Computing by robust transience: How the fronto-parietal network performs sequential category-based decisions

Warasinee Chaisangmongkon[1,2], Sruthi K. Swaminathan[3], David J. Freedman[3,4], and Xiao-Jing Wang[1,5,6]

[1]Yale University School of Medicine, Department of Neurobiology and Kavli Institute for Neuroscience, New Haven, CT

[2]Institute of Field Robotics, King Mongkut's University of Technology Thonburi, Bangkok, Thailand

[3]The University of Chicago, Department of Neurobiology, Chicago, IL

[4]Grossman Institute for Neuroscience, Quantitative Biology, and Human Behavior, Chicago, IL

[5]New York University, Center for Neural Science, New York, New York

[6]NYU-ECNU Joint Institute of Brain and Cognitive Science, NYU-Shanghai, Shanghai, China

[*] Correspondence: xjwang@nyu.edu

Summary

Decision making involves dynamic interplay between internal judgements and external perception, which has been investigated in delayed match-to-category (DMC) experiments. Our analysis of neural recordings shows that, during DMC tasks, LIP and PFC neurons demonstrate mixed, time-varying, and heterogeneous selectivity, but previous theoretical work has not established the link between these neural characteristics and population-level computations. We trained a recurrent network model to perform DMC tasks and found that the model can remarkably reproduce key features of neuronal selectivity at the single-neuron and population levels. Analysis of the trained networks elucidates that robust transient trajectories of the neural population are the key driver of sequential categorical decisions. The directions of trajectories are governed by network self-organized connectivity, defining a 'neural landscape', consisting of a task-tailored arrangement of slow states and dynamical tunnels. With this model, we can identify functionally-relevant circuit motifs and generalize the framework to solve other categorization tasks.

Introduction

Many human behaviors can be viewed as a series of category-based computations (Roelfsema et al., 2003; Rabinovich and Varona, 2011). For example, shopping requires determining categories of desired items then using that category information to guide our navigation of the store. Neurophysiological studies have investigated the neural basis of such behavior using delayed match to category (DMC) tasks in which monkeys indicate whether a test stimulus is a categorical match to a previously presented sample stimulus. Previous work revealed that, in

animals trained to perform DMC tasks, neural responses in the lateral intraparietal area (LIP) and prefrontal cortex (PFC) can exhibit robust selectivity for the category of the sample stimulus which persists across the memory delay period (Freedman et al., 2001; Freedman and Assad, 2006; Swaminathan and Freedman, 2012).

Few computational models have addressed the neural mechanisms underlying sequential category computations as in the DMC task. One class of models suggests that serial categorical decisions rely on an interplay between multiple subpopulations, each encoding specific task parameters, such as stimulus features, categories, rules, and choices (Amit et al., 1994; Ardid and Wang, 2013; Engel and Wang, 2011). These subpopulations are endowed with strong mutual excitations among neurons that prefer the same stimulus feature, driving a stable memory of that feature (Wong and Wang, 2006; Wang, 2001). These models elucidated candidate mechanisms of delay-period persistent selectivity, but they are not designed to address the diversity or temporal variability in neural responses, which are apparent in neural data. In particular, the majority of neurons are responsive to multiple task variables (Raposo et al., 2014; Rigotti et al., 2013; Ibos and Freedman, 2014; Mante et al., 2013; Ibos and Freedman, 2016; Freedman and Assad, 2016) and neural encoding often shows baffling temporal variability (Brody et al., 2003a; Crowe et al., 2010; Jun et al., 2010; Shafi et al., 2007). To address these phenomena, recent studies considered an alternative hypothesis that information is distributed randomly within the neural network (Rigotti, et al. 2010; Raposo et al., 2014). The idea can be implemented with a network with random connectivity and, to generate different behaviors, downstream circuits can read out relevant information through optimized synaptic weights (Jaeger, 2001; Maass et al., 2002; Rigotti et al., 2010). However, random networks generally do not capture task-specific representations, which can only be acquired through learning. In this realm, we lack a unified framework that can recapitulate all these diverse experimental findings.

To address this problem, we trained a recurrent network model to solve DMC tasks and compared the dynamics of the model network to LIP and PFC data from monkeys performing the DMC task (Freedman and Assad, 2006; Swaminathan and Freedman, 2012). We found that appropriately trained networks reproduce key features of category-dependent responses in the neural data not accounted for by previous models.

Results

Delayed-match-to-category task and model architecture

We analyzed neural recordings from the studies of Freedman and Assad (2006) and Swaminathan and Freedman (2012), where Macaque monkeys were trained on delayed-match-to-category (DMC) tasks. In each DMC trial, the stimulus sequence consists of a fixation spot, a sample stimulus, a delay period, and a test stimulus (Figure 1A). Both sample and test stimuli were randomly drawn from a set of random-dot motion directions evenly spaced from 0° to 360° and divided arbitrarily

into two categories (marked by red and blue color in Figure 1B). Subjects learned to report whether the sample and test stimuli belong to the same category (match) or different categories (non-match). The first and second study used 12 and 6 evenly spaced motion directions (30° and 60° apart), respectively. For the first study, we have 156 lateral intraparietal (LIP) from two monkeys. For the second study, we have 74 LIP and 380 prefrontal neurons (PFC) in two other monkeys. We refer to the LIP populations from the first and second experiments as the LIP1 and LIP2, respectively.

Previous studies showed that firing rates of LIP and PFC neurons are markedly tuned to the learned stimulus categories, although the strengths and latencies of categorical signals may differ across areas (Freedman and Assad, 2006; Swaminathan and Freedman, 2012; Swaminathan et al., 2013). The broad similarity in category-related responses suggest that they play overlapping roles in solving the DMC task (Goodwin et al., 2012; Merchant et al., 2011). Instead of stressing the difference between the two regions, this work focuses on understanding the common response patterns observed in both areas.

We trained a recurrent neural network to solve the DMC task. The recurrent network represents a cortical microcircuit in either prefrontal or parietal region, which receives sensory information from visual areas and sends signals to trigger movements in motor areas (Andersen et al., 1990; Cromer et al., 2011; Lewis and Van Essen, 2000; Miller et al., 2002). The network is sparsely connected to noisy input neurons that encode the direction of the sample and test stimuli, mimicking direction-tuned activity in area MT (Figure 2B, top panel, Freedman and Assad, 2006; Born and Bradley, 2005). A subset of the recurrent population is connected to two action neuron pools, whose activities reflect match or non-match decisions. All connections (input, output, and recurrent) are trained with a supervised method (a Hessian-free algorithm, Martens and Sutskever, 2011; Mante et al., 2013), which adjusts synaptic weights to minimize the difference between the network outputs and specified target responses (i.e. to minimize errors). We instructed the match neuron to hold activity at zero from the beginning of the trial through the delay period, then reach a value of 5 (in arbitrary units) at 200 ms after test stimulus onset on match trials or remain at zero throughout non-match trials. The analogous pattern holds for the non-match neurons. To determine the model's choice, the action neurons' activity is passed through a nonlinear threshold function (Figure 2B, bottom panels; see also STAR Methods). The match (or non-match) choice is selected when the function value of match (or non-match) neuron is higher than a threshold of 0.85. We added other output neurons to help stabilize the networks during resting and post-choice period (see Figure S1A-B).

There are many network configurations that can produce the appropriate output given the specified sensory input (Mante et al., 2013; Barak et al., 2013; Sussillo, 2014). We guide the algorithm to find a subset of solutions that comply with biological constraints by employing additional training strategies (see Figure S1 and STAR Methods). First, the activity of recurrent neurons is restricted to positive values, as is true for neuronal firing rates. Second, the network is trained to not only

minimize errors but also to attain sparse synaptic connections. This is achieved by constraining the norm of synaptic weights and eliminating weak synapses iteratively. The target probability of connection is approximately 12%, comparable to measurements from mammalian cortical circuits (Song et al., 2005). Third, single neurons should exhibit low spontaneous firing rates when the network is not performing the task. To satisfy this requirement, we instructed the network to hold the sum of all neurons' activity to a small value for 1 s before the trial onset (mean activity $= 0.01$). Fourth, we employed a progressive training protocol similar to that used to train monkeys on the DMC task (Freedman and Assad, 2006; Swaminathan and Freedman, 2012), whereby the network first started by learning the easiest version of DMC task with only two stimuli, then intrinsic noise and more stimuli are gradually introduced. Lastly, the delay durations varied slightly from trial to trial during training ($0.9 - 1.1$ ms), resulting in a model that can perform the task with a larger range of delays (Barak et al., 2013) (Figure S1C). These constraints and modifications greatly enhanced success rate and the quality of training outcomes.

Training was terminated when the accuracy of the model matched the average performance of animal subjects (88.76%). Both the model and monkeys are less accurate when categorizing near-boundary stimuli (e.g. 15° degrees away from the boundary, Figure 2C).

The resulting trained networks provide a candidate dynamical mechanism for solving DMC tasks. To test whether the trained network uses a similar mechanisms as real neuronal networks, the model activity must be compared to the recorded experimental neural data.

Heterogeneity in temporal profiles of category selectivity

We compared the temporal profiles of category selectivity of LIP and PFC neural recordings to those of trained model networks. To quantify the temporal properties of category selectivity, we tested whether the firing rates at each time window are significantly modulated by stimulus categories (t-tests, $p < 0.05$, Bonferroni corrected). A 'category selectivity phase' is defined as a series of consecutive time windows where neural activity shows significant modulation by stimulus categories. The neuron's category selectivity duration is the duration of its longest selectivity phase. The strength of category encoding for each time bin within selectivity phases was quantified by the sensitivity index or d' (see STAR Methods).

Figure 3A illustrates variability in the temporal profiles of category-dependent firing rates for LIP and PFC neurons. Many neurons showed persistent category-dependent responses during the delay period (Figure 3A, first panel). In the majority of neurons, the category-selective firing rates undergo marked changes through the delay period. For instance, the category-dependent firing pattern may decay before the end of the delay (Figure 3A, second panel) or may commence in the middle of delay (Figure 3A, third panel). Furthermore, some neurons switch their category preference in the middle of a trial (Figure 3A, last panel). In congruence to the neural data, the trained model exhibits heterogeneity in category selectivity

profiles, in which persistent, transient, and switching selectivity patterns are observed (Figure 3B). We also quantified fractions of neurons with persistent, transient and switching selectivity and found that neural data and model show similar trends (Figure S2A).

To visualize the heterogeneity at the population level, we plotted d' of all neurons sorted by the onset of category selectivity from earliest to latest (Figure 3C-D). All heat maps of neural data confirmed two important observations (LIP2 dataset, Figure 3C; LIP1 and PFC datasets, Figure S2B-C). First, neurons can become selective to categories at any time point during the trial. The category selectivity phase does not necessarily align with or overlap the sample stimulus, which originates the category memory. Second, we observed heterogeneity in the duration of category selectivity across the population. The distribution of category selectivity durations shows a long tail, whereby most neurons exhibit selectivity over short durations but a small fraction of persistent neurons are consistently detected in all datasets (Figure S2D-G). The pattern of category selectivity in model populations reproduces all main features of the neural data.

Furthermore, our trained networks reproduce mixed category and match selectivity, which is evident in our neural data (Figure S2H-I) and other studies (Ibos and Freedman, 2014; Ibos and Freedman, 2016; Park et al., 2014; Rishel, 2013; Mante et al., 2013; Rigotti et al., 2013). This suggests that a large portion cortical neurons and model neurons participate in more than one computation.

Population response trajectories

We investigated whether neural data and the model exhibit similar patterns of population response trajectories. To this end, we first analyzed the dynamics of neural population responses in LIP and PFC. A neural state is a point in high-dimensional state space, where each dimension corresponds to the average firing rate of a neuron at a given time. As neural activity changes over time, a sequence of neural states at consecutive time points forms a population trajectory through state space. For each population, we visualized neural trajectories in a low-dimensional subspace that is most responsive to task conditions using demixing principal component analysis (Machens, 2010; Brendel et al., 2011; Machens et al., 2010) (DPCA), which finds a small set of orthogonal axes that not only capture the most variance in data (like standard PCA), but also segregates response variability due to different task variables onto separate axes. In our case, DPCA yields population response components ranging from one that captures the most variance due to task conditions to one that captures the most variance due to changes in time. We applied DPCA to the mean population response during the sample ($-100 - 650$ ms relative to sample onset), delay ($800 - 1550$ ms) and test ($1600 - 2150$ ms) epochs separately (see STAR Methods). Note that task conditions were defined by sample motion directions during the sample and delay periods, and by sample categories and test directions during the test period.

For the sample period, we applied DPCA, removed the most time-sensitive component and represented the remaining components on a 2D axis by a multidimensional scaling analysis (Figure 4A-C; see STAR Methods). At the beginning of the sample epoch, population trajectories originated at the same baseline for all stimuli (black dots), then they fan out radially, discriminating different sample directions. At the end of stimulus presentation (colored dots), neural states for all stimulus directions appear in an elliptical configuration, where the stimuli at the middle of both categories (dark blue and red dots) elicit more distinct population responses than stimuli close to category boundaries (light blue and red dots). Overall, LIP and PFC populations show a mixed encoding of sample directions and categories, consistent with our earlier report in Engel et al. (2015).

For the delay epoch, we applied DPCA to delay responses ($800 - 1550$ ms) and projected responses of the whole trial ($-250 - 1900$ ms) onto DPCA axes (see STAR Methods). Here we show two components that participate in the maintenance of categorical working memory (Figure 4E-G; see also Figure S3). Components in the first column capture the most variance in the delay response due to changes in stimuli (35.5%, 66.3%, and 53.2% for LIP1, LIP2, and PFC, respectively), which constitute a much larger proportion than the second largest component (4.9%, 4.0%, and 10.2%). The components in the first column depict strong and stable encoding of sample categories, representing the main mode of working memory (Figure 4E-G, left panels). The second column shows components with the largest mixture of variance due to changes in time and stimuli, and the neural traces show time-varying category working memory that switches categorical preference mid-delay (Figure 4E-G, right panels). Notably, neural trajectories during the sample and delay epochs show that LIP and PFC populations encode categories by several independent components with different temporal profiles.

For the test epoch, the procedure was similar to the sample epoch, except that the neural response was averaged across trials that share the same sample category and test direction. The neural trajectories are grouped into 4 conditions (Figure 4I-K); BB (dark blue color) corresponds to trials with blue sample category and blue test category; RR (dark red) to red sample and red test; BR (purple) to blue sample and red test; RB (orange) to red sample and blue test. At the beginning of the test period, neural trajectories are clustered according to the sample categories (black dots). As the test period evolves, the neural traces diverge into four separate clusters encoding sample and test category combinations (BB, RR, BR, and RB conditions). Finally, the trajectories corresponding to match conditions (BB and RR) travel toward the same location and analogously so for non-match conditions (BR and RB). Overall, the test-related trajectories encode sample-test category combinations and form states corresponding to match and non-match decisions.

The population trajectories of the trained model networks remarkably reproduce the main features of those from LIP and PFC for all task epochs, when the same analyses are applied (Figure 4D,H,L). Mixed direction and categorical encoding is apparent during the sample period (Figure 4D). The population dynamics during the delay encode categories in both stable and time-varying manners (Figure 4H). Lastly

population responses encode sample-test category conditions and converge toward match or non-match states during the test period (Figure 4L). Note that although the model incorporates a large amount of noise and heterogeneity, the neural data tend to show more variability, especially variance due to changes in time within the trial, which may reflect a timing signal not incorporated in our model (Figure S3). Furthermore, the model tends to be more category-selective than the data, perhaps because the model is a smaller network exclusively trained on DMC tasks.

Robust transient dynamics underlie delayed-match-to-category computations

We characterized the dynamical mechanisms of the model focusing on two objectives. First, the existence and abundance of mixed selectivity neurons (Ibos and Freedman, 2014; Ibos and Freedman, 2016; Mante et al., 2013; Park et al., 2014; Raposo et al., 2014; Rigotti et al., 2013; Rishel, 2013) and the time-varying selectivity for task-relevant variables at both single-neuron (Brody et al., 2003a; Jun et al., 2010; Shafi et al., 2007) and population levels (Machens, 2010; Crowe et al., 2010; Meyers et al., 2008; Wohrer et al., 2013) have sparked a debate on the dynamical nature of working memory (Druckmann and Chklovskii, 2012; Goldman, 2009; Sarma et al., 2016; Savin and Triesch, 2014; Singh and Eliasmith, 2006). Since the model captures all these features, it is possible now to pinpoint the working memory dynamics that give rise to these patterns of selectivity. Second, we sought to understand how the sequential categorical computations are carried out, i.e. how the sample category information is encoded, maintained, and combined with the test category to generate appropriate behavioral choices.

Figure 5A-B illustrates the overall trajectories of the model network during DMC tasks, visualized by plotting the largest three principal components of the neural activity. The state space contains key attracting fixed points or unstable saddle points (at which neural activity has near-zero velocity). While performing the task, the network undertakes slow and reliable transitions through these key regions. The directions of movement are determined by the current state location and input. Neural states evolve from the resting state (gray cross, Figure 5A-B) to states associated with sample categories (red and blue dots mark the end of sample period, Figure 5A-B), then category working memory (red and blue stars mark the end of delay period), sample-test category combinations (dark blue, dark red, orange, and purple lines), and finally match or non-match decision states (green and brown crosses). The whole series of state transitions solves the DMC task.

We characterized the network dynamics at each time epoch in more detail. During rest and fixation periods, the network trajectories are confined within the resting state's basin of attraction despite the noisy sensory signals the network receives (gray traces, Figure 6A inset). When the system receives a stimulus-selective input, stimuli in different categories propel neural states to separate directions (Figure S4A) and, over time, out of the resting state basin (Figure 6A and Figure S4B). If the network inputs stay on for a long period, the network would converge to stimulus-dependent steady states (red and blue crosses, Figure 6A), which are clustered based on stimulus categories. The arrangement of input-dependent stable states

(Rabinovich et al., 2001, 2008) results in directional neural trajectories that distinguish between stimulus categories.

At the end of the sample period, the network's slow and transient states are distributed within two regions in state space (red and blue dots, Figure 6B; variability in locations is a result of noise and network's low velocity is shown in Figure S4C). Using these states as initial conditions, we simulated network activity without input and noise. We observed that network states relax along two narrowing tunnels, one for each sample category, maintaining category memories in a dynamic manner (Figure 6B). The velocity vector field near one of the tunnel centers is plotted in Figure 6C, whereby arrow lengths indicate relative velocity magnitude. The plot shows that the neural state moves more slowly as it approaches the end of the tunnel (see also Figure S4D) and arrow directions point toward the middle of the tunnel funneling the system's state to a specific region near the end of the delay. This dynamical analysis revealed that categorical working memory is maintained by robust trajectories, which explains why we observed time-varying selectivity at both single-neuron (Figure 3) and population (Figure 4) levels.

Note that states associated with stimuli near the category boundary are closer to each other at the beginning of the delay, compared to states of stimuli further away from the boundary (pale red and blue lines in Figure 6A and Figure 6B). Misclassification occurs when the end-of-sample states stray outside of the tunnel under the influence of noise and end up in the wrong categorical tunnel or in the basins associated to rest state or choices (Figure S4E). This takes place more often for near-boundary stimuli, leading to poorer performance as shown in Figure 2C.

The end-of-delay regions are in the proximity of saddle or stable points (stability of this state varies across networks trained by an identical protocol), leading to low network velocity and keeping the memory of categories for an extended period. This allows the networks to perform well even when delay durations vary (accuracy $>$ 70% in the range of $0.7 - 1.3$ s delay) (Figure S1C). Note that if the delay is prolonged much longer than 1 s, our simulation shows two possible outcomes. First, if stable states associated to categorical working memory emerged during training, neural states simply rest in stable states. We observe that category-related fixed points are likely to emerge if the network is trained with variable delay duration randomly drawn from a larger range ($0.8 - 2$ s, Figure S5). Second, prolonged delays may lead to a gradual decay of working memory, whereby the network collapses to fixed points associated to resting state or random choices. The neural datasets we investigate cannot distinguish between these two scenarios.

At the onset of the test period, neural states are distributed between two regions of state space associated to red and blue categories of working memory (red and blue dots in Figure 6D); variability in state locations is due to noise. When the test stimulus is introduced, the direction-selective input shifts the landscape, in the same fashion as in the sample period (Figure 6A), directing the network towards stimulus-dependent stable states. However, since trajectories are launched from two possible initial locations depending on the sample category, the neural paths

split into four separate streams encoding sample-test category combinations (dark blue, dark red, purple, and orange lines in Figure 6D), bringing neural states to four separate clusters of transient states. This dynamical picture provides a concrete example of state-dependent computations, where the same stimulus can be interpreted differently or lead to different behavioral outcomes depending on the prior experience of the network (Buonomano and Maass, 2009).

Soon after the test stimulus appears, the match (or non-match) output neuron can read out from the recurrent neural states and ramp up to response threshold during the match (or non-match) trial. The response time is usually within a few hundred milliseconds after the test stimulus onset. Finally, after the response is committed and test stimulus is removed, the network relaxes along its natural landscape. The four regions in state space (dots in Figure 6E) are mapped onto two steady states (crosses in Figure 6E); RR and BB traces go to one point (match attractor), while RB and BR traces go to a separate point (non-match attractor). These stable states complete the sequence of DMC computations.

The dynamics of the model reveal that a single cortical network can carry out a series of computations by utilizing different regions of state space to perform different computations. This idea is well-supported by recent work investigating sensory encoding (Rabinovich et al., 2001, 2008), decision processes (Raposo et al., 2014; Mante et al., 2013; Murakami and Mainen, 2015), and movement execution (Churchland et al., 2012; Hennequin et al., 2014). To understand computational mechanisms, one must consider the population dynamics as a whole. Observing this network-level phenomenon through the activity of a single neuron amounts to watching a moving object in three-dimensional space through its one-dimensional projection. The projected image may miss salient information (such as categorical discrimination) at some moments, or it may contain information from more than one process. Therefore, mixed and time-varying selectivity are expected and observed at the single neuron level.

Structural and functional connectivity of trained networks

We performed a series of analyses to understand the connectivity structure that governs robust transient dynamics. We compared our trained recurrent networks and randomly connected networks (RCNs), which were previously investigated as a source of mixed time-varying selectivity (Rigotti et al., 2010; Barak et al., 2013). We found that although RCNs encode a mixture of stimulus- and time-dependent variability, they do not exhibit the self-generated categorical neuronal coding during the sample and/or delay periods. In particular, in both neural data and networks with trained recurrent connections, the majority of neurons with persistent selectivity are strong categorical discriminators, whereas in RCNs persistent representation is not category-specific (Figure S6A-E). However, units in our recurrent neural network do display mixed selectivity, therefore this work extend the work of Rigotti et al. to networks with wiring structures that emerge from training to perform a cognitive task.

In trained networks, the distribution of synaptic connections is sparse and unimodal with mean weight equals to zero (Figure S6G), but exhibits clear hierarchical structure not present in RCNs. To reveal the hierarchy, we computed degree centrality, defined as the total number of connections each neuron sends (out-degree) and receives (in-degree). All trained connectivity exhibits heavy-tailed degree distributions, i.e. few neurons are connected to large numbers of neighbors acting as network hubs (Figure 7A, mean kurtosis for all ten networks $= 18.341$, $p = 0.005$). Furthermore, trained networks also exhibit a high correlation between in-degree and out-degree (Figure 7B; Spearman rank correlation, $N = 150$, $\rho = 0.546$, $P < 10^{-7}$), suggesting that hub neurons aggregate information from and broadcast it to large numbers of neighbors. Neurons with high degree also tend to have larger positive incoming connections (large in-strength, Figure S6H) and larger average activity (Figure S6I) than low-degree counterparts, which means these neurons have greater influence on neural state trajectories.

The robust dynamics underlying sequential decisions result from ongoing competition and cooperation among neurons within the circuit. We measured neural response similarity ($r_{response}$), defined as the covariance between synaptic currents of neural pairs across task conditions (see STAR Methods), and structural coupling, defined as the sum of synaptic connections from neuron i to j and from j to i. $r_{response}$ is correlated with structural coupling in all task epochs (Pearson correlation, average $r = 0.176$, $P < 10^{-4}$; see statistical test against null models in Figure S6J), suggesting that neurons with similar category or match selectivity tend to have strong positive synaptic couplings, while neurons with opposite encoding have strong negative couplings. This gives rise to competitive dynamics between subpopulations that encode different concepts (Wong and Wang, 2006; Wang, 2002). To further investigate neural couplings, we divided the neural population into four groups based on their noiseless activity at the end of the delay period: (1) neurons that are active when a stimulus belongs to the red category, but silent for the blue category (denoted as R group, average 10.2% of population); (2) neurons that are active exclusively for the blue category (B, 14.7%); (3) neurons that are responsive for both red and blue (BOTH, 14.5%); (4) neurons that are not responsive at all (NR, 60.6%). Then we assessed average connections within and between these subclasses. We found strong within-group excitation and between-group mutual inhibition for R and B groups (Figure 7C), mediating competition between the two categories. Furthermore, we found that neurons in BOTH group have high degrees and activity, but are less sensitive to categories, compared to R and B groups (Figure S7A-C). The BOTH group receives net excitatory connections from itself as well as from R and B groups (Figure 7C). The activity of BOTH neurons tends to increase over the delay, while that of R and B neurons tends to decrease (Figure S7D). BOTH neurons' activity drives correlations between neural states associated to red and blue categories, which is apparent in both the neural data and model (Figure S7E-F). Overall, these findings show the cooperation between two categories of neural pools through BOTH neurons. This co-activation is likely responsible for the temporal dynamics which brings neural states to the end-of-

delay regions, where match / non-match decisions can be made separately from the categorical decision.

Lastly, the model yields a testable prediction that functional and structural coupling among neurons with persistent selectivity and prefer the same category tend to be larger relative to all connections in the network. For a given pair of neurons with the same category preference, we measured their average category selectivity duration (CSD, defined in Figure 3) and noise correlation (r_{noise}), i.e. the correlation coefficient between a neuron pair's rate fluctuations averaged across all task conditions. Neural pairs that contain non-selective neurons are removed from the analysis. We found that persistent neurons in the model tend to have far larger functional couplings than non-persistent neurons (Figure 7D), whereby neural pairs with average CSD larger than 90th percentile have larger average noise correlation ($m_1 = 0.237$) compared to other pairs ($m_2 = 0.031$, t-test, $p < 10^{-7}$). This result holds for any time window at which r_{noise} is measured. The effect remains significant when the same analysis is performed on synaptic coupling instead of r_{noise} ($m_1 = 0.234, m_2 = -0.031$ $p < 10^{-7}$) and when controlled for average neural activity (ANCOVA, $F = 326.69, p < 10^{-7}$, see STAR Methods).

Neuronal representation during flexible categorization with multiple rules

Recent studies have shown that single neurons in LIP (Fitzgerald et al., 2011) and PFC (Cromer et al., 2010) are multitaskers, as they encode categorical information for different sets of stimuli (e.g. differentiating between Dogs versus Cats for animal classification task and Sports versus Sedans in car classification). These studies found that: (1) multitasking neurons were the strongest category discriminators (Cromer et al., 2010); (2) neurons' tuning strengths for different stimulus sets were correlated (Fitzgerald et al., 2011). We refer to these tasks as 'independent-input' paradigms, as the two categorical schemes involve independent stimulus sets with likely non-overlapping sensory representations. In contrast, another set of studies employed a different paradigm where subjects were instructed to categorize the same set of stimuli under two different schemes (e.g. categorizing the same images of animals into dogs versus cats, or big versus small depending on the active rule) (Roy et al., 2010; Goodwin et al., 2012). We refer to these tasks as 'shared-input' paradigms, because both categorical schemes share the same sensory representation. These experiments show that: (1) rule-dependent responses emerged as soon as the rule cue was presented (Goodwin et al., 2012); (2) multitasking neurons were more commonly observed, whereas specialized neurons (i.e. neurons that encoded categories exclusively for one scheme) were less common in the independent-input paradigm compared to the shared-input paradigm (Roy et al., 2010; Cromer et al., 2010).

We asked if different task paradigms elicit different dynamical landscapes and can the discrepancy in dynamical structures alone account for these experimental observations?

To investigate this question, we trained recurrent neural networks, using the same protocol we used for the standard DMC task, to solve either independent-input or shared-input categorization tasks (see STAR Methods). For the independent-input paradigm, one input neuron group encodes motion directions (scheme A, red and blue categories), while another group encodes spatial locations of a circle stimulus (scheme B, pink and green categories, Figure 8A). The network's task is to categorize stimuli according to the boundary associated with each stimulus set (dashed black lines in Figure 8A). For the shared-input paradigm, the model learned to categorize motion directions by two different boundaries (Figure 8D). Prior to the fixation epoch, the model receives a 500 ms input pulse from two separate input neurons (colored squares in Figure 8D) signifying whether the horizontal (scheme A) or vertical boundary (scheme B) is in effect.

The two task paradigms result in markedly different landscapes. In the independent-input case, the trained networks form two working memory tunnels during the delay, similarly to those in Figure 6B, but these tunnels are shared between the two categorical schemes (Figure 8B). In particular, one tunnel corresponds to the red category of scheme A and the green category of scheme B while another tunnel corresponds to the blue and pink categories. Note that the opposite configuration (red/pink and blue/green) is also possible. Only two tunnels are required to solve the independent-input task, since the entrances of the tunnels can be mapped onto appropriate stimuli by modifying separate sets of input weights from different sensory neuron groups (Figure S8A) and the ends of the tunnels are mapped onto appropriate choices by a similar mechanism (Figure S8B). Hence, the two categorical schemes can share the same categorical discrimination machinery via appropriate mapping. Neurons that participate in driving trajectories along the tunnels must be active in both schemes, leading to strong correlation between the CTI of the two schemes (Figure 8C; Pearson correlation, $N = 150, r = 0.85, P < 10^{-7}$). Furthermore, neurons with persistent contribution to tunnel trajectories tend to be those with the strongest categorical selectivity (Figure S6C-D); therefore, multitasking neurons are the most robust category discriminators (Figure 8C, see statistical test in Figure S8C).

Tunnel sharing is only possible when sensory inputs contain rule information, i.e. motion direction stimuli entail a horizontal boundary and dot location stimuli entail a vertical boundary. In such case, the recurrent network does not need to memorize the categorization rule through the delay to map the content of working memory to the appropriate choices during the test period. This strategy would fail in the shared-input paradigm, where the same set of stimuli must be mapped to different choices during the test period depending on the active rule. Instead, we observed that neural trajectories diverge to encode categorization rules after the rule cue is displayed through the fixation period (Figure 8E), resembling rule representations observed in PFC (Wallis et al., 2001; Goodwin et al., 2012). When the sample stimulus is presented, the dynamic representations encode both rules and categories, which persist through the delay (Figure 8F). Since the two category schemes no longer share the same tunnels, the correlation between category tuning

strengths reduces or vanishes (Figure S8D). The overall effects from ten instances of networks trained with either paradigm show that the shared-input paradigm leads to lower correlation in category tuning strengths between the two schemes (Figure 8G; Pearson correlation, $N = 5$ realizations of network. Independent-input, average $r = 0.76$, $P < 10^{-6}$ in all networks. Shared-input, average $r = 0.08$, only one out of 5 realizations has significant correlation, $P < 0.01$). Consequently, we observed a significantly smaller number of multitasking neurons and a larger number of specialized neurons compared to the independent-input paradigm, in accordance with experimental findings (Figure 8H; t-test, $N = 5$, $P = 0.001$). Collectively, this comparison between independent-input and shared-input paradigms illustrates how dynamical landscapes can adapt to various categorical structures, and difference in landscapes alone can explain a lot of experimental findings.

Discussion

Our results contribute four important insights. First, our model suggests that robust transient dynamics, equipped with stimulus-dependent attracting states and robust trajectory tunnels, underlie delayed associative computations in cortical circuits. Second, we show that networks endowed with reproducible trajectories capture statistics of the heterogeneous and time-varying category selectivity at both the single-neuron and population levels, thus bridging the robust transience framework to neurophysiology of the primate parieto-frontal network. Third, we reveal the features of structural and functional connectivity that support robust transience, and suggest a testable prediction about the relationship between the temporal profiles of selectivity and interneuronal correlations. Fourth, our model explains observations from experiments that incorporate multiple categorization rules through the idea of shared state space landscape.

Much emphasis has been put on the reward-dependent learning mechanism that explains the emergence of categorical representation (Roelfsema and Ooyen, 2005; Engel et al., 2015; Rombouts et al., 2012; Savin and Triesch, 2014). Though providing valuable insights on synaptic plasticity, many of these studies have not focused on the temporal profiles of category selectivity and none have evaluated whether the end results of training resemble the neural dynamics in the brain. Our study focuses on the dynamical properties of networks that successfully solve the task and exhibit similar response features compared to neurophysiological data. The accrued insights provide essential foundations for future generative models. Note that this study assumed that although neurons were recorded at different times and in different animals, their activities represents sampled firing rates from a single working population.

To the extent possible, our model parameters were calibrated by experimental measurements such as the sparse connectivity of the trained networks (Song et al., 2005), the neural time constant (Murray et al., 2014), the width of sensory tuning (Albright, 1984). Other parameters where set in the same range as those in previous modeling studies such as the initial recurrent network connections and noise in the

networks (Mante et al., 2013). The changes in these parameters do not affect our overall findings, but can impact training. For example, longer neural time constant will make it easier to train networks on longer delay epochs and higher noise in the network will reduce the chance of training success.

Importantly, our results suggest that time-varying patterns of category working memory result from a slow dynamic transition from one location in state space to another mediated by a dynamical tunnel that constrains the course of trajectories. This is distinct from purely feedforward models (Goldman, 2009; Savin and Triesch, 2014) or models that utilize rapid transitions to stable states (Wong and Wang, 2006; Wang, 2002). One accompanying feature of such a mechanism is the reliable emergence of persistently selective neurons among other neurons with heterogeneous temporal dependence. This gives rise to the category-selective population code whose dominant mode is stable, yet also exhibits a time-varying secondary mode. Similar population dynamics have been observed in other tasks (Machens, 2010; Raposo et al., 2014), but have not been accounted for by other models.

Despite their time-varying dynamics, networks utilizing robust transience support and advance the central idea of strong reverberatory dynamics underlying working memory and decision making (Goldman-Rakic, 1990; Wang, 2001; Wang, 2002; Wong and Wang, 2006; Murray et al. 2017). The persistent neurons in our model, albeit few in quantity, are the main drivers of delay dynamics as they are among the strongest category discriminators and form large connections among each other. The circuit motifs proposed in classical models, such as strong local excitation and mutual inhibition among dominant neuron groups, are apparent in the current framework, although they are embedded in more heterogeneous circuits, which allows them to flexibly partake in sequential computations and generate mixed representations in accordance with experimental evidence. The presence of multiple stable states is also the key constituent of robust dynamics in our model. Furthermore, structural organization in local circuits may vary in a continuum from random networks, to robust dynamics, to stable attractors, depending on the extent of training (Barak et al., 2013). In particular, our model predicts that networks trained on protocols in which delay durations vary across trials tend to develop more temporally stable persistent activity (see Figure S5). Future animal experiments can test this hypothesis.

Our work contributes to a growing line of research on robust transient dynamics and its role in complex neural computations. The principle has been proposed for the spatiotemporal sensory encoding (Rabinovich et al., 2001), movement generation (Hennequin et al., 2014), and other cognitive processes (Rabinovich and Varona, 2011; Rabinovich et al., 2008, 2014), which speaks to its prevalence in neural circuit processing across brain regions and species. Most importantly, through detailed comparison between neurophysiological data and model, our contribution provides compelling evidence that robust transience governs sequential categorical decisions in primate cortical circuitry.

Author Contributions

WC and XJW designed research, performed model simulations and analyzed data. DJF and SKS designed and performed experiments. WC, DJF and XJW wrote the paper.

Acknowledgements

This work was supported by NIH R01MH062349, R01MH092927, NSF-NCS 1631571, STCSM grants 14JC1404900 and 15JC1400104. We thank John Assad for valuable contributions during all phases of the neurophysiological studies, which produced the data examined here. We thank John Murray, Francis Song, and William Gaines for intellectual and helpful discussions. The Authors declare that there is no conflict of interest.

References

Albright, T. D. (1984). Direction and orientation selectivity of neurons in visual area mt of the macaque. Journal of neurophysiology, 52, 1106–1130.

Amit, D. J., Brunel, N., and Tsodyks, M. (1994). Correlations of cortical hebbian reverbera- tions: theory versus experiment. The Journal of neuroscience, 14, 6435–6445.

Andersen, R., Asanuma, C., Essick, G., and Siegel, R. (1990). Corticocortical connections of anatomically and physiologically defined subdivisions within the inferior parietal lobule. Journal of Comparative Neurology, 296, 65–113.

Ardid, S. and Wang, X.-J. (2013). A tweaking principle for executive control: neuronal circuit mechanism for rule-based task switching and conflict resolution. J Neurosci, 33, 19504–17.

Barak, O., Sussillo, D., Romo, R., Tsodyks, M., and Abbott, L. F. (2013). From fixed points to chaos: three models of delayed discrimination. Prog Neurobiol, 103, 214–22.

Barak, O., Tsodyks, M., and Romo, R. (2010). Neuronal population coding of parametric working memory. J Neurosci, 30, 9424–30.

Borg, I. and Groenen (1997). Modern multidimensional scaling: theory and applications. (New York: Springer).

Born, R. T. and Bradley, D. C. (2005). Structure and function of visual area mt. Annu. Rev. Neurosci., 28, 157–189.

Brendel, W., Romo, R., and Machens, C. K. (2011). Demixed principal component analysis. In Advances in Neural Information Processing Systems, pp. 2654–2662.

Brody, C. D., Herna´ndez, A., Zainos, A., and Romo, R. (2003a). Timing and neural encoding of somatosensory parametric working memory in macaque prefrontal cortex. Cereb Cortex, 13, 1196–207.

Brody, C. D., Romo, R., and Kepecs, A. (2003b). Basic mechanisms for graded persistent activity: discrete attractors, continuous attractors, and dynamic representations. Current opinion in neurobiology, 13, 204–211.

Buonomano, D. V. and Maass, W. (2009). State-dependent computations: spatiotemporal processing in cortical networks. Nat Rev Neurosci, 10, 113–25.

Bullmore, E. and Sporns, O. (2009). Complex brain networks: graph theoretical analysis of structural and functional systems. Nat Rev Neurosci, 10, 186–98.

Churchland, M. M., Cunningham, J. P., Kaufman, M. T., Foster, J. D., Nuyujukian, P., Ryu, S. I., and Shenoy, K. V. (2012). Neural population dynamics during reaching. Nature.

Cohen, M. R. and Kohn, A. (2011). Measuring and interpreting neuronal correlations. Nature neuroscience, 14, 811–819.

Cromer, J. A., Roy, J. E., Buschman, T. J., and Miller, E. K. (2011). Comparison of primate prefrontal and premotor cortex neuronal activity during visual categorization. J Cogn Neurosci, 23, 3355–65.

Cromer, J. A., Roy, J. E., and Miller, E. K. (2010). Representation of multiple, independent categories in the primate prefrontal cortex. Neuron, 66, 796–807.

Crowe, D. A., Averbeck, B. B., and Chafee, M. V. (2010). Rapid sequences of population activity patterns dynamically encode task-critical spatial information in parietal cortex. J Neurosci, 30, 11640–53.

Druckmann, S. and Chklovskii, D. (2012). Neuronal circuits underlying persistent represen- tations despite time varying activity. Current Biology.

Engel, T. A., Chaisangmongkon, W., Freedman, D. J., and Wang, X.-J. (2015). Choice-correlated activity fluctuations underlie learning of neuronal category representations. Na- ture Communication.

Engel, T. A. and Wang, X.-J. (2011). Same or different? a neural circuit mechanism of similarity-based pattern match decision making. J Neurosci, 31, 6982–96.

Fitzgerald, J. K., Freedman, D. J., and Assad, J. A. (2011). Generalized associative represen- tations in parietal cortex. Nat Neurosci, 14, 1075–9.

Freedman, D. J. and Assad, J. A. (2006). Experience-dependent representation of visual categories in parietal cortex. Nature, 443, 85–8.

Freedman, D. J. and Assad, J. A. (2016). Neuronal mechanisms of visual categorization: an abstract view on decision making. Annual review of neuroscience, 39, 129–147.

Freedman, D. J., Riesenhuber, M., Poggio, T., and Miller, E. K. (2001). Categorical repre- sentation of visual stimuli in the primate prefrontal cortex. Science, 291, 312–6.

Goldman, M. S. (2009). Memory without feedback in a neural network. Neuron, 61, 621–34.

Goldman-Rakic, P. S. (1990). Cellular and circuit basis of working memory in prefrontal cortex of nonhuman primates. Prog Brain Res, 85, 325–35; discussion 335–6.

Goodwin, S. J., Blackman, R. K., Sakellaridi, S., and Chafee, M. V. (2012). Executive control over cognition: stronger and earlier rule-based modulation of spatial category signals in prefrontal cortex relative to parietal cortex. J Neurosci, 32, 3499–515.

Hennequin, G., Vogels, T. P., and Gerstner, W. (2014). Optimal control of transient dynamics in balanced networks supports generation of complex movements. Neuron, 82, 1394–1406.

Ibos, G. and Freedman, D. J. (2014). Dynamic integration of task-relevant visual features in posterior parietal cortex. Neuron, 83, 1468–1480.

Ibos, G. and Freedman, D. J. (2016). Interaction between spatial and feature attention in posterior parietal cortex. Neuron, 91, 931–943.

Jaeger, H. (2001). The “echo state” approach to analysing and training recurrent neural networks-with an erratum note. Bonn, Germany: German National Research Center for Information Technology GMD Technical Report, 148, 34.

Jun, J. K., Miller, P., Herna´ndez, A., Zainos, A., Lemus, L., Brody, C. D., and Romo, R. (2010). Heterogenous population coding of a short-term memory and decision task. J Neurosci, 30, 916–29.

Lewis, J. W. and Van Essen, D. C. (2000). Corticocortical connections of visual, sensorimotor, and multimodal processing areas in the parietal lobe of the macaque monkey. Journal of Comparative Neurology, 428, 112–137.

Maass, W., Natschla¨ger, T., and Markram, H. (2002). Real-time computing without stable states: A new framework for neural computation based on perturbations. Neural computa- tion, 14, 2531–2560.

Machens, C., Romo, R., and Brody, C. (2010). Functional, but not anatomical, separation of what and when in prefrontal cortex. The Journal of Neuroscience, 30, 350–360.

Machens, C. K. (2010). Demixing population activity in higher cortical areas. Front Comput Neurosci, 4, 126.

Machens, C. K., Romo, R., and Brody, C. D. (2005). Flexible control of mutual inhibition: a neural model of two-interval discrimination. Science, 307, 1121–4.

Mante, V., Sussillo, D., Shenoy, K. V., and Newsome, W. T. (2013). Context-dependent computation by recurrent dynamics in prefrontal cortex. Nature, 503, 78–84.

Martens, J. and Sutskever, I. (2011). Learning recurrent neural networks with hessian-free optimization. In Proc. 28th Int. Conf. on Machine Learning.

Maslov, S. and Sneppen, K. (2002). Specificity and stability in topology of protein networks. Science, 296, 910–913.

Merchant, H., Crowe, D. A., Robertson, M. S., Fortes, A. F., and Georgopoulos, A. P. (2011). Top-down spatial categorization signal from prefrontal to posterior parietal cortex in the primate. Frontiers in systems neuroscience, 5.

Meyers, E. M., Freedman, D. J., Kreiman, G., Miller, E. K., and Poggio, T. (2008). Dynamic population coding of category information in inferior temporal and prefrontal cortex. J Neurophysiol, 100, 1407–19.

Miller, E. K., Freedman, D. J., and Wallis, J. D. (2002). The prefrontal cortex: categories, concepts and cognition. Philos Trans R Soc Lond B Biol Sci, 357, 1123–36.

Miller, E. K., Nieder, A., Freedman, D. J., and Wallis, J. D. (2003). Neural correlates of categories and concepts. Curr Opin Neurobiol, 13, 198–203.

Murakami, M. and Mainen, Z. F. (2015). Preparing and selecting actions with neural popu- lations: toward cortical circuit mechanisms. Current opinion in neurobiology, 33, 40–46.

Murray, J. D., Bernacchia, A., Freedman, D. J., Romo, R., Wallis, J. D., Cai, X., Padoa-Schioppa, C., Pasternak, T., Seo, H., Lee, D., et al. (2014). A hierarchy of intrinsic timescales across primate cortex. Nature neuroscience, 17, 1661–1663.

Murray J, Bernacchia A, Roy N, Constantinidis C, Romo R, Wang X-J (2017). Stable subspace coding for working memory coexists with population dynamics in prefrontal cortex. *Pro. Natl. Acad. Sci*., 141, 394-399

Park, I. M., Meister, M. L., Huk, A. C., and Pillow, J. W. (2014). Encoding and decoding in parietal cortex during sensorimotor decision-making. Nature neuroscience.

Rabinovich, M., Huerta, R., and Laurent, G. (2008). Neuroscience. Transient dynamics for neural processing. Science, 321, 48–50.

Rabinovich, M., Volkovskii, A., Lecanda, P., Huerta, R., Abarbanel, H., and Laurent, G. (2001). Dynamical encoding by networks of competing neuron groups: winnerless competi- tion. Physical Review Letters, 87, 068102.

Rabinovich, M. I., Sokolov, Y., and Kozma, R. (2014). Robust sequential working memory recall in heterogeneous cognitive networks. Frontiers in systems neuroscience, 8.

Rabinovich, M. I. and Varona, P. (2011). Robust transient dynamics and brain functions. Frontiers in computational neuroscience, 5.

Raposo, D., Kaufman, M. T., and Churchland, A. K. (2014). A category-free neural population supports evolving demands during decision-making. Nature neuroscience.

Rigotti, M., Barak, O., Warden, M. R., Wang, X.-J., Daw, N. D., Miller, E. K., and Fusi, S. (2013). The importance of mixed selectivity in complex cognitive tasks. Nature, 497, 585–90.

Rigotti, M., Ben Dayan Rubin, D., Wang, X.-J., and Fusi, S. (2010). Internal representation of task rules by recurrent dynamics: the importance of the diversity of neural responses. Front Comput Neurosci, 4, 24.

Rishel, C. A., Huang, G., and Freedman, D. J. (2013). Independent Category and Spatial Encoding in Parietal Cortex. Neuron, 77, 969–979.

Roelfsema, P. R., Khayat, P. S., and Spekreijse, H. (2003). Subtask sequencing in the primary visual cortex. Proceedings of the National Academy of Sciences, 100, 5467–5472.

Roelfsema, P. R. and van Ooyen, A. (2005). Attention-gated reinforcement learning of internal representations for classification. Neural Comput, 17, 2176–214.

Rombouts, J., Roelfsema, P., and Bohte, S. M. (2012). Neurally plausible reinforcement learning of working memory tasks. In Advances in Neural Information Processing Systems, pp. 1871–1879.

Roy, J. E., Riesenhuber, M., Poggio, T., and Miller, E. K. (2010). Prefrontal cortex activity during flexible categorization. The Journal of Neuroscience, 30, 8519–8528.

Rubinov, M. and Sporns, O. (2010). Complex network measures of brain connectivity: uses and interpretations. Neuroimage, 52, 1059–1069

Sarma, A., Masse, N. Y., Wang, X.-J., and Freedman, D. J. (2016). Task-specific versus gen- eralized mnemonic representations in parietal and prefrontal cortices. Nature neuroscience, 19, 143–149.

Savin, C. and Triesch, J. (2014). Emergence of task-dependent representations in working memory circuits. Front Comput Neurosci, 8, 57.

Shafi, M., Zhou, Y., Quintana, J., Chow, C., Fuster, J., and Bodner, M. (2007). Variability in neuronal activity in primate cortex during working memory tasks. Neuroscience, 146, 1082–1108.

Singh, R. and Eliasmith, C. (2006). Higher-dimensional neurons explain the tuning and dynamics of working memory cells. J Neurosci, 26, 3667–78.

Song, S., Sjo¨stro¨m, P. J., Reigl, M., Nelson, S., and Chklovskii, D. B. (2005). Highly nonrandom features of synaptic connectivity in local cortical circuits. PLoS Biol, 3, e68.

Stokes, M. G., Kusunoki, M., Sigala, N., Nili, H., Gaffan, D., and Duncan, J. (2013). Dynamic coding for cognitive control in prefrontal cortex. Neuron, 78, 364–75.

Sussillo, D. (2014). Neural circuits as computational dynamical systems. Current opinion in neurobiology, 25, 156–163.

Sussillo, D. and Barak, O. (2013). Opening the black box: low-dimensional dynamics in high-dimensional recurrent neural networks. Neural Comput, 25, 626–49.

Swaminathan, S. K. and Freedman, D. J. (2012). Preferential encoding of visual categories in parietal cortex compared with prefrontal cortex. Nat Neurosci, 15, 315–20.

Swaminathan, S. K., Masse, N. Y., and Freedman, D. J. (2013). A comparison of lateral and medial intraparietal areas during a visual categorization task. J. Neurosci., 33, 13157–13170.

Wallis, J. D., Anderson, K. C., and Miller, E. K. (2001). Single neurons in prefrontal cortex encode abstract rules. Nature, 411, 953–6.

Wang, X. J. (2001). Synaptic reverberation underlying mnemonic persistent activity. Trends Neurosci, 24, 455–63.

Wang, X.-J. (2002). Probabilistic decision making by slow reverberation in cortical circuits. Neuron, 36, 955–968.

Wohrer, A., Humphries, M. D., and Machens, C. K. (2013). Population-wide distributions of neural activity during perceptual decision-making. Prog Neurobiol, 103, 156–93.

Wong, K.-F. and Wang, X.-J. (2006). A recurrent network mechanism of time integration in perceptual decisions. J Neurosci, 26, 1314–28.

Figure Legends

Figure 1 | Delayed-match-to-category (DMC) task and neurophysiological recordings. (**A**) Time course of a DMC experiment. A sample stimulus is followed by a short delay and a test stimulus. To receive reward, subjects must respond whether the sample and test stimuli belong to the same (match) or different (non-match) categories. (**B**) Sample and test stimuli are randomly drawn from a set of dot motion stimuli, divided into two categories (red and blue arrows). We analyzed neural recordings from two experiments. The first experiment used 12 motion directions,

and LIP neurons were recorded (denoted LIP1). The second experiment used 6 motions directions, and neurons from LIP (denoted LIP2) and PFC were recorded.

Figure 2 | Model structure and training protocol. (**A**) A set of recurrently connected neurons were trained to solve the DMC paradigm. The recurrent population is connected to direction-selective input units, and approximately 1/5 of recurrent neurons (blue hatched circles) are connected to two output units representing match and non-match choices. All synapses are updated with a supervised learning algorithm. (**B**) Activity of the input neurons encode directions of sample and test stimuli (top panel). x-axis, time; y-axis, input neurons labeled by preferred directions; neural activity is color-coded. After training, the model generates appropriate decision output. The match output unit (middle panel) shows ramping activation during match trials (green trace) and remains silent during non-match trials (brown trace). The opposite pattern holds in the non-match output unit (bottom panel). x-axis, time; y-axis, neural activity; shaded areas, s.d. across trials. Red lines mark the activity threshold where behavioral choice is registered. (**C**) The psychometric functions of animal subjects (left panel) compared to that of model networks (right panel). Error bars indicate s.d. across all recording sessions for and across 10 network realizations for model.

Figure 3 | Both neural recordings and model networks demonstrate heterogeneity in the temporal profiles of category selectivity. (**A**) Examples of different classes of category selectivity profiles from LIP and PFC populations. Average firing response as a function of time, color-coded by stimulus directions from red to blue category. x-axis, time from sample onset; y-axis, average firing rate. The colored bar on top shows category-selective period. Color (red or blue) indicates neuron's category preference, while color intensity indicates category tuning strength. (**B**) Average neural response of model units with same plotting convention as in (**A**). Overall, both neural data and model demonstrate heterogeneity in selectivity time course such as persistent, transient, and switching selectivity. (**C-D**) Colored heat maps showing category selectivity profiles of all neurons in the LIP2 dataset ($N = 61$, **C**) and of a trained model ($N = 122$, **D**) with same color-coding convention as colored bars in (**A-B**). Neurons with no selectivity phase are excluded. Across both neural data and model, neurons' category selectivity latencies and durations are highly variable. x-axis, time; y-axis, neurons sorted by category selectivity latency.

Figure 4 | The model captures essential features in population response patterns of LIP and PFC neurons. Neural response trajectories during sample (left column, **A**-**C**), delay (two middle columns, **E**-**G**), and test periods (right column, **I**-**K**) for three neural data sets: LIP1 (top row), LIP2 (second row), and PFC (third row). Colors of traces encode stimulus identities by the same convention as in Figure 3. During the sample period (left column), trajectories begin at roughly the same location for all stimulus directions (black dots correspond to the onset of sample epoch), then fan out into elliptical shapes encoding the directions of stimuli and some category information. During the delay (middle columns), population trajectories have two main components encoding stimulus categories in stable and

time-varying manners, respectively. x-axis, relative rate (principal component score); y-axis, time relative to sample onset. At the beginning of test period (right column), trajectories encode sample categories (black dots), then neural traces diverge into four separate clusters encoding the four possible sample-test category combinations. BB (dark blue lines) corresponds to blue sample category / blue test category condition and similarly RR (dark red) to red sample / red test; BR (purple) to blue sample / red test; RB (orange) to red sample / blue test. Finally, the traces for match conditions (BB and RR) unfold along the same direction and analogously so for non-match conditions (BR and RB). (**D**, **H**, **L**) Population trajectories of the a representative model instance analyzed by the same procedures. The model reproduces population response patterns of neural data in all task epochs.

Figure 5 | Overall dynamical landscape of the trained network. (**A**) A conceptual schematic portraying the series of transitions between behavioral epochs to solve DMC task. (**B**) Neural trajectories of the model implementing the computational process in (**A**). Neural states evolve serially from the resting state (gray cross) to states associated with sample categories (red and blue dots), then category working memory (red and blue stars), sample-test category combinations (dark blue, dark red, orange, and purple lines), and finally match/non-match choices (green and brown crosses, respectively). Task epoch labels (sample, delay, test) indicate neural states at the beginning of the epoch. All crosses denote stable states; stars denote slow or fixed points associated to working memory; dots denote transient states.

Figure 6 | The trained network forms a dynamical landscape that gives rise to ro- bust trajectories and executes category-based computations. (**A**) Neural trajectories during sample categorization. At the beginning of a trial, neural states stay within the basin of attraction of the resting state fixed point (gray cross), even upon receiving prolonged (1 s) noisy input (inset, three noisy trajectories plotted in gray). Under the influence of direction-tuned inputs (due to sample stimulus presentation), a set of stable fixed points appear in state space (red and blue crosses, colors denote stimulus directions), propelling the states towards areas associated to red or blue categories. (**B**-**C**) Neural landscape associated to category working memory. (**B**) Red and blue dots mark possible locations of the neural states at the end of the sample period (with noise). Red and blue lines show noiseless trajectories during the delay originating from these positions. Black arrows mark flow directions. (**C**) The line shows an example neural state path selected from the trajectories in (**B**). Black dots mark the neural states at different time points during the delay in 150 ms increments. Arrows show a velocity vector field at states nearby the trajectory (with norms scaled down for clarity). The working memory landscape resembles tunnels that force neural states to flow along two possible routes (arrows in C show the movement flow within the tunnel), generating robust time-varying memory of sample categories. The neural state movements slow down or stop at the end of delay (marked by shorter distance between dots near the end of delay in (**C**)). (**D**-**E**) Neural trajectories during match decisions. (**D**) Red and blue dots mark locations that neural states occupy at the end of the delay (with noise). When the test stimuli appear, neural states move towards the same input-dependent fixed

points as shown in (**A**). Since there are two possible starting regions (associated to red or blue sample categories), trajectories diverge along four separate streams, encoding sample-test category conditions. (**E**) Continuing from (**D**), when test stimuli are removed, neural states fall into the basins of attraction corresponding to match or non-match stable states. Data plotted in (**B**-**E**) come from a representative model instance.

Figure 7 | Structural and functional connectivity that supports robust transient dynamics. (**A**) Degree distribution (total number of connections) in a representative sample of trained network. All trained networks exhibit long-tail degree distribution, showing existence of hub neurons. (**B**) Scatter plot shows the number of incoming connections (in-degree, x-axis) versus outgoing connections (out-degree, y-axis) for all neural units in a trained network. Strong correlations between in-degree and out-degree are observed in trained networks, which is unexpected if their connectivities are random (Pearson correlation, $N = 150, r = 0.482, P < 10^{-7}$). (**C**) Average synaptic weights between neuron groups that are active only for red stimuli (R group), only for blue (B group), for both red and blue (BOTH group), and not responsive (NR group). Colors indicate average synaptic connection (purple, inhibitory connections; green, excitatory connections) from a presynaptic group (x-axis) to a postsynaptic group (y-axis). R and B groups exhibit within-group excitation and across-group inhibition, while BOTH group receives excitatory connections from itself and from R and B groups. NR group receives inhibitory connections from all groups. (**D**) Average noise correlation (r_{noise}, y-axis) of neuron pairs grouped by percentile ranks of their average category selectivity durations. x-axis, the center of each rank bin (bin width $= 10$ percent). Neurons with persistent category-selective activity tend to form large functional connections (high r_{noise}). All 10 network realizations demonstrate similar features; data plotted come from a representative sample.

Figure 8 | Robust transience framework explains neural selectivity during flexible categorization involving multiple rules. (**A**) An independent-input categorization paradigm. Networks learn to categorize two separate sets of stimuli (motion directions, scheme A; stationary dots at different spatial locations, scheme B). The two stimulus sets are represented by two separate groups of sensory neurons and subject to different categorization rules. (**B**) Noiseless trajectories during the delay from networks trained with independent-input paradigm (black dots mark the beginning of the delay epoch). Networks form two working memory tunnels, which are utilized by both stimulus sets to maintain category working memory. (**C**) We measured neurons' category tuning index (CTI), which measures the strength of categorical sensitivity to any preferred category. Tunnel sharing results in a large correlation between CTIs for scheme A (x-axis) versus scheme B (y-axis) during the delay period (Pearson correlation, $n = 150, r = 0.85, P < 10^{-7}$). Data plotted in **B-C** come from a representative model instance. (**D**) A shared-input categorization paradigm. Networks must categorize a single stimulus set by two different boundaries, signaled by colors of the rule cue. (**E**) Noiseless trajectories during the rule cue and fixation periods for a network trained with shared-input

paradigm. Black and gray dots mark the beginning of the rule cue and fixation period, respectively. Trajectories split into two streams corresponding to different rules. (**F**) Network forms four separate tunnels to maintain category-rule combination. Categorization rules coded by dot colors. Data plotted in **E-F** come from a representative model instance. (**G**) Correlations between category tuning indices for scheme A and scheme B across 5 realizations of networks trained with independent- or shared-input paradigm. Error bars indicate minimum and maximum correlations within each group. Independent-input paradigm results in large positive correlations between CTIs of the two schemes, while shared-input paradigm does not. (**H**) The independent-input paradigm produces a significantly smaller number of specialized neurons but a larger number of multitasking neurons compared to the shared-input paradigm (t-test, $n = 5$, stars indicate $P < 0.05$).

STAR METHODS

CONTACT FOR REAGENT AND RESOURCE SHARING

Further information and requests for resources should be directed to and will be fulfilled by the Lead Contact, Xiao-Jing Wang (xjwang@nyu.edu).

EXPERIMENTAL MODEL AND SUBJECT DETAILS

All neural data were from Freedman and Assad (2006) and Swaminathan and Freedman (2012), where experimental protocol and recording procedures were described in detail. In summary, four male monkeys (*Macaca mulatta*, weighing from 8.0 - 14 kg) were trained to indicate whether a test stimulus was in the same category as a previously presented sample stimulus by releasing a lever. Stimuli were high-contrast, 9.0° diameter random dot movies, which moved at 12° per s with 100% coherence. In the first experiment (Freedman and Assad, 2006), the stimulus set comprised 12 motion directions (30° apart) and neural responses were recorded from 156 lateral intraparietal (LIP) neurons in two monkeys (Figure 1B, left panel). In the second study (Swaminathan and Freedman, 2012), 6 motion directions were used (60° apart), and neural data were collected from 76 LIP and 447 prefrontal neurons (PFC) in two other monkeys (Figure 1B, right panel). To avoid confusion, we refer to the LIP populations from the first and second experiments as the LIP1 and LIP2 datasets respectively. In the first experiment, four test stimuli closest (15°) to the category boundary were removed before recording from monkey H. The second experiment incorporated stimuli at category boundary, which were removed from this analysis. To combine data from four monkeys, all stimulus directions were rotated so that the category boundary corresponds to $0° - 180°$ axis.

The monkeys were implanted with a head post and recording chamber(s), at coordinates determined by magnetic resonance imaging. In the first experiment, the chamber was centered approximately at P3, L10 to allow access to both the intraparietal sulcus (IPS) and superior temporal sulcus by means of a dorsal approach. The recording chamber was centered approximately 3.0 mm posterior to the intraural line, and 10.0 mm lateral from the midline. For the second experiment, PFC chambers were centered on the principal sulcus and anterior to the arcuate sulcus at $\sim$ 27.0 mm anterior to the intra-aural line, while LIP chambers were positioned over the intraparietal sulcus (IPS) centered $\sim$ 3.0 mm posterior to the intra-aural line. The first study was conducted at Harvard University and all experimental procedures followed Harvard Medical School and National Institutes of Health guidelines. The second study was conducted at University of Chicago, where all procedures were in accordance with the University of Chicago's Animal Care and Use Committee and US National Institutes of Health guidelines.

METHOD DETAILS

Model architecture and training

We trained a recurrent network model to perform delayed-match-to-category task. The single-unit dynamics is governed by the following equation:

$$\tau\dot{\mathbf{x}}(t) = -\mathbf{x}(t) + \mathbf{W_{rr}}\mathbf{r}(t) + \mathbf{W_{ur}}\mathbf{u}(t) + \boldsymbol{\eta}(t),$$

where $x_i(t)$ is a synaptic current variable of neuron i at time t and neural activity (firing rate, r_i) is a rectified nonlinear function of x: $r_i = tanh_+(x_i)$. This constrained firing rates to be positive. The recurrent network has 150 units, and the connectivity matrix ($\mathbf{W_{rr}}$) is initialized to have 10% probability of connections, where non-zero weights are drawn from a normal distribution of zero mean and s.d. $= 0.28$. The neural time constant (τ) is 100 ms. Each recurrent neuron receives an independent white noise input ($\boldsymbol{\eta}(t)$) with zero mean and the final $\sigma_r = 0.6$ (see progressive protocol below for σ_r value during training). The input to the network at a given time ($\mathbf{u}(t)$) is fed into the recurrent network through synaptic weight $\mathbf{W_{ur}}$, which is initialized similarly to $\mathbf{W_{rr}}$. The model parameters were set to be in the same range as those in Mante et al. (2013). The changes in these parameters do not affect our qualitative findings.

There are 33 input units. The first 32 units have direction-tuned activity with equally spaced preferred directions from 0° to 360°. When a motion stimulus of direction ϕ appears, the mean activity of input units depends of the unit's preferred direction, θ:

$$u(\theta, \phi) = a\exp(-(\phi - \theta)^2/2\sigma^2),$$

with $\sigma = 43.2°$ and $a = 0.8$. The last input unit signals the appearance of fixation dot with a mean activity $= 0.05$ during the fixation period. At each time step, input units' activity contain contribution from white noise with $\sigma_u = 0.6$. A standard trial

consists of 1 s resting period, followed by 500 ms fixation, 650 ms sample stimulus, 1 s delay, 250 ms test stimulus, and 1 s choice period. The total trial duration $T = 4.4$ s. In the model, the duration of test stimulus was clipped to the approximate reaction time of the monkeys at 250 ms (instead of 650 ms in task protocol) to mimic the fact that visual stimulus was removed after monkeys make movement response. Model dynamics are simulated using Euler update with $\Delta t = 10$ ms.

Five output units linearly read out the synaptic currents of the recurrent circuit:

$$\mathbf{y} = \mathbf{W_{ro}x}.$$

To train the network, we specified the desired target activity for each output unit and iteratively adjusted all plastic synapses to minimize the discrepancy between the readout activity and the target output. Figure S1 illustrates and explains the target outputs of all five units. To assess the performance of the model, we passed activity of action units to a saturating nonlinear function:

$$\tilde{y}_i = 0.5\ tanh(y_i - 3.0).$$

Reaction time is defined as the time window at which $\tilde{y}$ passes a threshold of 0.85 (in arbitrary units). The network responds match (or non-match) if $\tilde{y}_1$ (or $\tilde{y}_2$) passes the threshold within 1.2 s after the test stimulus onset and the activity of the opposite choice does not pass the threshold.

All plastic synapses (input, recurrent, and output) as well as the initial conditions of the network activity, $\mathbf{x}(t = 0)$, are updated with a supervised training technique called Hessian-free (HF) algorithm Martens and Sutskever (2011; Mante et al. 2013), which is designed to minimize the error defined as:

$$e = \sum_{k=1}^{K} \sum_{t=0}^{T} \sum_{i=1}^{N_o} (\hat{y}_i - y_i)^2.$$

The error is effectively the square of the difference between target output, $\hat{y}_i$, and network output, y_i, summed over all N recurrent neurons from first time step $t = 0$ to the end of the trial $t = T = 4.4$ s and across all K trials in each training batch. HF belongs to the family of truncated Newton methods, which identifies update directions using second-order curvature and combines geometric insight and optimization heuristics to find solution with relatively low computational resource. In addition to standard regularizations in HF method, we imposed L_1 regularization, which simultaneously minimize the L_1 norm of parameters, $\alpha \sum_{j=1}^{n} |w_j|$. This constrains the algorithm to find sparse synaptic matrix solutions. $\alpha = 0.001$ controls the contribution of the L_1 regularization term on the objective function.

We employed a progressive training protocol, which started with the simplest version of DMC task and gradually increased task difficulty as model's performance passed criteria (C). This yielded an overall higher success rate and faster training. In the first step, the task involved only two stimuli at the middle of categories (effectively a delayed-match-to-sample task) and individual neurons receive no

independent noise (the number of trials per batch, K, is 200; $C = 99\%$). Second, noise is gradually added to the system. We increased units' noise, σ_r, by 0.05 at training batches where the network performance passed the criterion, $C = 87\%$, until desired level of noise is reached ($\sigma_r = 0.6$). Third, synapses with near-zero weight are gradually removed, whereby 5% of smallest synapses are set to zeros and their future updates are set to zeros on training batches with performance larger than the criteria ($C = 87\%$). Synaptic clipping is repeated until the probability of connections in the recurrent weight matrix equals 12%. Finally, the number of stimuli (as well as the number of training trials per iteration) was progressively increased, until the network can perform DMC with 12 motion directions ($K = 2160$, $C = 87\%$).

Model for multi-scheme categorization tasks

The training method for multi-scheme categorization tasks is identical to the standard DMC task except for the structure of inputs. In the independent-input paradigm, the input population consists of 65 units. One unit encodes the fixation dot and 32 units encode motion directions as described for standard DMC. Another set of 32 units encode the angular locations of a dot stimulus, modeled with periodic Gaussian current profiles as described in *Model architecture and training*. The outputs for match and non-match trials are identical to the standard task, but provide appropriate match or non-match answers corresponding to the new task rule. In the shared-input paradigm, another 500 ms task epoch was added before the fixation period to provide categorization rule signal. The input population consists of 35 units. The first 33 units are identical to standard DMC. The two additional units represent the task rules (horizontal or vertical categorization boundaries) by a pulse current with a magnitude $= 0.3$ when the corresponding rule is active and a magnitude of zero otherwise.

Analysis of model dynamics and connectivity

We trained 10 instances of the model and performed the same analyses on them. The results shown reflect behaviors observed across all network realizations.

Noiseless trajectories are simulated by setting σ_r (noise in the firing rate) and σ_u (noise in the input current) to zeros (see *Model architecture and training* section for definition of σ_r and σ_u). Neural states in Figure 5-6 are defined by the synaptic currents, $\mathbf{x}$, at each time point to allow both excitatory and inhibitory (subthreshold) dynamics to be observed; neural states defined by firing rates yield a similar picture. The stable resting state is defined as the terminal steady state when network is simulated with the initial condition obtained from training, $\mathbf{x}(t = 0)$, without noise and input. Stable states associated to choices are determined by running dynamics to terminal states for match or non-match trials without noise or input. All stable states are confirmed to have zero velocity and the eigendecomposition of linearized dynamics around these locations yields only negative eigenvalues, indicating

attracting states. Velocity vector fields were computed from the dynamic equation in *Model architecture and training*. The magnitude of velocity is defined as the norm of velocity vectors, $\| \dot{\mathbf{x}} \|^2$. All network trajectories and vector field plots reflect the first 2-3 largest principle components of all the data in the graph, except in Figure 4 where the procedures mirror the analysis of neural data. The locations of saddle points are determined by the optimization methods in ref. Sussillo and Barak (2013; Mante et al. 2013).

To compare the trained networks with random networks, we trained six randomly connected networks (RCNs) of N units, whose $N \times N$ synaptic matrix has n non-zero elements per row on average. The non-zero synaptic weights are randomly drawn from a Gaussian distribution of zero mean with variance $1/n$ ($N = 1{,}500, n = 100$) Barak et al. (2013). The input population (33 units) is structured as in the standard network. Approximately 30% of RCN units receive currents from one of the input units. On each trial, the initial synaptic variable, x_i, was drawn independently from a Guassian distribution to generate variability in neural response. Firing rate is defined as $r_i = tanh(x_i)$. We simulated approximately 17,000 trials of DMC task and collected average neural activity during the time window of $250 - 500$ ms after the test stimulus onset. A support vector machine (SVM) is trained with least squares method to decode match decision from the RCN's activity using half of the data. The model performance is defined as the accuracy of the trained SVM on decoding another half of the data.

We utilized two measures of functional connectivity, neural response similarity and noise correlation. Neural response similarity ($r_{response}$) quantifies the similarity or difference in neural encoding during task performance. For example, neurons that strongly prefer red category are very similar to each other ($r_{response} >> 0$), but not similar to neurons that are weakly selective to categories ($r_{response} \sim 0$) and markedly different from neurons that prefer blue category ($r_{response} << 0$). $r_{response}$ is defined by the covariance (unnormalized measure of correlation) between synaptic currents of a neural pair under noiseless condition. Note that this is different from r_{signal} in other literature, which is defined as a Pearson correlation between averaged firing rates of a neural pair across stimuli Cohen and Kohn (2011). Noise correlation, r_{noise}, is the Pearson correlation between the rate fluctuations of a neuron pair, averaged across all stimulus directions. Both $r_{response}$ and r_{noise} are functions of time. We calculated them at every 250 ms time window, from the beginning of the sample epoch to the end of test period. Results shown generally hold for all time windows. The range of $r_{response}$ reported are averaged across all time windows and across all model realizations.

For Figure 7D, we also calculated the mean category selectivity duration (CSD) of each neural pair. Neurons with no category tuning (CSD $= 0$) were removed from the analysis. To investigate the dependence of r_{noise} on CSD, neural pairs were segregated into two groups: one group with mean CSD above 90th percentile and another group with CSD below 90th percentile. Then we compared r_{noise} between the two groups with t-test. The same procedure was performed using the synaptic

coupling, $c_{ij} = w_{ij} + w_{ji}$, between two neurons instead of r_{noise} and the same results were obtained. Also, since r_{noise} and average neural activity are correlated and this correlation can drive the result, we controlled for average neural activity using ANCOVA. The mean r_{noise} of the two groups is adjusted based on the fitted correlation between r_{noise} and average neural activity across all task conditions before the difference between groups is assessed.

QUANTIFICATION AND STATISTICAL ANALYSIS

Analyzing temporal properties of selectivity

Neurons with average firing rates less than 1 Hz in each and every epoch were removed from all analyses (leaving $N = 156, 74, 380$ for LIP1, LIP2, and PFC, respectively).

Spike trains of individual trials were filtered with a sliding Gaussian kernel of 90 ms width and 30 ms time step. We focused our analysis on the time period from the onset of the fixation epoch to 250 ms after the test stimulus onset. Only correct trials are included in the analysis.

To analyze selectivity time course, we combined different measures. First, for each time window, we tested whether neurons' firing rates are significantly modulated by stimulus categories (t-tests, $P < 0.05$, Bonferroni corrected) and defined 'a category selectivity phase' as a series of consecutive time windows where neurons are significantly selective to categories. Any selectivity phases shorter than 150 ms were removed from the analysis. The category selectivity duration (CSD) is defined as the duration of its longest category selectivity phase. Categories selectivity phases are plotted in Figure 3A-D and Figure S2B-C. CSD appears in Figure 7D and Figure S2D-G.

In addition, we used 'stimulus selectivity' measure which is very similar to 'category selectivity' measure described above. The only difference is for each time window we tested whether firing rates depend significantly on stimulus directions (one-way ANOVA, $P < 0.05$, Bonferroni corrected) rather than stimulus categories. Then we defined 'stimulus selectivity phase' as a series of consecutive time windows where neurons are significantly selective to directions. Stimulus selectivity duration is plotted in Figure S6C-E.

For category selectivity, we also calculated category selectivity preferences and magnitude at each time window. To achieve this, we assessed the strength of category tuning with a d' measure for each time window in category selectivity phase, defined as:

$$d' = \frac{\mu_1 - \mu_2}{\sqrt{\frac{1}{2}(\sigma_1^2 - \sigma_2^2)}},$$

where μ_i and σ_i is the mean and standard deviation of firing rates in response to stimuli in category i. d' is unbounded and varies from positive values (neuron preferring category blue) to negative values (preferring category red). d' measure is color-coded in Figure 3A-D.

Neurons are classified into three classes based on the properties of their category selectivity phases. (1) Persistent selectivity describes neurons that have only one selectivity phase, which overlaps sample, delay, and test epochs for at least 150 ms. (2) Partial selectivity refers to neurons that are selective to only one category, but not included in the persistent selectivity group. (3) Switching selectivity covers neurons that switch their category preference at least once during the trial. We calculated the proportion of neurons in each group relative to the total number of category-selective neurons. This classification method is used in Figure S2A.

Another measure of category tuning strength we used is category tuning index (CTI). We identified groups of stimulus pairs with the same distance (direction pairs of 30°, 60°, 90°, 120° apart) and within each group split the pairs into two subgroups (same category v.s. different categories). The CTI measured the difference in firing rate (averaged across all trials for each direction) for each neuron between pairs of directions in different categories (a between category difference) and the difference in activity between pairs of directions in the same category (a within category difference). The CTI was defined as the difference between the within category and between category differences divided by their sum. Values of the index could vary from 1 (strong differences in activity to directions in the two categories) to -1 (large activity differences between directions in the same category, no difference between categories). A CTI value of 0 indicates the same difference in firing rate between and within categories.

Population response analysis

To extract population response patterns, we applied demixing principal component analysis (DPCA) to the firing rate traces averaged by task conditions. The algorithmic details and mathematical justification are outlined in ref. Brendel, Romo, and Machens (2011; C. K. Machens, Romo, and Brody 2005; C. K. Machens 2010; C. Machens, Romo, and Brody 2010). In brief, DPCA computed marginalized covariance matrices, denoted C_ϕ, that account for neural response variance due to a subset of task variables, $\phi \in \{t, \theta, \{t, \theta\}\}$. Matrices C_ϕ can be computed by first calculating the marginalized average:

$$\begin{aligned} \overline{y}_t &= \langle \mathbf{r}(t,\theta) \rangle_\theta, \\ \overline{y}_\theta &= \langle \mathbf{r}(t,\theta) \rangle_t, \\ \overline{y}_{t,\theta} &= \mathbf{r}(t,\theta) - \overline{y}_t - \overline{y}_\theta, \end{aligned}$$

and finding marginalized covariance through equation:

$$C_\phi = \langle \overline{y}_\phi \overline{y}_\phi^T \rangle$$

The variance captured by any subset of weight vectors, $\mathbf{W} = [\mathbf{w}_1, \ldots, \mathbf{w}_j, \ldots]$, due to a variable subspace ϕ is denoted by $v_\phi^2(\mathbf{W}) = \sum_j \mathbf{w}_j^T C_\phi \mathbf{w}_j$. DPCA algorithm searches for a set of orthogonal weight vectors, $\mathbf{W}^*$, that maximizes the cost function, $\mathrm{L}=\frac{\sum_\phi v_\phi^2}{\sum_\phi v_\phi}$. Maximizing L will optimize the trade-off between two objectives: enlarging the numerator that represents the overall variance captured by $\mathbf{W}^*$ and downsizing the denominator to ensure that each component only captures variance in a single variable subspace Brendel, Romo, and Machens (2011).

Average neural activity was divided into three portions based on the task epochs: sample ($-100 - 650$ ms relative to sample onset), delay ($800 - 1550$ ms), and test ($1600 - 2150$ ms). For the model, the activity during test epoch came from a slightly wider time window ($1600 - 2650$ ms) to capture the convergence to match and non-match states, which occurs slightly slower in the model. DPCA analysis was performed on simulated activity of the model in presence of noise, using comparable number of trials as in the data (~ 400 trials). We first de-noised the neural responses by applying regular principal component analysis and focusing on the subspace of M largest components that explain 95% of variance in the data (M may vary from dataset to dataset). Then the activity was passed through the DPCA algorithm, which yields M components ranging from the most stimulus-dependent component ($\mathbf{w}_\theta$), to the component that captures most combined variance of stimulus and time ($\mathbf{w}_{\theta,t}$), to the most time-varying component ($\mathbf{w}_t$). For the sample and test period, we removed $\mathbf{w}_t$ from the $\mathbf{W}^*$ matrix and derived two-dimensional representation of activity within the remaining subspace using classical multidimensional scaling Borg and Groenen (1997) (MDS). Stimuli are represented as vectors in an $M - 1$ dimensional space; each dimension corresponds to a principal component. The MDS algorithm searches for 2D coordinates of stimuli that preserve their pairwise Euclidean distances. For the delay period, we projected activity during $-250 - 2000$ ms onto $\mathbf{W}^*$, $\mathbf{z}(t) = \mathbf{W}^*\mathbf{Y}(t)$ to visualize the overall neural activity within a subspace spanned by delay-related components. Figure E-H plotted projected activity on $\mathbf{w}_\theta$ and $\mathbf{w}_{\theta,t}$.

Note that since we do not have neurophysiological recordings from one monkey for trials where test stimuli are 15° away from the boundary (see EXPERIMENTAL MODEL AND SUBJECT DETAILS), we need to remove all trials with test stimuli 15° away from the boundary to construct population trajectories shown in Figure 4I. This step was not applied to other analyses in this paper.

DATA AND SOFTWARE AVAILABILITY

Software for modeling and data analysis is written in MATLAB. Requests for source code and data should be directed to our Lead Contact.

KEY RESOURCES TABLE

REAGENT or RESOURCE	**SOURCE**	**IDENTIFIER**
Software and Algorithms		
Hessian-Free algorithm	Martens and Sutskever, 2011	http://www.cs.toronto.edu/~ilya/pubs/
DPCA algorithm	Machens et al., 2010; Brendel et al., 2011	https://github.com/machenslab/dPCA

Figure 1

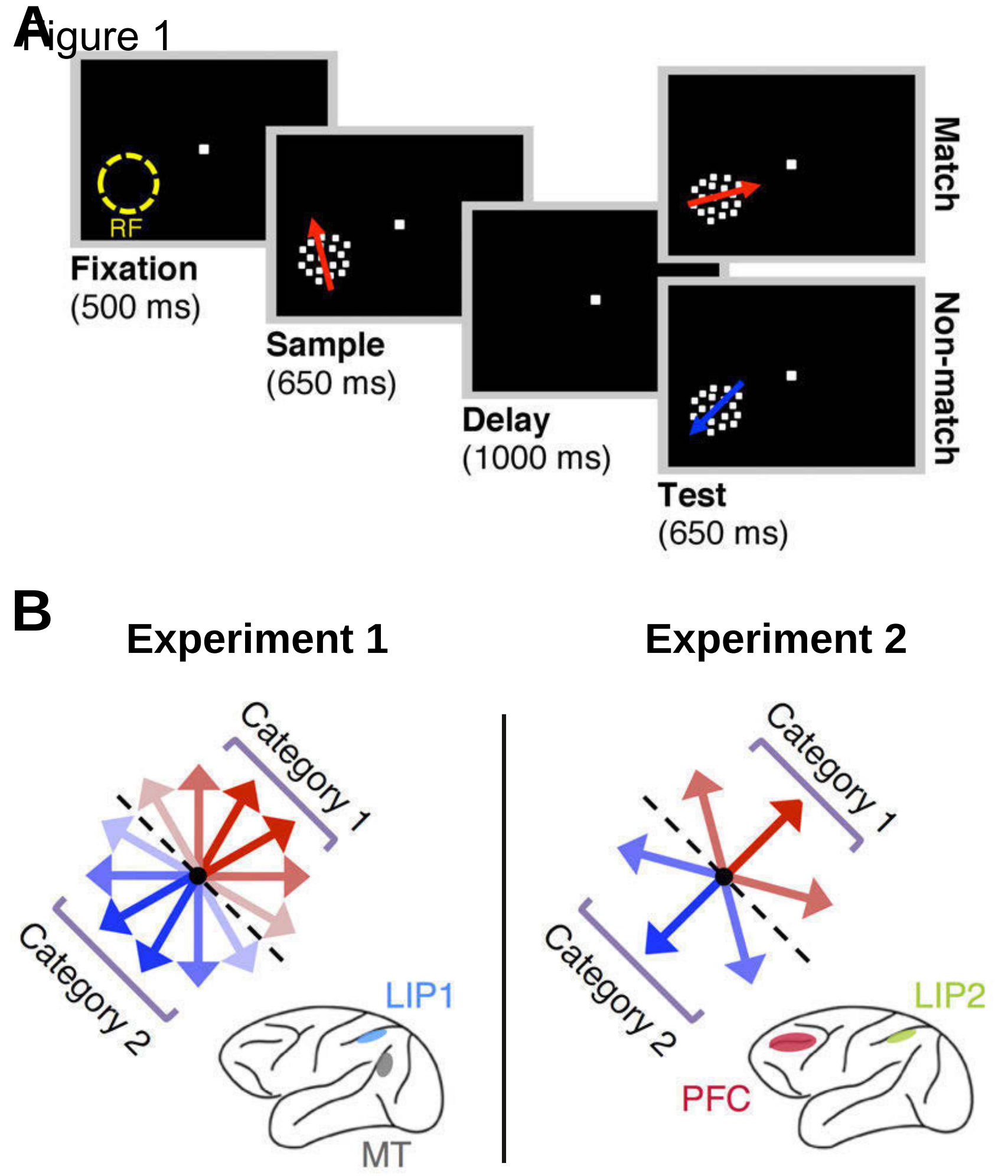
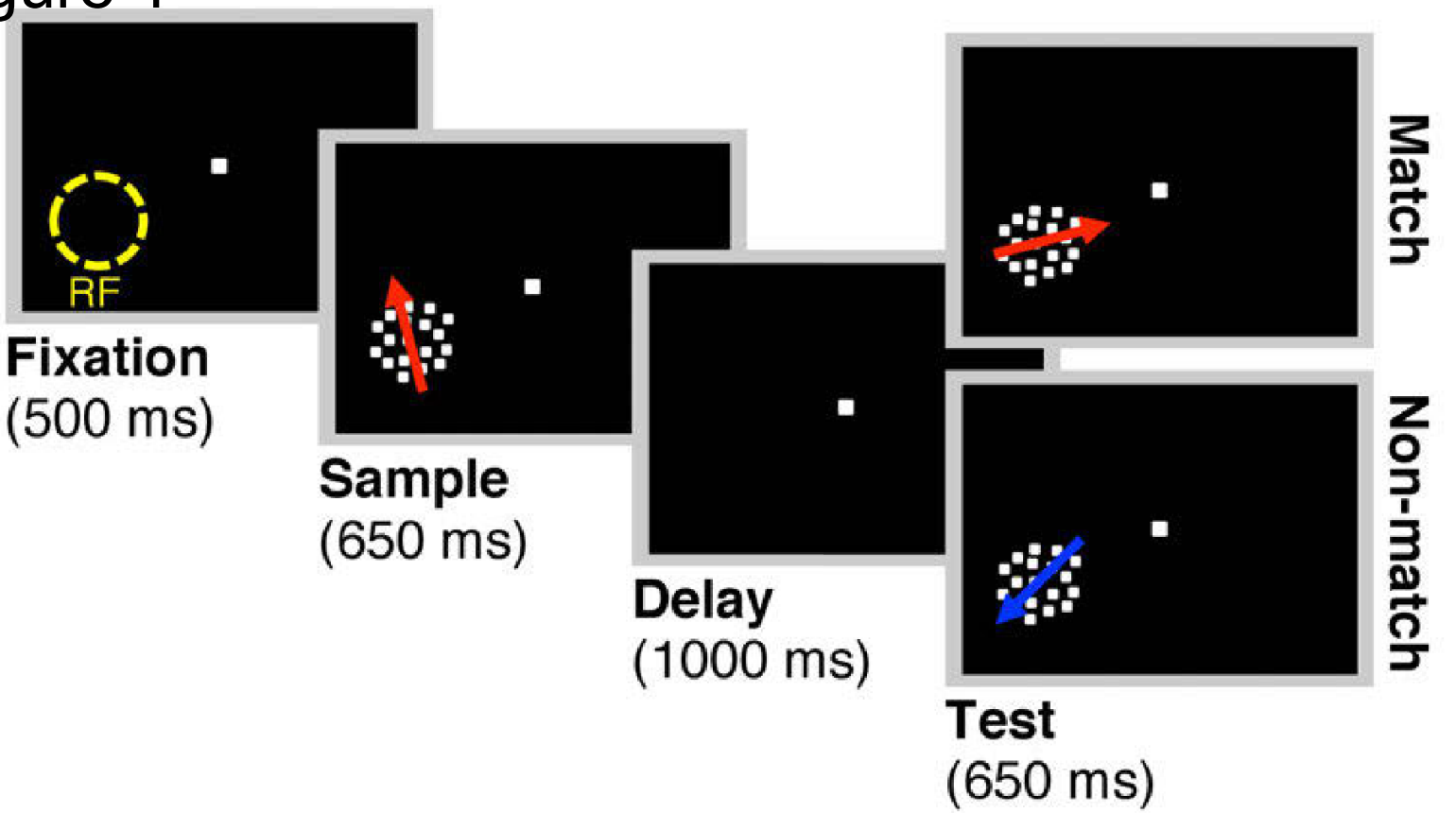

A
RF
Fixation
(500 ms)
Sample
(650 ms)
Delay
(1000 ms)
Test
(650 ms)
Match
Non-match
B
Experiment 1
Experiment 2
Category 1
Category 2
LIP1
MT
Category 1
Category 2
LIP2
PFC

Figure 1


A
RF
Fixation
(500 ms)
Sample
(650 ms)
Delay
(1000 ms)
Test
(650 ms)
Match
Non-match

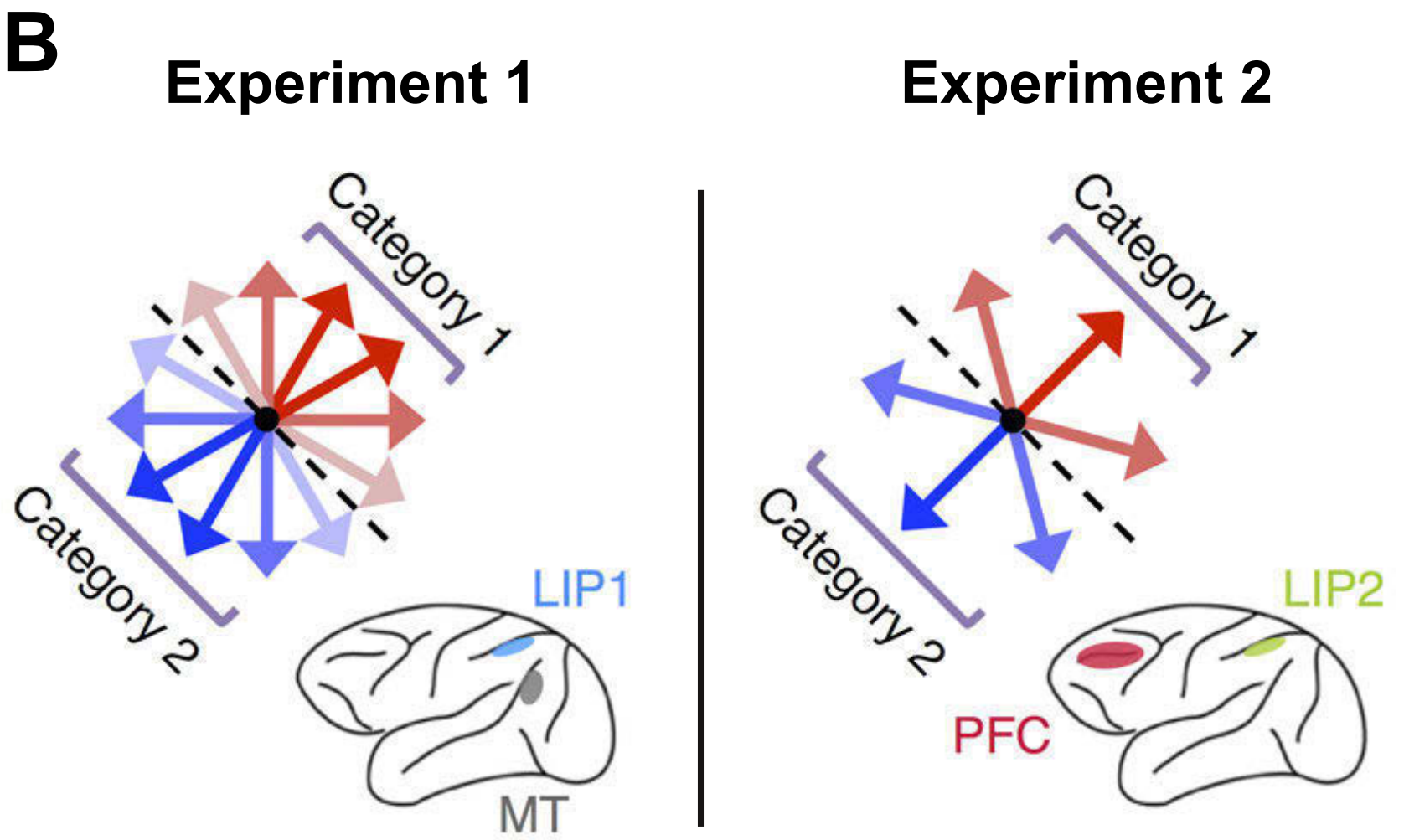

B
Experiment 1
Experiment 2
Category 1
Category 2
LIP1
MT
Category 1
Category 2
LIP2
PFC

Figure 2

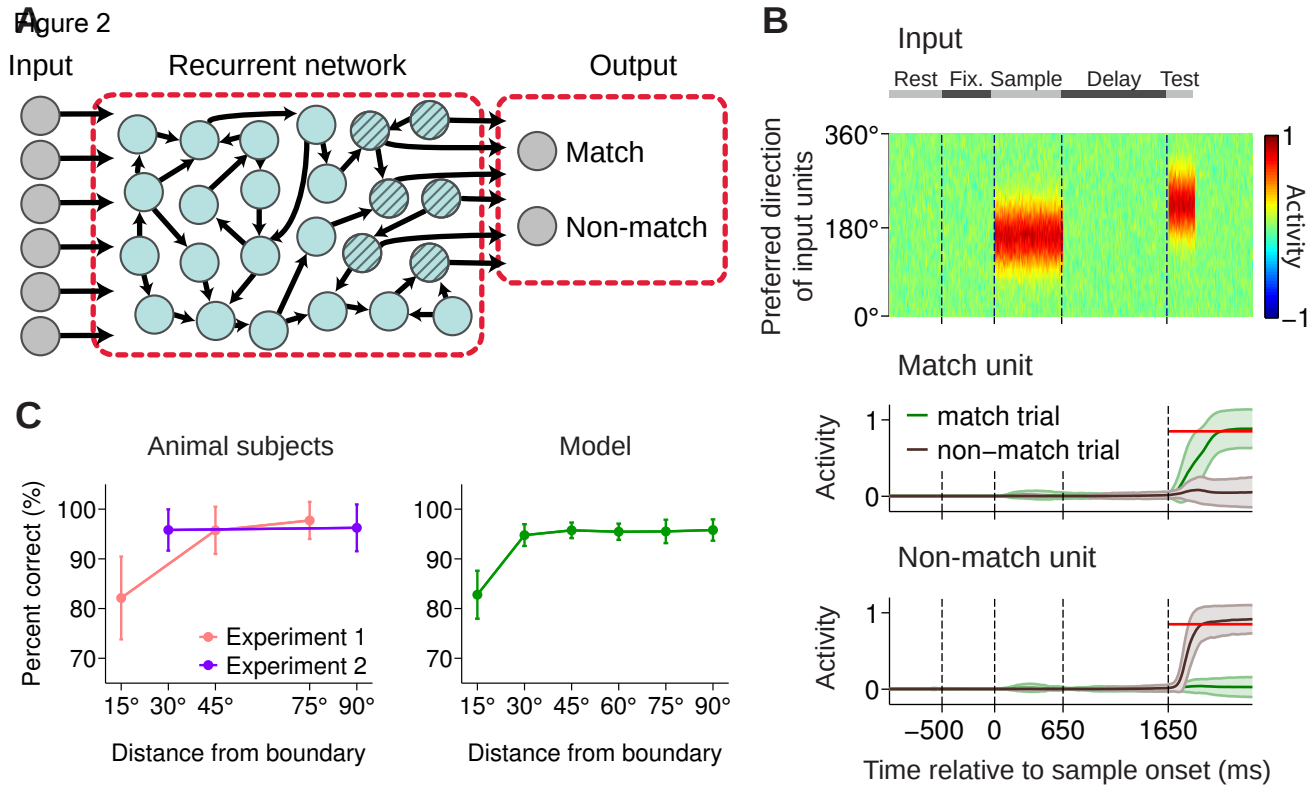

A
Input
Recurrent network
Output
Match
Non-match
B
Input
Rest
Fix.
Sample
Delay
Test
Preferred direction of input units
360°
180°
0°
1
Activity
−1
Match unit
Activity
1
0
match trial
non−match trial
Non-match unit
Activity
1
0
−500
0
650
1650
Time relative to sample onset (ms)
C
Animal subjects
Model
Percent correct (%)
100
90
80
70
Experiment 1
Experiment 2
15°
30°
45°
60°
75°
90°
Distance from boundary
Distance from boundary

Figure 2

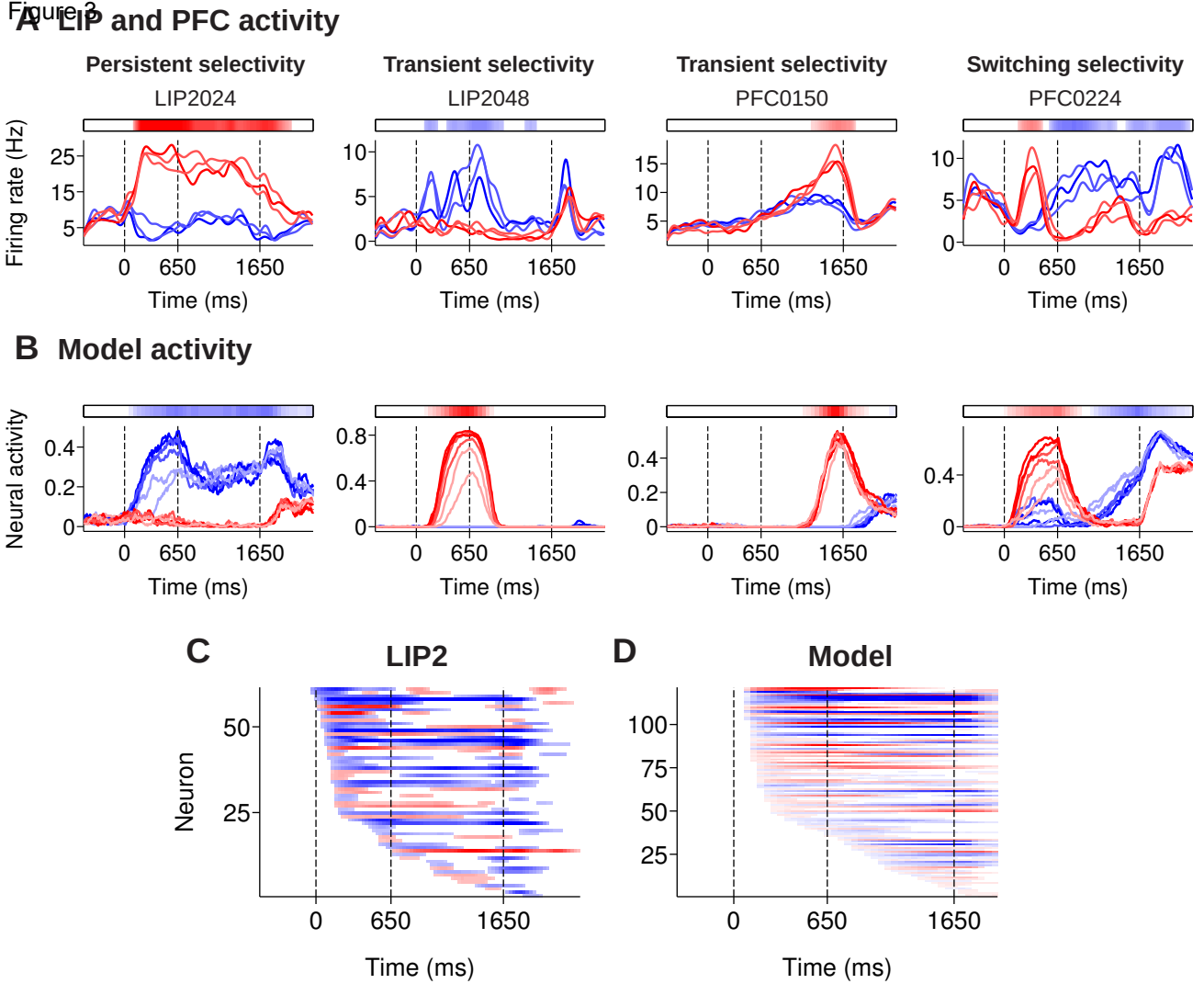

A LIP and PFC activity
Persistent selectivity
LIP2024
Transient selectivity
LIP2048
Transient selectivity
PFC0150
Switching selectivity
PFC0224
Firing rate (Hz)
Time (ms)
B Model activity
Neural activity
C
LIP2
D
Model
Neuron

Figure 4

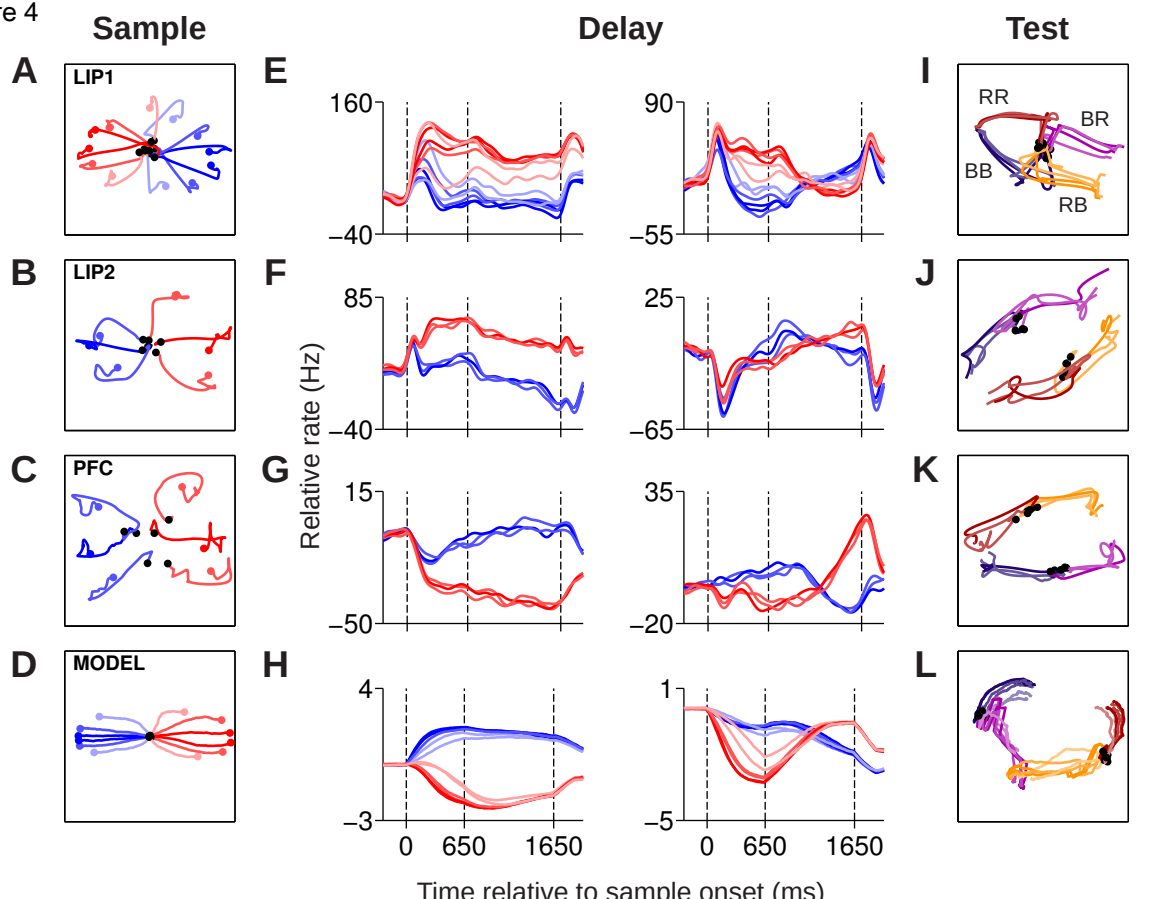
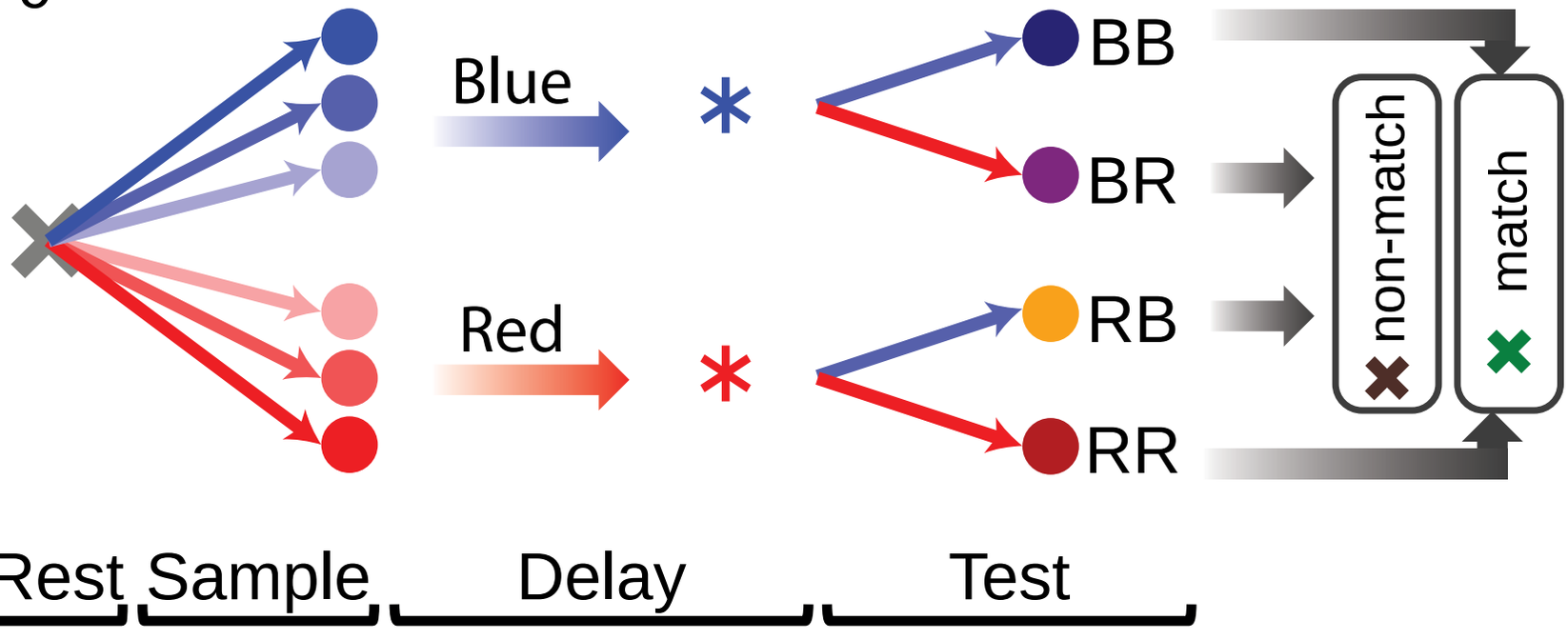

Sample
Delay
Test
A
LIP1
B
LIP2
C
PFC
D
MODEL
E
F
G
H
160
−40
90
−55
85
−40
25
−65
15
−50
35
−20
4
−3
1
−5
Relative rate (Hz)
0
650
1650
0
650
1650
Time relative to sample onset (ms)
I
J
K
L
RR
BR
BB
RB

Figure 5


A
Blue
Red
BB
BR
RB
RR
non-match
match
Rest
Sample
Delay
Test

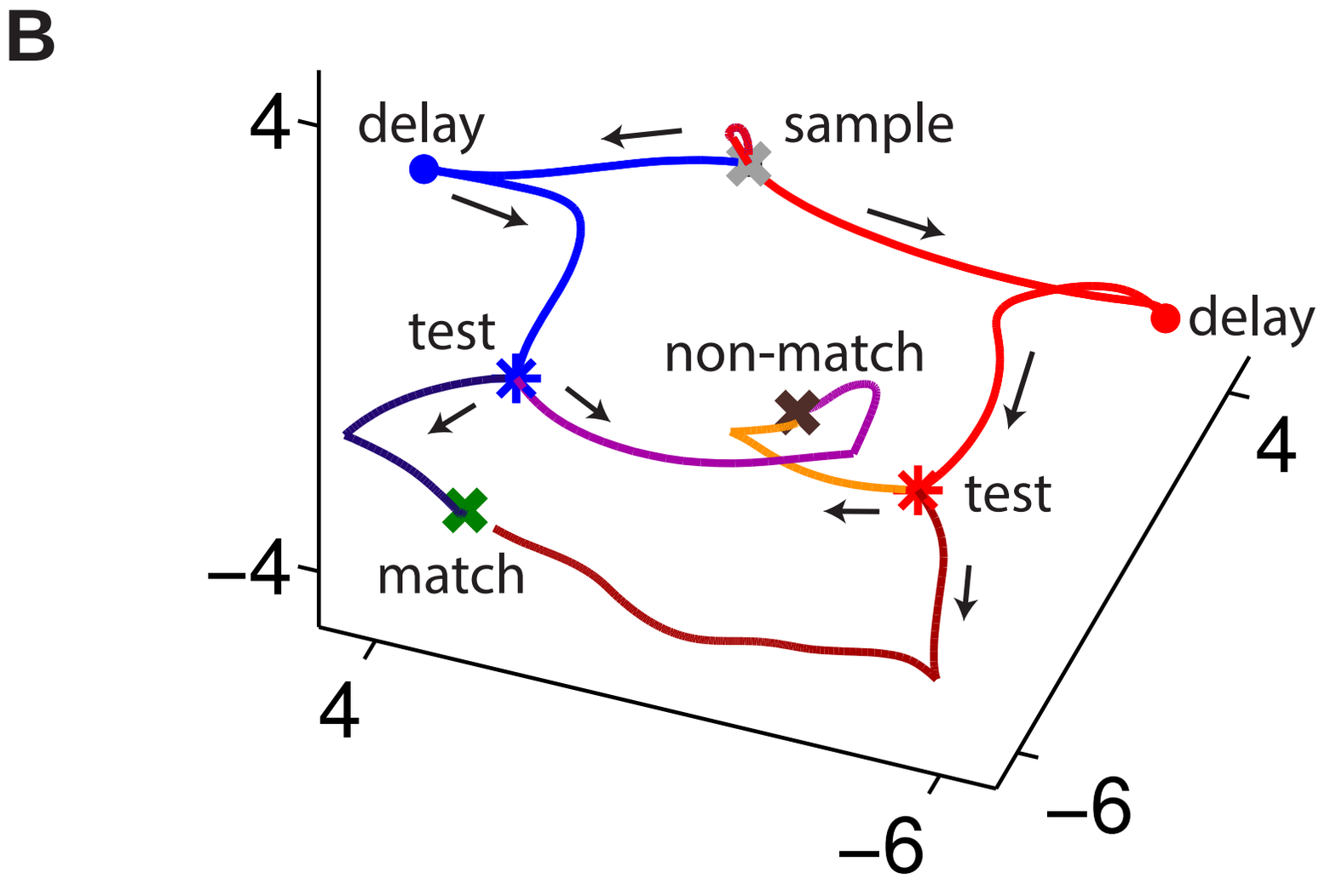

B
4
delay
sample
test
non-match
delay
test
match
−4
4
4
−6
−6

Figure 6

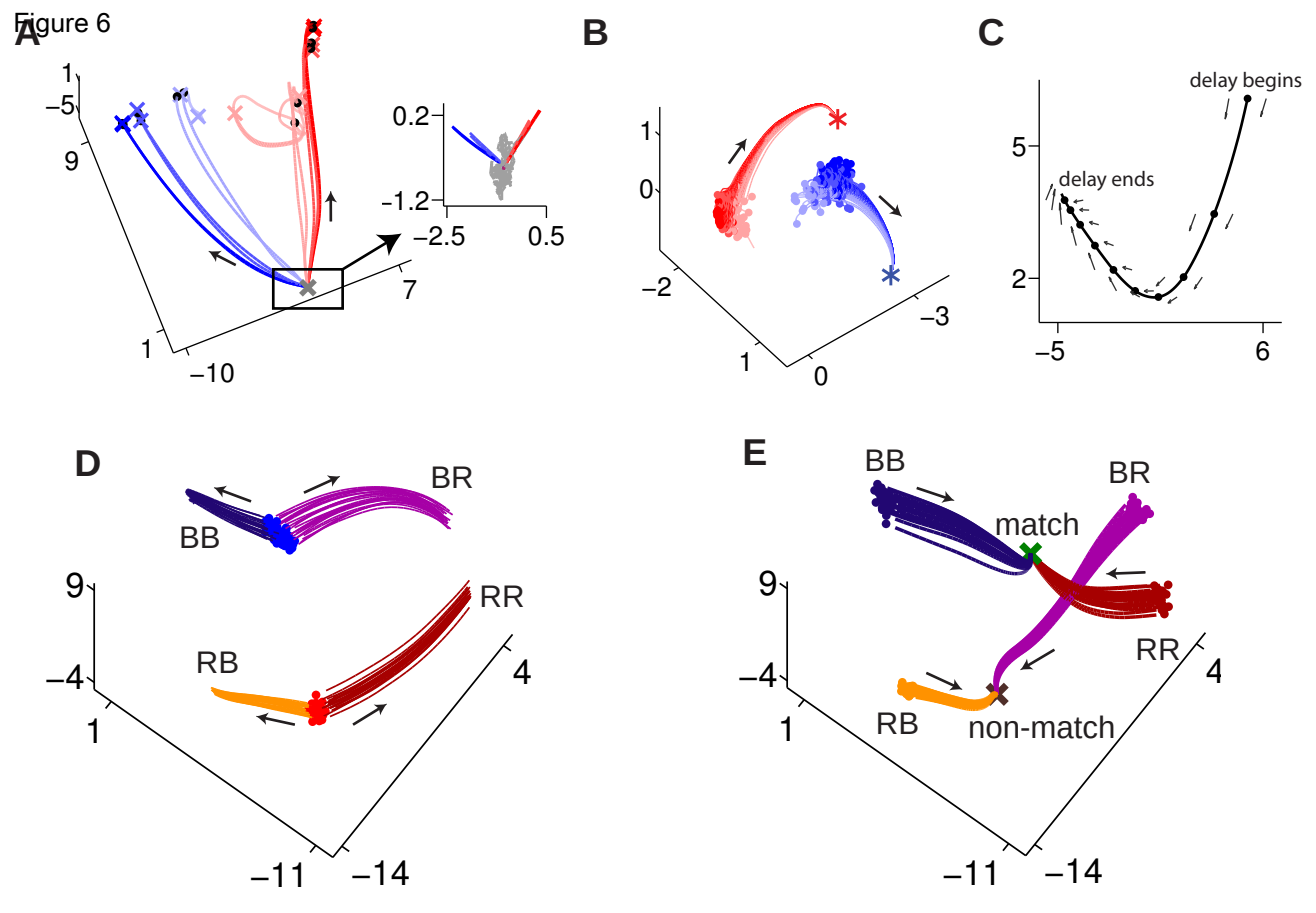

A
1
−5
9
1
−10
7
0.2
−1.2
−2.5
0.5
B
1
0
−2
1
0
−3
C
delay begins
delay ends
5
2
−5
6
D
BB
BR
RB
RR
9
−4
1
−11
−14
4
E
BB
BR
match
RB
RR
non-match
9
−4
1
−11
−14
4

Figure 7

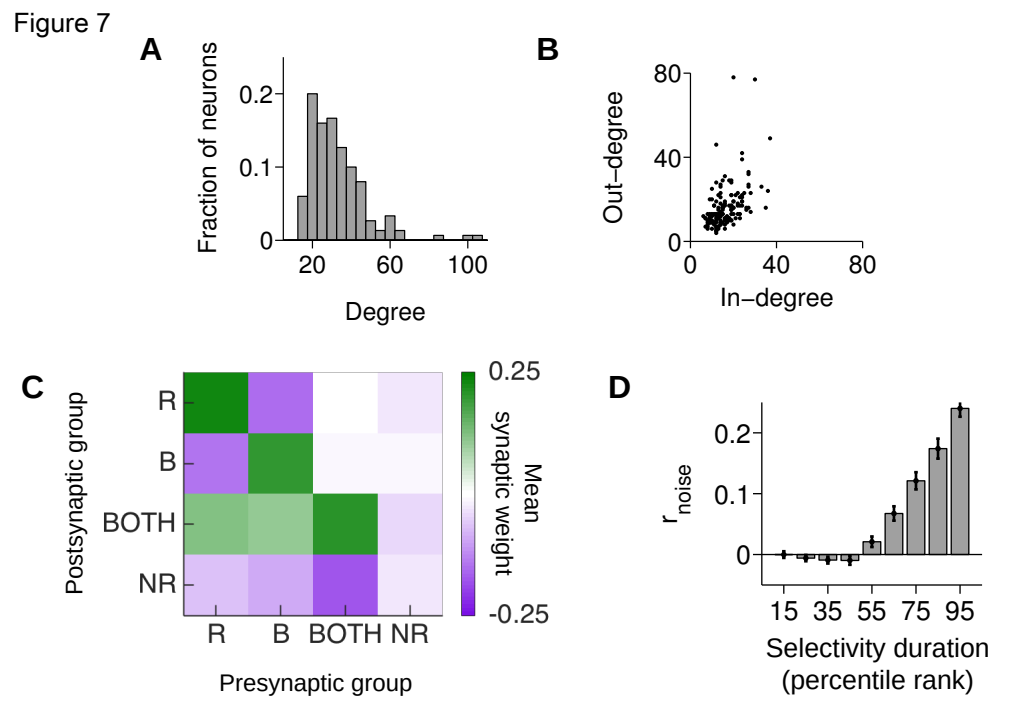
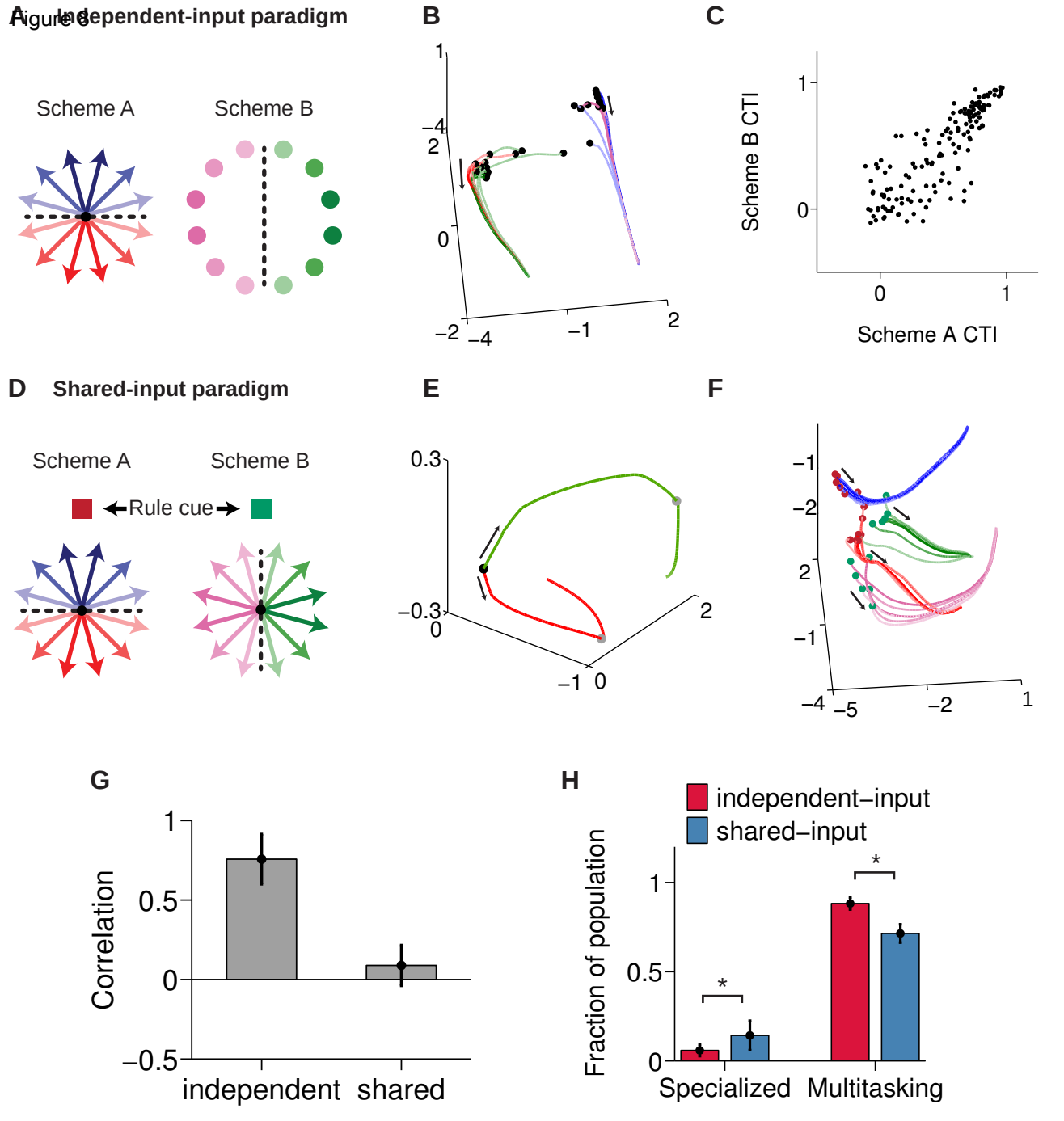

A
Fraction of neurons
0.2
0.1
0
20
60
100
Degree
B
Out–degree
80
40
0
0
40
80
In–degree
C
Postsynaptic group
R
B
BOTH
NR
R
B
BOTH
NR
Presynaptic group
0.25
Mean synaptic weight
-0.25
D
r noise
0.2
0.1
0
15
35
55
75
95
Selectivity duration (percentile rank)

Figure 3


A Independent-input paradigm
Scheme A
Scheme B
B
C
Scheme B CTI
Scheme A CTI
D Shared-input paradigm
Scheme A
Scheme B
Rule cue
E
F
G
Correlation
independent
shared
H
independent–input
shared–input
Fraction of population
Specialized
Multitasking

Supplemental Figures

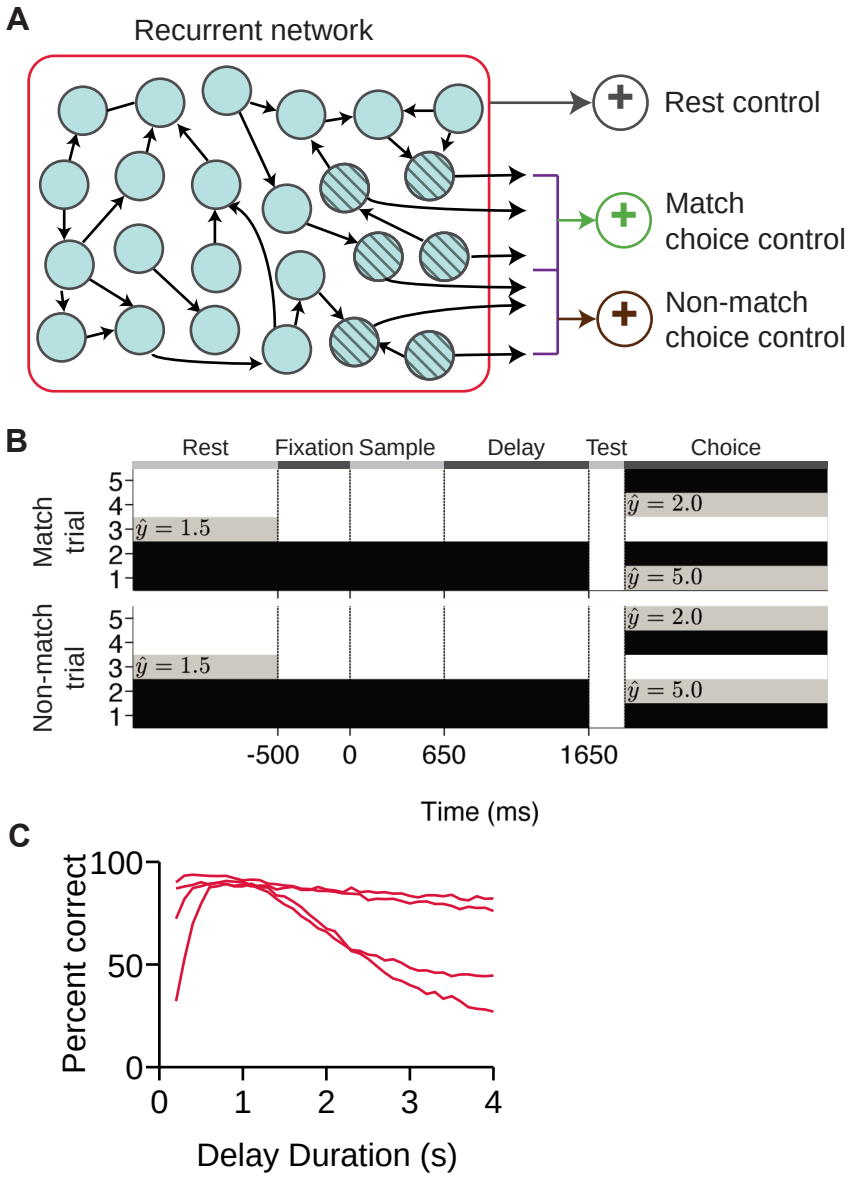


Figure S1 Related to Figure 2. Recurrent network training strategies. (**A**) A schematic showing additional output units of the model. Input units are excluded for clarity. In total, there are five output units; two units were drawn in Figure 2a and three other units are illustrated here. These additional units help stabilize the network during the resting and post-response period; without them the trained networks may end up in oscillating states before and after the task is performed. (**B**) Target activity, $\hat{\mathbf{y}}$, of all output units during match trial (top panel) and non-match trial (bottom panel). Black color indicates $\hat{y}_i(t) = 0$; gray, $\hat{y}_i(t)$ has the value written on the graph. White color indicates that $\hat{y}_i(t)$ is not specified; therefore the corresponding network output $y_i(t)$ is not taken into account during optimization. The five output units can be broken down into three types. (**1**) Action units (drawn in Figure 2a; match and non-match neurons correspond to unit 1 and 2 in (**B**), respectively). The target activity of action units is zero from resting period to the end of delay. During the choice period, the match unit has value of 5 for match trials and value of 0 for non-match trials. The opposite pattern holds for the non-match unit. The action units were instructed to keep their activity at zero prior to the test period to prevent premature response before the test stimulus is revealed. Action units receive plastic connections from 30 recurrent neurons, where all connections are initialized to have zero value. (**2**) A rest control unit (gray circle in (**A**); unit 3 in (**B**)) receives non-plastic synapses of weight = 1 from all recurrent neurons, effectively computing the sum all recurrent activity and holds a target activity of 1.5 during the rest period. (**3**) Choice control units (match or non-match units correspond to green or brown circles in (**A**) and unit 4 or 5 in (**B**), respectively). The match unit computes the sum of activity from the 15 neurons that are connected to action units. During the choice period, it has target activity of 2.0 for match trials and zero for non-match trials. The opposite pattern holds for the non-match choice control unit. Note that if the output units were not constrained to hold stable activity for a long time window after test stimulus disappears, some networks may develop oscillating end states, which might be hard to interpret. In the cortical circuit, there might be an additional mechanism to shut down or reset the dynamics after the trial is complete. (**C**) Percent correct responses as a function of delay durations. Four lines correspond to four different networks that were successfully trained starting from different initial conditions. To get a percent correct measure, we collected the network choice on each trial. For example, if the match output neuron has activity over a certain threshold (0.85) while the non-match neuron activity stays under the threshold, we considered that a match choice. We varied the delay duration from 0.2 to 4 seconds with 100 ms interval. For each delay duration, we simulated 500 trials of DMC tasks and measured percent correct trials. After training, the model can perform DMC task for a good range of delays ($0.7 - 1.3$ s).

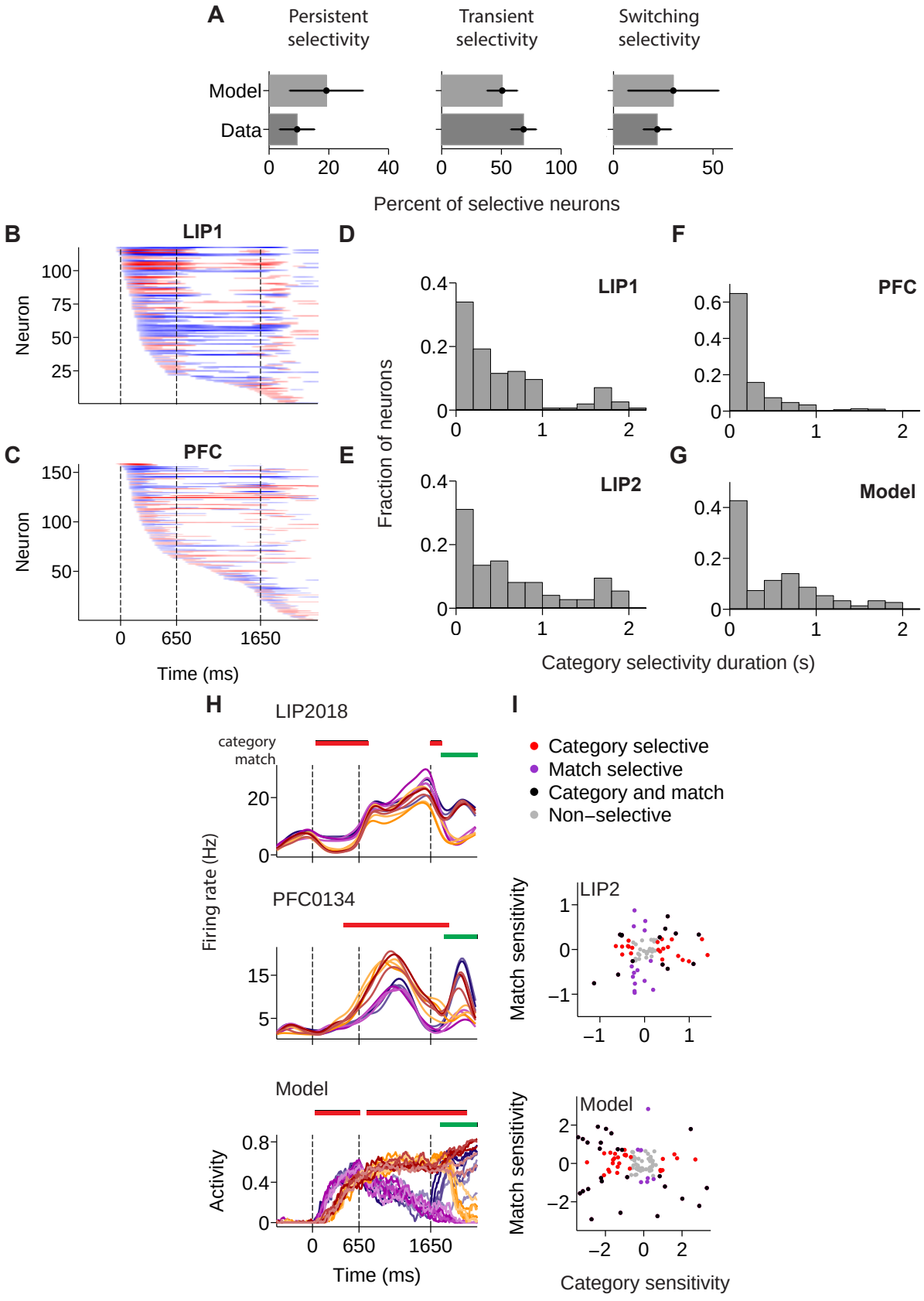

A
Persistent selectivity
Transient selectivity
Switching selectivity
Model
Data
0
20
40
0
50
100
0
50
Percent of selective neurons
B
LIP1
Neuron
100
75
50
25
C
PFC
Neuron
150
100
50
0
650
1650
Time (ms)
D
LIP1
Fraction of neurons
0.4
0.2
0
0
1
2
E
LIP2
0.4
0.2
0
0
1
2
F
PFC
0.6
0.4
0.2
0
0
1
2
G
Model
0.4
0.2
0
0
1
2
Category selectivity duration (s)
H
LIP2018
category
match
20
0
Firing rate (Hz)
PFC0134
15
5
Model
Activity
0.8
0.4
0
0
650
1650
Time (ms)
I
Category selective
Match selective
Category and match
Non–selective
LIP2
Match sensitivity
1
0
–1
–1
0
1
Model
Match sensitivity
2
0
–2
–2
0
2
Category sensitivity

Figure S2 Related to Figure 3. Model replicates the temporal profiles of category selectivity and the mixture of category and match/non-match selectivity observed in the data. (**A**) We classified neurons into three categories. (1) "persistent selectivity" group are neurons that have only one selectivity phase that lasts through the delay period and overlaps sample and test epochs for at least 150 ms. (2) "transient selectivity" group are neurons whose selectivity phases overlap with one or two task epochs. (3) "switching selectivity" are neurons that switch their category preference in the middle of a trial. Both recorded neural populations and model populations contain comparable fractions of neurons in each class. The majority of neurons have transient selectivity (68.6% for data, 50.7% for model), followed by switching group (22.0% for data, 30.0% for model) and persistently selective group (9.4% for data, 19.2% for model). Note that there is a quantitative variation in percentages across different instances of the model trained by the same protocol (s.d. shown), but the overall trend is similar to that observed in the LIP and PFC neural populations. (**B**-**C**) Colored heat maps showing category selectivity phases of all neurons in LIP1 (**B**) and PFC (**C**) datasets. Same plotting convention as in Figure 3C-D is used. (**D**-**G**) The distribution of category selectivity durations for all three neural datasets and model. Small portion of neurons with persistent category selectivity is consistently detected in all populations, leading to heavy-tail distributions. (**H**) Average firing rate traces of neuron samples from LIP (top), PFC (middle), and model (bottom). Colors code task conditions during test period: dark red corresponds to RR condition (sample in red category, test in red category); dark blue to BB; orange, RB; purple, BR. These neurons show both category-selective response during sample and/or delay periods and match-selective response during the test period. The bars above the traces indicate the time windows where firing rates are significantly selective to categories (top bar) and match decisions (bottom bar, t-tests, $P < 0.05$, Bonferroni corrected). (**I**) Category (x-axis) and match (y-axis) sensitivity measured by d' are independent and, for both neural data (top, LIP2) and model (bottom), we observed neurons that exhibit both category and match selectivity (black dots), in addition to those showing only category (red), match (purple), or no (gray) selectivity (paired t-tests, $P < 0.05$, Bonferroni corrected).

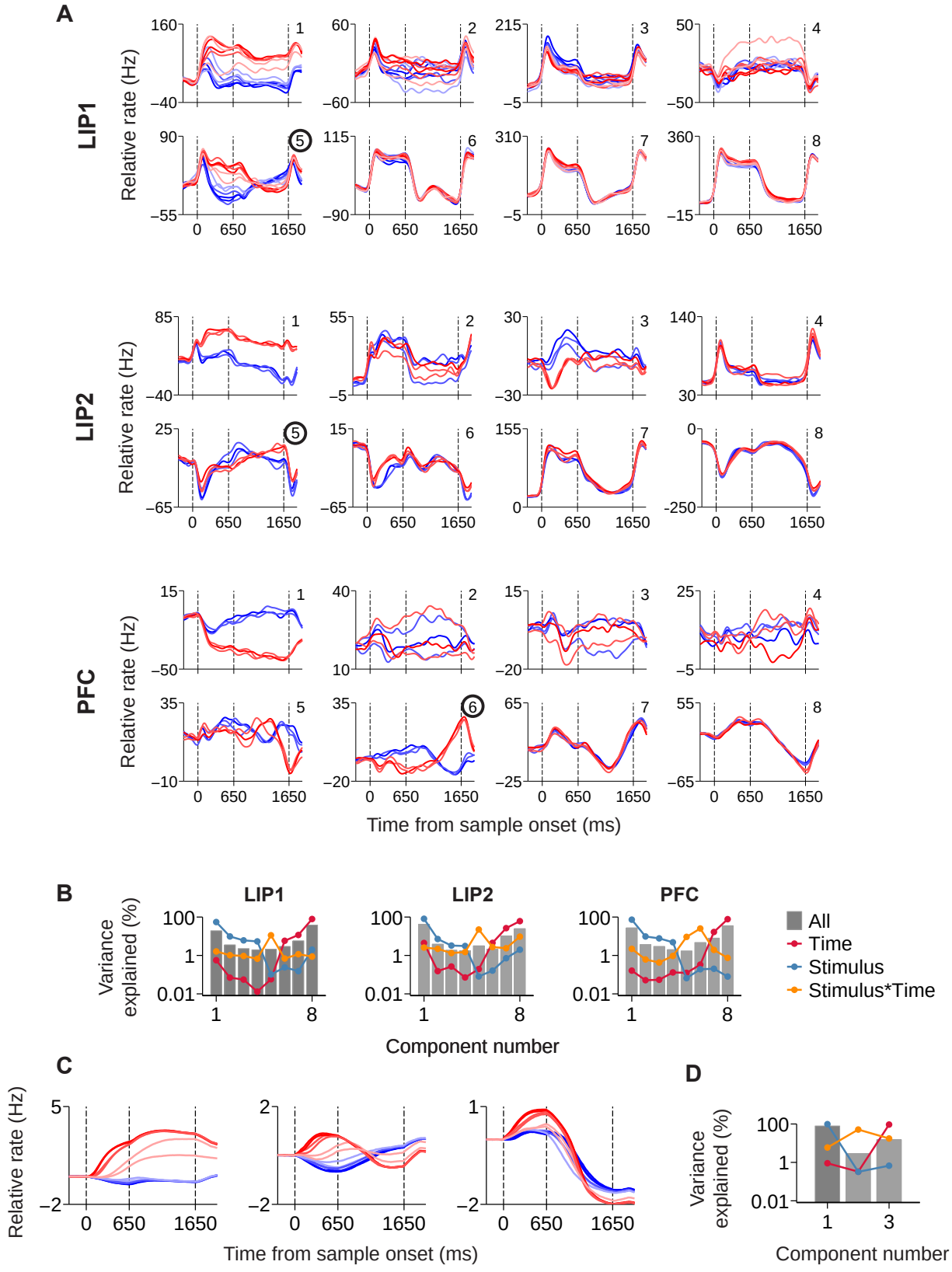

A
LIP1
LIP2
PFC
Relative rate (Hz)
Time from sample onset (ms)
B
LIP1
LIP2
PFC
Variance explained (%)
Component number
All
Time
Stimulus
Stimulus*Time
C
Relative rate (Hz)
Time from sample onset (ms)
D
Variance explained (%)
Component number

Figure S3 Related to Figure 4. Demixed principal components during the delay period.(**A**) Demixed principal components from the delay activity of LIP1, LIP2 and PFC datasets, arranged from the most stimulus-dependent component (number 1) to the most time-varying component (number 8). Only the largest 8 components (based on overall variance captured) are plotted here. Components marked with black circles around the label numbers capture the largest covariance in responses due to time and stimuli. (**B**) Variability in the neural responses captured by the 8 largest components in each population (plotted in log scale). Gray bars indicate overall variance; red lines, time-dependent variance; blue, stimulus-dependent variance; orange, stimulus-time covariance. (**C**) Demixed principal components from the delay activity of a representative model, which displays three components with either a stable category code, a time-varying category code, or a time-varying component with little stimulus encoding. (**D**) Percent variance captured by three components in (**C**). Plotting convention is the same as in **B**. In the model, the time-varying component accounts for relatively small overall variance in the delay response (15%), compared to the neural data (25%-38%). This is expected because our model is a small network extensively trained to perform only DMC tasks where timing information is not relevant. The model network, therefore, has larger category selectivity and little time selectivity. On the contrary, category-related activity is only one of many functions of the real neural networks and timing information might be important for other tasks they perform.

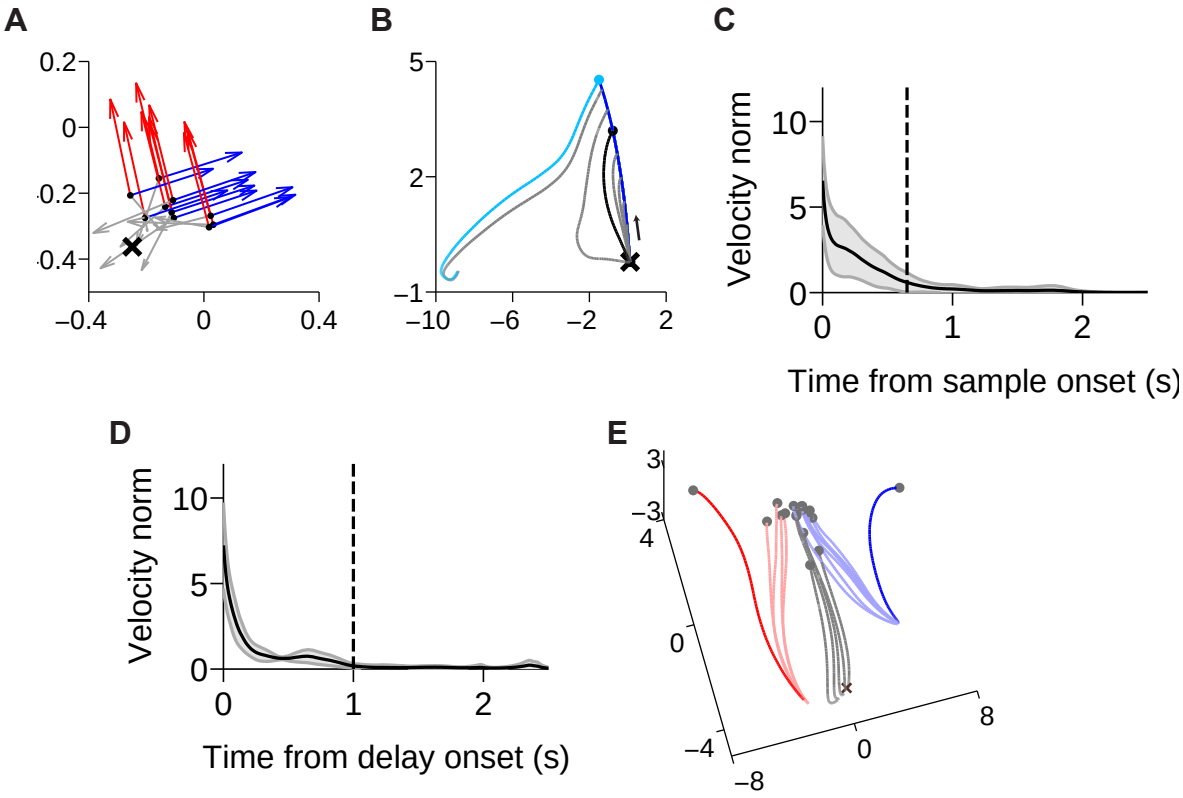

A
0.2
0
0.2
0.4
−0.4
0
0.4
B
5
2
−1
−10
−6
−2
2
C
Velocity norm
10
5
0
0
1
2
Time from sample onset (s)
D
Velocity norm
10
5
0
0
1
2
Time from delay onset (s)
E
3
−3
4
0
−4
−8
0
8

Figure S4 Related to Figure 6. Model dynamics during categorization and category working memory. (**A**) We probed the dynamical landscape around the resting fixed point (black cross, $\boldsymbol{x_0}$), by perturbing the state with random vectors $\boldsymbol{\delta}$, $\delta_i \sim \mathcal{U}(0, 0.09)$, and plotting directions of local velocity. Note that we used excitatory perturbations to provoke the network to leave stable states. Black dots mark the perturbed locations ($\boldsymbol{x_0} + \boldsymbol{\delta}$). With zero input, velocity vectors always point back to the resting state (gray arrows). When inputs represent motion directions in category red (red arrows) or blue (blue arrows), velocity vectors point to separate paths depending on stimulus categories. This demonstrates that local velocity is largely influenced by direction-selective inputs. (**B**) If the network is exposed to consistently strong input for an extended period of time, the neural states eventually escape the basin of resting attractor. We simulated neural trajectories starting at the resting state (black cross) while providing direction-selective input of various durations. The stimulus direction is 255°, which is near the middle of category blue, and stimulus durations range from 10-150 ms. The dark blue line shows the trajectory during stimulus presentation. If the stimulus duration is too short, neural trajectories are attracted back to the resting state upon stimulus removal (black dot marks the neural state at the end of 100 ms stimulus; black line shows the trajectory). Continuous input of consistent direction brings neural state further from the resting state. If the input duration is sufficiently long, relaxed trajectories converge to a different region in state space (light blue dot marks the offset of 150 ms stimulus; light blue line shows the trajectory). (**C**) The norm of network velocity ($\|\dot{\boldsymbol{x}}\|^2$) when the network is exposed to direction-selective input over a period of 2.5 s. Shaded area indicates s.d. of norm across 12 stimuli. Dashed line marks the end of sample period in our standard DMC protocol. Network velocity drops rapidly during stimulus presentation. (**D**) Same plot as in (**C**) over the delay period when the network relaxes in the absence of input. Dashed line marks the end of delay period, where state velocity is relatively low. (**E**) Noiseless neural trajectories under perturbation. We simulated neural dynamics under large perturbation while network categorizes near-boundary stimulus (187°, which is in the red category; $\sigma_r = 2.5$). Here we plot noiseless trajectories as the network relaxes during the delay. Gray dots mark the end-of-sample states. For reference, dark blue and dark red lines show noiseless trajectories during the delay for stimulus directions at the middle of categories. Network can correctly classify the motion direction in some trials, settling at the end-of-delay state associated with red category (pale red lines). Errors occur in cases where neural states land close to and flow along the blue category tunnel (pale blue lines) or fall into basins of attraction associated to resting, match, or non-match states (gray lines; as an example, here we plot trajectories that converge to the non-match state, brown cross).

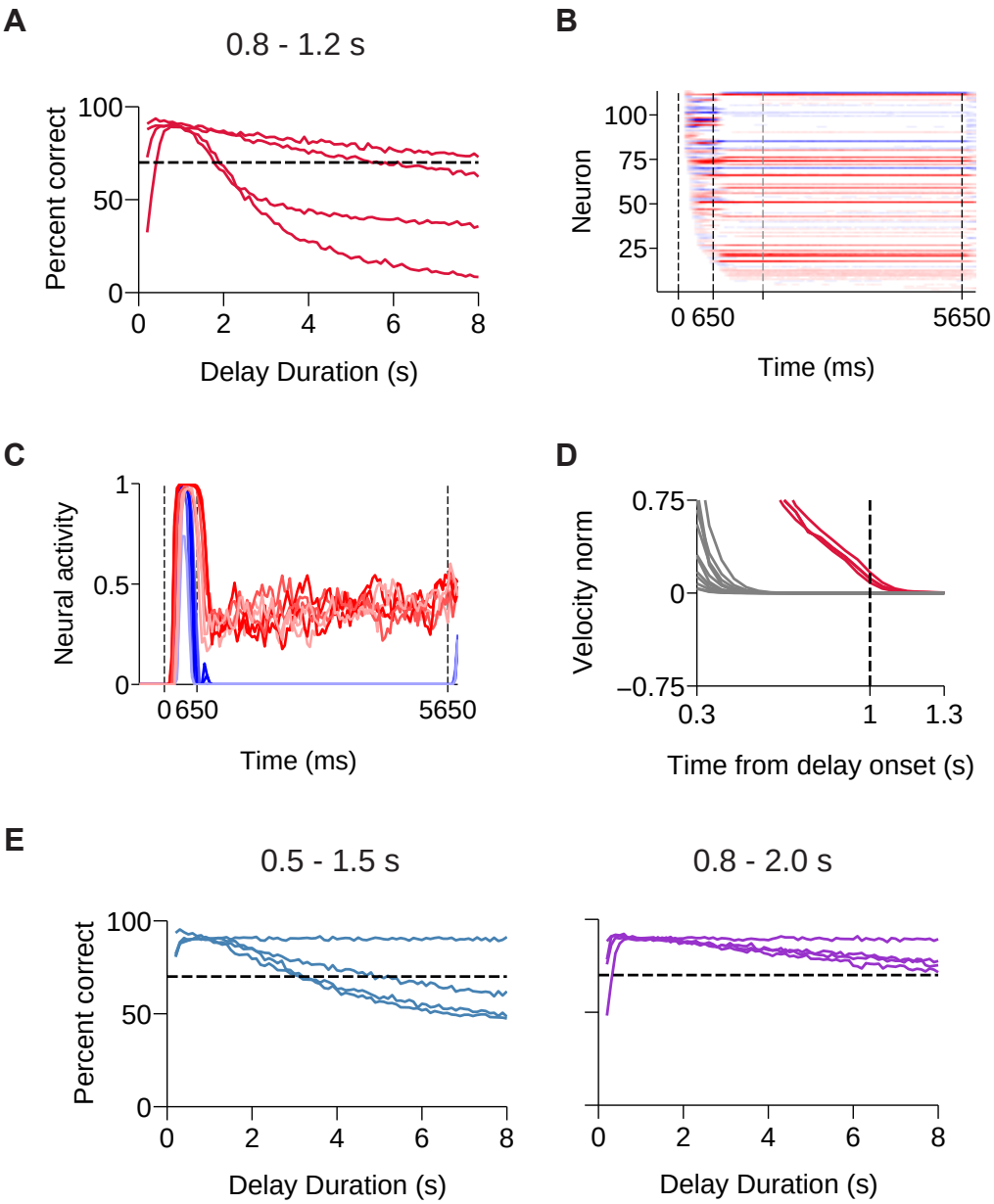

A
0.8 - 1.2 s
Percent correct
100
50
0
0
2
4
6
8
Delay Duration (s)
B
Neuron
100
75
50
25
0 650
5650
Time (ms)
C
Neural activity
1
0.5
0
0 650
5650
Time (ms)
D
Velocity norm
0.75
0
−0.75
0.3
1
1.3
Time from delay onset (s)
E
0.5 - 1.5 s
0.8 - 2.0 s
Percent correct
100
50
0
0
2
4
6
8
Delay Duration (s)
Delay Duration (s)

Figure S5 Related to Figure 6. Networks trained with variable delay durations develop more stable representation. When the delay durations vary from trial to trial and sample categories must be recalled at arbitrary time point, networks are more likely to develop stable fixed points associated to stimulus categories (Barak et al., 2013). To study how the training protocol impacts the dynamics of representation, here we experiment with variants of DMC task in which delay durations spread across larger range of values. (**A**) Four networks were trained to perform the DMC task with delay durations ranging from $0.8 - 1.2$ s. The plot shows percent correct response as a function of delay duration. Black dashed line marks 70% performance for reference. Some networks learn to perform the task at delays much longer than those used during training, while others fail at longer delays. (**B**) Networks that can perform DMC at long delays have stable category representation. The heat map of categorical selectivity strength (d') over 5 s delay (same plotting convention as in Figure 3c-d; the network shown is the most stable sample from (**A**)). (**C**) Neural activity of a sample neuron within the stable network, color coded by sample directions. (**D**) The magnitude of network velocity as a function of time for networks that have stable performance (gray) versus networks that do not (red). The graph focuses on the velocity during the end of delay period. Networks with stable performance reach fixed points before the end of delay period. Note that even networks that reach fixed points at the end of delays may demonstrate a slow decay in performance as the mnemonic epoch extends, because prolonged exposure to random perturbations drive networks to escape working memory steady states. (**E**) Training protocols utilizing larger delay range tend to generate more stable recurrent models. Percent correct response as a function of delay duration for two training protocols using variable delay lengths drawn from ranges of $0.5 - 1.5$ s (left panel) and $0.8 - 2.0$ s (right panel). Networks' performances are more stable as the delay ranges are wider.

Our main findings suggest that the time-varying categorical selectivity results from robust transient dynamics. This framework does not require the network to reach stable states during the delay period, as the decrease in state velocity allows the model to endure a sufficient variation in delay durations as shown in Figure S1. The model's behavior is reasonable compared to monkeys'. In animals trained with fixed delay duration of 1 s, the response accuracy diminishes if the delay period is unexpectedly extended to 1.5 s (unpublished behavioral data in Freedman laboratory). This entails that, when delay period is extended, working memory is disrupted by other signals such as premature behavioral plan. For sufficiently short delay durations, working memory can in principle be sustained by purely transient dynamics.

One question that arises is: What happens if neural networks or animals must perform DMC task at longer or variable delays? Although our datasets do not allow us to study this question directly, this modeling experiment suggests that one candidate mechanism for coping with longer delays is the formation of steady states associated to category working memory. Here we show that such steady states are more likely to occur if training protocols incorporate larger delay range. Consistent with this notion, a previous study that employs variable delay protocol showed that prefrontal networks occupy low-velocity working memory across the

delay (Stokes et al., 2013), in contrast to the current study and others (Meyers et al., 2008; Crowe et al., 2010) with fixed delay that show time-varying representation. To directly test this hypothesis, future studies could compare the temporal variability in the neural representations when animals are trained with fixed versus variable delay protocols.

However, one should be cautious when interpreting this result in the context of neurophysiological experiments. First, animals' performance and neural circuits are impacted by other signals associated with timing, anticipation, impatience, etc. Even if a subset of population in higher cortical areas maintains stable representation, others can contain temporal variations due to other signals. Thus, it is difficult to distinguish memory-related activity from other activity. Second, the current model does not incorporate other mechanisms to mediate timing, therefore the formation of fixed points may arise simply due to lack of alternatives. Another study shows that, in monkeys, the ability to generalize to different delays may rely on changes in speed of neural dynamics (Brody et al., 2003a), which could be mediated by synaptic mechanisms (Barak et al., 2010) not integrated in our model. Future experiments should compare stable versus variable delay protocols to establish this issue.

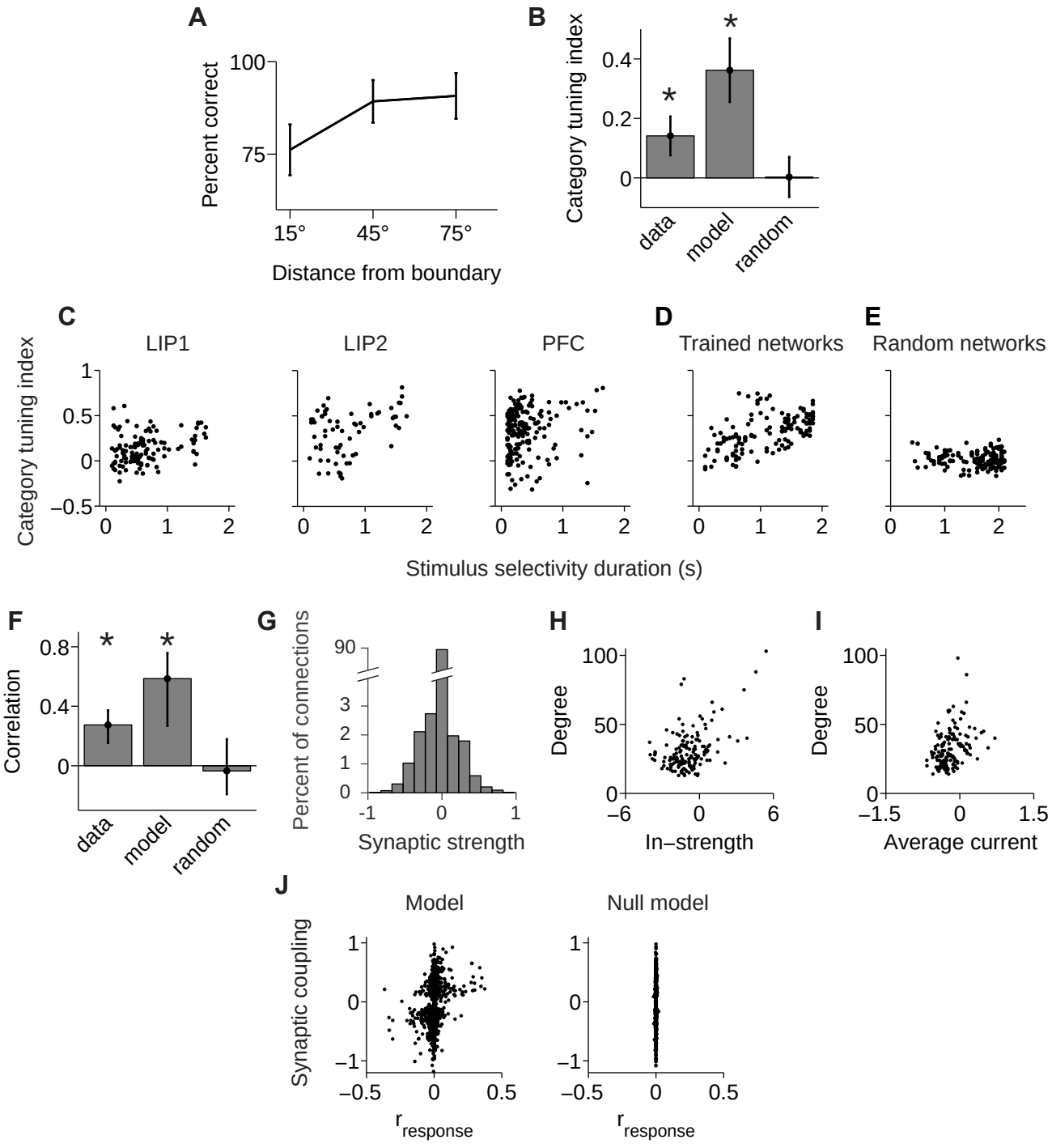

A
Percent correct
100
75
15°
45°
75°
Distance from boundary
B
Category tuning index
0.4
0.2
0
*
*
data
model
random
C
LIP1
LIP2
PFC
Category tuning index
1
0.5
0
−0.5
0
1
2
D
Trained networks
E
Random networks
Stimulus selectivity duration (s)
F
Correlation
0.8
0.4
0
*
*
data
model
random
G
Percent of connections
90
3
2
1
0
-1
0
1
Synaptic strength
H
Degree
100
50
0
−6
0
6
In−strength
I
Degree
100
50
0
−1.5
0
1.5
Average current
J
Model
Null model
Synaptic coupling
1
0
−1
−0.5
0
0.5
r response
r response

Figure S6 Related to Figure 7. Comparison between trained networks and random networks. Randomly connected networks (RCNs) represent an easy way to generate mixed selectivity by virtue of random connections. They can be constructed with sparse connectivity, whereby non-zero weights are randomly drawn from a Guassian distribution of zero mean, and the output weights are trained by a linear classifier to perform DMC tasks. We trained six RCNs until they can perform DMC task a at comparable performance as trained networks and animals. (**A**) Percent correct response is plotted as a function of distance from stimuli to category boundary. Error bars indicate s.d. across six instances of RCNs. (**B**) We measured the average strength of category tuning during the selective time windows using a category tuning index (CTI). Neurons with more positive CTI values have stronger category selectivity regardless of which category they prefer (see Experimental Procedures). Bar graphs show average CTIs during the delay from all three datasets (first bar), ten instances of trained networks (second bar), and six instances of RCNs (third bar). Both data and model demonstrate robust category tuning (stars indicate that average CTI are significantly different from zero for all datasets or networks), while the neural representation of RCNs contains no categorical coding. (**C**) We assessed the stimulus selectivity duration of each neuron – the duration of the longest time window where firing rates are significantly modulated by stimuli (ANOVA, P¡0.005). The correlations between CTI and stimulus selectivity duration are significant in LIP1 (Pearson correlation, $N = 124$, $r = 0.297$, $P < 10^{-3}$), LIP2 ($N = 68$, $r = 0.497$, $P = 10^{-5}$) and PFC ($N = 182$, $r = 0.153$, $P = 0.039$). (**D**) We found strong correlations between CTI and stimulus selectivity durations in the trained model as well ($N = 131$, $r = 0.558$, $P = 10^{-7}$). In both neural data and trained model, it is rare to observe persistent activity without category selectivity. (**E-F**) RCNs do not capture such regularity. Correlations can be zero, positive, or negative in random networks (**E**; $N = 1500$, $r = 0.022$, $P = 0.39$). The average correlation is close to zero (mean $r = -0.035$; **F**). Stars indicate that correlations in all datasets or networks are positive and significant. RCNs do have neurons with persistent stimulus selectivity, but these neurons do not necessarily show categorical coding. (**G-J**) Nonrandom features in structural and functional connectivity in trained networks. (**G**) Synaptic strength of the trained network has unimodal distribution with zero mean (average strength across 10 instances of the model $= -0.008$). (**H**) High-degree neurons tend to have more positive in-strength (Spearman rank correlation, $N = 150$, $\rho = 0.318$, $P < 10^{-4}$). (**I**) High-degree neurons tend to have larger neural activity (Spearman rank correlation, $N = 150$, $\rho = 0.402$, $P < 10^{-7}$). x-axis, average synaptic current ($\bar{x}$) across task conditions and time. (**J**) The relationship between neural response similarity (x-axis) and synaptic coupling (y-axis).

Following practice in graph theory (Bullmore and Sporns, 2009), we compare interneuronal interactions in trained model to null models, generated by randomly re-wiring the connectivity matrix while preserving degree statistics. These null models are generated by function `randmio_dir` from Brain Connectivity Toolbox (Rubinov and Sporns, 2010), which randomly pick pairs of connections and swap them (Maslov and Sneppen, 2002). This method preserves distribution of in-degree, out-degree and out-strength, but not in-strength. Significant deviations from null models reflect the meaningful structure of trained networks beyond degree hierarchy. We found that the responses of trained models have higher variance across stimuli

than its random counterparts. As a result, trained networks demonstrate a larger degree of functional segregation compared to null models throughout DMC epochs. This means some neural pairs are highly similar in their encoding while some pairs are highly distinct, and this is so to a larger extent than would be expected by chance. Average range of $r_{response} = 0.43$ in trained models (left panel) versus 0.04 in null models (right panel). $r_{response}$ is correlated to synaptic coupling (regression, $\beta = 0.015$, $P < 10^{-5}$) and the effect is significantly larger in trained networks compared to random networks (t-test, $N = 100$ shuffles, $P < 10^{-4}$, average $\beta \sim 10^{-3}$). The relationship between two variables during the last 250 ms time window of delay epoch is shown here, but the same conclusion holds for all time windows from the onset of sample to the end of trial.

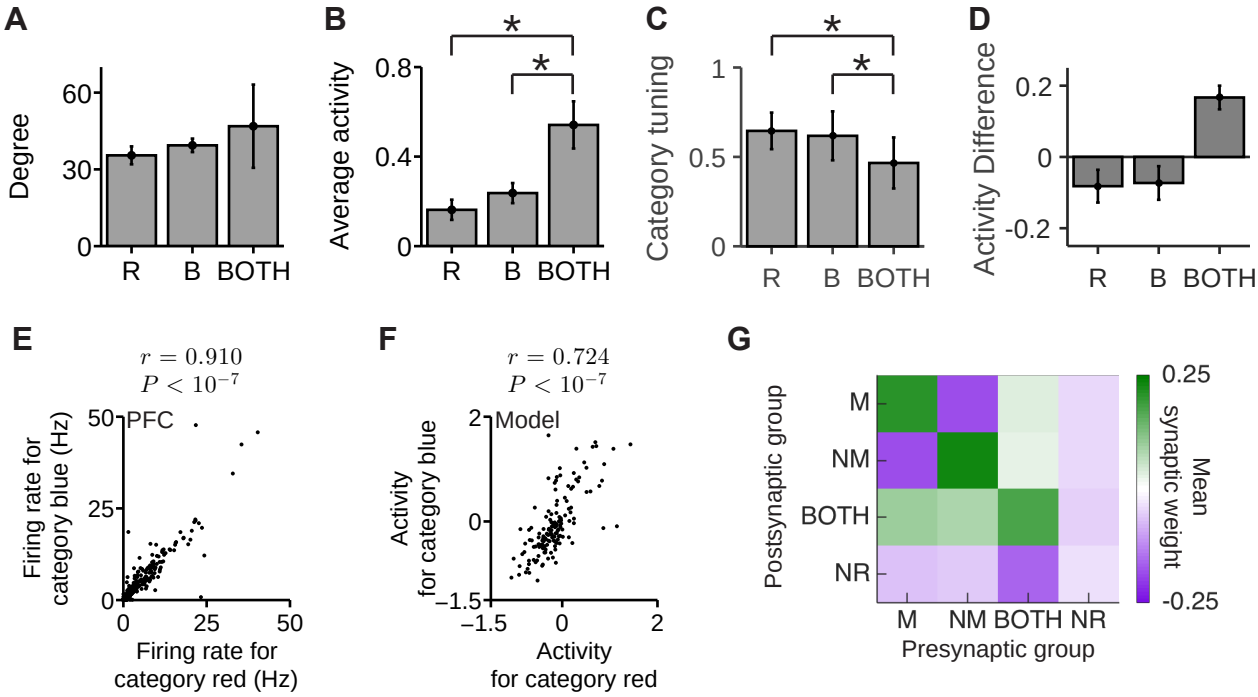

A
Degree
60
30
0
R
B
BOTH
B
Average activity
0.8
0.4
0
*
*
C
Category tuning
1
0.5
0
D
Activity Difference
0.2
0
-0.2
E
$r = 0.910$
$P < 10^{-7}$
PFC
Firing rate for category blue (Hz)
Firing rate for category red (Hz)
50
25
F
$r = 0.724$
$P < 10^{-7}$
Model
Activity for category blue
Activity for category red
2
−1.5
G
Postsynaptic group
M
NM
BOTH
NR
Presynaptic group
Mean synaptic weight
0.25
-0.25

Figure S7 Related to Figure 7. Competition and cooperation between decision-related neuron pools. (**A**-**D**) Average degree, neural activity, category tuning strength and the average changes in the activity through the delay of three neuron groups; activated at the end of delay for red sample stimuli only (denoted R group), for blue sample stimuli only (B group), and for both red and blue (BOTH). Error bars indicate s.d. across 10 instances of trained networks. Neurons in BOTH group have similar degrees (**A**), larger activity (**B**), and weaker category sensitivity (**C**) compared to R and B groups. Note that BOTH neurons have non-zero firing rates to red and blue categories, but the magnitude of firing rates can differ across categories. In (**D**), we plotted the average changes in neural response (responses at the end of the delay subtracted by those at the beginning). The activity of BOTH neurons tends to increase during the delay, while the activity of R and B neurons tends to decrease. To simplify the analysis at (**B**) and (**D**), we simulated the network activity with no noise, responding to stimulus directions at the middle of categories. (**E**) Average firing rates in response to category red (x-axis) and blue (y-axis) in PFC population are highly correlated (Pearson correlation, $N = 380$, $r = 0.910$, $P < 10^{-7}$). (**F**) Average neural responses to category red (x-axis) and blue (y-axis) are highly correlated in trained networks (Pearson correlation, $N = 150$, $r = 0.724$, $P < 10^{-7}$). Activity is defined as synaptic variable, x, but same conclusion holds when the same analysis is performed on rate, r, ($r = 0.679$, $P < 10^{-7}$). In a categorical decision network that relies solely on competition, firing responses to category red and blue stimuli should be anti-correlated. However, we found large positive correlations in both neural data and model, which are mainly driven by BOTH neurons, suggesting cooperative mechanism. (**G**) Average synaptic weights between neuron groups that are active only for match trials (M group), only for nonmatch (NM group), for both match and nonmatch (BOTH group), and not responsive (NR group). Activation is determined by firing rates at the match or non-match stable states (simulated without noise). Same plotting convention as in Figure 7e. M and NM groups exhibit strong within-group excitation and between-group inhibition. BOTH group receives net positive connections from itself as well as from M and NM groups. NR group receives net negative connections from all other groups.

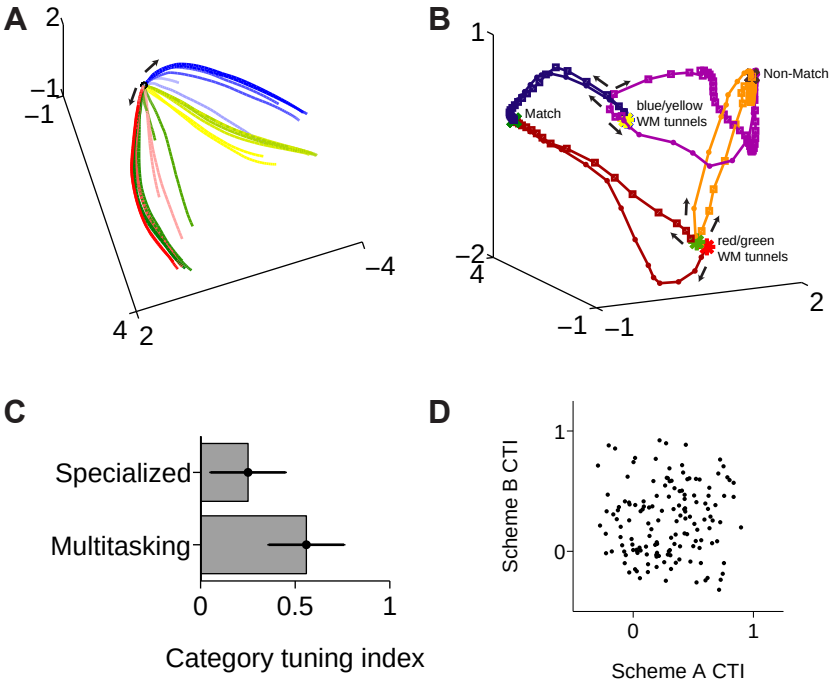

A
2
−1
−1
−4
4
2
B
1
Non-Match
Match
blue/yellow
WM tunnels
red/green
WM tunnels
−2
4
−1
−1
2
C
Specialized
Multitasking
0
0.5
1
Category tuning index
D
Scheme B CTI
1
0
0
1
Scheme A CTI

Figure S8 Related to Figure 8. Neural representations during task switching experiments. (**A**) Neural trajectories during the sample period of a network trained with independent-input flexible categorization paradigm. Same network as plotted in Figure 8c. Stimuli in the category red of scheme A and the category green of scheme B drive neural states to the same direction where the entrance of the shared working memory tunnel is located, and analogously so for blue and yellow categories. Black dots mark the beginning of sample period. (**B**) Neural trajectories during test period for the independent-input paradigm show that the exits of working memory tunnels in Figure 8c are mapped onto appropriate choice-related stable states by test stimulus inputs. Stars indicate the trajectories at the beginning of test period, which are locations of end-of-delay states color coded by categories. The colors of trajectories traces (dark blue, dark red, purple, and orange) follow the convention in standard DMC in Figure 5f-g (dark blue, BB and YY; dark red, RR and GG; purple, BR and YG; orange, RB and GY). Green and brown crosses indicate locations of fixed points associated to match and non-match choice, respectively, which is the final states of the network. Trajectories marked by dots are from trials when category scheme A was active, whereas trajectories marked by squares correspond to scheme B. (**C**) Comparison between category tuning strengths of multitasking (selective to categories in both schemes) and specialized (selective to categories in one scheme) neurons. Individual neurons' category strengths are accumulated from all 5 realizations of networks trained on independent-input protocol (multitasking, $N = 554$; specialized, $N = 78$). Error bars indicate s.d. across all neurons. Multitasking neurons are significantly more category selective than specialized neurons (t-test, $P < 10^{-7}$). (**D**) CTI of category scheme A (x-axis) versus that of scheme B (y-axis) for networks trained with shared-input paradigm. There is no correlation between the two indices ($N = 150$, $r = 0.053$, $P = 0.52$).

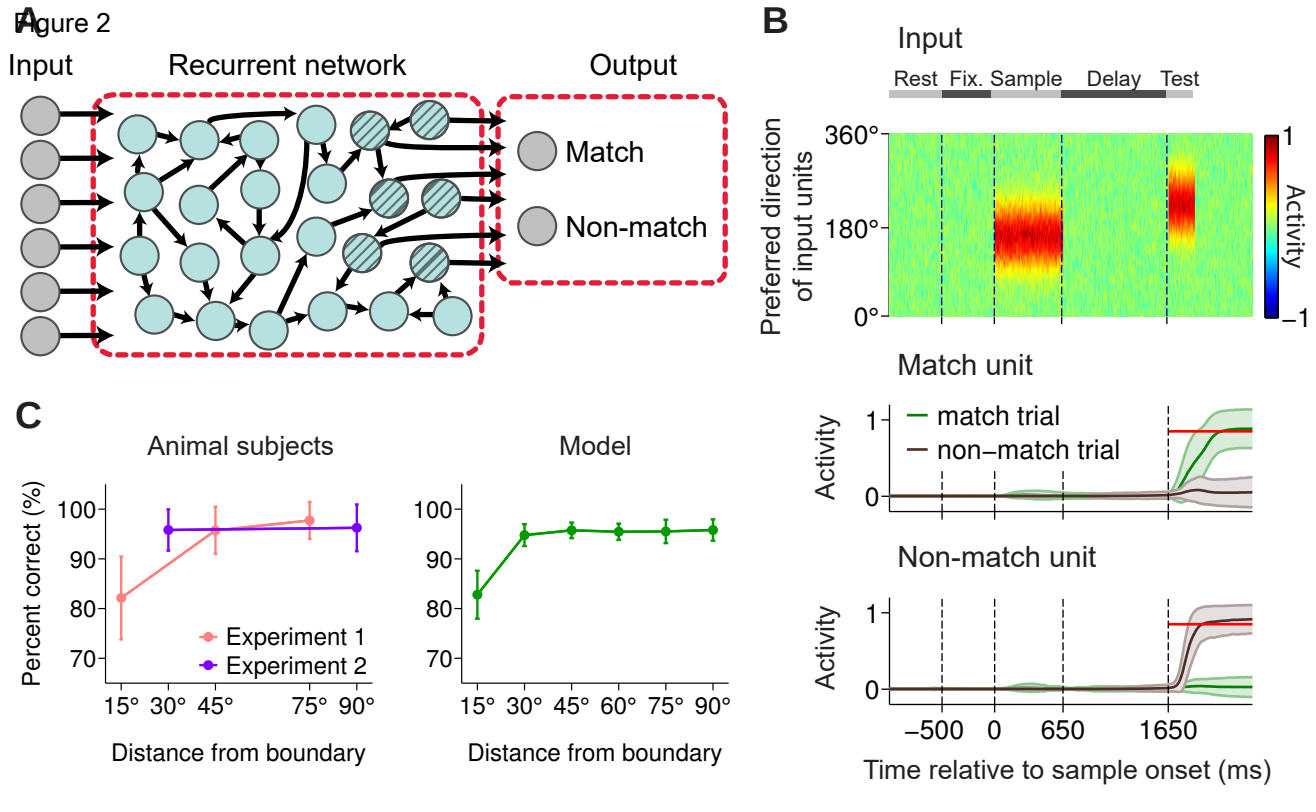


Figure 2
A
Input
Recurrent network
Output
Match
Non-match
B
Input
Rest
Fix.
Sample
Delay
Test
Preferred direction of input units
360°
180°
0°
1
Activity
−1
Match unit
Activity
1
0
match trial
non–match trial
Non-match unit
Activity
1
0
−500
0
650
1650
Time relative to sample onset (ms)
C
Animal subjects
Model
Percent correct (%)
100
90
80
70
Experiment 1
Experiment 2
15°
30°
45°
60°
75°
90°
Distance from boundary
Distance from boundary

Figure 7

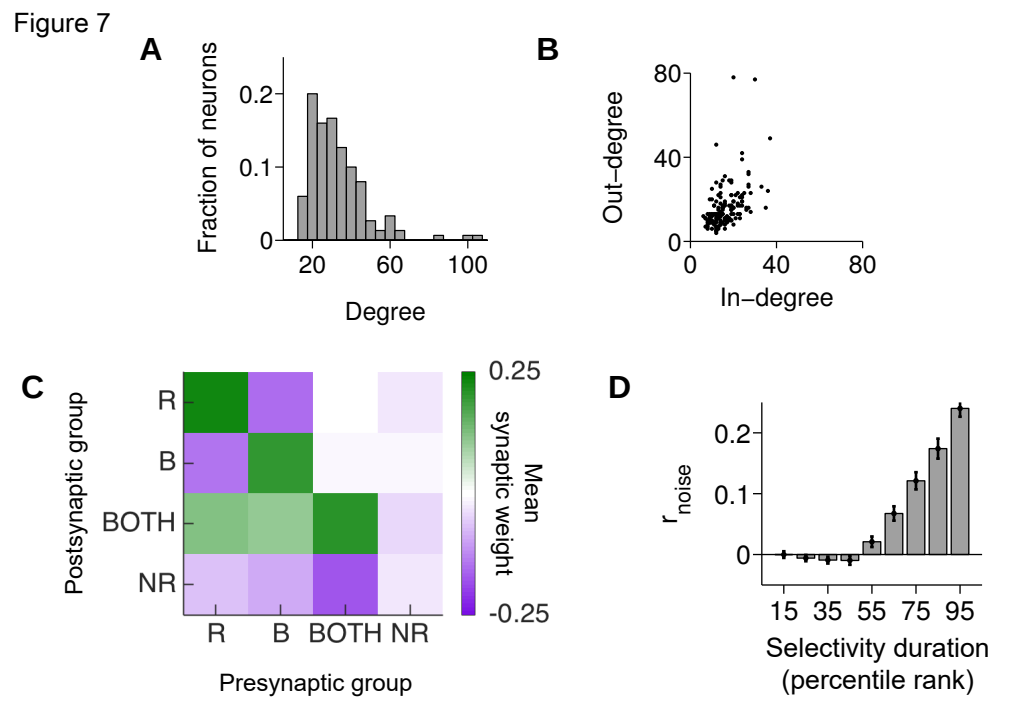
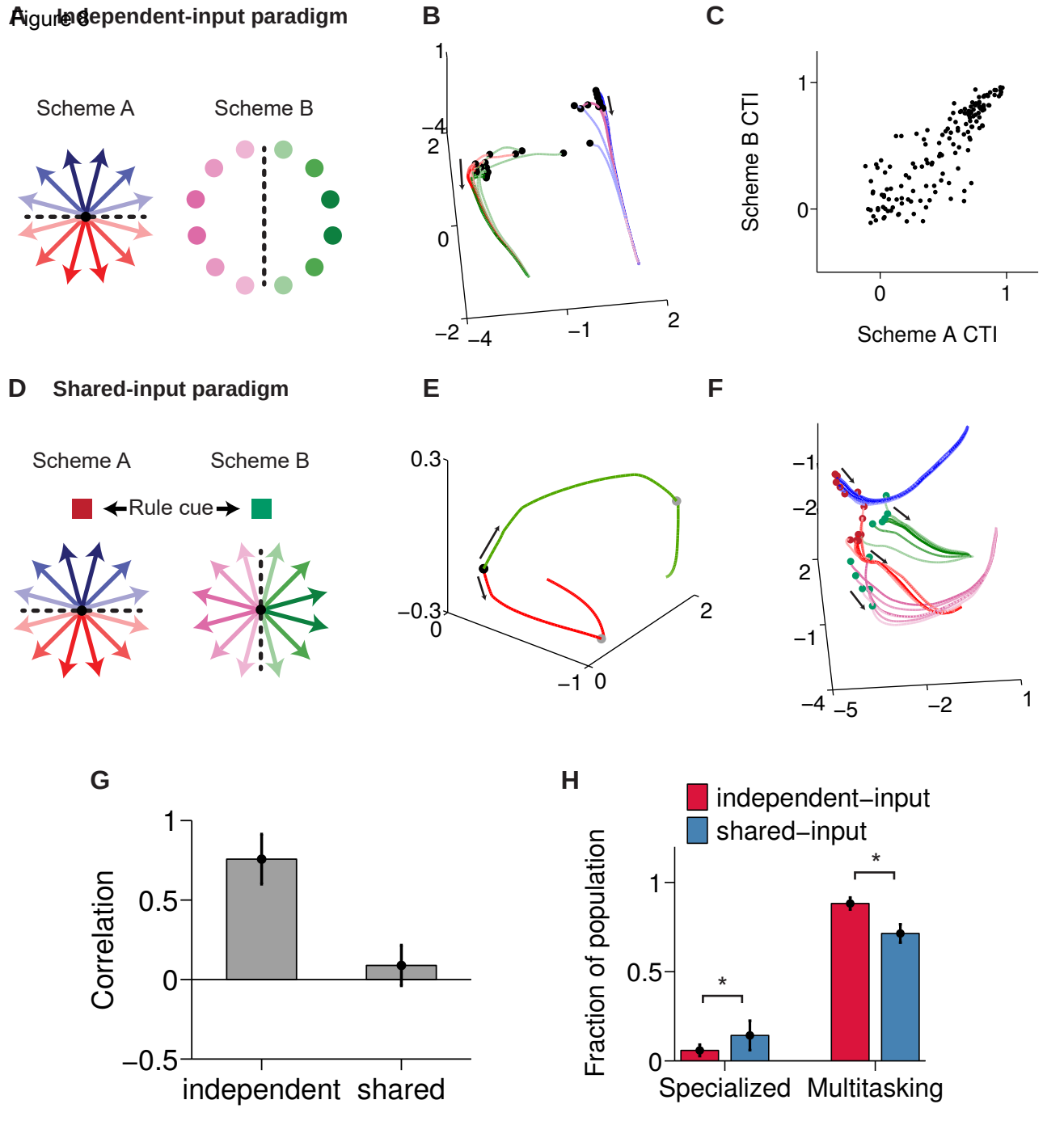
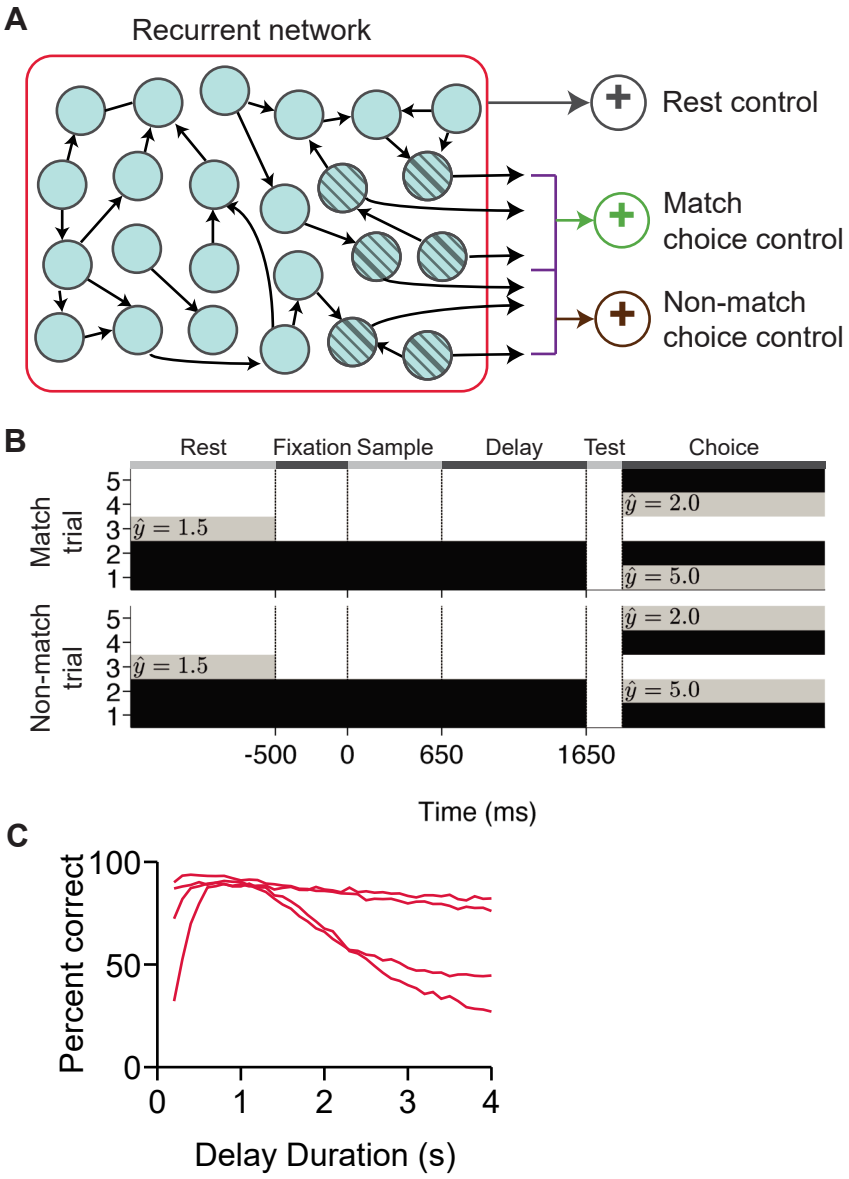
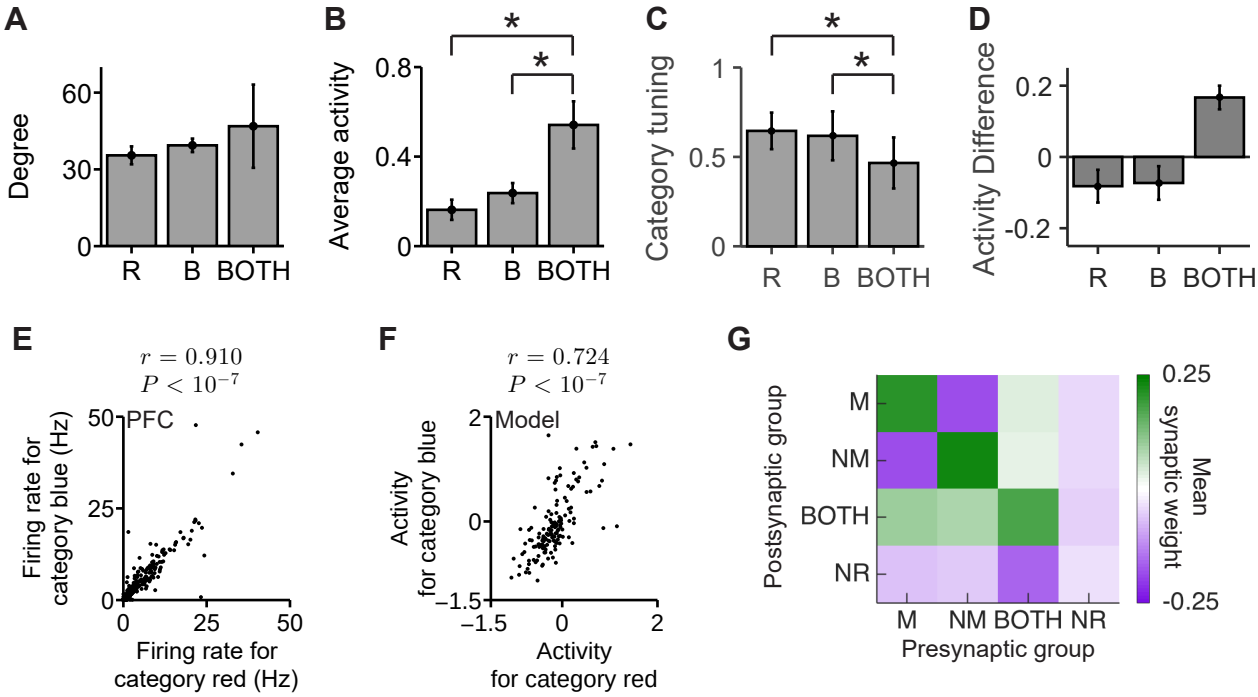

A
Fraction of neurons
0.2
0.1
0
20
60
100
Degree
B
Out–degree
80
40
0
0
40
80
In–degree
C
Postsynaptic group
R
B
BOTH
NR
R
B
BOTH
NR
Presynaptic group
0.25
Mean synaptic weight
-0.25
D
r noise
0.2
0.1
0
15
35
55
75
95
Selectivity duration (percentile rank)

Figure 3


A Independent-input paradigm
Scheme A
Scheme B
B
C
Scheme B CTI
Scheme A CTI
D Shared-input paradigm
Scheme A
Scheme B
Rule cue
E
F
G
Correlation
independent
shared
H
independent–input
shared–input
Fraction of population
Specialized
Multitasking

Supplemental Figures




A
Degree
60
30
0
R
B
BOTH
B
Average activity
0.8
0.4
0
*
*
C
Category tuning
1
0.5
0
D
Activity Difference
0.2
0
-0.2
E
$r = 0.910$
$P < 10^{-7}$
PFC
Firing rate for category blue (Hz)
Firing rate for category red (Hz)
50
25
F
$r = 0.724$
$P < 10^{-7}$
Model
Activity for category blue
Activity for category red
2
−1.5
G
Postsynaptic group
M
NM
BOTH
NR
Presynaptic group
Mean synaptic weight
0.25
-0.25