foundation under Grant No. NSF PHY11-25915 and DMS-0847749 . K.W. acknowledges funding from the German Research Foundation (research fellowship Wi 3767/1-1). D.Q.N. was supported by the Generalitat de Catalunya (PIV-DGR 2010 program). C.C. received support from the Tab Williams Family Endowment. The work was carried out at the Esther Koplowitz Centre, Barcelona, and partly at the Kavli Institute for Theoretical Physics, Santa Barbara CA, and at the Centre de Recerca Matemàtica, Barcelona. We thank Mark Goldman, Jaime de la Rocha, and Alex Roxin for fruitful discussions. We wish to acknowledge the late Patricia Goldman-Rakic, in the laboratory of whom experimental data were collected.

Author contributions

A.C. conceived and designed the study; C.C. collected the experimental data; K.W. and A.C. analyzed the data; D.Q.N. and A.C. implemented the computational models; A.C., C.C., D.Q.N., and K.W. wrote the manuscript.

Figure legends

Figure 1: Behavioral and neural fingerprints of spatial working memory. **(a)** Saccade endpoints reporting cue location after a 3-s delay (fixation on black cross) for one monkey. White crosses indicate mean saccade locations for each of 8 cues. Color indicates angular deviation from mean responses. **(b)** Angular deviations pooled over cues and monkeys. Triangles indicate plus/minus median absolute deviation. **(c)** PFC neurons represented cue location in selective sustained delay activity. Population activity (100 ms sliding window, $n = 204$) for preferred (solid) and non-preferred cues (dashed). Cue and response periods marked in gray. **(d)** Population delay tuning curve for all neurons (thick line, $n = 204$), and for $n = 102$ neurons stronger (thin line) and weaker tuning (dotted line). **(e)** Distribution of mean delay firing rates (upper histogram, mean 9 Hz, s.d. 9.2 Hz), of tuning strength T (right histogram, mean 0.21, s.d. 0.15) and their correlation (central plot, $P < 0.0001$, $n = 204$). **(f)** Distribution of a rate modulation index (Online Methods) did not deviate significantly from a Gaussian (Lilliefors test, $P = 0.27$, $n = 204$) with zero mean (t -test, $t = 0.23$, $P = 0.82$, $n = 204$). Negative (positive) values correspond to a firing rate decrease (increase) during the delay. Filled bars mark neurons with significant changes (Wilcoxon test, $P < 0.05$). **(g)** Distribution of modulation indexes for preferred location (top) and tuning strength T (bottom). Filled bars indicate neurons with significant changes (permutation test, $P < 0.05$). Gray shades are bootstrap-estimated s.e.m.

Figure 2: Bump attractor dynamics during the delay can explain behavioral inaccuracies. **(a)** Spatio-temporal representation of network activity during the delay period in an individual trial. Gray levels and z-axis elevations schematize neuronal firing rates. **(b)** Same trial as in panel **a**, but represented on the time-network plane. Gray scale represents firing rate elevations. The black triangle shows the location of the initial cue, right before the beginning of the delay (Encoded population activity in bottom panel). The white triangle indicates the behavioral response decoded from network activity at the end of the delay (Decoded population activity in bottom panel). Leftward displacement of the white relative to the black triangle indicates a CW behavioral response deviation in this trial. **(c)** Same as in panel **b** but for a different trial. Rightward displacement of the white relative to the black triangle indicates a CCW, inaccurate trial.

Figure 3: Tuning curves computed from CW and CCW behavioral trials show model-predicted shift in the delay period. **(a)** Representation of saccade endpoints for one session. For each cue, trials are separated in half based on their relative CW (red) and CCW (blue) saccadic responses. **(b)** Sample neuron delay-period responses in the CW and CCW conditions. Triangles indicate the circular mean of the responses, an estimate of the preferred cue for each condition. The distance between these two circular means is the *tuning bias*. **(c)** Population average of the tuning bias for all neurons across time shows significantly positive values by the end of the delay. Cue (C) and response (R) periods are indicated with gray areas. Tuning curves were estimated over 1 s sliding windows. Error bars (shaded area) indicate s.e.m. Thick lines indicate periods of significantly positive tuning bias (permutation test, $P < 0.05$).

Figure 4: Responses to flank stimuli (1 and 2 locations from preferred) become correlated with upcoming behavior during the delay. **(a)** Sample neuron responses to flank stimuli superimposed on its tuning curve (data from last second of delay). Each circle represents a single trial: y is the neuron's firing rate and x is the angle of saccadic endpoint (also ticked on the line above). Dashed lines indicate cue locations, and saccade deviations are marked with horizontal stems. Black (gray) circles indicate behavioral imprecision towards (away from) the neuron's preferred location. **(b)** Firing rate deviations from tuning curve correlate positively with saccadic deviations from cue location. Saccadic deviations toward the preferred cue have positive sign. Same data as in **a**. **(c)** Population average of correlation coefficients computed as in **b** for tuned ($n = 204$, solid line) and non-tuned neurons ($n = 523$, dashed line). Rate-behavior correlations were computed over 1 s sliding windows. Shaded areas indicate bootstrapped s.e.m. Thick lines mark periods of significantly positive correlation (permutation test, $P < 0.05$). **(d)** Average rate-behavior correlation over the last 2 s of the delay for tuned (solid) and non-tuned neurons (dashed) with low, medium, and high tuning strength T (**Supplementary Fig. 2**). **(e)** For the same neuron as in **a**, R^2 values for the linear regression of firing rate and behavioral deviations (solid line) are highest for individual flank stimuli. Dashed line is mean R^2 for shuffled surrogates, R^2_{shuffle} . **(f)** Population average of corrected R^2 ($R^2 - R^2_{\text{shuffle}}$) is significantly positive for flank stimuli in tuned neurons. Dotted lines are 95% confidence intervals for R^2_{shuffled} . **(g)** Corrected R^2 averaged over cues is significantly positive for tuned but not non-tuned neurons (tuned: $P = 0.045$, non-tuned: $P = 0.36$, $n = 204$, one-tailed permutation test).

Figure 5: As predicted by the bump attractor model, neuronal variability correlates with upcoming behavioral responses in a cue-dependent way: maximally for flank stimuli, and earlier in the delay for cues closer to the neuron's preferred location. **(a)** Absence of rate-behavior correlation for trials in which the presented cue coincided with the neuron's preferred location θ_{pref} . **(b)** Significant rate-behavior correlation (permutation test, $P < 0.05$), as early as 2 s before saccade, for trials with cues presented just next to θ_{pref} . **(c)** Late-delay rate-behavior correlation for cues presented two locations away from θ_{pref} . **(d)** Absence of rate-behavior correlation for cues opposing θ_{pref} . Correlation was computed in 1-s sliding windows. In **(b,c)** we evaluated the center of mass (CM, black triangles) of time points with significantly positive correlation (one-sided permutation test, $P < 0.05$). CM standard errors were evaluated with a bootstrap procedure. The curve shown in **Fig. 4c** is the result of combining trials used in panels **b** and **c** in this figure.

Figure 6: Fano factors (FFs) follow the predictions of the bump attractor model. **(a)** Depending on the response properties of recorded neurons, cues can be classified as preferred, flank and tail cues. **(b)** Left: Schematic representation of 4 different late-delay population activity profiles in response to the same 90° cue. Orange lines represent trials with behavior closer to the target (accurate trials) and green lines trials with behavior farther from the target (inaccurate trials). The range of neural responses for these types of trials are marked with vertical rectangles for specific neurons in the network, those for which the presented cue represents a preferred, a flank or a tail cue. Right: According to the bump attractor hypothesis, neural response variability should be higher for inaccurate than accurate trials, selectively for flank stimuli. **(c)** In the data, when flank stimuli responses are separated into trials with behavioral responses farther or closer from the mean saccadic endpoint for that cue, FF dynamics separate by the end of the delay with inaccurate responses showing higher FFs than accurate responses (one-sided permutation test, $P < 0.05$). **(d)** The difference between FFs in inaccurate and accurate trials at the end of the delay (averaging counts in the last 500 ms) depends on the cue condition. The FF difference between accurate and inaccurate trials is significant for preferred and flank stimuli (permutation test, $P < 0.05$, $n = 181$). FF was computed in 100 ms windows.

Figure 7: Noise correlations between pairs of neurons depend on the stimulus as predicted by the bump attractor model. **(a)** Scheme of delay population activity profiles in response to 3 repeated presentations

of 180° cue. We focus on two neurons in this schematic network (\circ and \square) lying at opposite sides of the activity bump. **(b)** Only in this configuration, trial-to-trial correlations between the two neurons are expected to be negative because a displacement of the bump, illustrated in **a**, leads to an increase of firing rate in one neuron and a decrease in the other neuron (\circ and \square in **a**). **(c)** Delay-period tuning curves for a sample pair of PFC neurons with the same preferred cue. Relevant cue conditions are indicated below the x -axis. Spike count correlations were computed for each pair in the final 200 ms of the delay. Inset: scatter plot of spike counts for this sample pair. **(d)** Same as **c**, for a pair of neurons whose preferred cues were separated by one cue. Inset: scatter plot of spike counts for the “In-flank” stimulus. **(e)** Pairwise correlation for same-tuning pairs ($n = 15$) computed separately for Peak, Flank, and Tail cue conditions. Significant deviation from zero mean was tested combining the last two 200 ms bins in the delay period (two-tailed t -test, $P < 0.05$, $n = 30$). **(f)** Same as **e** for different-tuning pairs ($n = 10$). Negative correlation for in-flank stimuli was tested with one-sided t -test ($t = 2.42$, $P = 0.026$, $n = 20$).

Figure 8: Comparison of three alternative memory representations in three mechanistic models: a bump attractor model maintained by continuous topographic recurrent excitation (left), a discrete attractor model with 8 populations (middle), and a non-attractor decaying bump model sustained by a slow intrinsic current (right). **(a)** Sample simulated delay activity in one trial for each model (see also **Supplementary Videos**). Triangles mark decoded response. Insets display recurrent excitatory connectivity patterns. **(b)** Distributions of behavioral responses over 16,000 simulations for each model reveal similar behavioral variability. **(c)** Tuning bias analysis as in **Fig. 3** for neural and behavioral data obtained from each model reveal significant positive bias only for the bump attractor model. **(d)** Correlation between rate and behavioral deviations towards the neuron's preferred location (as in **Fig. 4**) becomes gradually positive in the delay only for the bump attractor model.

Supplementary Figure 1: Schematic illustration of the first prediction from the bump attractor model regarding the relation between single-neuron delay tuning curves and the direction of small behavioral imprecisions. **(a)** Scheme of 2 possible delay population activity profiles (red and blue lines) in response to the repeated presentation of a 90° cue. Repeated presentation of any given cue results in

variable behavior, with saccades directed CW (red) and CCW (blue) from the correct cue location. Neural responses are marked for one single neuron in the network (circles, preferred cue 144°). **(b)** Same as **a** for a different cue, presented at 180°. Again two network responses for two different trials are plotted, which give rise to CW (red) and CCW (blue) behavioral imprecisions. Responses from the same neuron with preference at 144° are indicated by filled circles. **(c)** Tuning curves computed for the single neuron indicated in **a** and **b**, separately recorded from trials with CCW and CW behavior. The tuning curve is constructed by plotting the neuron's response to different stimuli locations (x -axis). Here panels **a** and **b** contribute two points to each tuning curve in panel **c**, the rest of the points are obtained by presenting all other possible cues to the network (not shown). Notice that CW tuning curves are displaced CCW relative to CCW tuning curves: the tuning bias defined as the distance from the CCW to the CW tuning curve centers is predicted to be positive.

Supplementary Figure 2: Comparison of delay tuning properties of the databases of tuned and non-tuned neurons. **(a)** Mean tuning for tuned neurons (solid) is higher than for non-tuned neurons (dashed). Note that the alignment to preferred locations induces a spurious tuning by itself. **(b)** Histograms of tuning strength T for tuned (black) and non-tuned neurons (gray) differ significantly in their medians (triangles, 0.0832 vs. 0.1757, $P < 1e-6$, Wilcoxon test) but have a significant overlap. **(c)** Neurons with similar tuning strength T in the two databases (tuned, black and non-tuned, gray) have lower mean firing rate for non-tuned neurons, thus explaining non-significant tuning in cells with high tuning strength T in the non-tuned neuron database. Low, medium and high tuning was defined using tertiles obtained from the tuned neuron database and splitting the non-tuned database using the same intervals ($n = 353, 111$ and 59 non-tuned neurons for low, medium and high tuning, respectively).

Supplementary Figure 3: Late-delay Fano factor difference between accurate and inaccurate trials for flank cues increases for more extreme behavioral deviations. As the fraction of trials used to build accurate and inaccurate response classes is reduced (i.e. the two classes differ more in their mean response accuracy) the difference between Fano Factors computed from neural responses at the end of the delay (last 500 ms) for flank cue stimuli in each of these classes of trials grows parametrically. Significance was assessed with a permutation test ($P < 0.05$).

Supplementary Figure 4: Validation of ANOVA models. Graphical representation of the ANOVA residuals corresponding to the analyses in **Fig. 6a-d** and **Fig. 7e-h**. Data deviated mildly from the normality assumption due to long tails (panels **a, e**) and some outlier points (panel **e**), and from the homoscedasticity assumption in panels **b, d, g, h** (Levene's test $P < 0.05$). Note that despite this significant heteroscedasticity, there is not more than a factor 3 between the various dispersions. Heteroscedasticity is the most worrisome aspect of the data, and it is particularly severe in the case of unequal sample sizes. In our case the sample size corresponding to the different monkeys (**d**) and the different neuronal pair tunings (**h**) were different. However, the cases with the smaller sample size had smaller variance and this condition is known to reduce the power of the test rather than inflate the false positive rate. Thus, for both ANOVA tests, non-normality and heteroscedasticity do not question the rejection of the hypotheses, since the corresponding corrections are unlikely to cancel the strong significance of the effects found ($P = 0.0013$ for the interaction cue \times accuracy in **Fig. 6** and panels **a-d**, and $P < 0.001$ for main effect of selectivity in **Fig. 7** and panels **e-h**).

Supplementary Figure 5: Coding models that match experimental data quantitatively. The models in Fig 8 were in qualitative agreement with experimental behavioral and neural data, but the quantitative match of neuronal firing rates was limited by the network implementation. To evaluate quantitative predictions that took into account the specific firing rates and heterogeneity present in the data (**Fig. 1**), we tested two coding models for inaccurate memory-guided behavioral responses: the bump attractor model and the decaying bump model. **(a)** In the bump attractor model, dispersion of behavioral responses is due to noise-induced diffusion of the location of a rigid bump representation. In different trials (red and blue) in response to the same cue the bump diffuses during the delay to different locations (lower panel), giving rise to different read-outs and inaccurate, off-target behavioral responses. **(b)** In the decaying bump model, the bump is established at cue presentation, but begins to decay once the cue is removed. During the delay, the coding slowly decays away, with different decay rates in different trials. The read-out is inaccurate at the end of the delay due to the decreased selectivity. **(c-d)** Generation of multiple firing rate trials (8 equidistant cues, 10 trials per cue) and surrogate Poisson spike trains for $n = 200$ neurons from the models in corresponding panels **a** and **b**

above generates behavioral and neural data in quantitative agreement with the experimental data (**Fig. 1**). Behavioral responses from computational models (panel **c**, left, and panel **d**, right) are represented as in **Fig. 1b**. The memory dependence of these behavioral inaccuracies is revealed by comparing with the models' behavioral responses for a brief 0.5 s delay (black dashed curve). Average tuning curves of model neurons (panel **c**, right and panel **d**, left) in the delay period computed as in **Fig. 1d**.

Supplementary Figure 6: Model-derived surrogate data mimicking our experimental dataset supports a bump attractor representation, not a decaying bump representation. (**a,b**) Tuning curve bias analysis (**Fig. 3c**) performed on surrogate data from the diffusing and decaying bump models (**Supplementary Fig. 5**), respectively. For the bump attractor, but not the decaying bump model, the tuning curve bias computed from trials with CW versus CCW behaviors (**Fig. 3**) becomes increasingly positive through the delay. (**c,d**) Rate-behavior correlation analysis (**Fig. 4c**) applied to the surrogate data of **Supplementary Fig. 5**. Only for the bump attractor model spike counts in response to flank stimuli correlated with behavioral deviations in the direction of the neuron's preferred cue as delay progressed. (**e,f**) For surrogate data generated from the bump attractor model, but not the decaying bump model, Fano Factors computed separately for trials when a flank stimulus was presented are larger when considering inaccurate trials with large behavioral deviations (solid line) as compared to accurate trials (dashed line) (c.f., **Fig. 6d**). This is shown for average activity in the last 500 ms of delay period, when considering cues presented at different points of the neuronal tuning curve. (**g,h**) Only surrogate data derived from the bump attractor model mimics the experimental results (**Fig. 7e,f**) that neuron pair noise correlation is negative for those pairs with dissimilar tuning, when responding to a middle flank stimulus (*In-flank* condition). Solid lines correspond to labels "Peak," "Flanks," and "Tails" as in **Fig. 7e**. Dashed lines correspond to labels "In-flank," "Peaks," and "Out-flanks" as in **Fig. 7f**.

Supplementary Code 1: Matlab code to run the bump attractor network model of **Fig. 8**. The model is described in Online Methods generically, and this Matlab code specifies all the parameters necessary to run the simulation.

Supplementary Code 2: Matlab code to run the discrete attractor network model of **Fig. 8**. The model is described in Online Methods generically, and this Matlab code specifies all the parameters necessary to run the simulation.

Supplementary Code 3: Matlab code to run the decaying bump network model of **Fig. 8**. The model is described in Online Methods generically, and this Matlab code specifies all the parameters necessary to run the simulation.

Supplementary Video 1: Video demonstrating network activity and behavioral read-out in the course of one simulation of the bump attractor network model of **Fig. 8**. The upper left panel shows the temporal progress along the task timeline (F=fixation, C=cue, D=delay, R=response) with a red time-stamp marker. The upper right panel shows in black the simulated visual scene: fixating cross and square visual cue, together with the instantaneous network readout during the delay in red. The lower panel shows the activity of excitatory neurons in the network. The 8 possible locations of cue stimuli are indicated on the lower panel with gray dots. When external stimuli are applied to the network, these are indicated with a thick black line on the location of corresponding neurons on the x -axis. This occurs in the cue period of the trial (C in upper panel), and is implemented as a depolarizing current to excitatory neurons around the cue location at $\theta = 0^\circ$, and in the response period (R in upper panel), where a hyperpolarizing current is injected to all excitatory neurons in the network. The location encoded instantaneously in the excitatory population activity is read out continuously with a population vector algorithm (Online Methods and **Supplementary Code 1**) and it is indicated with a black triangle in the lower panel and with a red radial line in the upper right panel. Parameters for this simulation are as detailed in **Supplementary Code 1**.

Supplementary Video 2: Video demonstrating network activity and behavioral read-out in the course of one simulation of the discrete attractor network model of **Fig. 8**. Same layout and symbols as for **Supplementary Video 1**. Parameters for this simulation are as detailed in **Supplementary Code 2**.

Supplementary Video 3: Video demonstrating network activity and behavioral read-out in the course of one simulation of the decaying bump network model of **Fig. 8**. Same layout and symbols as for **Supplementary Video 1**. Parameters for this simulation are as detailed in **Supplementary Code 3**.

Online Methods

Behavioral task and recordings

Two adult, male rhesus monkeys (*Macaca mulatta*) were trained in an oculomotor delayed response task requiring them to view a visual stimulus on a screen and make an eye movement after a delay period. During execution of the task, neurophysiological recordings were obtained from the lateral prefrontal cortex. Detailed methods of the behavioral task, training, surgeries and recordings, as well as descriptions of neuronal responses in the task have been published before^{3,31,51,52} and are only summarized briefly here. Visual stimuli were 1° squares, flashed for 500 ms at an eccentricity of 14°. Stimuli were presented randomly at one of 8 possible locations around the fixation point. A delay period lasting 3 s followed the presentation of the stimulus, at the end of which the fixation point turned off, and an eye movement towards the location of the remembered stimulus was reinforced with liquid reward. Eye position was monitored with a scleral eye coil system. Neuronal activity was monitored using tungsten electrodes of 1–4 M Ω impedance at 1 kHz. All experiments were conducted in accordance with the guidelines set forth by the National Institutes of Health, as reviewed and approved by the Yale University Institutional Animal Care and Use Committee.

Data analysis

A total of 822 neurons were recorded from two monkeys with the 8-target oculomotor delayed response task. Sample size was determined by the fact that it had been collected previously^{3,31,51,52}. Neuronal responses in this task have been characterized previously³, but their trial-to-trial relationship to behavioral parameters has not been investigated before. We compiled our dataset with neurons that showed significant tuned delay period activity, evaluated using a bootstrap method as explained in ref. 3 and for which end points of saccadic eye movements were available. Our database thus consisted of 204 *tuned neurons*, 172 from monkey COD and 32 from monkey MAR. For some analyses in **Fig. 4** we also compiled a dataset with the remaining neurons that did not show significant tuned delay activity and for which behavioral response data was available (*non-tuned neurons*, $n = 523$: 388 from monkey COD and 135 from monkey MAR).

In parallel, we built an additional database with pairs of neurons recorded simultaneously, from the same or separate electrodes 0.2 - 1 mm apart, both of which showed tuned delay activity as per the

criteria described above. This database consisted of $n = 64$ pairs, 53 of which were recorded in monkey COD and 11 of which came from monkey MAR. All our results subsisted if we eliminated pairs recorded from the same electrode from our correlation analyses (7 out of 15 same-selectivity pairs, and 2 out of 10 different-selectivity pairs came from a single electrode). This affected especially the power of the statistical tests concerning same-selectivity pairs (**Fig. 7e**), but none of the essential effects reported here.

For each neuron in our database, preferred location was determined by computing the circular mean of the cue angles (0° to 315° , in steps of 45°) weighted by the neuron's mean spike count over the delay period (3 s) upon each cue presentation. To this end, we computed for each neuron the complex quantity $\mathbf{T} = \left(\sum_{j=1}^8 n_j e^{i\theta_j} \right) \left(\sum_{j=1}^8 n_j \right)^{-1}$, where n_j is the mean spike count during the delay period in response to the cue θ_j ($j = 1..8$), and we extracted its modulus T and angle θ_{pref} . $\mathbf{T} = T e^{i\theta_{pref}}$. The angle θ_{pref} constitutes our estimate of the neuron's preferred location during the delay, and the tuning strength T is our estimate of the delay tuning quality. T can reach a maximal value of 1 when the neuron responds exclusively to one cue during the delay (i.e. $n_j = 0$ for all $\theta_j \neq \theta_{pref}$) and a minimum value of 0 when the neuron's response is evenly balanced around the circle, such as the case when the neuron responds with equal number of spikes to all cues (all n_j equal).

Spike trains for each neuron and condition (cue in one of 8 possible locations) were analyzed in time windows of various lengths (typically 1 s for tuning curve and firing rate estimations, 0.1 s for Fano Factor estimations, 0.2 s for spike count correlations) that slid over the duration of the trial in steps of 0.1 s to estimate trial-related time evolution. To test stationarity through the delay we computed modulation indices for preferred rate and tuning as the difference in measures obtained in the first and last second of the delay divided by their sum (**Fig. 1f,g**). The modulation index for preferred location θ_{pref} (**Fig. 1g**) was computed as the circular distance between θ_{pref} estimated in the first and last seconds of the delay, divided by the maximal possible distance of 180° . All these modulation indices share the property that they are bounded between -1 and 1 and stationarity is characterized by a concentration around 0.

Displacements of tuning curves computed for clockwise (CW) and counter-clockwise (CCW) behavior trials (**Fig. 3**) were evaluated with the *tuning bias*. We defined this tuning bias as $\theta_{pref}^{CW} - \theta_{pref}^{CCW}$, where

θ_{pref}^{CW} (θ_{pref}^{CCW}) is the neuron's preferred location calculated as detailed above but restricting it to CW (CCW) behavior trials (see **Fig. 3a**). Notice that a positive value of the tuning bias results when the preferred location for CW trials is displaced CCW relative to the preferred location for CCW trials. A positive tuning bias is expected for tuning curves generated from displaced population activity bumps (**Supplementary Fig. 1**), as described in previous computational studies^{53,54}. Thus, the attractor bump hypothesis predicts a significantly positive tuning bias $\theta_{pref}^{CW} - \theta_{pref}^{CCW}$ at the end of the delay period. Since this analysis depends on the estimation of two centers of tuning curves, derived each from half the typical number of trials, we expect the tuning bias to be better estimated in well-tuned neurons. To emphasize the tuning properties of neurons with better tuning, we computed the population tuning bias as an average weighted by our tuning measure T .

The population Fano Factor was estimated as:

$$FF = \frac{N_T^{-1} \sum N_i FF_i}{N_T^{-1} \sum N_i} = \langle FF_i \rangle_{ij} = \left\langle \left(\frac{\sqrt{N_i} n_{ij} - \bar{n}_i}{\sqrt{N_i - 1} \sqrt{\bar{n}_i}} \right)^2 \right\rangle_{ij} \quad (1)$$

where angular brackets $\langle \rangle_{ij}$ indicate average over all trials of all conditions (neuron and cue), N_i is the number of trials in condition i (corresponding to a specific neuron and cue), N is the total number of conditions, N_T is the total number of trials in all conditions ($N_T = \sum N_i$), \bar{n}_i is the mean spike count in condition i , and n_{ij} is the spike count in trial j of condition i . This calculation yields an average of the Fano Factors of individual conditions FF_i in the database, weighted by the total number of trials in each condition N_i , and is therefore a more robust estimator of the population Fano Factor FF than the unweighted average over conditions.

Spike count correlations were calculated as the Pearson correlation coefficient between the spike counts of pairs of neurons, and then averaged over neurons.

All analyses were carried out in Matlab, using the CircStat toolbox⁵⁰ to perform circular statistics.

Statistical methods

We applied parametric tests (t-test, ANOVA, or Harrison-Kanji test for circular data) to validate our hypotheses whenever our data satisfied the assumptions of normality and homoscedasticity (Lilliefors

test of normality, $P > 0.05$, and Levene's test for equality of variances, $P > 0.05$, respectively). For two ANOVA analyses, we accepted mild deviations from the assumptions since they were unlikely to affect the strong effect reported by the ANOVA ($P < 0.005$, **Supplementary Fig. 4**). When tests comparing one or two samples failed to meet the parametric test assumptions, we applied the non-parametric Wilcoxon test or a permutation test when normality was not met, and Welch's t -test for heteroscedastic data. Fano-Factors and tuning strength T were Box-Cox transformed to correct skewness (exponent range [0.22, 0.28]) prior to testing. Pearson correlations were Fisher-transformed prior to population tests. Pairwise comparisons critically predicted by the model were also re-tested with permutation tests to validate near-threshold t -test hypothesis rejections (**Figs. 6,7**). When the model prediction identified the sign of the comparison we used one-sided tests, and this is explicitly indicated in the text. Paired tests are used to compare measurements within neurons, in ANOVAs this is accomplished by using a mixed-effects design with neuron as random factor, nested in the monkey factor. Bootstrap estimates of the standard error of the Fano Factor are calculated as the standard deviation of Fano Factors evaluated in 1,000 bootstrap samples obtained by randomly resampling with replacement from the spike counts for all cues, independently for each neuron. In all analyses, outliers beyond 3 standard deviations of the population mean were removed for population tests and descriptive statistics are indicated by mean \pm s.e.m.

Computational models

We tested different network representations for memory maintenance in three computational network models, the bump attractor network, the discrete attractor network and the decaying bump network. We provide Matlab code for these models, including all relevant parameters, as **Supplementary Notes**. In brief, the models were firing-rate network models with $N = 512$ excitatory neurons and 512 inhibitory neurons labeled by an angle θ_i used for decoding, characterized by an input-output function $r = \phi(I)$, mutually coupled via all-to-all connectivity matrices W_{EE} , W_{EI} , W_{IE} , W_{II} and subject to a white noise Gaussian input $\xi(t)$ of standard deviation σ .

$$\tau_E \frac{dr_i^E}{dt} = -r_i^E + \phi\left(\sum_j (W_{EE}^{ij} r_j^E - W_{EI}^{ij} r_j^I) + I_m + I_0^E\right) + \sigma_E \xi(t)$$

$$\tau_I \frac{dr_i^I}{dt} = -r_i^I + \phi\left(\sum_j (W_{IE}^{ij} r_j^E - W_{II}^{ij} r_j^I) + I_0^I\right) + \sigma_I \xi(t)$$

Connectivity matrices were all homogeneous ($W^{ij} = W_0$) except for connectivity among excitatory neurons, which had different patterns depending on the network model: for the bump attractor network W_{EE}^{ij} was a circular Gaussian function of $i - j$ (inset in **Fig. 8a**, left), for the discrete attractor network W_{EE}^{ij} took one of 5 possible values depending on which two of 8 populations neurons i and j belonged to (inset in **Fig. 8a**, middle), for the decaying bump network W_{EE}^{ij} was zero so no recurrent excitation was included. Instead, the decaying bump network included one intrinsic current in excitatory neuron that provided slowly decaying depolarization during the delay period:

$$\tau_m \frac{dI_m}{dt} = -I_m + \alpha_m \frac{r^E}{1 + e^{-2(r^E - 2)}},$$

for the other two network models $I_m = 0$. We simulated a task mimicking the behavioral task in the data: 1 s pre-stimulus, 500 ms stimulus presentation (increased current input to subset of neurons θ_i near stimulus location θ_s), and 3 s delay period where models evolved autonomously based on their dynamics. At the end of the delay period, the behavioral response was decoded using a population vector decoder ($\hat{\theta} = \arg\left(\sum_{j=1}^N r_j^E e^{i\theta_j}\right)$) and a global hyperpolarizing injected current erased selective network activity. Sample trials for each network model are illustrated in the **Supplementary Videos**.

For each model, we obtained a data set of neuron firing rates that mimicked our experimental data set: 200 neurons, with 8 different cues ($\theta_s = 0^\circ, 45^\circ, 90^\circ, 135^\circ, 180^\circ, 225^\circ, 270^\circ, 315^\circ$) and 10 independent trials per cue were sampled from a total of 16,000 independent trial simulations, and the behavioral response decoded from the population activity at the end of the delay was recorded. We applied the analyses of **Figs. 3,4** to these datasets in **Fig. 8**.

While these network models replicated the qualitative features of the experimental dataset, specific firing rate values, heterogeneity of tuning and near-Poisson spiking statistics could not be matched quantitatively in our simple firing rate network models (**Fig. 8**). For this reason, we further tested our analyses on surrogate data from two computational coding models that could match quantitatively the experimental data of **Fig. 1**, the bump attractor model and the decaying bump model (**Supplementary Figs. 5,6**). In both coding models, spike trains for individual trials were derived as inhomogeneous Poisson processes with time-varying firing rate $\lambda(t)$ described by an evolving bump of activity. In both cases, the bump shape was based on a von Mises distribution (a circular Gaussian) so that the firing rate

of neuron i during one trial took the form

$$\lambda_i(t, \theta_i) = r_{min} + (r_{max} - r_{min}) \frac{e^{\kappa(t)\cos(\theta_i - \theta(t))} - e^{-\kappa_0}}{e^{\kappa_0} - e^{-\kappa_0}},$$

where θ_i is the neuron's preferred location and $\theta(t)$ is the location of the bump center. The parameter $\kappa(t)$ determines the width of the bump. For large $\kappa(t)$, the standard deviation of the von Mises distribution is approximately $w(t) = 180/(\pi\sqrt{\kappa(t)})$, so we refer to $w(t)$ as the bump width. When $w(t) = w_0$, where $w_0 = 180/(\pi\sqrt{\kappa_0})$, the firing rate equation is normalized so that the parameters r_{min} and r_{max} determine each neuron's lowest and highest firing rates, respectively.

In the bump attractor model, the bump width was fixed at $w(t) = w_0 = 35^\circ$ for all trials. The variation across trials was due to diffusion in the bump center $\theta(t)$. In each trial, the bump center was initialized at the location of the cue presentation θ_s at the beginning of the delay period. During the delay period, the bump center evolved according to

$$\frac{d\theta}{dt} = \sigma\eta(t),$$

where $\eta(t)$ is Gaussian white noise and the diffusion magnitude $\sigma = 4.04$ was chosen so that the bump center had a standard deviation of 7° at the end of a 3-s delay period, similar to experimental data (**Fig. 1b**). For each neuron, the minimum and maximum firing rates were randomly drawn from Gaussian distributions so that $r_{min} = 7 \pm 1$ Hz and $r_{max} = 14 \pm 3.4$ Hz (mean \pm s.d.), with the constraint that $r_{min} < r_{max}$.

In the decaying bump model the bump center was fixed at the cue position $\theta(t) = \theta_s$ and its width was initially the same as in the bump attractor model $w(t) = w_0 = 35^\circ$. However, in each trial the bump width increased linearly with time,

$$\frac{dw}{dt} = \alpha,$$

where the speed α was chosen randomly for each trial from a gamma distribution with mean of 17.5 °/s and a standard deviation of 5.83 °/s. The parameters r_{min} and r_{max} were randomly drawn from Gaussian distributions so that $r_{min} = 7 \pm 1$ Hz and $r_{max} = 14 \pm 3.4$ Hz (mean \pm s.d.), with the constraint that

$r_{min} < r_{max}$. The parameters r_{min} and r_{max} specify the maximum and minimum firing rates only for the initial bump width when $w(t) = w_0$; the widening of the bump decreases the range of firing rates. This choice of parameters mimicked the experimental neural data and behavioral reports of **Fig. 1** quantitatively.

For each of the two models, we obtained a data set of neuron spike trains that mimicked our experimental data set: 200 neurons, with 8 different cues ($\theta_s = 0^\circ, 45^\circ, 90^\circ, 135^\circ, 180^\circ, 225^\circ, 270^\circ, 315^\circ$) and 10 independent trials per cue were sampled from a total of 16,000 independent trial simulations. Trials consisted in a fixation period of 1 s (homogeneous Poisson at rate r_{min}), a cue period of 0.5 s (formulas above with fixed $\theta(t) = \theta_s$ and $w(t) = w_0$), and a delay period of 3 s (time-varying formulas above).

For each individual trial that we simulated, we could also extract a behavioral response θ_{out} based on the firing rate at the end of the delay. To this end, we extracted Poisson spike counts for $N = 4,000$ neurons with evenly-spaced angles θ_i . Fixing the firing rates to those calculated at the end of the delay, we counted spikes for 0.5 s. From these spike counts $\{n_i, i = 1..N\}$ we computed the population vector $V = \sum_{i=1}^N n_i e^{i\theta_i}$, from which we extracted the decoded behavioral response θ_{out} : $V = |V| e^{i\theta_{out}}$. We treated this response similarly to the behavioral response of the monkey in our experimental data set.

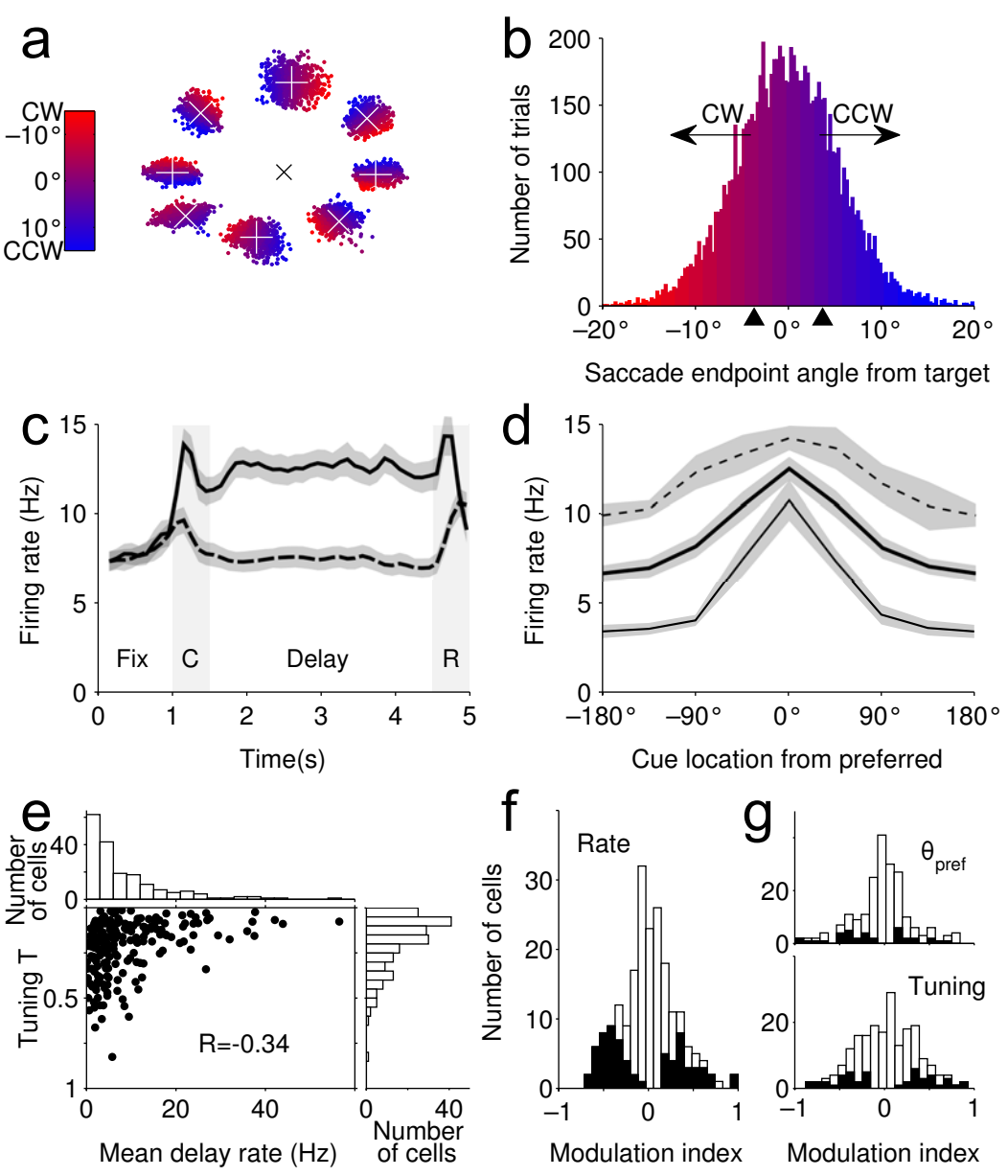
References

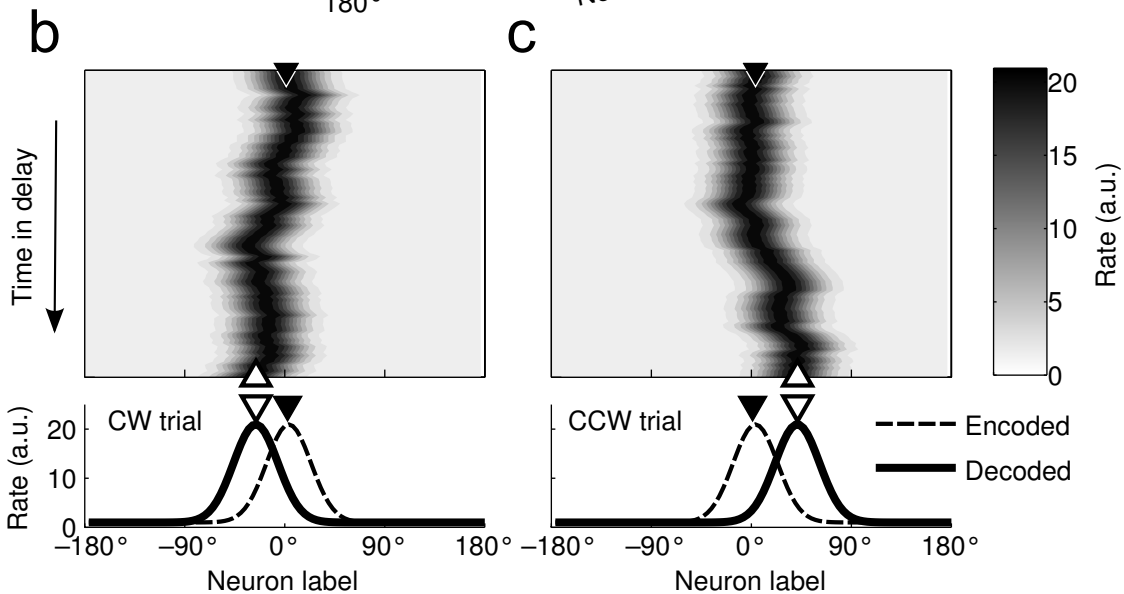
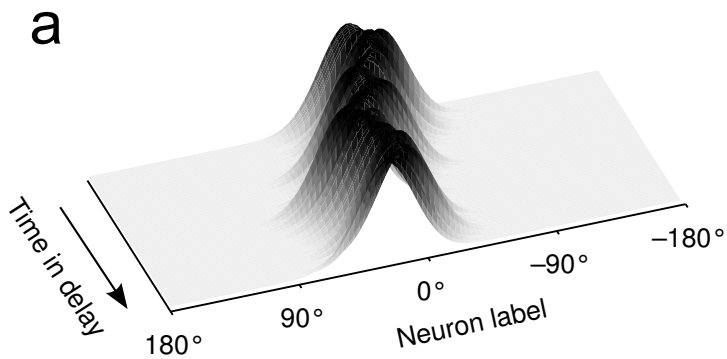
1. Gnadt, J. W. & Andersen, R. A. Memory related motor planning activity in posterior parietal cortex of macaque. *Exp Brain Res* **70**, 216–20 (1988).
2. Funahashi, S., Bruce, C. J. & Goldman-Rakic, P. S. Mnemonic coding of visual space in the monkey's dorsolateral prefrontal cortex. *J Neurophysiol* **61**, 331–49 (1989).
3. Constantinidis, C., Franowicz, M. N. & Goldman-Rakic, P. S. Coding specificity in cortical microcircuits: A multiple-electrode analysis of primate prefrontal cortex. *J Neurosci* **21**, 3646–3655 (2001).
4. Wilson, H. R. & Cowan, J. D. A mathematical theory of the functional dynamics of cortical and thalamic nervous tissue. *Kybernetik* **13**, 55–80 (1973).
5. Amari, S. Dynamics of pattern formation in lateral-inhibition type neural fields. *Biol Cybern* **27**, 77–87 (1977).
6. Ben-Yishai, R., Bar-Or, R. L. & Sompolinsky, H. Theory of orientation tuning in visual cortex. *Proc Natl Acad Sci U S A* **92**, 3844–3848 (1995).
7. Pouget, A., Zhang, K., Deneve, S. & Latham, P. E. Statistically Efficient Estimation Using Population Coding. *Neural Comput* **10**, 373–401 (1998).
8. Compte, A., Brunel, N., Goldman-Rakic, P. S. & Wang, X. J. Synaptic mechanisms and network dynamics underlying spatial working memory in a cortical network model. *Cereb Cortex* **10**, 910–23 (2000).
9. Brody, C. D., Romo, R. & Kepecs, A. Basic mechanisms for graded persistent activity: discrete attractors, continuous attractors, and dynamic representations. *Curr Opin Neurobiol* **13**, 204–211 (2003).
10. Constantinidis, C. & Wang, X.-J. A neural circuit basis for spatial working memory. *Neuroscientist* **10**, 553–65 (2004).
11. Miller, P. Analysis of Spike Statistics in Neuronal Systems with Continuous Attractors or Multiple, Discrete Attractor States. *Neural Comput* **18**, 1268–1317 (2006).
12. Mongillo, G., Barak, O. & Tsodyks, M. Synaptic Theory of Working Memory. *Science* **319**, 1543–1546 (2008).
13. Goldman, M. S. Memory without Feedback in a Neural Network. *Neuron* **61**, 621–634 (2009).
14. Barak, O., Tsodyks, M. & Romo, R. Neuronal Population Coding of Parametric Working Memory. *J Neurosci* **30**, 9424–9430 (2010).
15. Jun, J. K. *et al.* Heterogenous population coding of a short-term memory and decision task. *J. Neurosci.* **30**, 916–929 (2010).
16. Hussar, C. R. & Pasternak, T. Memory-guided sensory comparisons in the prefrontal cortex: contribution of putative pyramidal cells and interneurons. *J. Neurosci.* **32**, 2747–2761 (2012).
17. White, J. M., Sparks, D. L. & Stanford, T. R. Saccades to remembered target locations: an analysis of systematic and variable errors. *Vision Res* **34**, 79–92 (1994).
18. Ploner, C. J., Gaymard, B., Rivaud, S., Agid, Y. & Pierrot-Deseilligny, C. Temporal limits of spatial working memory in humans. *Eur. J. Neurosci.* **10**, 794–797 (1998).
19. Wang, X. J. Synaptic reverberation underlying mnemonic persistent activity. *Trends Neurosci* **24**, 455–63 (2001).
20. Wu, S., Hamaguchi, K. & Amari, S. Dynamics and Computation of Continuous Attractors. *Neural Computation* **20**, 994–1025 (2008).
21. Burak, Y. & Fiete, I. R. Fundamental limits on persistent activity in networks of noisy neurons.

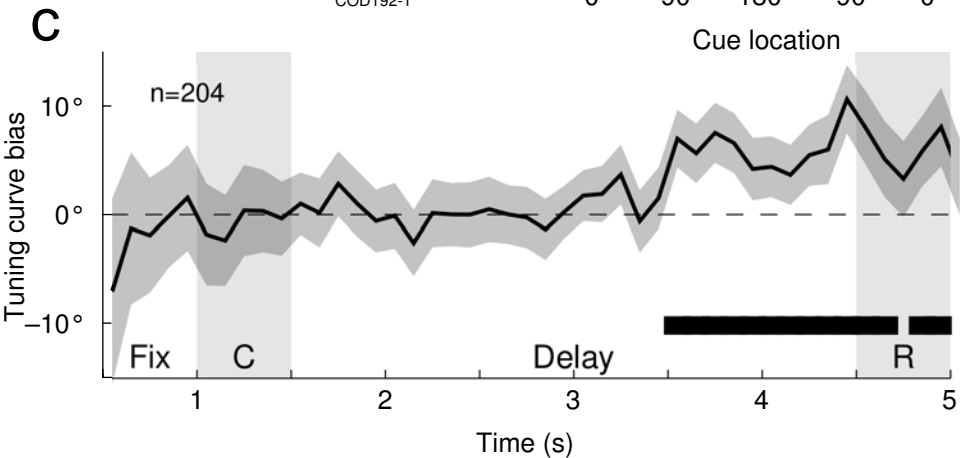
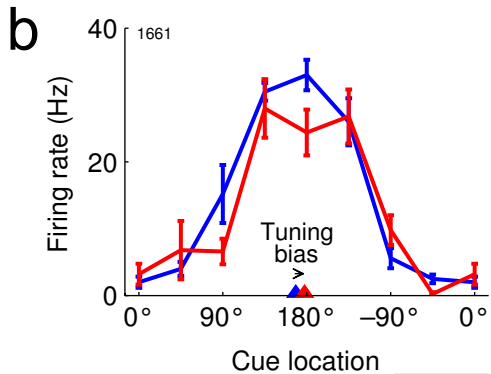
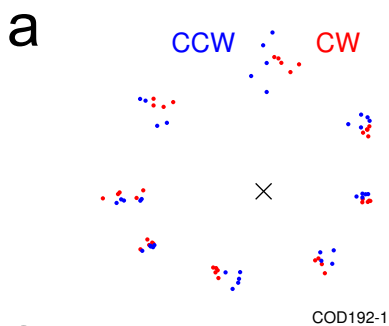
- PNAS* **109**, 17645–17650 (2012).
22. Shafi, M. *et al.* Variability in neuronal activity in primate cortex during working memory tasks. *Neuroscience* **146**, 1082–108 (2007).
 23. Constantinidis, C., Franowicz, M. N. & Goldman-Rakic, P. S. The sensory nature of mnemonic representation in the primate prefrontal cortex. *Nat. Neurosci.* **4**, 311–316 (2001).
 24. Shadlen, M. N. & Newsome, W. T. The Variable Discharge of Cortical Neurons: Implications for Connectivity, Computation, and Information Coding. *J. Neurosci.* **18**, 3870–3896 (1998).
 25. Churchland, A. K. *et al.* Variance as a signature of neural computations during decision making. *Neuron* **69**, 818–831 (2011).
 26. Goldman-Rakic, P. S. Cellular basis of working memory. *Neuron* **14**, 477–85 (1995).
 27. Durstewitz, D., Seamans, J. K. & Sejnowski, T. J. Neurocomputational models of working memory. *Nat. Neurosci* **3 Suppl**, 1184–91 (2000).
 28. Brody, C. D., Hernández, A., Zainos, A. & Romo, R. Timing and neural encoding of somatosensory parametric working memory in macaque prefrontal cortex. *Cereb. Cortex* **13**, 1196–1207 (2003).
 29. Zaksas, D. & Pasternak, T. Directional signals in the prefrontal cortex and in area MT during a working memory for visual motion task. *J Neurosci* **26**, 11726–42 (2006).
 30. Meyers, E. M., Freedman, D. J., Kreiman, G., Miller, E. K. & Poggio, T. Dynamic population coding of category information in inferior temporal and prefrontal cortex. *J Neurophysiol* **100**, 1407–1419 (2008).
 31. Compte, A. *et al.* Temporally irregular mnemonic persistent activity in prefrontal neurons of monkeys during a delayed response task. *J Neurophysiol* **90**, 3441–54 (2003).
 32. Renart, A., Song, P. & Wang, X.-J. Robust spatial working memory through homeostatic synaptic scaling in heterogeneous cortical networks. *Neuron* **38**, 473–485 (2003).
 33. Hansel, D. & Mato, G. Short-Term Plasticity Explains Irregular Persistent Activity in Working Memory Tasks. *J. Neurosci.* **33**, 133–149 (2013).
 34. Barbieri, F. & Brunel, N. Irregular persistent activity induced by synaptic excitatory feedback. *Front Comput Neurosci* **1**, 5 (2007).
 35. Britten, K. H., Newsome, W. T., Shadlen, M. N., Celebrini, S. & Movshon, J. A. A relationship between behavioral choice and the visual responses of neurons in macaque MT. *Visual Neuroscience* **13**, 87–100 (1996).
 36. Jun, J.K., Miller, P., Hernández, A., Zainos, A., Lemus L., Brody, C.D. & Romo, R. Heterogeneous population coding of a short-term memory and decision task. *J. Neurosci.* **30**, 916–929 (2010).
 37. Machens, C. K., Romo, R. & Brody, C. D. Functional, But Not Anatomical, Separation of ‘What’ and ‘When’ in Prefrontal Cortex. *J. Neurosci.* **30**, 350–360 (2010).
 38. Kilpatrick, Z.P., Ermentrout, B. & Doiron, B. Optimizing working memory with heterogeneity of recurrent cortical excitation. *J. Neurosci.* **33**, 18999–19011 (2013).
 39. Miller, E. K. The prefrontal cortex and cognitive control. *Nat. Rev. Neurosci.* **1**, 59–65 (2000).
 40. Meyer, T., Qi, X.-L. & Constantinidis, C. Persistent Discharges in the Prefrontal Cortex of Monkeys Naive to Working Memory Tasks. *Cereb Cortex* **17**, i70–i76 (2007).
 41. Kohn, A. & Smith, M. A. Stimulus dependence of neuronal correlation in primary visual cortex of the macaque. *J Neurosci* **25**, 3661–73 (2005).
 42. Bair, W., Zohary, E. & Newsome, W. T. Correlated Firing in Macaque Visual Area MT: Time Scales and Relationship to Behavior. *J. Neurosci.* **21**, 1676–1697 (2001).
 43. Cohen, M. R. & Newsome, W. T. Context-Dependent Changes in Functional Circuitry in Visual Area MT. *Neuron* **60**, 162–173 (2008).

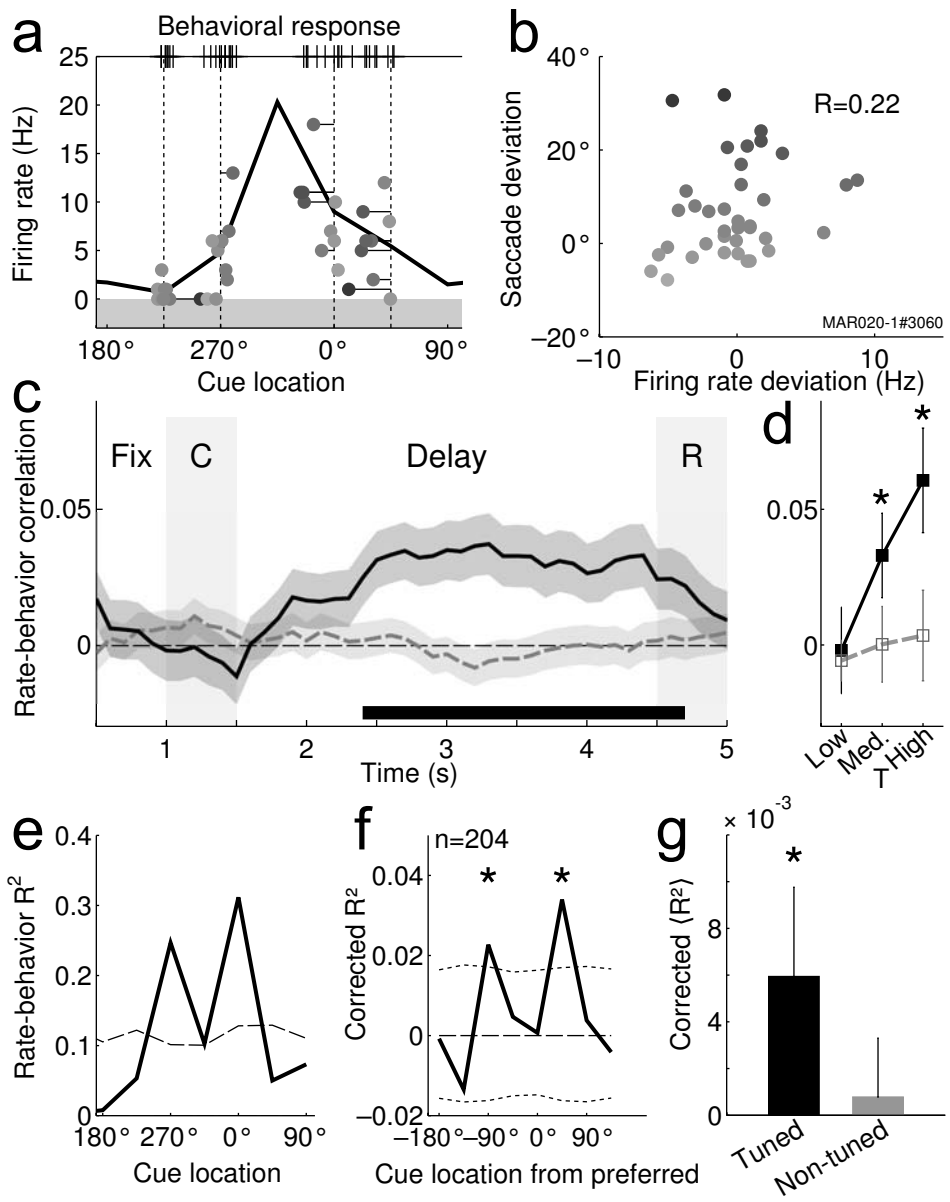
44. Cohen, M. R. & Maunsell, J. H. R. Attention improves performance primarily by reducing interneuronal correlations. *Nat Neurosci* **12**, 1594–1600 (2009).
45. Mitchell, J. F., Sundberg, K. A. & Reynolds, J. H. Spatial Attention Decorrelates Intrinsic Activity Fluctuations in Macaque Area V4. *Neuron* **63**, 879–888 (2009).
46. Amit, D. J., Fusi, S. & Yakovlev, V. Paradigmatic working memory (attractor) cell in IT cortex. *Neural Comput* **9**, 1071–1092 (1997).
47. Leutgeb, J. K. *et al.* Progressive transformation of hippocampal neuronal representations in ‘morphed’ environments. *Neuron* **48**, 345–358 (2005).
48. Wills, T. J., Lever, C., Cacucci, F., Burgess, N. & O’Keefe, J. Attractor dynamics in the hippocampal representation of the local environment. *Science* **308**, 873–876 (2005).
49. Durstewitz, D., Vitoz, N. M., Floresco, S. B. & Seamans, J. K. Abrupt transitions between prefrontal neural ensemble states accompany behavioral transitions during rule learning. *Neuron* **66**, 438–448 (2010).
50. Balaguer-Ballester, E., Lapish, C. C., Seamans, J. K. & Durstewitz, D. Attracting Dynamics of Frontal Cortex Ensembles during Memory-Guided Decision-Making. *PLoS Comput Biol* **7**, e1002057 (2011).
51. Constantinidis, C., Williams, G. V. & Goldman-Rakic, P. S. A role for inhibition in shaping the temporal flow of information in prefrontal cortex. *Nat Neurosci* **5**, 175–180 (2002).
52. Constantinidis, C. & Goldman-Rakic, P. S. Correlated discharges among putative pyramidal neurons and interneurons in the primate prefrontal cortex. *J Neurophysiol* **88**, 3487–3497 (2002).
53. Jin, D. Z., Dragoi, V., Sur, M. & Seung, H. S. Tilt Aftereffect and Adaptation-Induced Changes in Orientation Tuning in Visual Cortex. *J Neurophysiol* **94**, 4038–4050 (2005).
54. Compte, A. & Wang, X.-J. Tuning curve shift by attention modulation in cortical neurons: a computational study of its mechanisms. *Cereb Cortex* **16**, 761–78 (2006).

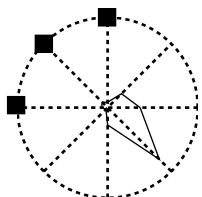
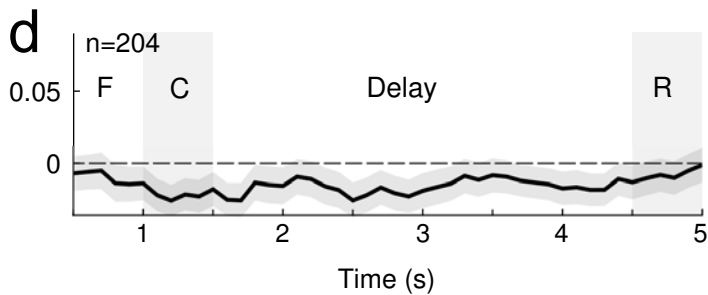
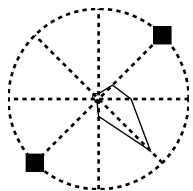
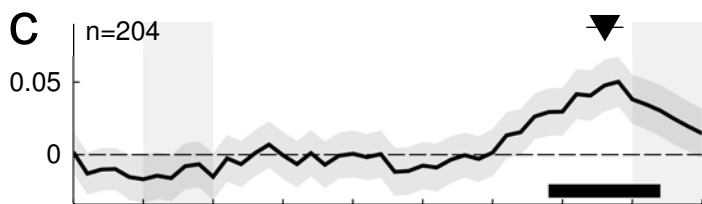
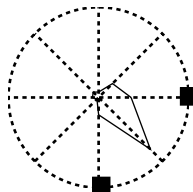
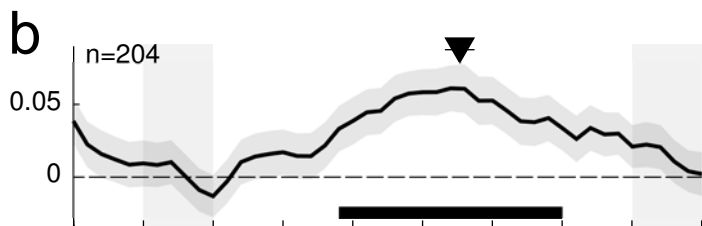
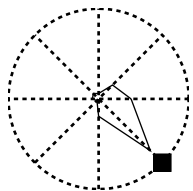
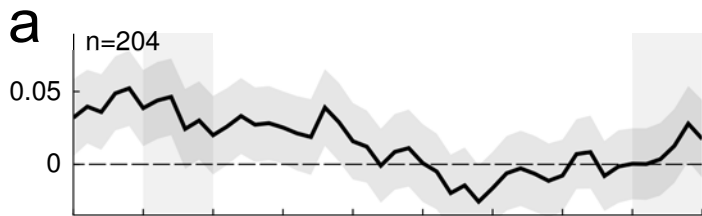
References in blue are cited only in Online Methods.

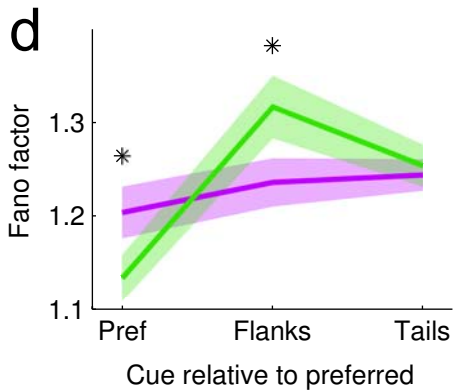
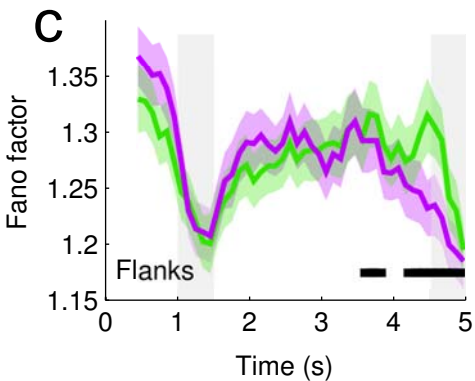
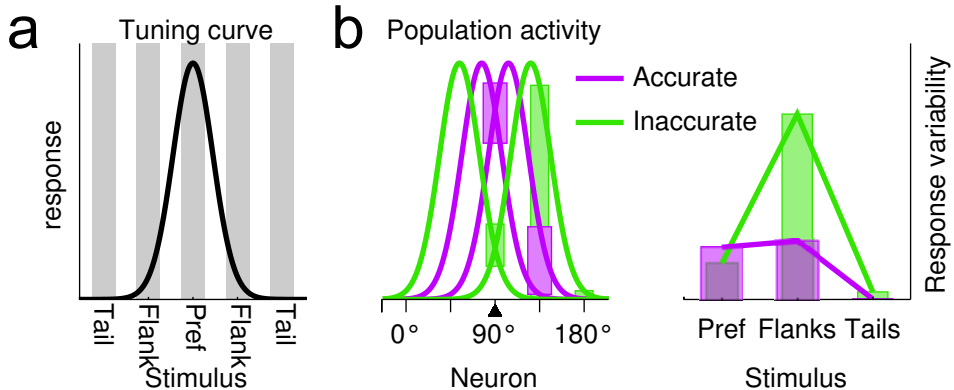


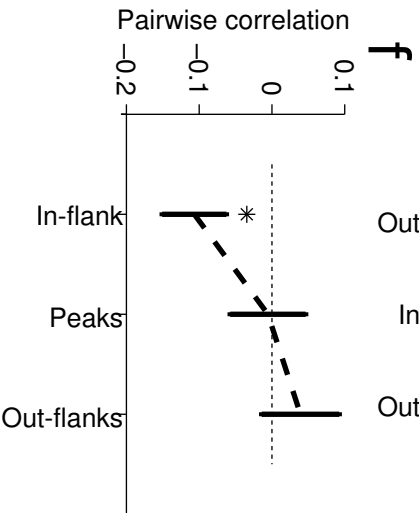
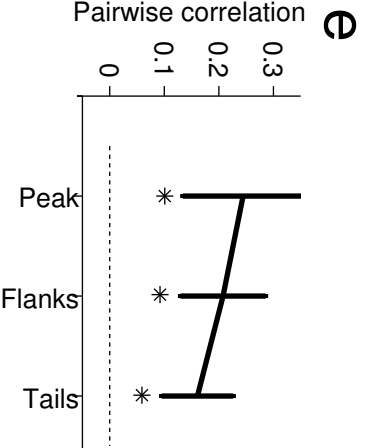
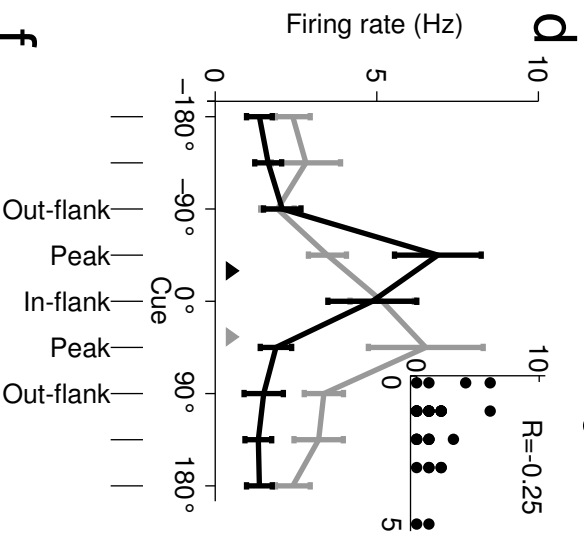
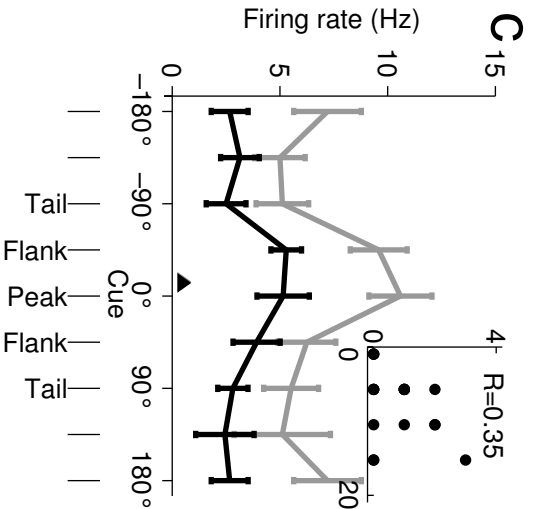
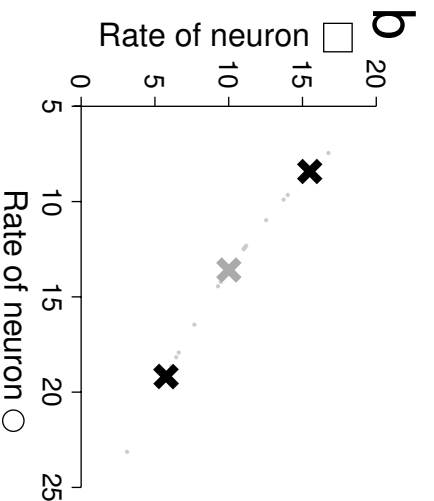
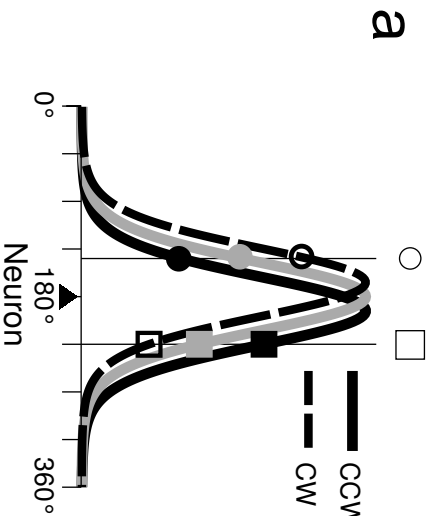








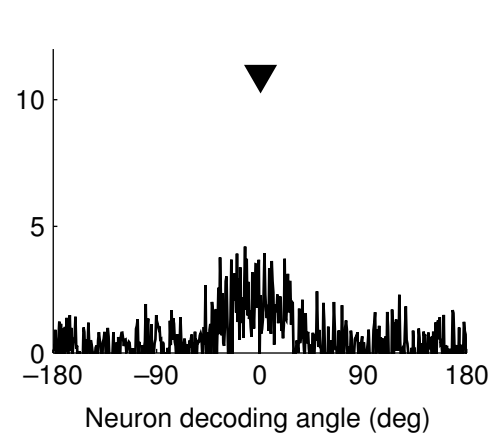
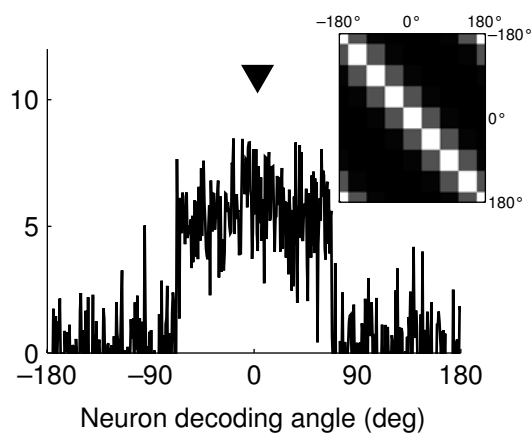
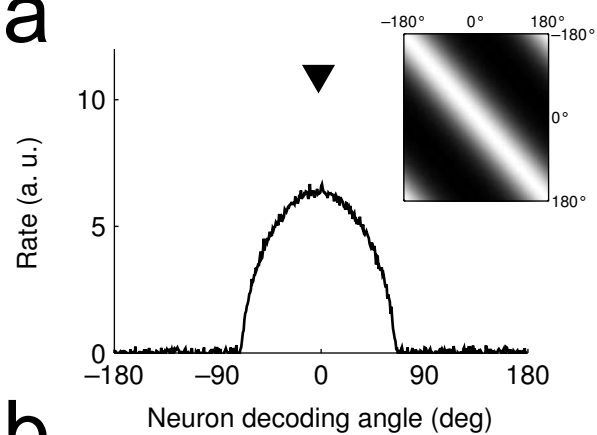
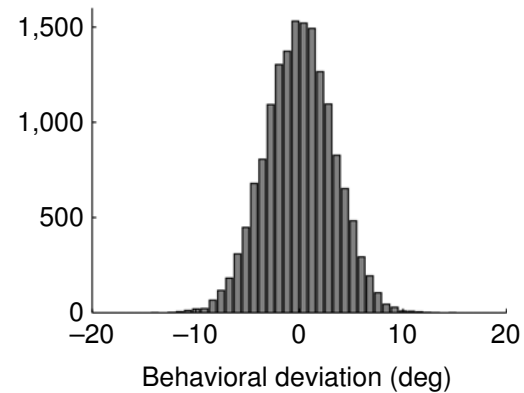
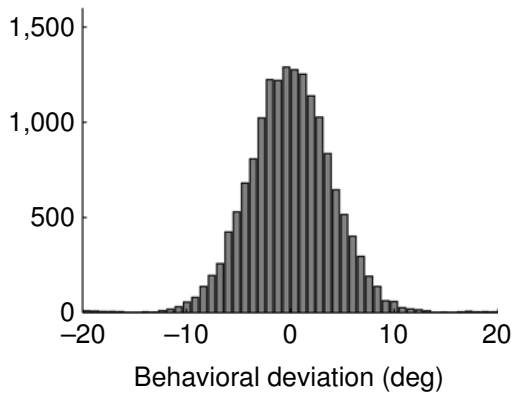
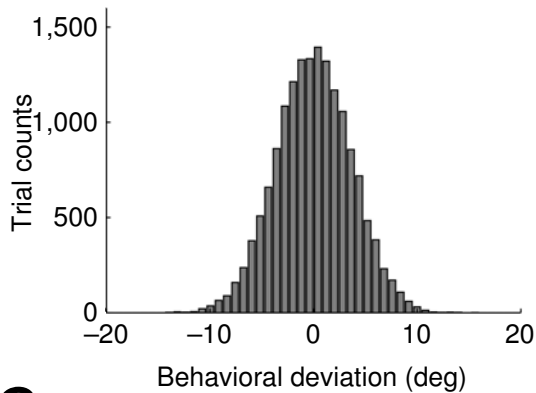
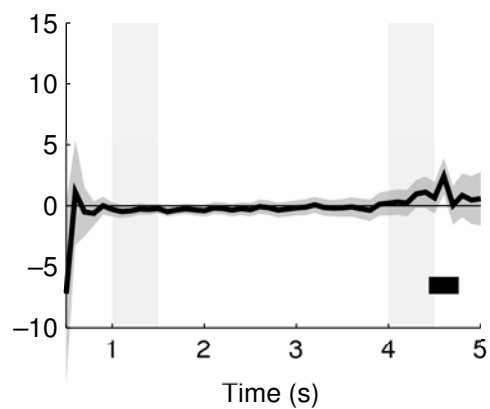
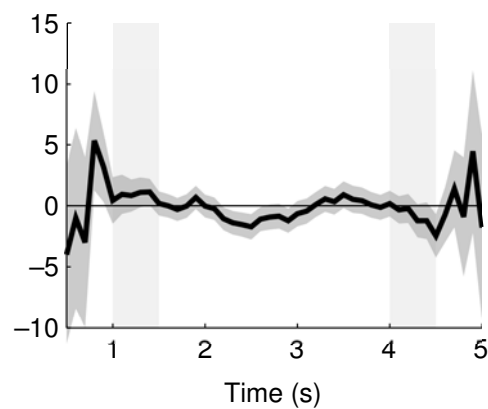
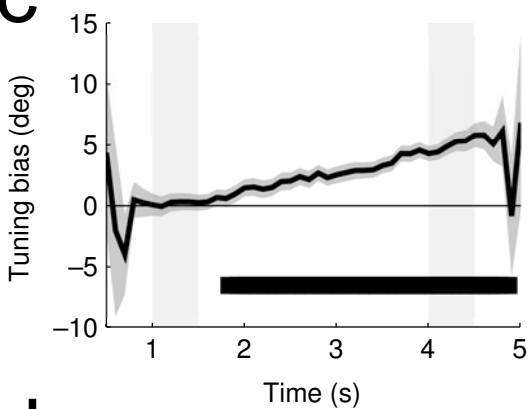




Bump attractor network

Discrete attractor network

Decaying bump network

a**b****c****d**