Published in final edited form as:

Nat Neurosci. 2021 May ; 24(5): 715–726. doi:10.1038/s41593-021-00821-9.

Rotational Dynamics Reduce Interference Between Sensory and Memory Representations

Alexandra Libby[1], **Timothy J. Buschman**[1,2,*]

[1]Princeton Neuroscience Institute, Princeton University

[2]Department of Psychology, Princeton University

Abstract

Cognition depends on integrating sensory percepts with the memory of recent stimuli. However, the distributed nature of neural coding can lead to interference between sensory and memory representations. Here, we show the brain mitigates such interference by rotating sensory representations into orthogonal memory representations over time. To study how sensory inputs and memories are represented, we recorded from auditory cortex of mice as they implicitly learned sequences of sounds. We found the neural population represented sensory inputs and the memory of recent stimuli in two orthogonal dimensions. The transformation of sensory information into a memory was facilitated by a combination of 'stable' neurons, that maintained their selectivity over time, and 'switching' neurons, that inverted their selectivity over time. Together, these neural responses rotated the population representation, transforming sensory into memory. Theoretical modeling showed this rotational dynamic was an efficient mechanism for generating orthogonal representations, thereby protecting memories from sensory interference.

One Sentence Summary:

Sensory representations dynamically rotate into an orthogonal memory representation, reducing interference with new sensory inputs.

Introduction

Maintaining the short-term memory of recent stimuli is critical to cognition. Memories provide the context needed for perception and decision-making[1,2] and are particularly important for learning to predict the future. Predictions are based on expectations; given the current context, one can predict which stimulus is likely to occur next. Expectations reflect previously learned statistical regularities between stimuli (i.e., 'A, B' is usually followed by 'C'). These are learned by forming associations through time, between the memory of recent

*Corresponding author: tbuschma@princeton.edu.

Author Contributions

TJB and AL conceived of the project and designed the experiment. AL did surgery on the animals, collected the data, constructed computational models, and analyzed the data, with supervision from TJB. AL and TJB wrote the paper.

Competing Interests

The authors declare no competing interests.

stimuli (‘A, B’) and the representation of current sensory inputs (‘C’). Once learned, these associations facilitate predictions (i.e., expecting C when you see A-B), which improves sensory processing[3,4] and facilitates decisions by allowing them to be made earlier[5–7].

While maintaining both sensory and memory information can facilitate cognition, it is unknown how the brain maintains both representations without interference. Previous work has shown neural networks have a limited capacity, unable to accurately encode multiple stimuli[8] or maintain multiple short-term memories[9,10]. The limited capacity of neural networks is thought to be due to interference that arises when simultaneously encoding sensory and memory representations in the same population of neurons. Theoretical work has shown interference can be reduced by orthogonalizing representations[11–13], possibly by having neurons with non-linear or random selectivity[14]. However, it remains unclear if, or how, such orthogonalization occurs in the brain. To address this, we investigated the mechanisms used by the brain to avoid interference between sensory and memory representations.

To study interference between sensory and memory representations, we used an implicit sequence learning paradigm to build associations between sensory stimuli[15,16]. As we detail below, this facilitated predictions of upcoming stimuli but also created interference between the short-term memory of recent stimuli and the sensory representation of new stimuli. We show the brain mitigates this interference by dynamically rotating sensory representations into an orthogonal memory representation. This rotation was supported by dynamics in the selectivity of individual neurons; we found populations of ‘stable’ neurons, which maintained their selectivity over time, and ‘switching’ neurons, which inverted their selectivity from the sensory to memory time periods. The combination of these dynamics facilitated the rotation of the population representation, allowing the same network of neurons to efficiently represent both sensory information and short-term memories.

Results

To study how sensory and short-term memory representations interact in sensory cortex, we exposed mice to sequences of four auditory chords (see methods). Statistical regularities in the transitions between chords created learnable predictions within the sequences (Fig. 1a). Sequences began with a pair of contextual chords: either an ‘A’ and ‘B’ stimulus pair (the AB context) or an ‘X’ and ‘Y’ stimulus pair (the XY context). These contexts predicted what chord would follow: on 68% of trials, the AB context was followed by a C chord and the XY context was followed by a C* chord. However, on a subset of trials (20%), the context was unexpectedly followed by the other C/C* stimulus (i.e., ABC* and XYC; the remaining 12% of trials were ambiguous stimuli, see methods). All sequences ended with a D chord. Importantly, the sequences were designed to balance the overall likelihood of each sensory stimulus across conditions. Therefore, before the start of each sequence, the animal had no expectation as to what chords it would experience. Only after the presentation of the contextual stimulus (A/X) could the animal predict the upcoming C/C* stimulus. Beginning naïve, the animals experienced 1500 sequences per day for 4 consecutive days (Fig. 1a, see methods). No behavioral task was required by the animal, allowing us to study how unsupervised learning impacts sensory processing and short-term memory[17–20].

To measure sensory and short-term memory representations in the brain, we recorded 522 neurons from the auditory cortex of 7 mice with an average of 130 neurons per day (across mice; see methods). Although electrodes were chronically implanted, individual neurons were not tracked across recording sessions. Neurons responded selectively to the presentation of the context chords (A vs. X and B vs. Y) and the predicted chord (C vs. C*; see Fig. 1b for example neurons). To capture how the population of recorded neurons represented each stimulus in the sequence, we trained linear SVM classifiers to discriminate the population firing rate responses to each pair of stimuli (A/X, B/Y, C/C*; responses were averaged over 10–110 ms after stimulus onset, see methods and Supplementary Fig. 1 for classifier performance). To ensure the classifier was unbiased, all trial types were balanced during training (i.e., there was an equal number of ABCD, ABC*D, XYCD, and XYC*D trials). Separate classifiers were trained for each recording session, using simultaneously recorded neurons. All analyses were performed on withheld data.

Each classifier defined an “encoding axis” for a pair of stimuli (the axis is the vector normal to the classifier’s hyperplane). By projecting the firing rate of the neural population onto the encoding axis, we could estimate stimulus information in the population at each moment in time (see Fig. 1c for schematic and methods for details). As expected, the population encoded each sensory stimulus when it was presented in the sequence on all four days (A/X: Fig 1d–f, B/Y: Extended Data Fig. 1, C/C*: Fig 1g–h; accuracy of decoding is shown in Extended Data Fig. 2).

Alignment of Encoding Axes Facilitates Prediction and Postdiction

Experience over days led to associative learning between stimuli in the sequence. These associations facilitated predictions in auditory cortex: on day 4, during the presentation of A/X, there was predictive encoding of the expected C/C* stimulus. This can be seen as the neural population representing C or C* when A or X was presented, respectively (Fig. 1g–h, black box; see also Extended Data Fig. 3). This predictive effect was relatively weak on days 1 through 3 before increasing on day 4, suggesting experience strengthened the prediction (Fig. 1i; Day 1 = 0.13±0.022, $p<1/5000$, Day 2 = 0.07±0.023, $p=0.0036$, Day 3 = 0.076±0.023, $p<1/5000$, Day 4 = 0.19±0.022, $p<1/5000$, bootstrap tests. Slope mean±SEM over days = 0.02±0.01, $p = 0.022$, two-sided bootstrap tests; Day 4 – Day 1 = 0.07, $p = 0.015$, one-sided permutation test; see Supplementary Fig. 2b for similar trends within a day).

The relationship between the associated A/X and C/C* stimuli can be highlighted by projecting the activity of the neural population into a 2D state space defined by the A/X sensory and C/C* sensory encoding axes (Fig. 2a). This state space tracks the co-evolution of information along both sensory axes during the sequence. On day 4, the A/X stimulus evoked an oblique response in this state space, indicating the A/X stimulus induced both its own representation and that of the predicted stimulus (A-C and X-C*, respectively). These predictions increased with experience, reflected by an increase in the angle of the neural trajectories in this 2D space over days (Extended Data Fig. 4a–b).

The oblique response in the sensory encoding state space suggests experience caused the A/X and C/C* representations to align. If true, then the A/X and C/C* encoding axes should

become more similar over time. To test this, we measured the angle between the A/X and C/C* axes (see methods). On day 1, the average angle across blocks was 84±7.4 degrees. This near orthogonality is consistent with sensory cortex independently representing different, unassociated, stimuli before learning. With experience, the angle significantly decreased (Fig. 2b; slope mean±SEM = −0.89±0.18 degrees/block of trials, p<1/5000) such that, by day 4, the angle was significantly less than orthogonal (67±8.2 degrees, p=0.0057, both one-sided bootstrap tests). This trend started within day 1 (Fig. 2b, p=.12, one-sided bootstrap test), suggesting alignment starts immediately with experience. Note, none of these results depended on the classifier type or its hyperparameters (Supplementary Fig. 3, see methods).

Our results are consistent with previous experimental and modeling work, which show single neurons respond similarly to associated stimuli[15,16,21–23]. Similarly, we found the response of single neurons to A/X and C/C* became more correlated with experience (Fig. 2c and d, Day 1 slope between A/X and C/C* selectivity = −0.09±0.1, p=0.20, n=121; Day 4 slope = 0.25±0.087, p=0.0022, n=143; change of slope over days = 0.08±0.04, p=0.028, one-sided bootstrap tests). Together, these results suggest implicit associative learning increased the number of neurons with joint selectivity to A and C (or X and C*), which aligned the encoding axes and facilitated the neural prediction of future stimuli across days (Fig. 1g–i).

We also examined the relationship between the representation of B/Y and the representations of A/X and C/C* (Extended Data Fig. 1f–g). Unlike A/X, the B/Y and C/C* sensory axes did not align, possibly because B/Y did not add predictive value about which C/C* stimulus would occur (as it was already fully predicted by the A/X stimulus)[24].

It is important to note that the alignment of the A/X and C/C* axes does not define a directional relationship between the stimuli. So, similar to the A/X stimulus inducing a predictive response along the C/C* sensory axis, the presentation of the C/C* stimulus should also evoke a response along the A/X sensory axis. This can be seen in the A/X-C/C* state space, where the presentation of the C/C* stimulus drove neural activity along an angle, encoding the A-C/X-C* association (Fig. 3a; this effect increased with experience, Extended Data Fig. 4c–d). This is a postdiction: new sensory inputs inform the representation of past events. Postdiction is a common psychological phenomenon, that improves perception (and can cause illusions)[25–28]. Our results suggest postdiction arises from the same neural mechanism as prediction – the alignment of population representations of associated stimuli. Consistent with this, we found the angles between A/X and C/C* sensory axes, calculated per mouse, were correlated with the strength of the animals' pre/postdiction (Supplementary Fig. 4a–b).

Alignment of Encoding Axes Leads to Interference

The alignment of A/X and C/C* sensory representations facilitated pre/postdiction, but also led to interference between the current sensory inputs and the representation of the past. This interference occurred on unexpected trials, when the initial sensory representation of A/X was overwritten by an unexpected C*/C. Before the C/C* stimulus, the representation of the A/X stimulus was correct (Fig. 3a). However, the onset of the unexpected C*/C stimulus

caused the population encoding of A/X to reverse and cross the hyperplane to encode the incorrect context: ABC*D trials were encoded as X-C*, and XYCD encoded as A-C (Fig. 3b; Day 4, A/X encoding on unexpected trials: −0.16±0.033, p<1/5000, two-sided bootstrap test). This effect grew with experience; unexpected C*/C sounds led to a stronger reversal of the A/X representation over days (slope = −0.039±0.015, p=0.0068, one-sided bootstrap test).

An Orthogonal Memory Representation Avoids Interference

Maintaining an accurate account of stimulus history is critical for making decisions and learning associations across time[1,2]. Therefore, we were interested if auditory cortex maintained a memory representation of the A/X context that was resilient to interference from associative learning. To this end, we trained an 'A/X memory' classifier to discriminate the AB/XY context using the activity of neurons during the presentation of the C/C* stimulus (Fig. 3c; Extended Data Fig. 5). This A/X memory axis encoded the memory of A/X during the C/C* stimulus, but not during the A/X stimulus (Fig. 3d; Extended Data Fig. 6). This complemented the A/X sensory axis, which accurately encoded A/X during its presentation, but failed during the memory period (Fig. 3d). In this way, the transition between A/X sensory to A/X memory encoding reflects a change in the representation of A/X context during the sequence. This change occurred during the B/Y presentation and progressed earlier in the sequence with experience (Extended Data Fig. 6c).

Unlike the A/X sensory representation, the A/X memory representation of the A/X chord was not overwritten by the C/C* stimulus (Fig. 3e). Interference was avoided because the A/X memory axis was orthogonal to the C/C* sensory axis. By day 4, the angle between A/X memory and C/C* sensory was 90±9.8 degrees (p=0.49, difference from 90 degrees, one-sided bootstrap test). Again, this changed with experience; the angle between A/X memory and C/C* sensory began slightly obtuse on day 1 and decreased to become orthogonal over days (Fig. 3f). Figure 3g summarizes the angular relationships between all three axes on day 4, showing both the predictive alignment of the A/X sensory and C/C* sensory axes and the orthogonality between the A/X memory and C/C* sensory axes.

The reduced angle between the A/X sensory and C/C* sensory representations suggest they reflect a single latent variable (the AC/XC* association). If true, then neural activity should follow a low dimensional trajectory within the A/X-C/C* sensory state space (see methods). Indeed, during the presentation of the C/C* stimulus, the dimensionality of the response within the A/X-C/C* sensory state space was significantly lower than expected by chance and was lower on day 4 compared to day 1 (Fig. 3h, see methods). Consistent with this, the dimensionality of the full neural space trended towards decreasing over days (Extended Data Fig. 5e). In contrast, the dimensionality of the A/X memory – C/C* sensory state space increased from day 1 to day 4 (Fig. 3h). These results suggest A/X sensory and C/C* encoding are captured by a single latent variable, while A/X memory is orthogonal to this sensory representation.

Finally, we tested how A/X sensory and A/X memory representations influenced sensory processing of the C/C* stimulus (Extended Data Fig. 7). There was a significant trial-by-trial correlation between the strength of A/X encoding, measured 50 ms prior to the C/C*

stimulus, and the strength of the C/C* response. Yet, the relationship was dissociated between the two A/X encoding axes. On unexpected trials, the C/C* representation was positively correlated with the A/X memory representation, but negatively correlated with the A/X sensory representation (Extended Data Fig. 7d, f). The reverse trend was seen on expected trials (Extended Data Fig. 7c). Together, these results suggest A/X sensory and memory representations have different roles in prediction: the sensory representation facilitated responses to expected stimuli, while the memory representation magnified unexpected stimuli or 'prediction errors'.

Rotational Dynamics Transform Sensory Representations into Orthogonal Short-term Memory Representations

Together, our results show the memory representation of the A/X stimulus was orthogonal to sensory inputs. By becoming orthogonal, the memory representation avoids interference by 'getting out of the way' of subsequent inputs (i.e., C/C*). Orthogonal representations have significant computational advantages[29,30]. Theoretical work has found orthogonalization minimizes interference[13], increases the memory capacity of neural networks[12], and maximizes separability of representations (improving decoding)[14]. Next, we were interested in understanding how the activity of individual neurons allowed the population representation of the A/X stimulus to transform from the sensory axis to an orthogonal memory axis.

Two general mechanisms could lead to orthogonal sensory and memory representations. First, sensory and memory could be represented by independent populations of neurons (Fig. 4a). Second, sensory and memory could be represented in orthogonal dimensions within the same population of neurons (Fig. 4b–c). To distinguish between these hypotheses, we tested whether neurons carried information about the A/X stimulus during both the sensory and memory time periods. Figure 4d shows the distribution of A/X selectivity during both the sensory and memory time periods across all A/X selective neurons. This distribution contains two types of neurons: 'single' neurons, which are selective during only one time period (sensory or memory), and 'conjunctive' neurons, which are selective during both time periods. Under the independent mechanism, there should be more single neurons than expected by chance (Fig. 4a, insets). To test this, we generated a null distribution by permuting A/X selectivity in each time period across neurons, breaking any association of A/X selectivity between the sensory and memory time periods (see methods). Contrary to the prediction from the independent mechanism, we found fewer single neurons than expected by chance (Fig. 4e; single/n = 0.33, n = 522, $p<1/1000$, one-sided permutation test). This suggests both sensory and memory representations exist within the same population of neurons, but along orthogonal dimensions (Fig. 4b–c).

There is a spectrum of mechanisms by which sensory representations could be transformed into memory representations within the same population. These mechanisms range from relying on neurons with random selectivity (Fig. 4c, left) to relying on neurons with structured changes in their selectivity (Fig. 4c, right). Previous work has argued random selectivity can generate orthogonal representations[29,30]. However, in our dataset, a chi-squared test found the observed counts of conjunctive and single neurons was significantly

different from what would be expected by a random mechanism (chi-squared statistic = 120.48, p=2.6e-22, df = 8, see methods). Likewise, across all four days, we found more conjunctive neurons and a greater ratio of conjunctive/single neurons than expected by the random mechanism (Fig. 4f–g; conjunctive/n = 0.18, n = 522, p<1/1000; conjunctive/single ratio = 0.54, p<1/1000, one-sided permutation tests). These effects increased with experience: the proportion of conjunctive neurons and ratio of conjunctive/single neurons were not initially significant on day 1, but increased to reach significance by the end of day 1 and continued to increase across days (Fig. 4e–g). Together, these results suggest orthogonalization in the population was not just due to random changes in selectivity, but was facilitated by a structured mechanism that increased the proportion of conjunctively selective neurons.

To understand the nature of the structured mechanism, we examined how the selectivity of individual neurons changed from the sensory to memory time periods. To this end, we used an unsupervised clustering algorithm[31] to group neurons by their timecourse of A/X selectivity (see methods). Clustering revealed two functional clusters of conjunctive neurons: ‘stable’ and ‘switching’ neurons. Stable neurons maintained their contextual preference across the sequence; switching neurons switched their A/X contextual preference during the sequence (Fig. 5a; see Fig. 5b for full population across all 4 days).

Further analyses confirmed stable and switching dynamics captured the timecourses of selectivity in auditory cortex. First, we measured the pairwise similarity of the timecourses of A/X selectivity for each pair of neurons. As seen in Figure 5c, the four clusters of stable and switching neurons have high similarity within their cluster and low similarity between clusters. Second, these clusters were consistent within subsets of trials (C, C*, and ambiguous stimuli; see methods and Extended Data Fig. 8). This reflects the reliability of clustering and suggests the observed dynamics were not due to non-linear mixing with the stimulus presented during the memory period. Finally, the clusters were non-overlapping when projected into low-dimensional spaces and similar clusters were seen with other clustering approaches (see Extended Data Fig. 9).

To confirm the significance of these neuron groups outside of clustering, we used a binomial test to show the observed counts of stable and switching neurons were greater than expected by chance (see methods; stable proportion = 0.12, p=5.96e-15; switching proportion = 0.06, p=0.019; conjunctive proportion = 0.18, p=6.1e-14; all greater than chance; single proportion = 0.33, p=0.00019, less than chance, n=522, all binomial tests). Several lines of evidence suggested the increase in stable/switching neurons was not due to smoothing: the sensory and memory time periods are well separated within the sequence, our smoothing kernel was less than the space between the time periods, and smoothing random selectivity did not produce the same level of structure as our neural recordings (Supplementary Fig. 5–6; see methods).

Next, we investigated how stable and switching neurons supported the rotation of the sensory encoding axis to the memory encoding axis by examining the classifier weights of each neuron group. As expected by their conjunctive selectivity, stable neurons contributed significant weights to both the sensory and memory encoding axes (Fig. 6a). Similarly,

switching neurons significantly contributed to both axes, but with inverted contributions to the A/X sensory axis and A/X memory axis (Fig. 6a). This reflects how switching neurons reverse their preference over time. A similar pattern was seen in the classifier weights of individual neurons: sensory and memory classifier weights were positively correlated in stable neurons and negatively correlated in switching neurons (Fig. 6b, stable neurons slope = 0.51±0.069, $p<1/5000$; switching neurons slope = −0.38±0.089, p=0.0004, one-sided bootstrap test). Over days of experience, the correlation between the A/X sensory and A/X memory weights of stable neurons increased, consistent with learning playing a role in developing structure in the rotation (Fig. 6c).

To visualize the rotational dynamics in the population, we plotted the timecourse of A/X selectivity for stable and switching neurons (Fig. 6d and 6e for day 1 and 4, respectively). During the A/X period, both stable and switching neurons increased their selectivity to their (initially) preferred chord, thereby creating the A/X sensory axis. Then, over the sequence, switching neurons inverted their selectivity, rotating the sensory axis to the A/X memory axis. Thus, both stable and switching neurons work together to facilitate a structured rotation; either alone is insufficient to create a memory axis that is orthogonal to the C/C* axis and able to avoid interference. To directly show how the A/X rotational dynamics avoided interference with C/C*, we plotted the response of stable and switching neurons to the four conditions (Fig. 6f, limited to neurons with significant C/C* selectivity, see methods). Consistent with the alignment of sensory representations, responses to C and C* followed the initial sensory responses to A and X, respectively (as in Fig. 2b–d). Yet, because of the rotational dynamics in the population, the A/X memory axis was orthogonal to the C/C* sensory response, and thus avoided interference (as in Fig. 3f–g).

Structure in Rotation Increases Efficiency of Orthogonal Representations

Our results suggest sensory and memory representations are represented along orthogonal dimensions in the same population of neurons. Furthermore, we found evidence for structure in the rotational dynamics of these representations (Fig. 4). Previous work has highlighted the advantages of random projections[14,32], but the relative advantages of a more structured rotation have not been quantified. Therefore, to contrast random and structured representations, we developed analytical and computational models of rotational dynamics (see methods). The computational model consisted of an input layer connected to a recurrent network of neurons (Fig. 7a). The input layer signaled the A/X and C/C* stimulus during each sensory period. Inputs fed into a recurrent ‘representational’ layer, which acted as a read-out of sensory/memory information and was intended to capture the neural activity observed in auditory cortex. Because the model’s structure mirrors our neural recordings, we could perform the same analyses on both datasets: calculating selectivity and training classifiers to estimate the encoding axes during each time period and examining the dynamics of individual neurons. As with our neural recordings, we found aligning the A/X and C/C* representations in the model increased both prediction and postdiction (Supplementary Fig. 4a–b).

Using this neural network model, we parametrically varied the degree of structure in the rotation by adjusting the recurrent weights in the representational layer. This allowed us to

control the relationship between A/X sensory and A/X memory selectivity in single neurons, while creating network models with rotations ranging from a random rotation, which relied on random patterns of selectivity in individual neurons, to a structured rotation, which relied on stable and switching neurons exclusively (Fig. 7a, lower panel shows the spectrum of A/X selectivity produced by random to structured rotations; see also Supplementary Fig. 4c). Regardless of the level of structure or randomness, all neural network models generated orthogonal representations between A/X memory and C/C* sensory (Supplementary Fig. 4e). Note that some form of rotation was required to preserve A/X memory accuracy; models without rotational dynamics showed interference between C/C* sensory and A/X memory (Supplemental Fig. 7).

As noted above, our experiments suggest there is significantly more structure in the rotational dynamics than expected by a random mechanism (Fig. 4). Consistent with our experimental observations, increasing the degree of structure in the rotation in the neural network model generated more conjunctive neurons, which were selective to both A/X sensory and A/X memory (Fig. 7b; a similar prediction was made by the analytical model). The computational and analytical models suggest the increased structure may have several computational benefits.

First, a structured rotation requires fewer selective neurons than a random transformation, creating a more compact representation. Figure 7c shows adding structure to the model's rotation decreased the proportion of neurons selective to A/X during the sensory and memory periods, while maintaining memory accuracy. This is because, in a structured network, there are more conjunctive neurons, which carry twice the information of neurons selective during a single time period. To test this hypothesis in our recorded neural data, we randomly permuted sensory and memory selectivity across neurons in order to estimate the distribution of selection expected by a random mechanism. As predicted, we found the percentage of selective neurons in our neural recordings was less than expected in a random mechanism (0.5, $p<1/1000$, n=522, one-sided permutation test). Furthermore, we found the percentage of selective neurons decreased over days (Fig. 7d; slope = -0.12 ± 0.03, $p<1/1000$, one-sided bootstrap test), suggesting the structured rotation is learned. The advantage of a compact representation is it is more resistant to interference and is robust to changes in the population (e.g., due to learning changing the selectivity of neurons)[33,34].

Second, our model showed increasing structure led to a more efficient transformation from sensory to memory. Transitioning between states requires energy[35], and so minimizing the magnitude of state change allows for a more efficient transformation. To measure the efficiency of the transformation from sensory to memory, we calculated the cityblock distance between the sensory and memory classifier weights (see methods). A smaller cityblock distance indicates fewer changes in neural activity are needed to transform sensory representations to memories. Increasing the structure in the model's rotational dynamics reduced the cityblock distance, reflecting a more efficient transformation (while maintaining A/X memory accuracy, Fig. 7e). Similarly, the transformation in our neural recordings was more efficient than chance (cityblock distance = 0.21, $p<1/1000$, n=522, one-sided permutation test, see methods) and the efficiency increased across days (Fig. 7f, slope = -0.062 ± 0.035, p=0.043, one-sided bootstrap test).

Altogether, our results show structured rotation is a more compact and efficient mechanism for generating orthogonality compared to randomization. It is more compact, because it requires fewer neurons to represent both sensory and memory information. It is more efficient, because it requires fewer changes in the neural response to move from the sensory to the memory representation. In other words, less energy (e.g., from a control input or making a physical connection) is needed to switch from a sensory to a memory representation.

Discussion

Our study found the brain avoids interference between sensory and memory representations, by rotating the memory representation to become orthogonal to incoming sensory inputs. To study the interference between representations, we used an implicit learning paradigm, in which mice were repeatedly exposed to sequences of sounds. Experience with the sequences of sounds aligned the neural representations of associated stimuli in mouse auditory cortex. This is consistent with previous work in the temporal lobe of monkeys, where single neurons learned to respond to pairs of temporally associated stimuli[15,21–23]. Our results extend these findings by showing associative learning leads to the alignment of population representations. This alignment facilitated predictions of upcoming stimuli: when the contextual stimulus (A/X) was presented, the neural population encoded the predicted stimulus (C/C*; Fig. 1i).

The sensory alignment can also explain postdiction. An important cognitive phenomenon, postdiction allows new information to update the perception of previous events. This is particularly useful for stabilizing perception under noisy conditions[28], because an ambiguous past percept can be updated to match the most probable scenario given the present stimulus[2]. However, we found postdiction can also ‘overwrite’ history when the animal encounters an unexpected stimulus. In this way, associative learning can lead to interference between sensory inputs, thereby reducing a sensory classifier’s ability to accurately represent the history of recent stimuli (Fig. 3b).

We found the brain avoids such interference by rotating sensory information into a memory subspace (Fig. 3d). In our experiments, the A/X memory encoding existed on day 1, but became orthogonal to the C/C* sensory axis with experience. Thus, despite the associative learning between A/X and C/C* sensory inputs, new stimulus inputs did not interfere with the memory of the context (Fig. 3e and g). These population dynamics, which we observed in the auditory cortex of mice performing an unsupervised learning paradigm, are surprisingly similar to those found in prefrontal cortex of primates performing working memory tasks (e.g., Extended Data Fig. 6a shows cross-temporal correlation similar to previous work[36,37]). Similar dynamics have also been found in recurrent neural networks trained on serial order recall[13]. Our results show explicit training on a working memory task is not necessary to generate rotational dynamics. Instead, they may be a property of how the brain processes and maintains sensory inputs. While short-term memory representations have been studied in the context of reward-driven behavior, the majority of learning in an animal’s lifespan is unsupervised and so these dynamics may exist to avoid interference in those situations.

We examined several different mechanisms that could explain the observed rotational dynamics, finding evidence that rotations were structured. Previous work proposed orthogonal representations could emerge from neurons with random selectivity[12,29]. Our results build on this hypothesis, suggesting the individual neuron dynamics are not purely random, but enriched with two functional neuron types: a 'stable' group that maintained its stimulus selectivity and a 'switching' group that switched its selectivity over time (Fig. 5). Previous work in monkeys has found a similar dichotomy in working memory; some neurons stably represent the contents of working memory[38,39], while others dynamically change their representation[40–42]. The relative contribution of stable and dynamic representations to working memory has been debated[43–46]. Our results argue both response types are important – it is the combination of sustained and dynamic responses that facilitates the transformation of sensory representations into orthogonal memory representations, thereby reducing interference (Fig. 6).

Adding structured dynamics to the rotation creates a more compact and efficient mechanism for generating orthogonal representations (Fig. 7d and f). This has several potential advantages. Compact (sparse) representations maximize the amount of information held in short-term memory[34]. Similarly, increasing the efficiency of the rotation minimizes the energy needed to transition states. In addition, unlike randomization, a structured rotation is a functional transformation and so it can be easily implemented in a neural network. Future work is needed to understand whether structured rotation is common to all brain regions or if it is restricted to sensory cortices, while more 'cognitive' regions (e.g., prefrontal cortex) use different mechanisms to generate orthogonality.

Future work is also needed to understand the mechanisms generating stable and switching dynamics. In the current study, we did not find any consistent differences in the anatomical location or intrinsic properties of stable and switching neurons that might explain their functional differences (Extended Data Fig. 10). Of course, stable/switching dynamics may reflect other, untested, biophysical differences, such as differences in cell-type. Alternatively, the dynamics may reflect network interactions, whether from local recurrent connectivity or non-linear interactions with top-down inputs. The latter may be more likely given that stable and switching dynamics increased over days and so may be learned (Fig. 4e–g). Finally, future work is needed to understand whether the increase in structural dynamics over days was induced by the increasing interference caused by the learned association.

Methods

Implicit Learning Paradigm

Mice were exposed to an implicit sequence-learning paradigm for four consecutive days. On each day, mice were head-fixed and listened to 1500 sequences of four chords (ABCD, ABC*D, XYCD, and XYC*D). Mice were initially naïve to all chords and sequences. Recordings lasted about an hour and were done at the same time each morning (± 1.5 hours). As animals were on reverse light-cycle; recordings were during their active time.

Within a sequence, each chord lasted 100 ms, and was separated by a 75 ms inter-sound interval (ISI). Inter-trial intervals (ITI) lasted between 500 and 1000 ms (random uniform distribution). Each chord was a combination of 2 frequencies; all between 10 kHz and 65 kHz and spaced by 7/12 of an octave. A, B, X, and Y sounds were lower in frequency than C and C* chords. The frequency of D fell between context and C/C* chords. If the frequency of A was less than B, then the frequency of X was greater than Y, and vice versa (8/12 of an octave). The frequencies and chords were varied across mice. Sound waveforms were created in Matlab, with a sample rate of 140 kHz, and played through MF1-S speakers (range - 1kHz to 65kHz, Tucker Davis Technologies, Alachua, FL /USA). Speakers were calibrated with a CM16 microphone (Avisoft-Bioacoustics, Glienicke, Germany) and an Ultramic USB microphone (Dodotronic, Castel Gandolfo RM, Italy) to a sound pressure level (SPL) of 70 dB. Sounds were played to left ear. A light was presented 100 ms before all sequences, although it did not evoke a response in auditory cortex neurons.

Each sequence began with two chords that indicated one of two contexts. In the first context, the A chord was always followed by the B chord (the ‘AB’ context). In the second context, the X chord was always followed by the Y chord (the ‘XY’ context). Context AB was most frequently followed by C (rarely by C*), while context XY was most frequently followed by C* (rarely by C). All trials ended with D. Expected sequences (ABCD, XYC*D) occurred on 68% of trials (equal number of ABCD and XYC*D). Unexpected trials (ABC*D, XYCD) occurred on 20% of trials (equal number of ABC*D and XYCD). Overall, the AB and XY contexts occurred equally per day, as did the C/C* stimuli. This prevented any *a priori* expectation of any stimulus.

The remaining 12% of trials contained an ‘ambiguous’ third stimulus, which was created by combining the frequencies of the C and C* chords. The ambiguous stimuli were presented randomly and did not interfere with associative learning. They were introduced for reasons unrelated to the current manuscript and therefore are excluded from all analyses, with the exception of verifying the clustering of the temporal dynamics (as detailed in Extended Data Fig. 8 and section Testing Cluster Labels on Withheld Data).

Prior to and after the block of 1500 sequence trials, the C and C* chords were played in isolation for 300 trials, to measure the stability of representations. Each chord was played for 100 ms with a random 500 – 1000 ms delay between chords.

All animals experienced the same paradigm and therefore were not assigned to ‘groups’ and did not require blinding of experimenters. All trial types occurred randomly during the 1500 trials on a given day, according to their probabilities and ensuring equal numbers of trial types. Experimenter was blind to trials during preprocessing of the data (e.g., filtering, spike sorting, etc).

Neuronal Recordings

Animal Subjects—All animal procedures were approved by the Princeton IACUC and carried out in accordance with National Institute of Health standards. Seven adult male PV-Cre+/− C57BL6 mice were used for recording and passive learning experiments. Mice were between 13 and 19 weeks old at the start of recording. Animals had free access to food and

water and were housed in a reverse light cycle. The ambient temperature was 68–79°F and the humidity was 20–40%. Experiments were conducted in a sound proofed behavioral chamber.

Neural Recordings—Neural activity was recorded at 30kHz using the Intan RHD2000 system (Intan Technologies, Los Angeles, CA). Analog signals for the speakers were split and routed to the interface board, allowing for alignment of sound timing and neural activity.

Implant Surgery—While under anesthesia, 32 channel silicon recording arrays (NeuroNexus, Fairfield, CT) were implanted into auditory cortex. Six mice were implanted with a four shank probe (8 electrodes per shank), inserted along the A/P axis. One mouse was implanted with a one shank probe (32 electrodes total). All electrode probes were implanted in right auditory cortex, centered on stereotaxic coordinates −2.7 AP and 4.8 ML from bregma. Probes were lowered to 970–1400 um below cortical surface to target primary auditory cortex (see Extended Data Fig. 10g for approximate neuron locations), although dorsal contacts may have also recorded from secondary auditory cortex. Electrodes were stabilized with KwikSil (World Precision Instruments, Sarasota, FL). Three screws (miniature self-tapping screws made from #303 stainless steel; J.I. Morris, Oxford, MA) were used to keep the headpost (3D printed at Midwest Prototyping, Blue Mounds, WI) and electrode stable. Ground wires were wrapped around the screw on the opposite side of the brain. Metabond (Parkell, Edgewood, NY) was used to fix all implants to skull.

After surgery, mice were given several days to recover and buprenorphine was provided during recovery. Prior to recording sessions, mice were acclimated to handling by the experimenter and head fixation in increments of 15 minutes. Location of silicon probe was confirmed with histology (see Fig. 1a for example electrode placement from one animal). Electrode tracks were determined by labeling for astrocytes (GFAP - green).

Isolating Single Neurons—Single units were isolated from the raw 30 kHz signal using Plexon Offline Sorter. Raw data was imported into Plexon Offline Sorter, and filtered using a 350 Hz highpass, 4-pole filter. Next, we re-referenced all channels to the common average. Using these traces, we identified clusters of spikes. Animals were excluded from future analysis if they had fewer than 5 single units. We recorded from 10 animals, but only 7 had sufficient single unit activity to be included; otherwise no data were excluded from the study. No statistical methods were used to pre-determine sample sizes, but our sample sizes are similar to those reported in previous publications[4,17]From the remaining animals, we found 522 single units across the 4 days of recording (n=121 on Day 1, n=124 on Day 2, n=134 on Day 3, and n=143 on Day 4). See Supplementary Table 1 for a breakdown of neurons recorded per animal per day.

Firing Rate Calculation—The instantaneous firing rate of neurons was estimated at each time point by inverting the inter-spike interval. This trace was smoothed with a 1 ms boxcar and downsampled to 1000 Hz. Data was then segmented by trial start and end times. For sequence data, trials were taken from 70 ms prior to the A/X chord to 355 ms after the end of the D chord. For the C/C* chord alone, trials were taken to start 70 ms prior to chord onset and end 280 ms after the chord ended. Data was smoothed again with a 20 ms boxcar.

All time labels in figures indicate the leading edge of any time frame or window (i.e., including data prior to that labeled point). In example firing rate plots (Fig. 1b), the mean and SEM are shown; expected conditions were randomly downsampled to match the trial count of unexpected conditions (n=150 trails per condition). Preprocessing of firing rate data (segmentation and smoothing), computation of the z-scored firing rate difference, and phenograph clustering were performed in Matlab 2016 (Mathworks, Natick, MA). All other analyses were performed in Python 3.5.5. For the python analyses (jupyter notebook[47]), we utilized the scipy version 0.19.1[48,49], sklearn[50], numpy version 1.13.1[51] and pandas version 0.20.3[52] packages (specific functions referenced below). All plotting was done with matplotlib version 3.0.3[53].The network model (Supplementary Fig. 4 and 7, Fig. 7) was written in Python 3.7.3, using PyTorch version 1.0.1.

Encoding Axes (Classifiers)

Training Classifiers—All classifiers were trained using the same procedure. Classifiers only differed in their training period and condition groupings. Each classifier was trained on each day for each animal, using the vector of averaged firing rate of simultaneously recorded neurons. Firing rate was averaged over 100 ms time period, starting 10 ms after stimulus onset (to account for delay in sensory response). This resulted in a matrix of mean firing rates for each neuron and each trial (i.e., matrix size = neurons × trials). See Supplementary Table 2 for a list of classifiers and their details. When forming groups for all classifiers, trials were balanced across (ABCD, XYCD, ABC*D and XYC*D) conditions and a subset (10%) were withheld for testing. The resulting classifier is a hyperplane defined by its orthogonal vector (size = neurons) and an intercept (size = neuron).

Classifier Type: We used a standard linear SVM (support vector machine) classifier for all classification analyses. The linear classifier relates x (features) to y (output) via a linear equation, with weights (w) and an intercept (b).

$$f(x) = w^T x + b$$

Using the projection ($f(x)$), inputs (x) can then be classified into categories. Here, we used classifiers to map ND firing rate data onto a 1D encoding space. This allowed us to understand how the firing rate data encodes a given stimulus (i.e., A vs. X).

For cross-validation, a condition-balanced set of trials (10% of all trials) were withheld and used in all future analyses and figures. The remaining trials were used to train the classifier. To prevent bias in the classifier, we downsampled the expected trials to match the unexpected trial count. To ensure all of the training trials were incorporated into the classifier, we downsampled the expected data 100 times and trained a classifier on each sample. The final classifier was calculated by taking the mean (intercept (b) and weights (w)) of these 100 trained classifiers.

To train the classifier we used the stochastic gradient descent SGDClassifier function in the sklearn.linear_model package (sklearn version 0.19)[50] for Python3 (version 3.5.5). This

method fits the classifier weights (w) and intercept (b), by minimizing the error function, which is a combination of a loss function ($L(y, f(x))$ = hinge loss) and regularization ($R(w)$ = elastic net):

$$E(w, b) = \frac{1}{n} \sum_{i=1}^{n} L(y_i, f(x_i)) + \alpha R(w)$$

Classifier Regularization: To minimize overfitting of the classifiers, we used elastic net regularization ($R(w)$ in the above equation). Elastic-net regularization combines the L1 and L2 norms to increase the sparsity and decrease the length of the weights (w), respectively. The alpha (α) parameter determines the amount of regularization. Parameters were the same for all classifiers (neural data and model): L1 ratio (the elastic net parameter specifying the ratio of L1 to L2 penalties) was set to 0.65, alpha (the regularization amount) was set to 0.01, the learning rate was set to 0.00001, and the number of iterations was set to 1000.

Testing the Generalizability of Results across Classifier Hyperparameters: As shown in Supplementary Fig. 3, we varied the hyperparameters and classifier type to ensure that neither affected our results. We tested hinge, log and squared loss linear classifiers. Hyperparameter ranges were alpha (regularization level) = [0, 0.001, 0.01, .1] and L1/L2 ratio (elastic net ratio) = [0, 0.25, 0.5, 0.65, 0.75, 1].

Sensory Classifiers: Sensory classifiers distinguished between two sounds during their presentation. The A/X, B/Y, and C/C* sensory classifiers were trained using the average stimulus evoked firing rate activity (A/X:10–110 ms, B/Y:185–285 ms, C/C*:360–460 ms). Condition groups were the two possible stimuli (e.g., A/X classifier distinguished A and X trials). Note, an additional C/C* sensory classifiers was trained using the response to the C/C* chords (100 ms) presented in isolation (Extended Data Fig. 3). Similar results were found with both C/C* classifiers.

Memory Classifier: The A/X memory classifier was trained to classify trials by their context (AB or XY), based on neural activity during the presentation of the C/C* stimulus (360–460 ms).

Relationship between the B/Y classifier and other classifiers: The probabilistic transition from A/X to C/C* allowed us to balance the trials used in their classifiers and, thus, independently decode both the A/X and C/C* representations during the sequence. In contrast, the transition from A/X to B/Y is deterministic (all A stimuli were followed by a B stimulus and X by Y). Because of this, we cannot balance A/X and B/Y trials when constructing a B/Y classifier, and so the B/Y classifier is only differentiated from the A/X sensory and memory classifier by the time period of the response. Unsurprisingly, on day 1, A/X and B/Y sensory axes were aligned (consistent with their association), but this decreased with experience (Extended Data Fig. 1f) In contrast, the B/Y and C/C* sensory axes were not aligned (Extended Data Fig. 1g). This may reflect the fact that the C/C* stimulus was already fully predicted by the A/X stimulus, and so the subsequent B/Y

stimulus did not add predictive value[24]. Importantly, this does not impact the results relating A/X sensory, C/C* sensory and A/X memory axes.

Cross-temporal Classification: To study the evolution of A/X context information, we performed cross-temporal classification. We trained classifiers to distinguish between A and X conditions using 25 ms time bins of firing rate data, stepping by 10 ms, throughout the sequence. These classifiers were then tested on their ability to distinguish A/X trials on withheld trials (across the same set of 25 ms bins). This provided an estimate of how well classifiers generalized across time (Extended Data Fig. 6a).

Tracking Evolution of Sensory and Memory Information with a Day: In addition to training a single classifier for the entire day, we also trained classifiers within a day. This allowed us to follow the timecourse of learning within a session. To train within day, classifiers were trained on 6 blocks per day, with each block consisting of 500 trials, stepped by 200 trials. For each block of trials, we balanced conditions and used cross validation (withholding 10% of trials) to test performance and make projections.

Within-day classifiers were used to study the angular relationship between encoding axes during learning (Fig. 2b and 3g) and to measure the cityblock distance between the A/X sensory and A/X memory classifier (Fig. 7f). To combine across animals, classifier weights were length normalized by the number of neurons. Within-day classifiers showed similar trends as whole-day classifiers when decoding responses (Supplementary Fig. 2). Differences are likely due to the limited number of trials (limiting the statistical power) available for testing the classifier performance within a block of trials (12 trials per condition per animal).

Projection onto Encoding Axis—To study how the high dimensional population activity encodes variables within the sequence, we projected the firing rate activity on withheld trials into a 1D encoding space defined by the trained classifiers. The projections are signed, based on labels chosen during training. For example, on the A/X-sensory axis, a negative projection indicates A encoding and a positive projection indicates X encoding (see Fig. 1c for schematic of projection).

To examine how this encoding evolved over time, we projected neural activity over the timecourse of the sequence. For each trial (withheld from training, n=1064 trials), we calculated the average vector of neuron firing rates (FR) in 25 ms time bins, stepping by 10 ms, over the course of the sequence presentation. At each time bin (t), we took the dot product between each firing rate vector ($FR(t)$) and the relevant encoding axis (w, size = neuron), and then added its intercept (b):

$$Projection(t) = FR(t) \bullet w + b$$

Classifier training and projections are done for each trial (withheld from training) on a mouse-by-mouse basis (i.e., each day is separate, and we did not create a pseudo-population of neurons across days). Before combining trial projections across mice, we z-scored the projections across conditions (i.e., subtracted mean and divided by standard deviation,

calculated from all conditions). Once, z-scored we could combine trial projections across animals to study population encoding. Note, this z-score captures the relative separation between conditions across time and ignores any absolute drift in firing rates occurring over time.

To examine how encoding of context (AB vs. XY) and C/C* stimulus changes over time, we performed two-sided t-tests (function: ttest_ind from scipy.stats package) on each time bin (df=1062). Although we did not test for normality, the z-scoring procedure was intended to normalize the data to support the assumption of normality underlying the t-test. The neural population was said to be carrying significant information about a stimulus (or memory) if the associated p-value ≤ 0.001, Bonferroni corrected for multiple comparisons across time (e.g., Fig. 1d–e).

Strength of Encoding: We defined the 'strength of encoding' as a metric of classifier accuracy, which combines both the magnitude and accuracy of the trial projections. Recall, a trial's projection on a given axis indicates how much that trial's activity represents either the negative or positive label (e.g., A and X respectively on the A/X sensory axis). Therefore, in order to combine trial projections across all conditions (and animals), we inverted (i.e., flipped the sign of) the projection of trials with a negative label. For example, to test A/X encoding (Fig. 1f, n=1064 withheld trials per day), we inverted the projection of A condition trials (ABCD and ABC*D).

Measuring Classifier Accuracy with AUC: After training, each classifier defines a hyperplane in feature space (firing rate of neurons) that separates samples as belonging to one of two conditions (e.g., AB or XY). This label may be correct (True positive or True negative) or incorrect (False positive or False negative). To supplement analyses using the projection onto encoding axes, we calculated the area under the curve (AUC) of the receiver operator characteristic (ROC) curve to measure classifier performance (Supplementary Fig. 1; using sklearn.metrics package). The area under the curve (AUC) statistic measures how well the classifier is able to separate the two distributions: an AUC of 1 indicates perfect performance, an AUC of 0.5 indicates random chance, and an AUC < .5 indicates the classifier assigned samples the wrong condition label. AUC was calculated for fixed time periods and in a sliding window moved over the sequence timecourse (Extended Data Fig. 2). The AUC showed qualitatively similar effects as the projections onto the encoding axis. In the main manuscript, we focused on the projections, because they show on a trial-by-trial basis whether neural representations move in the predicted or unpredicted direction and provide a more direct measure of effect size.

Statistics and Reproducibility—Almost all analyses used one of three different forms of non-parametric permutation tests: 1) a bootstrap to estimate the distribution of a statistic, 2) a bootstrap to estimate the distribution of a linear regression and 3) a permutation test to test for differences between groups. Below we detail these three tests.

Bootstrapped Estimates of the Distribution of a Statistic: Bootstrapped distributions were used to estimate distributions for plotting and statistical tests. In general, the procedure involved resampling data (e.g., trials or neurons) with replacement 5000 times[54] and

recalculating a statistic (e.g., mean projection, angle or principle components). Next, this distribution was tested against a null hypothesis (e.g., tested against zero). The percent of the bootstrapped distribution above or below the null hypothesis value was taken as the likelihood that the neural data was greater than (or less than) the null hypothesis.

For example, we tested the significance of encoding strength (accurate vs. inaccurate) during each time period. We randomly resampled from the observed trials with replacement[54]. Using this bootstrapped distribution, we could determine if the encoding strength (i.e., accuracy) was significantly different from zero (i.e., significantly positive/correct or negative/incorrect). To do this, we estimated the probability the observed response was zero by measuring the percent of the bootstrapped distribution that was above or below zero, for positive and negative observed mean projections, respectively (in this example, we doubled the probability and report the two-sided p-value).

Bootstrapped Estimate of the Distribution of Linear Regressions: To measure changes across time (e.g., days or blocks) we calculated linear regressions (scipy.stats.linregress) across time points. For example, we measured the change of encoding strength across days (Fig. 1i and 3b), change in angle between classifiers (Fig. 2b and 3g) and the change in relative structure in the rotation (Fig. 4e–g). To estimate the distribution of values of the linear regression, we used a bootstrapped procedure, as described above (5000 resamples). Using this distribution, we calculated the mean and standard deviation of the slope (or r-values) and also plotted the mean trend line and their 95% confidence intervals (i.e., ±1.96*standard deviation of the trend lines). Finally, we tested if the observed slope was significant, by estimating the probability the observed slope was zero, by measuring the percent of the bootstrapped distribution that was above or below zero, for positive and negative observed slopes, respectively (one-sided p-value).

Permutation Test for Differences Between Groups: We used permutation tests to test the significance of an observed difference across groups. First, we calculated the observed difference between group means. Then, the group labels were permuted (4999 times) and the difference between the means of the shuffled groups were calculated. The shuffled and non-shuffled differences were combined to create a null distribution (size = 5000). This null distribution shows the differences expected given no actual difference between the two groups. The likelihood of our observed difference was then estimated as its percentile within the full null distribution.

For example, we used a permutation difference test to compare the A/X encoding strength along the A/X sensory axes versus the A/X memory axes (Extended Data Fig. 6b). First, we calculated the difference of mean encoding strength between axes. Then, we randomly permuted the classifier labels (e.g., A/X sensory or memory) across trials (4999 times) and recalculated the difference in mean encoding strength. This distribution was combined with the original observed difference to create the null distribution, which was used to calculate the significance of the observed difference.

Assumptions Underlying Statistical Tests: One advantage of non-parametric tests is that they do not make assumptions about the underlying distribution. Note, there is a lower

bound on non-parametric significance measures; exact p-values below 1/N cannot be reported, where N is number of shuffles/resamples. Parametric t-tests were used when measuring differences between trial projections onto a specified encoding axis (e.g., Fig. 1d). These data were z-scored before the t-test, supporting the underlying assumption of normality, although normality was not formally tested for each time bin. In addition, these results were confirmed with nonparametric tests.

For all statistical tests, all neurons, and trials from the four relevant conditions were used. As detailed above, the experimental paradigm was repeated in 7 animals.

Additional information on research design is available in the Life Sciences Reporting Summary.

Calculation of Angles between Axes—To examine the relationship between representations, we measured the angle between encoding axes (i.e., the angle between the trained classifiers/hyperplanes). To estimate the angle across animals, classifier weights were normalized to length 1 and combined across animals (per day or per block). Using these vectors (one per hyperplane; size=neurons), we calculated their angle:

$$angel = arccos\left(\frac{A \bullet B}{\|A\|\|B\|}\right)$$

To ensure that the angle was not biased by outliers within the neural population, we bootstrapped across neurons, re-calculating the angle on each of 5000 resamples; these distributions were then used when measuring changes in angles across days with linear regression. The reported angles within a day were calculated by taking the mean and standard deviation across all blocks within a day (Fig. 3g). These results were not qualitatively different from angles calculated between classifiers trained on all trials within a day.

When building the A/X and C/C* classifiers, we labeled conditions such that associated stimuli shared the same sign (A and C were negative, X and C* were positive). Therefore, angles between the classifiers that are less than 90 degrees correspond to the classifier responses aligning with the predicted, expected associations (e.g., neurons that prefer A [or X] also prefer C [or C*]). Angles greater than 90 degrees indicate the populations' selectivity is aligned with respect to unexpected pairings (e.g., neurons that prefer A [or X] also prefer C* [or C]).

Testing Correlation of Single Neuron Responses to A/X and C/C*—We found alignment between the A/X sensory and C/C* sensory axes (Fig. 2b). To relate this to the selectivity of single neurons, we calculated each neuron's selectivity to the A/X and C/C* stimuli (see section on Temporal Selectivity Profiles). We measured the correlation in A/X and C/C* selectivity across neurons with linear regression (variance estimated by bootstrapping across neurons). To test if this relationship changed across days (Fig. 2c–d), we performed a regression on the slope across days.

Neural Activity in 2D State Spaces (Dimensionality and Angle)—To understand how the neural population encoded two dimensions at once, we projected neural activity over time onto two encoding axes, creating a 2D space. For example, Figure 2a shows the encoding of A/X sensory and C/C* sensory information along the x and y axes, respectively. Projections were as described above in *Projection on Encoding Axis*.

Calculation of Principal Components and Dimensionality of Neural Trajectories in State Space: To understand the dimensionality within each 2D state spaces, we performed principle component analysis (sklearn.decomposition.PCA) on the distribution of the mean projections of all four conditions within the 2D state space (i.e., concatenating the timecourse from ABCD, ABC*D, XYC*D, and XYCD). This resulted in two principal components (PCs). Each PC captured a proportion of the variance in responses within the 2D space, which defined the explained variance ratio (EVR) of that PC. The EVR of PC1 was used to estimate the dimensionality of the neural trajectories within the 2D state spaces. A high EVR for PC1 indicates low dimensionality, because PC1 is explaining most of the variance. Meanwhile, if PC1 EVR equals ½, the dimensionality is high, because both PCs explain similar amounts of variance, which occurs when the trajectories move equally in all dimensions (Fig. 3h).

We used a permutation test to test if the observed dimensionality was lower than expected by chance (Fig. 3h). For this, we created a null distribution of projections into the 2D state space by randomly permuting (4999 shuffles) the time labels within the sequence of each point, separately in both the x and y dimensions. For each permutation, we recalculated the EVR. The null distribution (5000; shuffles plus the observed value) was used to estimate the probability of randomly observing a value greater-than-or-equal to the original EVR (one-sided test).

Statistics on EVR and PC Angle: We used a bootstrap procedure (as described above) to estimate the distribution of PC angles (i.e., violins in Fig. 3h) and the EVR of the PCs. The bootstrap process involved randomly sampling trials (5000 with replacement) within each condition group, projecting neural activity into the state space, calculating the mean trajectory per condition, and then recalculating the PCs (and respective EVRs). These distributions were used for calculating regressions (one-sided test) and estimating the variance of the distributions.

We use a permutation procedure (as described above) to compare the PC angles and EVRs across state spaces. For this, we shuffled (4999 permutations) mean data projections across state spaces and recalculated the difference in PCs, angles, and EVR (5000; shuffles plus the observed value). With this distribution we estimated the probability of observing the original difference across state spaces under the null hypothesis that there was no difference between state spaces (one-sided test).

Dimensionality in Full Neural Space: To study global changes in the neural space, we also calculated the dimensionality of neural responses in the full N-dimensional neural space (where N is the number of neurons). The analyses followed the same framework as the dimensionality calculations within the encoding state spaces. On each day, we combined

neurons across animals to create a pseudo-population. We averaged the firing rate per condition (ABCD, ABC*D, XYCD, XYC*D, balancing number of trials) within 25 ms bins, stepped by 10 ms. The average response was calculated for each condition around the presentation of C/C* (340 to 520 ms) and PCA was performed on the concatenated data (size = (condition × time) × N). The distribution of EVRs was estimated with a bootstrap; resampling neurons with replacement per day and then recalculating the PCs. These were then used to estimate the change in EVR across days (Extended Data Fig. 5e).

Estimating Trial-by-trial Correlations Between C/C* Encoding and A/X Sensory and Memory Encoding—To test whether the A/X sensory and memory representations impact sensory processing, we correlated the A/X sensory and memory responses on a given trial with the strength of C/C* response (Extended Data Fig. 7). To understand how A/X encoding strength influences *future* sensory processing, we took A/X encoding 50 ms prior to the C/C* stimulus and correlated it with C/C* encoding strength (taken during the C/C* stimulus: 360–460ms). The timing of A/X and C/C* encoding were separated to ensure any observed relationships did not simply recapitulate the alignment of the axes.

We examined expected and unexpected stimuli independently (n=532 trials for both groups; all trials withheld from classifier training). Responses to negatively coded trials (e.g., A trials and C trials) were inverted such that all positive values indicate correct encoding and negative values indicate incorrect encoding. We used bootstrapped linear regression to correlate the strengths of the C/C* representations and the A/X sensory or A/X memory representation.

Timing of Crossover from A/X Sensory to A/X Memory Encoding—To gain insight into the timing of the A/X rotation, we estimated the moment during the sequence when A/X encoding switched from A/X sensory to A/X memory encoding (Extended Data Fig. 6c). The crossing timepoint was defined as when A/X encoding was stronger along the A/X memory axis than the A/X sensory axis. The crossover time was restricted to 25 ms after sequence onset to avoid early spurious crossovers. A bootstrap procedure (5000 resamples of trials with replacement) estimated variance in timing and changes over days.

Rotation of A/X Sensory to A/X Memory

Temporal Selectivity Profiles—A/X selectivity changed over the sequence timecourse, from a ‘sensory’ representation to a ‘memory’ representation. To understand how the dynamics of individual neurons supported this transformation, we measured each neuron’s temporal selectivity (n = 522 neurons). Selectivity was measured as the difference in firing rate between the AB and XY trials (n = 600 trials). All four conditions were balanced (n = 150 trials), ensuring A/X selectivity did not reflect the C/C* stimulus response. The A/X firing rate difference was calculated in 25 ms time bins (stepping by 10 ms) over the entire trial (from −160 ms to 790 ms, relative to the onset of the A/X stimulus, creating 96 time bins). To normalize each neuron’s firing rate difference, we z-scored its firing rate difference (for each time bin, t) against a null distribution, which was created by randomly permuting the trial labels (n=1000 shuffles of AB and XY trial labels).

$$zFR\left(t\right) = \frac{FR\,diff(t) - mean(FR\,diff\ shuffles(t))}{std(FR\,diff\ shuffles(t))}$$

To measure changes in selectivity within a day, we calculated each neuron's z-scored firing rate difference in 6 blocks per day (again balancing trials by condition; 500 trials per block, stepped by 200 trials). A similar approach was used to calculate z-scored firing rate differences to the C/C* stimulus (grouping trials by C/C*).

To illustrate how A/X rotational dynamics avoid interference from the C/C* sensory input, we plotted how stable and switching neurons respond to the four conditions (Fig. 6f). Each neuron's z-scored firing rate response to each of the four conditions (ABCD, ABC*D, XYCD, XYC*D) was estimated by calculating the difference in response to that condition, relative to the mean response to all conditions (trials were balanced across conditions). In order to combine and average condition responses across neurons within a group (stable and switching), we inverted neurons with a preference to X. As the goal of this analysis was to plot the response to the C/C* stimulus, we only included neurons selective to C/C* (similar results were seen when including all neurons). To declutter the plot, we averaged the condition traces across A and X conditions prior to the onset of C/C*.

Testing for Structure in the Rotation: Measuring the Proportion of Conjunctive and Single Selectivity in the Neural Population—The three different mechanisms for rotation (Fig. 4a–c; Independent, Random, and Structured) make different predictions about the number of neurons selective for A/X in one time period ('single' neurons, selective during either sensory or memory) or both time periods ('conjunctive' neurons selective during both sensory and memory). The independent mechanism predicts more single neurons than expected by the random mechanism. In contrast, a structured rotation predicts more conjunctive neurons (and fewer single neurons). Therefore, to differentiate between these mechanisms, we determined each neuron's A/X selectivity during the sensory and memory time periods. A neuron was considered selective for a given time period if its z-scored firing rate difference was significant at any time point during that time period (abs(z-score) ≥ 1.96, or $p \leq 0.025$, Bonferroni corrected by number of time points). Significance was independently measured in both the A/X sensory time period (0–100 ms) and the A/X memory period (350–450 ms). In rare cases where multiple crossings occurred within the time range, the selectivity was determined by the mean response.

For each time period, a neuron belonged to one of three categories: it represented the A stimulus, X stimulus, or was not selective for either (a null, or '0', neuron). Combining across the two time periods (sensory and memory), neurons can be in nine different categories (Supplementary Table 3). These can be grouped into three categories: conjunctive neurons that are selective during both time periods (AA, XX, AX, XA), 'single' neurons that are selective during only one time period (A0, X0, 0X, 0A) and non-selective neurons that are not selective to A/X in either time period (00). On a given day (or within a block), we calculated the proportion of recorded neurons in each selectivity category (conjunctive or single). To correct for the overall degree of selectivity in the network, we used the conjunctive/single ratio as our main measure of structure in the rotation.

Nonparametric Test against the Random Mechanism: While the independent mechanism predicts more single neurons than the random mechanism, the structured rotation predicts fewer single and more conjunctive neurons than the random mechanism. By definition, the random mechanism argues changes in selectivity should have no relationship over time. Therefore, to test against the random mechanism, while controlling for overall selectivity, we created a null distribution by permuting selectivity across neurons within each time period (n=1000 permutations), breaking any relationships in selectivity across time periods. We determined the likelihood of our observed results by measuring the percentile of our observed neural proportions in the null distribution (Fig. 4e–g).

To examine changes over time (blocks and days), we first controlled for changes in number of selective neurons, by z-scoring. To this end, we subtracted the mean of the random chance distribution and divided it by the standard deviation of the random chance distribution (Fig. 4e–g). To estimate the distribution of z-scored values, we used a bootstrap procedure (5000 resamples of neurons, per block per day). For each bootstrap, we recalculated the z-scored proportions of conjunctive and single neurons and used the resulting distributions to estimate changes across days.

Testing for Structure in the Rotation: Chi-squared and Binomial Tests—We tested the full table of A/X temporal selectivity of neurons from all 4 days (n=522; Supplementary Table 3) against random chance, by using a probabilistic model followed by chi-squared and binomial tests. For each time period, a neuron has a probability of being selective to either A or X. The probability of being selective to A or X during the sensory or memory period can be written as p_A^{sen} or p_X^{sen} and p_A^{mem} or p_X^{mem}, respectively. We assumed the probability per stimulus is equal: $p_s = p_A^{sen} = p_X^{sen}$ and $p_m = p_A^{mem} = p_X^{mem}$. Here, p_s and p_m are the probabilities of selectivity during the sensory period and memory period, respectively. Therefore, the probability of non-selectivity during the sensory period and memory period can be written as $1 - 2p_s$ and $1 - 2p_m$, respectively. Because the random mechanism predicts selectivity will be independent across time, the probability of each of the nine categories can be estimated by multiplying the probabilities of selectivity in each time period. Supplementary Table 4 shows the probabilities of each of the nine A/X selectivity types as predicted by a random mechanism.

Because we were interested in comparing the levels of conjunctive neurons against random, we used the counts of single (A0, X0, 0A, 0X) and non-selective neurons (00) to fit p_s and p_m (using the scipy.optimize.minimize function, by minimizing the sum of squared errors between the predicted single and non-selective counts). Using the fitted values for p_s and p_m, we calculated the probabilities of each of the nine neural selectivity categories. Together, these nine probabilities can be compared to the observed proportions in the neural data with a chi-squared and individual binomial tests.

Testing for Structure in the Rotation: Generating and Testing Random Selectivity Data—To further test against a random mechanism and ensure our results were not due to smoothing over time, we generated random temporal selectivity profiles and performed the same set of analyses as were applied to our neural data. Random profiles

consisted of responses that ranged from a sustained response (profile 0) lasting the full sequence of 550 ms to a short 50 ms response (profile 20), with intermediate profiles decreasing linearly after the initial 50 ms response (Supplementary Fig. 5 and 6). Next, each response profile was multiplied by random values drawn from a standard normal distribution ($\mu = 0$, $\sigma = 1$) to generate A/X selectivity. Finally, we smoothed this random data using kernels between sizes 80 ms and 400 ms (lowess function in Matlab). To avoid smoothing artifacts, time points were padded by the size of the largest smoothest kernel. This process was repeated 1000 times for each profile and smoothing level to create a distribution of randomly generated temporal selectivity profiles.

The resulting random selectivity data was analyzed in the same manner as neural data, calculating selectivity during A/X (0–100ms) and C/C* (350–450ms). For each 'random' population, we calculated the number of neuron counts per A/X temporal selectivity pattern (Supplementary Table 3). Note, random data cannot yield similar selectivity as our data, as the percentage of selective neurons was much greater than expected by chance (36%). So, we had to set a 'selectivity' threshold to reproduce the observed neural selectivity. Using these random selectivity counts, we tested if random data could generate a structure rotation to the same degree as observed in the neural data. Supplementary Fig. 6a shows resulting chi-squared on random data and Supplementary Fig. 6b compares results to neural data. Supplementary Fig. 6d shows the percentage of random populations that were significant ($p \leq 0.05$) for the three binomial tests that were all significant in the neural data.

Using the same techniques applied to the neural data, for each 'random' population we obtained z-scores indicating whether the 'random' population's selectivity proportions (of single and conjunctive neurons) were unexpected by a random mechanism. Next, we compared the level of structure observed from our neural data to what was observed in the randomly generated data. Supplementary Fig. 6e,g,i shows how far away, in standard deviations, the neural results were from the results generated for each profile/smoothing combination of randomly generated data. Supplementary Fig. 6f,h,j compares neural data to the combined results from all randomly generated data.

Clustering of Temporal Selectivity Profiles

Phenograph: To explore the structure in the temporal selectivity profiles across our neural population, we used the unsupervised 'Phenograph' clustering algorithm[55] to cluster the profiles of all recorded neurons, across all days (n=522). Phenograph works by 1) forming a connected graph of data points, where the edge weight between two nodes is the Euclidean distance in their temporal profile of selectivity, and then 2) clustering points based on the community structure within this connected graph. Each community has a 'modularity', which compares the density of edges within and between identified communities. The Louvain Community algorithm[56] iteratively discovers communities within the K-nearest neighbors and collapses connected nodes into groups until modularity is maximized[55]. We chose the Phenograph algorithm because it does not require *a priori* specification of the number of clusters and is used in several fields[31].

Varying the parameter (K) in Phenograph: The Phenograph algorithm is unsupervised; the only parameter is the number of local connections (k) that are used to define the local communities. Previous work has shown the algorithm is robust to changes in this parameter[31]. Based on recommendations from this work, we initially chose k=40. Shown in Extended Data Fig. 9b, we confirmed the stability of clustering by systematically varying k and calculating two measures of clustering – the Phenograph's modularity statistic and the silhouette score of the discovered clusters[57].

Validating Phenograph Clustering: d-prime: We used d-prime to validate that the clusters were not overlapping in space. For each pair of clusters being compared, the distance between cluster means (μ_1 and μ_2) is calculated (using Euclidean distance in the full 96-dimensional space). This is divided by the square root of the average of the variances (σ_1^2 and σ_2^2) within each cluster:

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

Following methods described above, we used a permutation test (shuffling cluster labels 999 times) to test whether the observed d-primes were significantly greater than expected by chance Extended Data Fig. 9a.

Validating Phenograph Clustering: UMAP: To further verify our observed clusters, we used the Uniform Manifold Approximation and Projection algorithm (UMAP[58]) to project the temporal profiles of single neurons into two dimensions (Extended Data Fig. 9c).

Validating Phenograph Clustering: K-means Clustering: We also compared the Phenograph clustering to K-means clustering (sklearn.cluster.KMeans) using the same data. As K-means performs poorly in high dimensional spaces[59], we clustered points within the reduced dimensional UMAP space. For each fit, we used 1000 random restarts. All function parameters were set to the default. Unlike Phenograph, K-means requires the number of clusters to be pre-specified. To test how the number of clusters affected the clustering, we varied the number of clusters pre-specified and compared the resulting silhouette scores (Extended Data Fig. 9d). To facilitate the interpretation of the silhouette score, we performed the same clustering on different data sets (random data and C/C* temporal selectivity profiles). The random data was generated by smoothing Gaussian noise with the same number of neurons and timepoints as the neural data.

Validating Phenograph Clustering: Cosine Distance Matrix: To assess similarity of the temporal profiles of all recorded neurons, we measured the cosine distance (scipy.spatial.distance.cdist) between the vectors (Fig. 5c). Cosine similarity is defined as:

$$1 = \frac{u \bullet v}{\|u\|\|v\|}$$

where −1 is maximally dissimilar and 1 is maximally similar.

Testing Cluster Labels on Withheld data: We tested whether the dynamics of stable and switching neurons were consistent across condition groups (e.g., trials where C, C*, or Cmix stimuli were presented). For example, the dynamics of A/X selectivity could be calculated using only C trails (i.e., ABCD and XYCD). We applied the original phenograph labels to these selectivity profiles and measured the correlation between the selectivity in the subset of trials and the original data set (Extended Data Fig. 8a–b). Note that, because only trials with the same 3rd stimuli were used, A/X firing rate differences could not arise from interactions between A/X and the C or C* stimulus. Furthermore, the ABCmixD and XYCmixD datasets were never included in the original A/X z-scored firing rate difference or phenograph clustering. Nevertheless, stable and switching neurons showed similar A/X response profiles and averaged selectivity on these withheld trials (Extended Data Fig. 8a–b).

Measuring the Contribution of Stable and Switching Neurons to the A/X Encoding Axes (Classifiers)

Average Classifier Weight for Stable and Switching Neurons: To understand how the stable and switching functional neuron types contributed to the A/X sensory and memory axes, we calculated the distribution of classifier weights for each cell type (Fig. 6a). To reduce noise, we post-hoc identified and isolate neurons that were selective for ‘none’ of the time periods ($p>.025$, Bonferroni corrected). To combine weights from both A-preferring and X-preferring neurons, the weights of all initially A preferring neurons were inverted (i.e., multiplying their weights by −1). Therefore, the weight distributions in Fig. 6a reflect similarity with the neuron’s initial preference (either A or X). We combined weights of all neurons (across days and animals) for each neuron group (stable, switching and none).

Correlation between Sensory and Memory Classifier Weights: We tested how A/X selectivity of individual neurons changed between the A/X sensory and A/X memory time periods by linearly regressing the weights of the A/X sensory and A/X memory classifiers for each group of neurons (stable, switching or none; Fig. 6b). The weight vector within each animal was length-normalized before combining weights across all animals and days. Statistical significance of the linear regression was determined with a bootstrap test (as described above, 5000 resamples of neurons with replacement). To test whether experience changed the linear relationship between A/X sensory and A/X memory classifier weights, we calculated the linear regression between weights on each day (Fig. 6c; bootstrapping to assess significance). To test whether the linear relationship in weights was significantly different between the switching and stable neurons, we compared the observed difference in slope to a randomly permuted, null distribution created by shuffling stable and switching labels across neuron weights (4999 times; Fig. 6c).

C/C* Selectivity in Functional Neuron Types—The representation of A/X and C/C* became more similar with experience (Fig. 2). To test whether these changes were specific to either stable or switching neurons, we examined C/C* stimulus temporal selectivity of both neuron types for both expected and unexpected sequences (Extended Data Fig. 10e). For each neuron group (stable or switching), we plotted their A/X and C/C* z-scored firing rate differences. To combine across all preferences within a group, we inverted the A/X

selectivity and C/C* selectivity of neurons that initially preferred A. This means the A/X responses are relative to initial preference. Likewise, C/C* responses reflect whether the neuron group's average selectivity aligns with expected (AC/XC*, positive responses) or unexpected (AC*/XC, negative response) sequences. To test whether a neuron group carried the prediction, we averaged the C/C* response (350–450 ms), and used a bootstrap test across neurons (5000 resamples, two-sided test, Extended Data Fig. 10f).

Testing for Intrinsic Differences in Stable and Switching Neurons

Fano Factor: Fano factor measures the variability of neural responses. To compare the inherent variability of each neuron group (stable and switch), we measured the fano factor of single neurons over the sequence. First, we binned the raw spiking data (sample rate = 30kHz; 40 ms bins, stepping by 16.7 ms). In each bin (w), spikes were summed and fano factor (F) was calculated across all trials per neuron:

$$F\left(w\right) = \frac{\sigma_w^2}{\mu_w}$$

For each functional group (stable and switching neurons), we combined across neurons and days (Extended Data Fig. 10a). Extended Data Fig. 10b shows the average fano factor during the pre-stim period (−400 to 0 ms, relative to the A/X stimulus start), the stimulus presentation periods (A/X, B/Y, C/C* and D) and the inter-chord intervals.

Intrinsic Timescale of Neuron Types: Following previous methods, we estimated each neuron's intrinsic timescale, using the autocorrelation, which reflects the time duration over which the neuron integrates information[60]. Extended Data Fig. 10c shows the average autocorrelation of stable and switching neurons (on the pre-stimulus period: −400–0 ms). To quantify the decay time, we fit an exponential to the bootstrapped average autocorrelation: $y = Ae^{-\frac{x}{\tau}} + C$ (using scipy.optimize.curve_fit).

Analytical Model of Rotational Dynamics—Our data show A/X information rotates from a sensory representation to a memory representation over time via dynamics in the population response. As detailed in the manuscript, these dynamics can range from random to structured. To understand the relative benefits of a structured vs. random rotation, we derived analytical expressions for the efficiency of the representation. To facilitate closed-form solutions, this model is simplified and so is missing several characteristics of the real data or the neural network model (which is detailed below), such as noise in the responses and variability in the level of selectivity across neurons.

Random Network: In the random network, we define the probability of a neuron's significant response to sensory input A and X as p_A^{sen} and p_X^{sen}, respectively. The probability of a non-selective sensory neuron is: $p_0^{sen} = 1 - p_A^{sen} - p_X^{sen}$. Likewise, during memory period, the probabilities of memory-selective neurons are p_A^{mem} and p_X^{mem} to A and X memory,

respectively. The probability of non-selective neurons during the memory period is: $p_0^{mem} = 1 - p_A^{mem} - p_X^{mem}$.

Structured Network: A 'pure' structured rotational network consists of only stable and switching neurons. Similar to the random network, these neurons are either selective for A or X during the sensory time period, with probability p_A^{sen} and p_X^{sen}.

Stable neurons maintain their selectivity across A and X time periods, such that:

$$p_{\mathrm{A}}^{\mathrm{mem}} = \begin{cases} 1, & if\ neuron\ is\ selective\ for\ sensory\ input\ A \\ 0, & else \end{cases}$$

$$p_X^{mem} = \begin{cases} 1, & if\ neuron\ is\ selective\ for\ sensory\ input\ X \\ 0, & else \end{cases}$$

The opposite relationship exists for switching neurons:

$$p_A^{mem} = \begin{cases} 1, & if\ neuron\ is\ selective\ for\ sensory\ input\ X \\ 0, & else \end{cases}$$

$$p_X^{mem} = \begin{cases} 1, & if\ neuron\ is\ selective\ for\ sensory\ input\ A \\ 0, & else \end{cases}$$

To define an intermediate model, we portion the total available neurons (N) into two groups: structured and random, with proportion q. Here, qN neurons adhere to *structured* (stable/switching) rules, while $(1 - q)N$ adhere to *random* rules.

First, the model allows us to write the conjunctive/single ratio as a function of rotational structure. Second, using this model, we can highlight the efficiency of a structured rotation with two metrics: percent selectivity and rotational cost.

Ratio of Conjunctive/Single Selective Neurons: One key feature differentiating a structured from a random rotation are the relative proportions of conjunctive neurons to singly selective neurons. In the *random* network, the likelihood of a conjunctive neuron is $p_{conj} = p_A^{sen}\left(p_A^{mem} + p_X^{mem}\right) + p_X^{sen}\left(p_A^{mem} + p_X^{mem}\right)$ and the probability of single neuron is $p_{sing} = p_0^{sen}\left(p_A^{mem} + p_X^{mem}\right) + p_0^{mem}\left(p_A^{sen} + p_X^{sen}\right)$.

To simplify the algebra, we can assume that the probability of a neuron selectively representing sensory or memory are all equal. In other words, $p = p_A^{sen} = p_X^{sen} = p_A^{mem} = p_X^{mem}$. With this simplification, the proportion of conjunctive to singly-selective neurons in the *random* network is:

$$\frac{p_{conj}}{p_{sing}} = \frac{4p^2}{4(1-2p)p} = \frac{p}{1-2p}$$

In contrast, in a *structured* rotation, all selective neurons are conjunctive, with a likelihood:

$$p_{conj} = p_A^{sen} + p_X^{sen} = 2p\,.$$

We defined the conjunctive/single ratio in an intermediate network by linearly mixing between the models. Recall q defines the proportion of structured neurons in the network. Therefore, the proportion of conjunctive to singly-selective neurons can be written as:

$$\frac{p_{conj}}{p_{sing}} = \frac{4p^2(1-q)+2pq}{4(1-2p)p(1-q)} = \frac{2p(1-q)+q}{2(1-2p)(1-q)} = \frac{p}{(1-2p)} + \frac{q}{2(1-2p)(1-q)} = \frac{1}{2(1-2p)}\frac{1}{(1-q)} - \frac{1}{2}$$

Given that $\frac{1}{(1-2p)} \geq 0 \; \forall \; p \in \left[0, \frac{1}{2}\right]$, and that $q \in [0,1]$, the conjunctive/single ratio scales with $\frac{1}{(1-q)}$. In other words, increasing q (i.e., increasing the proportion of structured neurons) increases the conjunctive/single ratio.

Percent Total Selectivity: Next, we examined metrics of efficiency using this analytical framework. First, we calculated the percent of selective neurons. A lower percentage indicates more efficiency in the network's representations.

In the *random* network, the probability that a neuron is involved in either the sensory and/or memory representation is $p_{sel}^{rand} = 1 - \left(1 - p_A^{sen} - p_x^{sen}\right)\left(1 - p_A^{mem} - p_X^{mem}\right)$. Assuming all probabilities of selectivity are equal, this reduces to: $p_{sel}^{rand} = 4p - 4p^2$.

In the *structured* network, because a neuron's sensory selectivity determines its memory selectivity, the probability of selectivity is $p_{sel}^{struc} = 1 - \left(1 - p_A^{sen} - p_x^{sen}\right) = 2p$.

Importantly, for $0 \leq p \leq \frac{1}{2}$, $p_{sel}^{struc} \leq p_{sel}^{rand}$, showing the structured rotation is always more efficient than the random one.

From these equations, we can define the percent selectivity in intermediate models as a linear mixing of the structure and random network, such that the number of selective cells in an intermediate model is a linear function of q: $p_{sel}^{mix} = \left(4p - 4p^2\right)(1-q) + 2\mathrm{pq} = -2\mathrm{p}(1-2\mathrm{p})\mathrm{q} + \left(4\mathrm{p} - 4p^2\right)$. Given that $p \leq \frac{1}{2}$, percent selectivity will decrease with q, showing that increasing structure in the network reduces the number of neurons involved in the representation.

Efficiency of the Rotation: Finally, we estimated the efficiency of the rotation. For this, we scored neurons based on how much their selectivity (response) changed between time periods. Neurons with no change (e.g., non-selective or stable neurons) cost 0. Neurons that

are selective in one time period but not the other (e.g., single neurons) cost 1. Neurons that switch their selectivity cost 2.

The *random* network consists of conjunctive, single and non-selective neurons. The cost of a random rotation is $Cost_{rand} = 2(p_A^{sen} p_X^{mem} + p_X^{sen} p_A^{mem}) + p_0^{sen}(p_A^{mem} + p_X^{mem}) + (p_A^{sen} + p_X^{sen}) p_0^{mem}$. Assuming a single likelihood of selectivity (p), this can be reduced to: $4p - 4p^2$.

A *structured* rotation involves only stable neurons (which cost 0) and switching neurons (which cost 2). Assuming half the conjunctive neurons are switching neurons, the cost of the structured rotation is $Cost_{struc} = 2\frac{1}{2}\left(p_A^{sen} + p_X^{sen}\right) = 2p$.

Intermediate models linearly combine these costs, weighted by q, and so the relative proportion of structure in the network is $Cost_{mix} = (4p - 4p^2)(1 - q) + 2pq = -2p(1 - 2p)q + (4p - 4p^2)$. Again, this is a linear function with respect to q and, given $p \leq \frac{1}{2}$, cost decreases with increased structure.

Altogether, this simplified analytical model shows increasing the structure of rotation increases the ratio of conjunctive/single neurons, reduces the total number of neurons involved in the representations, and increases the network efficiency.

Neural Network Model of Rotational Dynamics—To compare structured and random rotations, we developed a neural network model of rotational dynamics. This model extended the analytical model as it included sensory variance, noise, and the observed associative learning between the A/X and C/C* sensory representations (and subsequent interference).

The network consisted of two layers: an 'input' layer (L_i) that represented external inputs and a 'representational' layer (L_r) that captured the recorded neural responses. While there is no learning in the model, we used PyTorch to take advantage of its network module structure. The input layer consisted of 4 different inputs capturing sensory inputs (A, X, C, and C*). The representational layer consisted of 150 neurons with selectivity of each neuron determined by the feedforward weights (W_{ir}) from the input layer. Weights between neurons in the representational layer (W_{rr}) defined the recurrent dynamics. Firing rate of the representational layer was a rectified linear function of the input (x): $ReLU(x) = max(0, x)$. The network ran in two time steps, with added Gaussian noise ($\varepsilon = N(0, a)$, where $a = 2$, unless otherwise noted). Therefore, activation in the representational layer (L_r) at a given time period (t) can be described by $L_r(t) = ReLU(L_i(t) W_{ir} + L_r(t - 1) W_{rr} + \varepsilon)$.

Each trial involved the sequential activation of A or X inputs, followed by C or C* inputs. As in the task, the four possible sequence trial types were AC, AC*, XC, and XC*. Each instance of the model consisted of 1000 trials with equal trial counts per condition. There is no learning in the model; the weights between the sensory layer and network are preset to reflect the association between A/X and C/C*. Without learning, there was no need for unequal trial counts to generate an association and so trial counts per condition were balanced to match our neural analysis.

Model Recurrent Weights: Recurrent weights (W_{rr}) were designed two general ways. First, control models without rotation had either no recurrence ($W_{rr} = 0$), decaying self-recurrence, or stable self-recurrence (i.e., $W_{rr} = \frac{1}{2}I$ and $W_{rr} = I$ for decaying and stable, respectively).

Second, rotation models had recurrent weights that rotated representations, with varying degrees of structure. Creating rotation models, with or without structure, required specifying the selectivity of each neuron. The neuron's selectivity can be understood as a point in two-dimensional space, (w_s, w_m), where the first dimension (w_s) defines selectivity during the sensory time period and the second dimension (w_m) defines selectivity during the memory time period. We created a relationship between the sensory and memory selectivity by drawing their values from a 2D Gaussian with a non-diagonal covariance matrix. The covariance matrix's diagonal elements represent the variance of selectivity within each time period (set to .451). The off-diagonal elements represent the covariance across time periods (i.e., the relationship between sensory and memory selectivity). Increasing the off-diagonal covariance increased the structure of the rotation (Supplementary Fig. 4c).

Depending on its specified selectivity, each neuron was assigned a positive firing rate response to A and X (giving it A/X selectivity) per time period. This defined the sensory response matrix (Sen_{AX}, $size = N \times 2$), and the memory response matrix (Mem_{AX}), where the first and second columns of each matrix indicate A and X selectivity, respectively. For example, an A preferring neuron (n) would be assigned a (w_s,0) sensory response in $Sen_{AX}[n, :]$. If the neuron had stable selectivity for A, its memory response in $Mem_{AX}[n, :] = (w_m,0)$. Meanwhile, an AX switching neuron would be assigned the sensory response: (w_s,0) in Sen_{AX} and the memory response: (0, w_m) in Mem_{AX}. The network contained an equal number of stable and switching neurons (N=25 of each, N=50 overall), although the model results were robust to changes in the ratio between stable and switching neurons (e.g., using a ratio of 2:1 of stable to switching neurons, similar to the experimentally observed ratio, gave qualitatively similar results). In addition, there was an equal probability of neurons preferring A or X. Given the sensory and memory selectivity matrices, the recurrent weight matrix could be determined as $W_{rr} = Mem_{AX} * inv(Sen_{Ax})$.

To vary the degree of structured rotation in the model, we varied the covariance of sensory and memory selectivity. When covariance=0, the rotation occurs by random changes in selectivity. By increasing the covariance (from 0 to 0.45 in 50 steps), we increased the number of conjunctively selective neurons, which increased the rotation's structure. All graphs show the normalized covariance as the level of structure in the rotation (e.g., Fig. 7b, x-axis: random→structured). Normalized covariance (valued between 0 and 1) is the selectivity covariance (off-diagonal), divided by the variance in selectivity (diagonal). Increasing covariance led to an increase in the conjunctive/single ratio in the network (Fig. 7b). Note, unlike when analyzing the neural data, we did not z-score our calculation of the conjunctive/single ratio, because selectivity was fixed across model runs. Given the predictions from our analytical model, we quantified the how increased structure (x) leads to an increased conjunctive proportion (y) by fitting the function: $y = \frac{A}{B - x} + C$, (fit with the scipy.optimize.curve_fit).

We did not directly fit the observed neural conjunctive/single ratio to our model, because it is not possible to disambiguate measurement noise in neural activity and differing levels of structure in the rotational dynamics (random vs. structured).

Model Input Weights: Controlling the Degree of Association Between A/X and C/C*: Associations between A/X and C/C* stimuli were built into the model through construction of the input weights (W_{ir}). All weights between the input and representational layer (W_{ir}) were drawn from the absolute value of a Gaussian distribution (i.e., zero input weights lead to non-selective neurons). We parametrically controlled the level of association between A-C and X-C* in the population by adjusting the number of neurons with combined A-C and X-C* selectivity. The association level was varied between 0 and .95. Importantly, the structure of the model allowed us to manipulate the level of association without affecting the metrics of efficiency or the rotational dynamics.

Analysis: Alignment of Axes and Prediction/Postdiction: Neural network activity was analyzed in the same way as experimental neural data, using activity from the network's representational layer (L_r). As above, we trained linear classifiers (same parameters) during both sensory and memory time periods to determine A/X sensory, C/C* sensory, and A/X memory axes. Likewise, we calculated classifier accuracy (using AUC) and the angle between them, to ensure our model recapitulated the observed angular relationships.

Also following neural analyses, we validated the relationship between alignment (AC/XC* prediction) and interference (postdiction). We measured the network's prediction by calculating the C/C* sensory classifier's accuracy in discriminating the A/X sensory input, which reflects the extent to which A/X stimulus encodes the expected stimulus C/C* (Supplementary Fig. 2b). To determine the amount of postdiction or interference, we calculated how accurately the A/X sensory classifier discriminated A/X on unexpected trials (A-C*, X-C), during the A/X memory period (Supplementary Fig. 2c). The same model structure was used to study rotational dynamics; the level of rotation was fixed to 0.31 for these simulations (although the level of rotation did not affect the prediction/postdiction results).

RNN Weight manipulation: Validation of Rotation: To test whether rotation avoids interference, we created three networks to serve as controls that had recurrent weight matrices without rotational dynamics. In the first control, non-rotating, network, we set all weights to zero ($W_{rr} = 0$); this also removes any sustained activity. The second and third control networks had diminishing and sustaining self-excitatory activity by setting the recurrent weights to $W_{rr} = \frac{1}{2}I$ and I, respectively. To test whether these networks could avoid interference, the association between A/X and C/C* sensory inputs was set to 0.95. After simulating the networks, we assessed rotation by calculating the change in A/X selectivity (z-scored firing rate difference) across sensory and memory time periods (Supplementary Fig. 7a). Networks with non-rotating weight matrices ($W_{rr} = 0$, $\frac{1}{2}I$ and I) did not rotate their A/X representations, showing that non-linear mixing of A/X sensory and C/C* responses was not sufficient to induce a rotation.

Next, we compared the A/X memory accuracy in these non-rotating networks to the memory accuracy from networks with rotation (random and structured). In addition, we calculated A/X memory accuracy over all trials and on unexpected trials (where interference is expected). Supplementary Fig 7b shows that the non-rotating networks ($W_{rr} = 0$, $\frac{1}{2}I$ and I) exhibited interference.

Metric of Efficiency: Total Selectivity—We measured the efficiency of rotational dynamics by quantifying how compactly the network stores information about the A/X stimulus. To measure this, we calculated the percentage of neurons involved in representing A/X at any time during the sequence (i.e., the 'total selectivity'). We tested the observed percent of selective neurons against expectations from a random mechanism using the permutation test described in *Nonparametric Test against the Random Mechanism* (Fig. 7d).

Using our network model, we varied the rotational structure to test its impact on efficiency. For networks with varying rotation structure, we calculated the percent total selectivity (i.e., number of A/X selective neurons / N). Based on predictions of the analytical model, we used linear regression to relate the level of structure in the rotation to the total selectivity (Fig. 7c).

Metric of Efficiency: Cityblock Distance—We measured the network's efficiency by quantifying the number of neurons that changed their preference between the sensory and memory representations. To estimate the energy of a change in representation, we used the cityblock distance (i.e., Manhattan distance or L1 norm):

$$cityblock\,(u, v) = \sum_{i=1}^{n} |u_i - v_i|$$

To measure the efficiency in the neural data's rotation, we measured the cityblock distance between the A/X sensory and A/X memory axes. Classifier weights were normalized within each animal, by dividing the weight vector by its norm, before combining weights across animals. To control for changes in number of recorded neurons across days, we divided the cityblock distance by the total number of neurons. A null distribution, representing the random mechanism, was created by permuting selectivity across neurons within the sensory and memory time periods. The null distribution was then used to calculate the z-scored cityblock distance per block across the four days of recording (Fig. 7f).

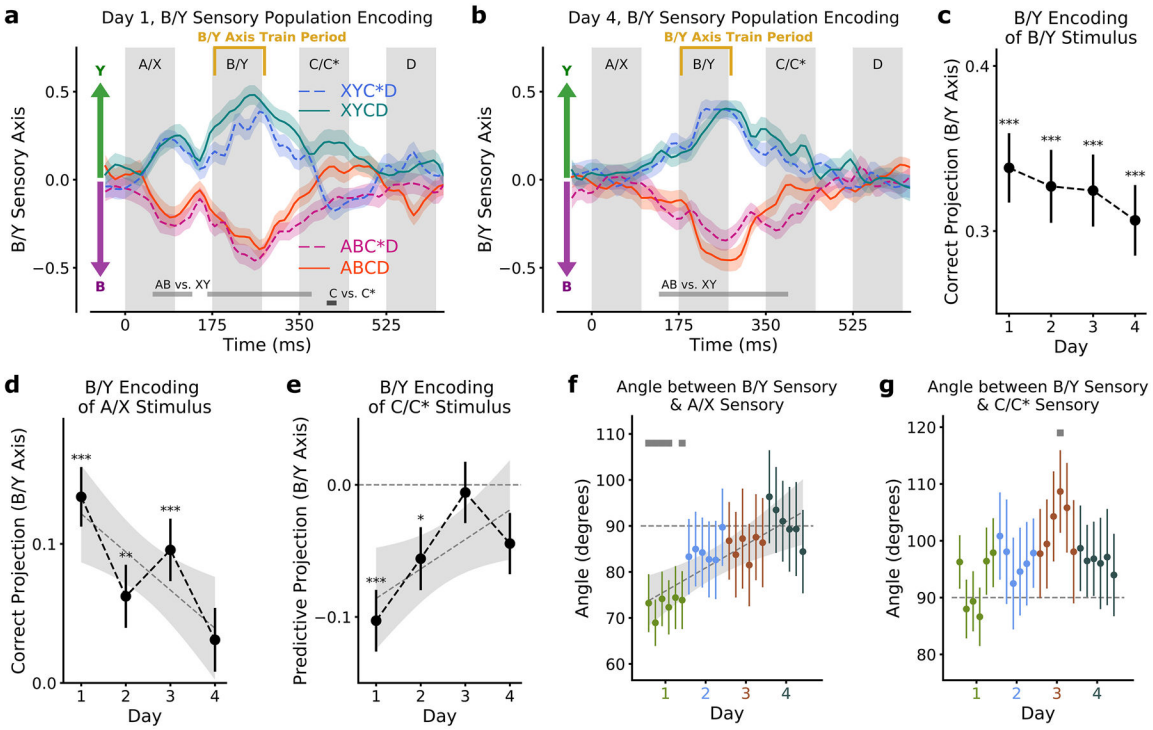
Using our network model, we tested if parametrically increasing the structure in the rotation decreased the cityblock distance between the axes, by using linear regression to relate the level of structure in the rotation to the cityblock distance between axes (Fig. 7e).

Data Availability—The data that support each main figure are included as Source Data. Original data is available upon reasonable request.

Code Availability—Code supporting the implementation and analysis of the neural network model is available on our lab GitHub repository (www.github.com/buschman-lab).

As the model was analyzed in the same way as the neural data, the same analysis code can be applied to neural data.

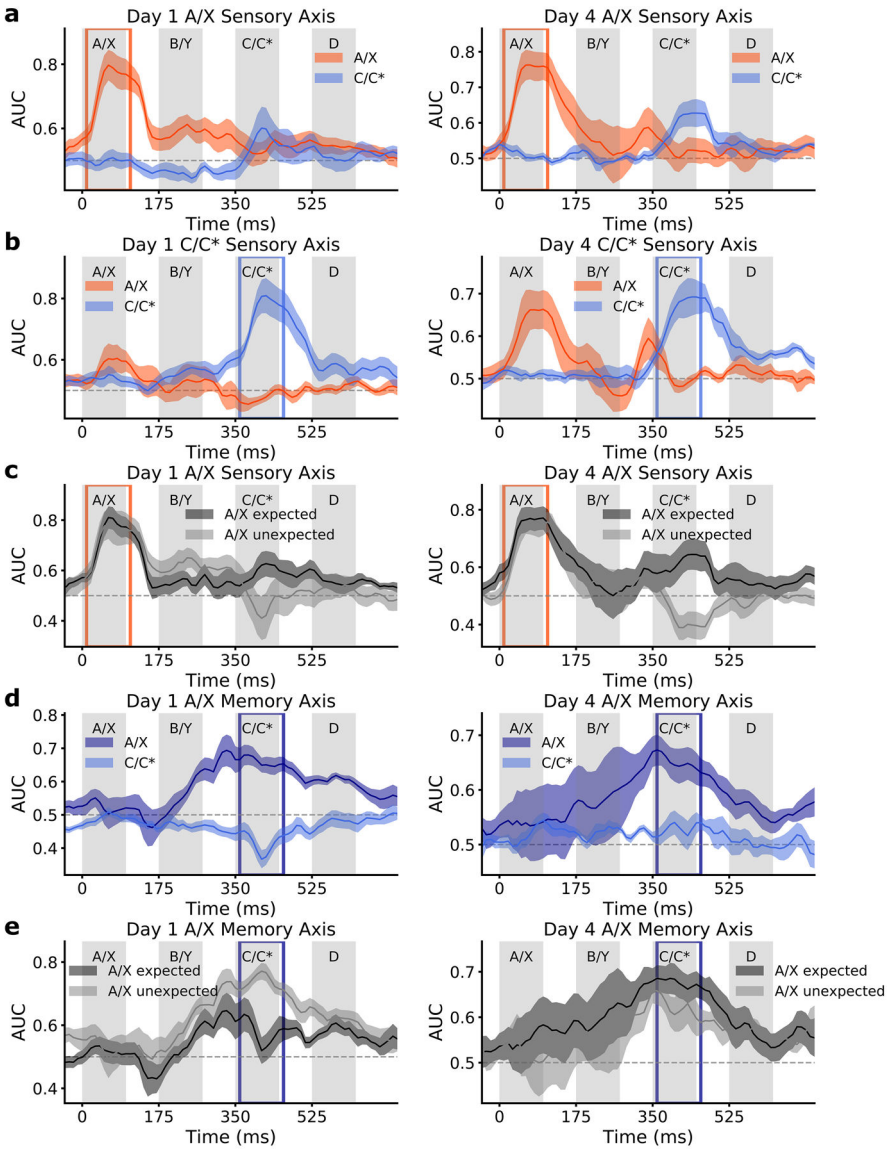
Extended Data



Extended Data Fig. 1: Encoding along the B/Y Sensory Axis.

(a) The neural population encoding of B/Y shown on **(a)** Day 1 and **(b)** Day 4. For each of the four conditions, the plot shows the mean±SEM of the population projection onto the B/Y sensory axis. Yellow outlines B/Y training period (185–285 ms). For panels a-e, n=1064 withheld trials, z-scored and then combined across animals per day. Positive and negative projections indicate Y (green) and B (purple) encoding, respectively. Light and dark grey horizontal bars mark significant differences for AB vs XY and C vs. C*, respectively (two-sided t-tests, $p \leq 0.001$, Bonferroni corrected). **(c)** Data show mean±SEM of B/Y stimulus encoding strength on the B/Y sensory axis. Negatively labeled conditions (i.e., B) were inverted, such that positive values on y-axis indicate B and Y trials are 'correctly' encoded as B and Y, respectively. Day 1 = 0.34±0.021, Day 2 = 0.33±0.22, Day 3 = 0.33±0.022, Day 4 = 0.31±0.021, all days $p<1/5000$ two-sided bootstrap tests. Slope across days mean±SEM = −0.01±0.01, p=0.16, one-sided bootstrap test. **(d)** Points show mean±SEM of A/X stimulus encoding strength on the B/Y sensory axis, during A/X stimulus presentation. For panels d-f, lines and shaded regions show mean and 95% CI of bootstrapped linear regressions. Positive values indicate correct A/X encoding: Day 1 = 0.13±0.021, $p<1/5000$, Day 2 = 0.062±0.023, p=0.0064, Day 3 = 0.096±0.022, Day 4 = 0.031±0.023, p=0.17, all two-sided bootstrap tests. Slope across days mean±SEM = −0.028±0.01, p=0.0016, one-sided bootstrap test. **(e)** Points show mean±SEM of C/C* stimulus encoding strength on the B/Y sensory axis. Positive values indicate correct encoding of C/C* association on the B/Y

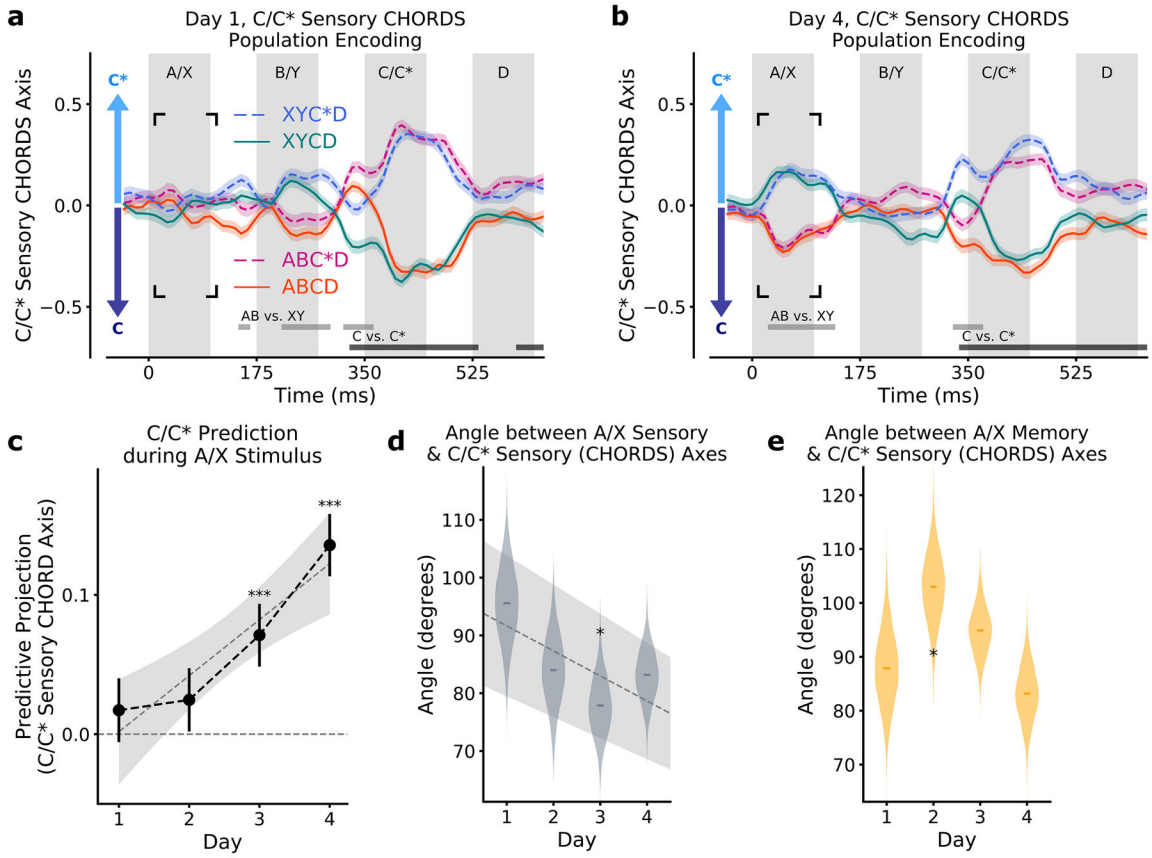
sensory axis (i.e., C and C* should go in B and Y direction, respectively). Day 1 = −0.10±0.023, p<1/5000, Day 2 = −0.056±0.024, p=0.016, Day 3 = −0.006±0.023, p=0.81, Day 4 = −0.044±0.023, p=0.055, all two-sided bootstrap tests. Slope across days mean ±SEM = 0.022±0.01, p=0.017, one-sided bootstrap test. Note, this trend does not appear in analysis of blocks of trials within a day (Supplementary Fig. 2f). **(f-g)** Points show mean ±SEM of angles between B/Y sensory axis and **(f)** A/X and **(g)** C/C* sensory axes (n=5000 resamples of neurons). Significant differences from 90 degrees shown by grey boxes (p≤0.01, one-sided bootstrap tests). Significant change in angle to A/X sensory axis over time is shown by grey line (shaded region is 95% confidence interval of bootstrapped linear regression). The B/Y and A/X sensory axes were initially aligned, but became orthogonal over days: change over days, slope = 0.84±0.24, p<1/5000, one-sided bootstrap test. B/Y and C/C* sensory axes were always orthogonal. Change over days: slope = 0.29±0.2, p=0.077, one-sided bootstrap test. For all panels, p-values: * ≤ 0.05, ** ≤ 0.01, *** ≤ 0.001.



Extended Data Fig. 2: Classifier Performance over Sequence Timecourse.

For each classifier, the accuracy (y-axis) was measured as the area under the curve (AUC; see methods). Accuracy was calculated using data from trials withheld from training (n=152 trials per animal) and was calculated in a sliding window fashion (25 ms windows, stepped 10 ms). Lines show mean±SEM of accuracy timecourses (n=7 animals). Day 1 and 4 shown in left and right panels, respectively. **(a)** A/X sensory classifier performance over time, shown for decoding A/X (orange) and C/C* (blue) stimuli. Orange rectangle indicates A/X training period (10–110 ms). **(b)** C/C* sensory classifier performance over time. Line colors follow panel a. Blue rectangle indicates C/C* training period (360–460 ms). Consistent with predictive coding shown by projections in Fig. 1g–h, on Day 4 the C/C* sensory classifier decoded A/X during the A/X stimulus and immediately before C/C*. **(c)** A/X sensory classifier performance shown for expected trials (black line - ABCD vs. XYC*D) and unexpected trials (grey line - ABC*D vs. XYCD). Consistent with postdiction results shown

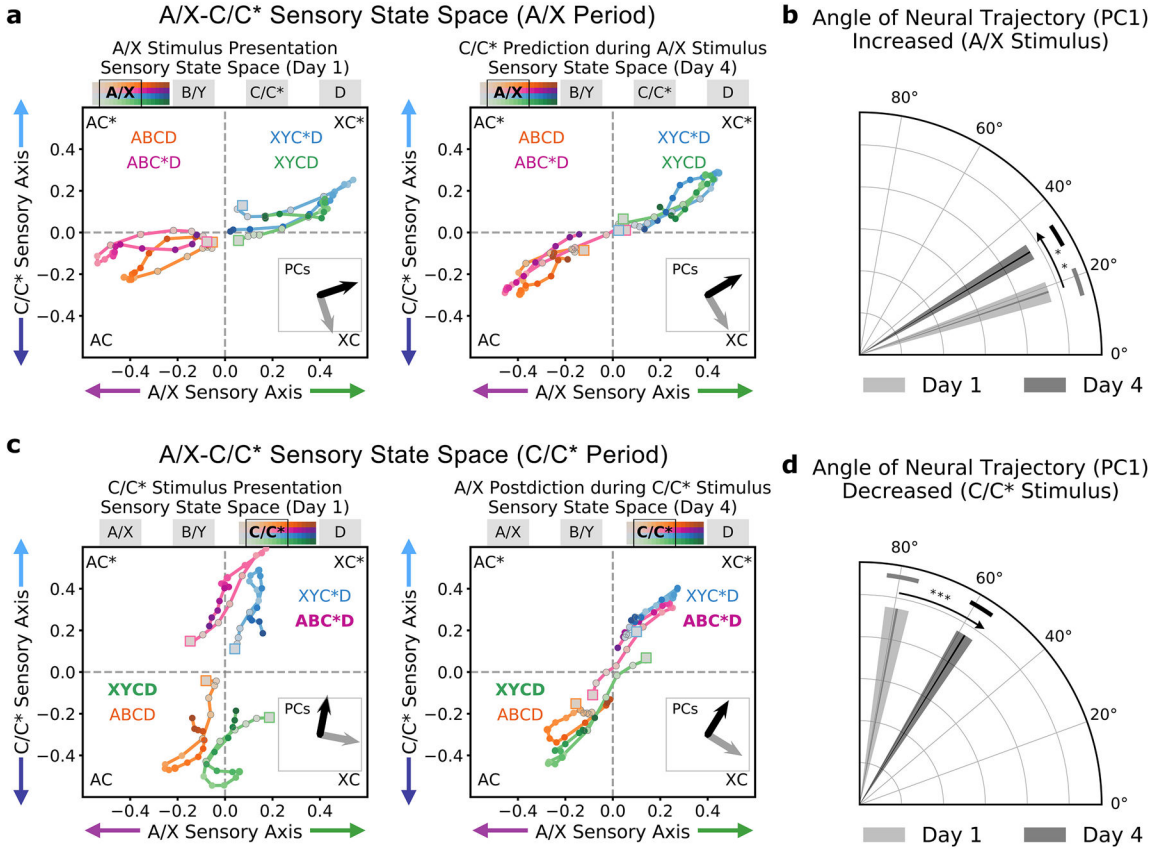
in Fig. 3b, the A/X sensory classifier performs well on expected trials, but incorrectly classifies A/X during unexpected C/C* trials. **(d)** A/X memory classifier performance over time, shown for A/X discrimination (dark blue) and C/C* (light blue). Dark blue rectangle indicates training period (360–460 ms). Consistent with projection results shown in Fig. 3c–d and Extended Data Fig. 5–6, A/X memory classifier can decode A/X near the end, but not beginning of the trial. A/X discrimination of C/C* is close to chance (AUC = 0.5), reflecting the fact that the two axes are independent. **(e)** A/X memory classifier performance, divided by expectation. Colors follow panel c. The A/X memory classifier performs well at discriminating A/X on both expected and unexpected trials.



Extended Data Fig. 3: Associative Learning Generalizes to C/C* Chords Presented Outside of Sequence.

(a-b) Lines show mean±SEM of neural activity (trials balanced across conditions) projected onto a C/C* chord encoding axis on **(a)** Day 1 (n=4204) and **(b)** Day 4 (n=4196). The C/C* chord encoding axis was trained using the firing rate response to the C/C* chord presented in isolation, outside of sequences (n=300 trials). Line colors indicate trial types (ABCD – orange, ABC*D – pink, XYCD – green, XYC*D – blue) and line style indicates 3rd chord type (C – solid, C* – dashed). Positive and negative projections indicate C* and C encoding, respectively. Light and dark grey horizontal bars mark significant differences for AB vs. XY and C vs. C*, respectively (two-sided t-test, $p \leq 0.001$, Bonferroni corrected). Results are consistent with projections onto the C/C* sensory axis (Fig. 1g–h). **(c)** Points show mean ±SEM of C/C* prediction strength during the A/X stimulus, which grew over days. Positive

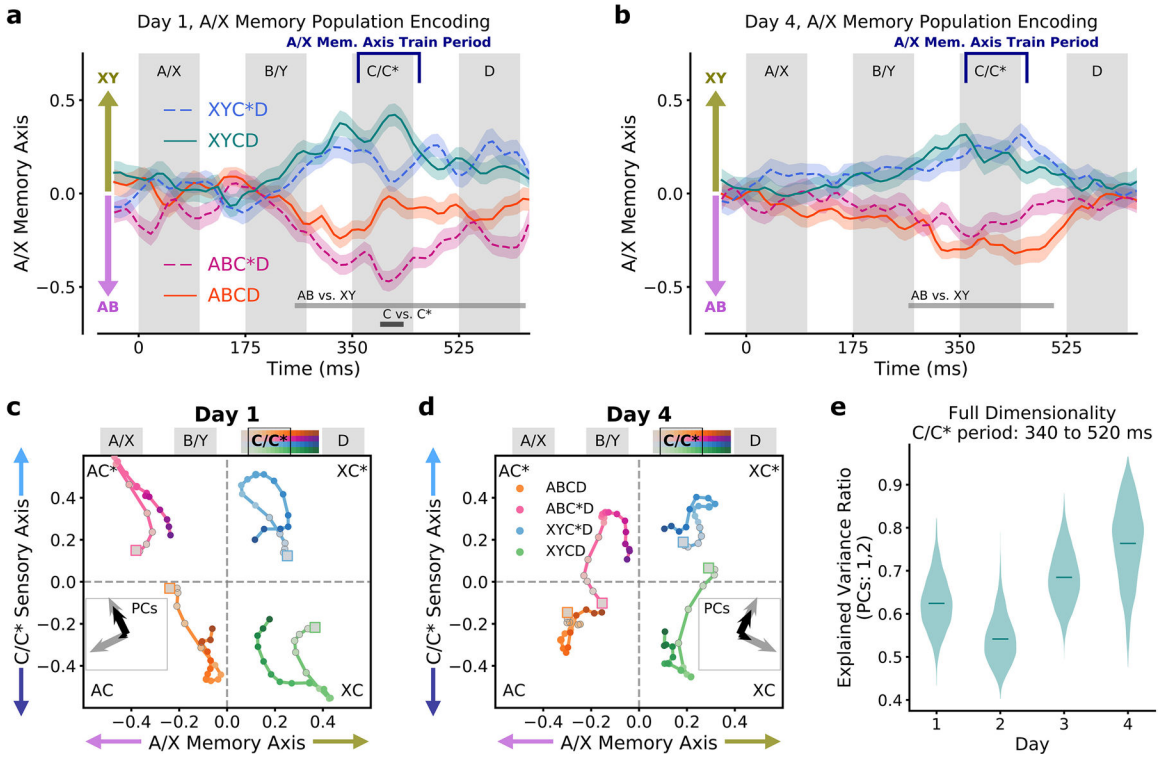
prediction (y-axis) indicates the C/C* chord sensory axis correctly encoded the association (i.e., C during A and C* during X) during the A/X stimulus (black outline in panels a-b, 10–110 ms). Day 1 = 0.017±0.023, p=0.44, Day 2 = 0.025±0.023, p=0.27, Day 3 = 0.071±0.023, p=0.0008, Day 4 = 0.14±0.022, p<1/5000, two-sided bootstrap tests. Trials used in other projection analyses were also used here (n=1064). For panels c-d, lines and shaded region show mean and 95% CI of bootstrapped linear regressions. Consistent with Fig. 1i, the prediction along the C/C* sensory chord axis increased across days; slope mean ±SEM = 0.04±0.01, p<1/5000, one-sided bootstrap test. **(d)** Violin plots show bootstrapped distributions of the angle between A/X sensory and C/C* chord sensory axes (n=5000 resamples of neurons). The mean±SEM angle between axes by day (degrees): Day 1 = 95±6.5, p=0.19; Day 2 = 84±5.9, p=0.16, Day 3 = 78±5.0, p=0.011, Day 4 = 83±4.4, p=0.064, one-sided bootstrap tests against 90 degrees. Regression across days: slope = −4.3±2.5, p=0.039, one-sided bootstrap test. **(e)** Angle between A/X memory and C/C* chord sensory axes. Angle (degrees) on Day 1 = 88±7.0, p=0.36; Day 2 = 103±6.3, p=0.019, Day 3 = 95±4.5, p=0.14, Day 4 = 83±5.2, p=0.10, one-sided bootstrap tests against 90 degrees. Regression across days: slope = −2.2±2.8, p=0.21, one-sided bootstrap test. For all panels, p-values: * ≤ 0.05, ** ≤ 0.01, *** ≤ 0.001.



Extended Data Fig. 4: Alignment of Neural Activity in A/X-C/C* State Space.

(a) Neural activity projected into A/X-C/C* state space for Day 1 (left) and Day 4 (right). Lines show mean projections of neural activity onto the A/X sensory axis (x-axis) and C/C*

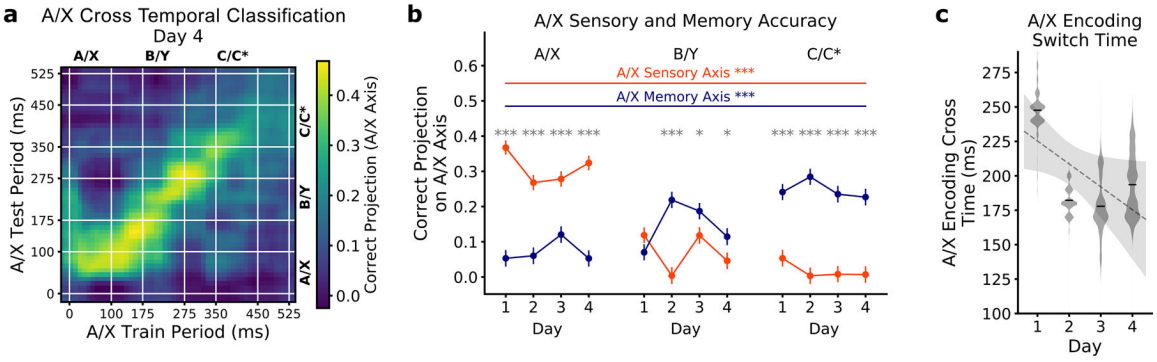
sensory axis (y-axis; n=1064 trials). Activity is shown during the A/X stimulus presentation (−10–170 ms) for each of four trial types, indicated by legend. Marker saturation increases with time, as shown in sequence timecourse legend above graph. Inset shows principal components (PCs) of neural trajectories in grey; black arrow size matches percentage of explained variance per PC (see methods). On day 1, the neural trajectory moved predominately along the A/X encoding axis (x-axis). By day 4, the neural trajectories followed an angle, encoding both A/X and the expected C/C* information (y-axis). **(b)** The angle of PC1 (relative to horizontal) during the A/X period increased across days. Radial lines show the circular mean±SEM of angle shown for Day 1 (light grey) and Day 4 (dark grey). Angle of PC1 per day (degrees): Day 1 = 18±2.9, Day 2 = 14±3.6, Day 3 = 11±4.7, Day 4 = 31±2.3 degrees (bootstrap, n=5000 resamples of neurons). Change in angle across days, slope = 3.7±2.4, p=0.0028, one-sided bootstrap test. **(c)** Neural activity during the C/C* stimulus period (340 to 520 ms) projected into A/X-C/C* state space, as in panel a. **(d)** The angle of PC1 (relative to vertical) during the C/C* period decreased across days. Format follows panel b. Angle of PC1 per day (degrees): Day 1 = 79±3.5, Day 2 = 74±3.8, Day 3 = 77±4.3, Day 4 = 58±2.7 (bootstrap); change in angle across days, slope = −6.0±1.4, p<1/5000, one-sided bootstrap test. For all panels, p-values: * ≤ 0.05, ** ≤ 0.01, *** ≤ 0.001.



Extended Data Fig. 5: A/X Memory Representation and Full Neural Dimensionality.

(a-b) The neural population encoding of A/X memory shown on **(a)** Day 1 and **(b)** Day 4. For each of the four conditions, the lines show the mean±SEM of the population projection onto the A/X memory axis (blue outlines A/X memory training period; for all panels, n=1064 withheld trials, combined across animals per day). Positive and negative projections

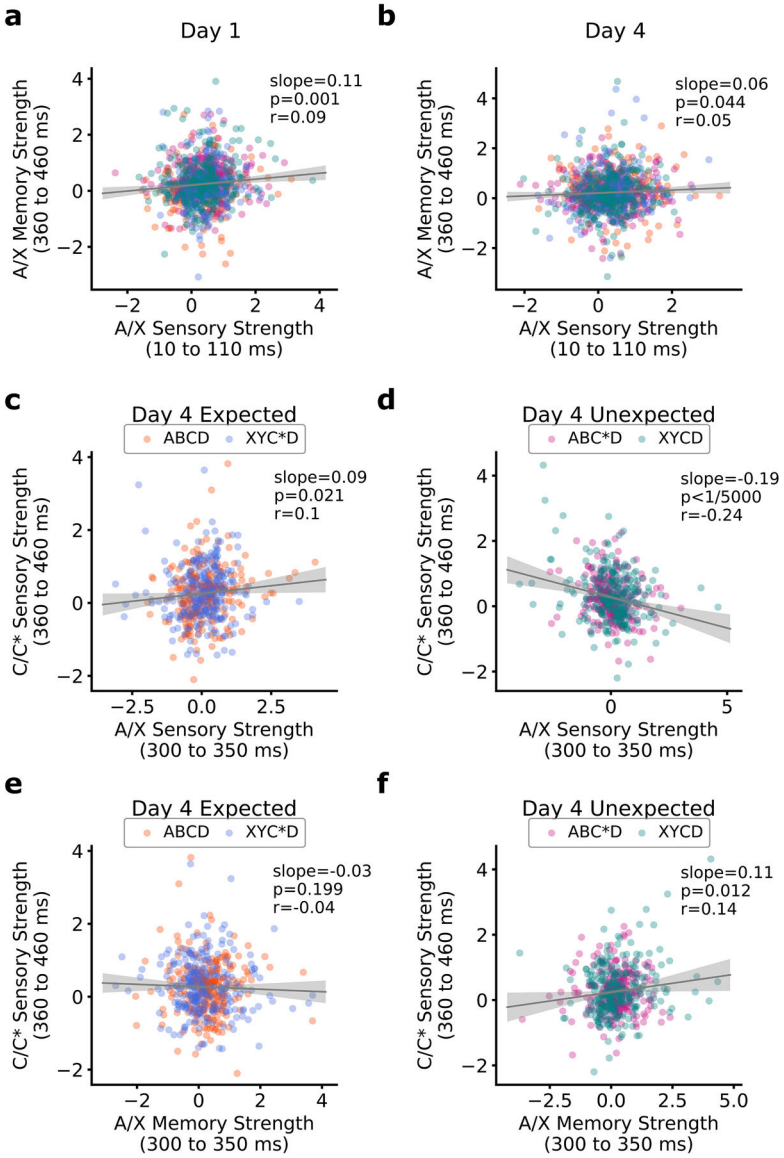
indicate XY and AB encoding, respectively. Light and dark grey bars mark significant differences for AB vs. XY and C vs. C* respectively (two-sided t-test, p≤0.001, Bonferroni corrected). **(c-d)** Neural activity projected into A/X memory - C/C* state space for **(c)** Day 1 and **(d)** Day 4, around the C/C* stimulus presentation (340–520 ms). The x-axis and y-axis are the projections of neural activity onto the A/X memory axis and the C/C* sensory axis, respectively. Marker saturation increases with time (shown across top). Inset shows PCs of neural trajectories in grey; black arrow size matches percentage of explained variance per PC (for distributions see Fig. 3h). **(e)** Violin plots show distribution of the dimensionality of the full neural response during the C/C* stimulus presentation. For each day, PCA was performed on the firing rate responses across a pseudo population (neurons were concatenated across animals; see methods). Similar to Fig. 3h, the dimensionality was estimated using the explained variance ratio (EVR) of the first two PCs (see methods). Dimensionality of the neural responses tended to decrease over days, as shown by the increased in the EVR of first two PCs: Day 1 = 0.63±0.062, Day 2 = 0.54±0.056, Day 3 = 0.68±0.064, Day 4 = 0.76±0.091 (bootstrap, n=5000 resamples of neurons). Change in EVR of first two PCs over days: slope mean±SEM = 0.053±0.034, p=0.065, one-sided bootstrap test.



Extended Data Fig. 6: A/X Sensory to Memory Transformation

(a) Cross-temporal performance of A/X classifiers. A series of A/X classifiers were trained across the sequence (x-axis; 25 ms windows, stepping by 10 ms) and then each classifier was tested across the sequence (y-axis). Color indicates the average correct projection on withheld data for all combinations of training times and test times. White bars indicated onset and offset of A/X, B/Y, C/C* stimuli. Note, the low cross-temporal decoding performance between the A/X and C/C* time periods reflects the temporal dynamics of the representation of A/X during the sequence. **(b)** Points show mean±SEM of correct projection along the A/X sensory axis (orange) and A/X memory axis (blue), during the first three stimuli in the sequence (A/X, B/Y, and C/C* columns). Positive values indicate correct encoding strength; negatively encoded conditions (i.e., A) were inverted before averaging. Horizontal bars indicate significant differences between A/X encoding during the A/X stimulus and C/C* stimulus. The A/X sensory axis had stronger A/X encoding during A/X sensory compared to the C/C* stimulus (differences per day: Day 1 = 0.31, Day 2 = 0.26, Day 3 = 0.27, Day 4 = 0.32, all p≤1/5000, one-sided permutation tests). The A/X memory axis had stronger encoded A/X encoding during the C/C* stimulus compared to the A/X stimulus (differences per day: Day 1 = −0.19, Day 2 = −0.22, Day 3 = −0.11, Day 4 = −0.17,

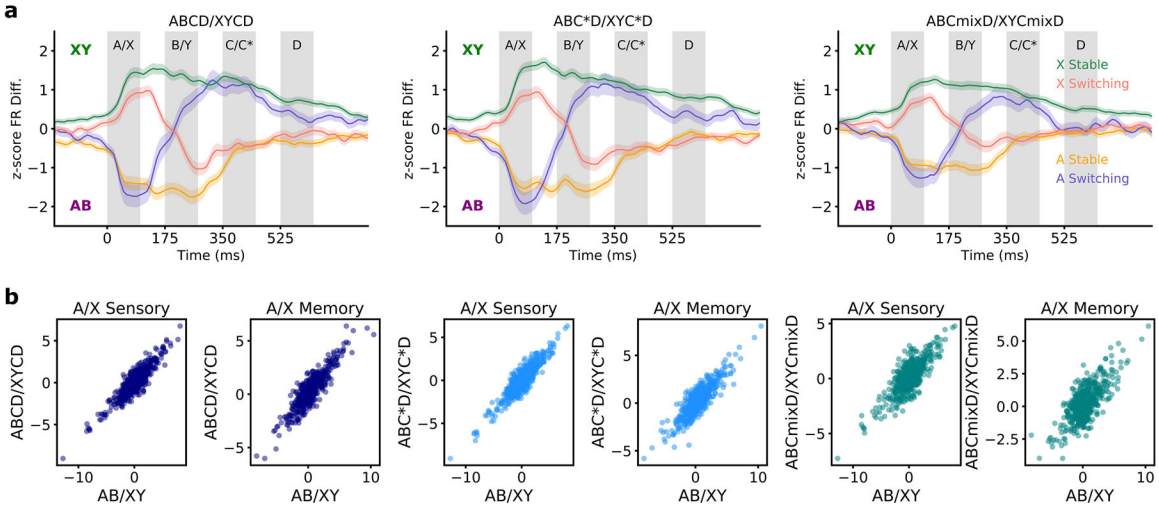
all p=0.0002, one-sided permutation tests). **(A/X Stimulus)** Projections of neural activity during the A/X stimulus (10–110 ms) onto A/X sensory axis (mean±SEM): Day 1 = 0.37±0.2, Day 2 = 0.27±0.022, Day 3 = 0.28±0.022, Day 4 = 0.33±0.021, all p<1/5000. Onto A/X memory axis: Day 1 = 0.053±0.023, p=0.022, Day 2 = 0.06±0.024, p=0.013, Day 3 = 0.12±0.023, p<1/5000, Day 4 = 0.053±0.023, p=0.019 (all two-sided bootstrap tests). During the A/X stimulus, A/X sensory encoding was stronger than A/X memory encoding on all days (Sen. – Mem. differences: Day 1 = 0.31, Day 2 = 0.21, Day 3 = 0.16, Day 4 = 0.27, all p≤1/5000, one-sided permutation tests). **(B/Y Stimulus)** Projections of neural activity during the B/Y stimulus (180–280 ms) onto the A/X sensory axis: Day 1 = 0.12±0.022, p<1/5000, Day 2 = 0.0038±0.024, p=0.87, Day 3 = 0.12±0.023, p<1/5000, Day 4 = 0.046±0.024, p=0.046. Onto A/X memory axis: Day 1 = 0.07±0.023, p=0.0044, Day 2 = 0.22±0.23, p<1/5000, Day 3 = 0.19±0.23, p<1/5000, Day 4 = 0.11±0.024, p<1/5000 (all two-sided bootstrap tests). During B/Y stimulus, A/X sensory encoding was slightly stronger than A/X memory on Day 1 (Sen. – Mem. diff. = 0.05, p=0.064), but after experience, A/X memory encoding of A/X information was significantly stronger than A/X sensory encoding (Day 2 = −0.21, p=0.0002, Day 3 = −0.07, p=0.017, Day 4 diff. = −0.07, p=0.02, all one-sided permutation tests). **(C/C* Stimulus)** Projections of neural activity during the C/C* stimulus (360–460 ms) onto A/X sensory axis: Day 1 = 0.053±0.023, p=0.021, Day 2 = 0.0034±0.023, p=0.87, Day 3 = 0.008±0.023, p=0.72, Day 4 = 0.0069±0.023, p=0.77. Onto A/X memory axis: Day 1 = 0.24±0.023, Day 2 = 0.28±0.023, Day 1 = 0.24±0.024, Day 4 = 0.23±0.024, all p<1/5000 (all two-sided bootstrap tests). During the C/C* stimulus, the A/X memory encoding was stronger than A/X sensory encoding on all days (Sen – Mem. differences: Day 1 = −0.19, Day 2 = −0.28, Day 3 = −0.23, Day 4 = −0.22, p=0.0002, all one-sided permutation tests). **(c)** Violin plots show distribution of when A/X memory encoding strength crossed A/X sensory encoding strength. Horizontal line indicates mean. Mean±SEM of switch times (ms) relative to sequence onset: Day 1 = 248±13, Day 2 = 182±10, Day 3 = 178±22, Day 4 = 194±22 (n=5000, bootstrap over trials). The switch time decreased over days (slope mean±SEM = −16±8.2, p=0.02, one-sided bootstrap test). The change in switch time decreased the most between days 1 and 3 and then stabilized by day 4 (Day 4–3 diff. = 16.11±31, p=0.31, one-sided bootstrap test). For all panels, p-values: * ≤ 0.05, ** ≤ 0.01, *** ≤ 0.001.



Extended Data Fig. 7: A/X Sensory and A/X Memory Encoding have Opposite Effects on C/C* Encoding.

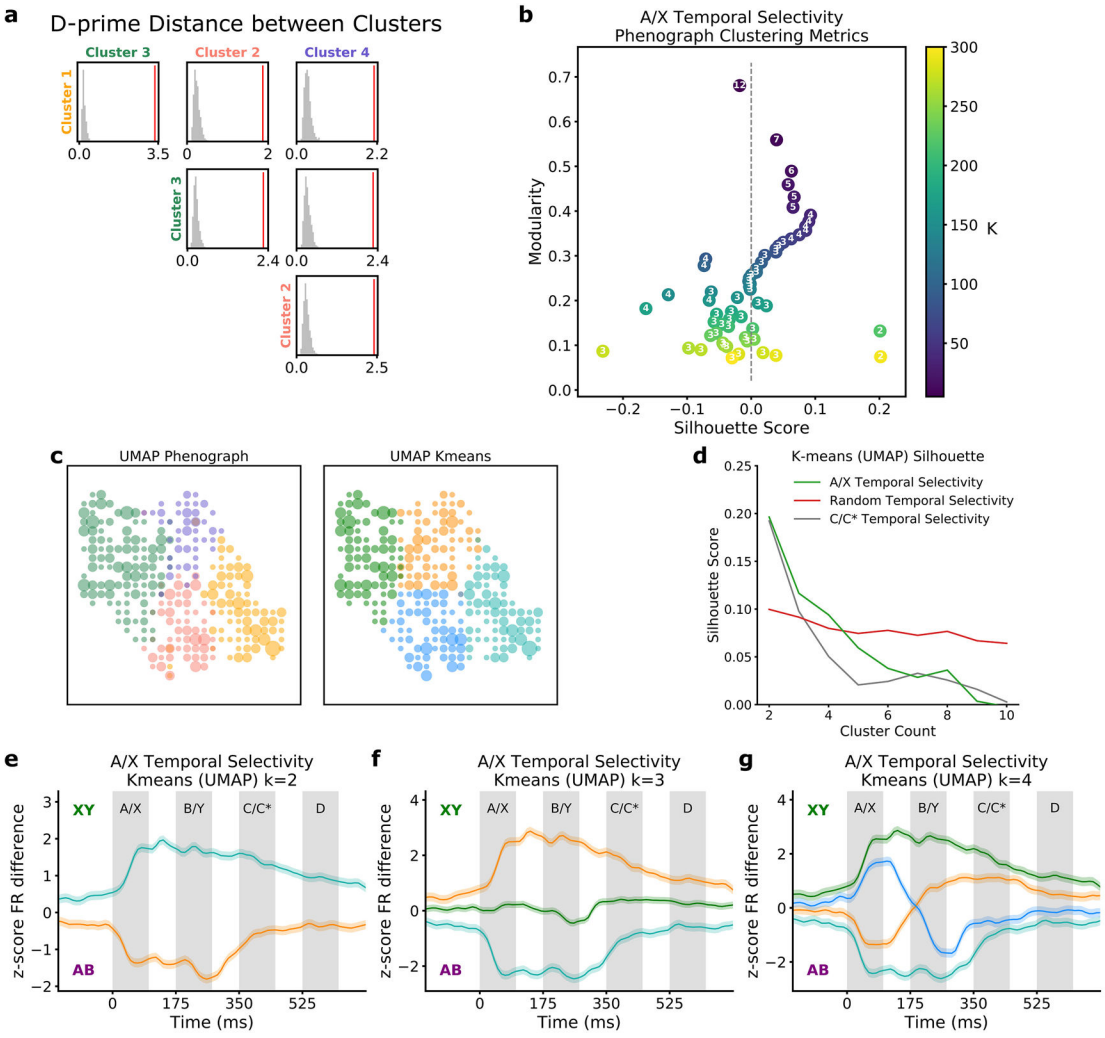
Trial-by-trial correlation of encoding strength along three relevant axes: A/X sensory, C/C* sensory, and A/X memory. Positive and negative values on each encoding axis indicate correct and incorrect projections, respectively. All lines show mean and 95% confidence interval of bootstrapped linear regressions; slope, correlation (r) and p-values (all one-sided bootstrap tests, uncorrected for multiple comparisons across panels) are listed in plots. **(a-b)** Correlation between A/X sensory encoding strength (x-axis; 10–110 ms) and A/X memory encoding strength (y-axis; 360–460 ms) on **(a)** Day 1 and **(b)** Day 4. Consistent with a transformation of A/X information from sensory to memory, there is a significant correlation on Day 1 and 4. **(c-f)** Relationship between A/X encoding strength (x-axis) and C/C* sensory encoding strength (y-axis). A/X encoding strength by the sensory and memory axes was estimated during the 50 ms prior to C/C* onset (300–350 ms). C/C* sensory encoding strength was estimated during C/C* (360–460ms). Panels show correlations between C/C*

representation and A/X sensory representation (c and d) or A/X memory representation (e and f). Correlations are shown for both expected stimuli (c and e; ABCD, XYC*D) and unexpected stimuli (d and f; ABC*D, XYCD). **(c)** On day 4, A/X sensory encoding was positively correlated with C/C* encoding accuracy on expected trials (ABCD, XYC*D). **(d)** On day 4, A/X sensory encoding was negatively correlated with C/C* encoding accuracy on unexpected trials (ABC*D, XYCD). **(e)** On day 4 there was no significant correlation between A/X memory encoding accuracy and C/C* encoding on expected trials. **(f)** On day 4, A/X memory encoding accuracy was positively correlated with C/C* encoding during unexpected trials.



Extended Data Fig. 8: Dynamics of A/X Selectivity Are Consistent across C/C* Stimuli.

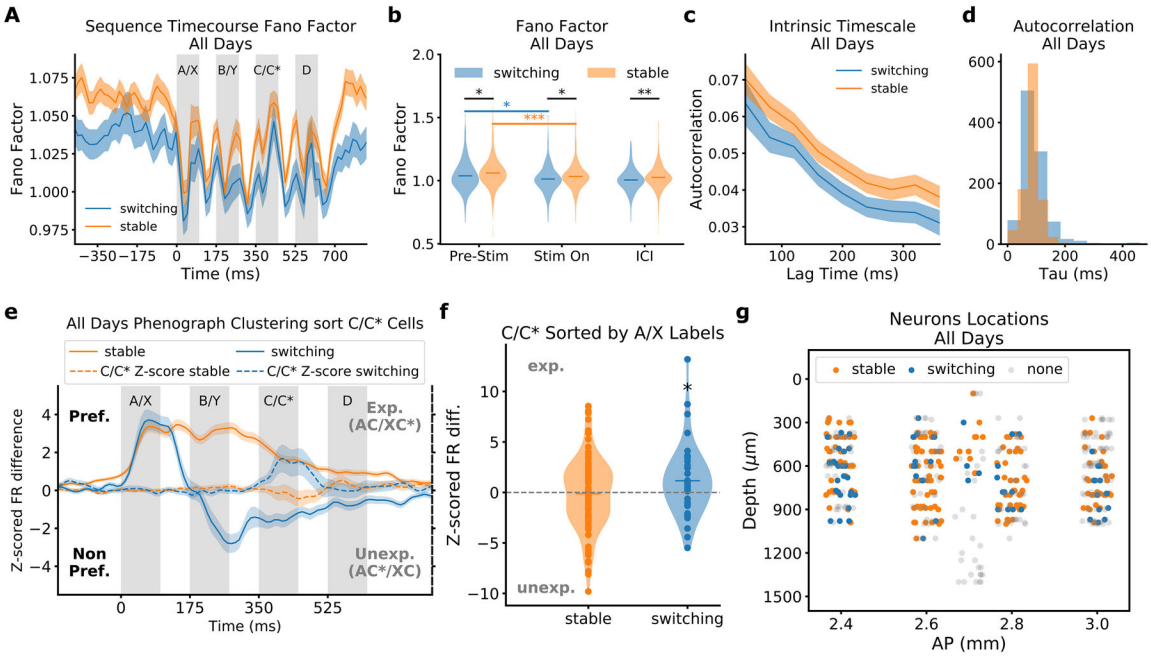
Phenograph clustering (Fig. 5a) was applied to z-scored firing rate differences calculated for specific C/C*/Cmix stimuli. **(a)** Z-scored differences were calculated for XYCD-ABCD (left), XYC*D-ABC*D (middle), and XYCmixD-ABCmixD (right) pairs of conditions. Lines show mean±SEM A/X selectivity over time per original Phenograph cluster (n=522). Note, Cmix trials involved presenting a novel stimulus that was a mix between the two chords making up C and C*; ABCmixD and XYCmixD sequences occurred on 12% of trials, randomly distributed throughout the day (see methods). **(b)** Data points show the individual neurons' original AB-XY z-scored differences (x-axis) were highly correlated with z-scored differences calculated on the 'C' trials (XYCD-ABCD), during sensory (dark blue, left; r=0.91±0.01) and memory (dark blue, right; r=0.89±0.02) time periods. Similarly, the correlation was high to 'C*' trials (XYCD-ABC*D) during sensory (blue, left; r=0.92±0.01) and memory (blue, right; r=0.87±0.02) time periods. Finally, this correlation was also seen on Cmix trials (XYCmixD-ABCmixD) during both sensory (green, left; r=0.8±0.02) and memory (green, right; r=0.73±0.02) time periods. All correlations were significant (p<1/5000, one-sided bootstrapped linear regressions, n=5000 resamples across neurons).



Extended Data Fig. 9: Stable and Switching Dynamics Capture the Temporal Dynamics of Single Neurons.

(a) Sensitivity index (d-prime) calculated between all pairs of the four Phenograph clusters (see methods). Red line shows observed d-prime; histograms show d-prime after permutation (1000 shuffles). All clusters were more separated than expected by chance (all $p \leq 0.001$, one-sided permutation tests). **(b)** Plot shows how systematically varying the number of neighbors in the Phenograph algorithm (K; color-axis) changed the goodness of clustering, as measured by the silhouette score (x-axis) and modularity (y-axis, see methods). White text shows the resulting number of identified clusters. A *k* value between 35 and 45, results in 4 clusters and high silhouette scores and modularity. Increasing the K-value beyond this recommended range leads to unstable clustering with highly variable silhouette scores and low modularity. **(c)** Density of UMAP projection of A/X temporal selectivity. Dot colors indicate Phenograph clustering (left) and K-means clustering (right, number of clusters = 4) labels. Area of circle indicates number of data points in region (max size = 8). **(d)** K-means silhouette score as a function of cluster number. K-means was performed on UMAP projections for timecourse of A/X selectivity, random selectivity, and C/C* selectivity. **(e-g)** A/X temporal selectivity profile clustered by K-means applied to

UMAP, as shown in panel c. Lines show mean±SEM of each cluster's selectivity timecourse, after each K-means run, when the number of clusters set to **(e)** k = 2, **(f)** k = 3, and **(g)** k = 4.



Extended Data Fig. 10: Properties of Stable and Switching Neurons.

(a) Intrinsic variability was higher in stable neurons. Lines show mean±SEM of fano factor of stable (orange, n=355) and switching (blue, n=167) neurons over the sequence (neurons combined days). **(b)** Violin plots show distribution of fano factor during pre-stimulus period (−400–0 ms), stimulus presentation (A/X, B/Y, C/C* and D/D* combined, 100 ms each) and inter-chord interval (ICI; 75 ms each). Fano factor was higher in stable neurons compared to switching neurons before the stimulus (stable, mean±SEM = 1.06±0.01, switching = 1.04±0.01; diff. = −0.02, p=0.02), during stimulus presentation (stable = 1.03±0.01, switching = 1.01±0.01; diff. = −0.02, p=0.016), and during the ICI (stable = 1.03±0.01, switching = 1.01±0.01; diff. = −0.02, p=0.01, all one-sided permutation tests). Difference between the pre-stimulus and stimulus periods were significant for both neuron types (stable = 0.03, p=0.001; switching = 0.03, p=0.03; one-sided permutation tests). **(c)** Line show mean±SEM of intrinsic autocorrelation of functional neuron types, calculated during the pre-stimulus period. The autocorrelation at lag zero was removed for clarity. **(d)** Histograms show distribution of time constants from autocorrelations. The time constant (tau; x-axis) provides a measure of each neuron types's intrinsic timescale; it was estimated by fitting an exponential function to the autocorrelation shown in panel c. No difference was observed between neuron types: switching mean±SEM = 94±51 ms, stable = 86±31 ms (bootstrapped exponential fit; n=1000 resamples with replacement). **(e)** Switching neurons carried slightly more of the A/X-C/C* association than stable neurons. Lines show mean±SEM of stable (orange) and switching (blue) neurons' A/X and C/C* temporal selectivity profiles. Neurons without significant C/C* selectivity were removed (stable n=123; switching, n=24, data combined across days). Selectivity of neurons is plotted with respect to their initial A/X

preference (i.e., initial selectivity is always positive). Dashed lines show C/C* selectivity of the same neurons. Responses to associated stimuli (AC/XC*) are positive, while responses to unassociated stimuli (AC*/XC) are negative. **(f)** Violin plots show distribution of average predictive selectivity during C/C* stimulus presentation (350–450 ms). Each dot is a neuron; all days included. The mean±SEM of prediction in stable neurons = −0.13±0.31, p=0.67; switching neurons = 1.16±0.53, p=0.022, two-sided bootstrap tests. **(g)** Data points show estimated locations of neurons in each functional cluster along recording array. Switching (blue), stable (orange), and none (grey) neurons are plotted according to their estimated electrode location (x-axis – AP, y-axis – depth (DV) based on implant coordinates; 6 probes had 4 shanks separated by 200 μm). Small, random jitter in anterior-posterior (AP) direction was added for clarity of presentation and does not reflect actual differences. For all panels, p-values: * ≤ 0.05, ** ≤ 0.01, *** ≤ 0.001.

Supplementary Material

Refer to Web version on PubMed Central for supplementary material.

Acknowledgements

The authors thank Camden MacDowell, Matt Panichello, Caroline Jahn, Flora Bouchacourt, Patricia Hoyos, and Sarah Henrickson for their detailed feedback during the writing of this manuscript. We also thank Brandy Briones for helping with histology and Britney Morea for helping with surgery. We thank the Princeton Laboratory Animal Resources staff for their support. This work was supported by NIMH R01MH115042, ONR N000141410681, and NIH DP2EY025446 to TJB.

References

1. Summerfield C & de Lange FP Expectation in perceptual decision making: neural and computational mechanisms. Nat. Rev. Neurosci 15, 745–756 (2014). [PubMed: 25315388]
2. Kiyonaga A, Scimeca JM, Bliss DP & Whitney D Serial Dependence across Perception, Attention, and Memory. Trends Cogn. Sci 21, 493–497 (2017). [PubMed: 28549826]
3. de Lange FP, Heilbron M & Kok P How Do Expectations Shape Perception? Trends Cogn. Sci 22, 764–779 (2018). [PubMed: 30122170]
4. Fiser A et al. Experience-dependent spatial expectations in mouse visual cortex. Nat. Neurosci 19, 1658–1664 (2016). [PubMed: 27618309]
5. Jaramillo S & Zador AM The auditory cortex mediates the perceptual effects of acoustic temporal expectation. Nat. Neurosci 14, 246–251 (2011). [PubMed: 21170056]
6. Chun MM & Jiang Y Contextual Cueing: Implicit Learning and Memory of Visual Context Guides Spatial Attention. Cognit. Psychol 36, 28–71 (1998). [PubMed: 9679076]
7. Dehaene S, Meyniel F, Wacongne C, Wang L & Pallier C The Neural Representation of Sequences: From Transition Probabilities to Algebraic Patterns and Linguistic Trees. Neuron 88, 2–19 (2015). [PubMed: 26447569]
8. Carandini M & Heeger DJ Normalization as a canonical neural computation. Nat. Rev. Neurosci 13, 51–62 (2012).
9. Buschman TJ, Siegel M, Roy JE & Miller EK Neural substrates of cognitive capacity limitations. Proc. Natl. Acad. Sci 108, 11252–11255 (2011). [PubMed: 21690375]
10. Sprague TC, Ester EF & Serences JT Reconstructions of Information in Visual Spatial Working Memory Degrade with Memory Load. Curr. Biol 24, 2174–2180 (2014). [PubMed: 25201683]
11. Bouchacourt F & Buschman TJ A Flexible Model of Working Memory. Neuron 103, 147–160.e8 (2019). [PubMed: 31103359]

12. White OL, Lee DD & Sompolinsky H Short-Term Memory in Orthogonal Neural Networks. Phys. Rev. Lett 92, 148102 (2004). [PubMed: 15089576]
13. Botvinick MM & Plaut DC Short-term memory for serial order: A recurrent neural network model. Psychol. Rev 113, 201–233 (2006). [PubMed: 16637760]
14. Rigotti M et al. The importance of mixed selectivity in complex cognitive tasks. Nature 497, 585–590 (2013). [PubMed: 23685452]
15. Sakai K & Miyashita Y Neural organization for the long-term memory of paired associates. Nature 354, 152 (1991). [PubMed: 1944594]
16. Miyashita Y & Chang HS Neuronal correlate of pictorial short-term memory in the primate temporal cortex. 3 (1988).
17. Gavornik JP & Bear MF Learned spatiotemporal sequence recognition and prediction in primary visual cortex. Nat. Neurosci 17, 732–737 (2014). [PubMed: 24657967]
18. Li N & DiCarlo JJ Unsupervised Natural Experience Rapidly Alters Invariant Object Representation in Visual Cortex. Science 321, 1502–1507 (2008). [PubMed: 18787171]
19. Maheu M, Dehaene S & Meyniel F Brain signatures of a multiscale process of sequence learning in humans. eLife 8, e41541 (2019). [PubMed: 30714904]
20. Kim R, Seitz A, Feenstra H & Shams L Testing assumptions of statistical learning: Is it long-term and implicit? Neurosci. Lett 461, 145–149 (2009). [PubMed: 19539701]
21. Yakovlev V, Fusi S, Berman E & Zohary E Inter-trial neuronal activity in inferior temporal cortex: a putative vehicle to generate long-term visual associations. Nat. Neurosci 1, 310–317 (1998). [PubMed: 10195165]
22. Griniasty M, Tsodyks MV & Amit DJ Conversion of Temporal Correlations Between Stimuli to Spatial Correlations Between Attractors. Neural Comput. 5, 1–17 (1993).
23. Amit D, Brunel N & Tsodyks M Correlations of cortical Hebbian reverberations: theory versus experiment. J. Neurosci 14, 6435–6445 (1994). [PubMed: 7965048]
24. den Ouden HEM, Friston KJ, Daw ND, McIntosh AR & Stephan KE A Dual Role for Prediction Error in Associative Learning. Cereb. Cortex 19, 1175–1185 (2009). [PubMed: 18820290]
25. Eagleman DM Motion Integration and Postdiction in Visual Awareness. Science 287, 2036–2038 (2000). [PubMed: 10720334]
26. Aru J, Tulver K & Bachmann T It's all in your head: Expectations create illusory perception in a dual-task setup. Conscious. Cogn 65, 197–208 (2018). [PubMed: 30212753]
27. Choi H & Scholl BJ Perceiving Causality after the Fact: Postdiction in the Temporal Dynamics of Causal Perception. Perception 35, 385–399 (2006). [PubMed: 16619953]
28. Fischer J & Whitney D Serial dependence in visual perception. Nat. Neurosci 17, 738–743 (2014). [PubMed: 24686785]
29. Elsayed GF, Lara AH, Kaufman MT, Churchland MM & Cunningham JP Reorganization between preparatory and movement population responses in motor cortex. Nat. Commun 7, 13239 (2016). [PubMed: 27807345]
30. Itskov PM, Vinnik E & Diamond ME Hippocampal Representation of Touch-Guided Behavior in Rats: Persistent and Independent Traces of Stimulus and Reward Location. PLoS ONE 6, e16462 (2011). [PubMed: 21305039]
31. Levine JH et al. Data-Driven Phenotypic Dissection of AML Reveals Progenitor-like Cells that Correlate with Prognosis. Cell 162, 184–197 (2015). [PubMed: 26095251]
32. Rigotti M, Ben Dayan Rubin DD, Wang X-J & Fusi S Internal Representation of Task Rules by Recurrent Dynamics: The Importance of the Diversity of Neural Responses. Front. Comput. Neurosci 4, (2010).
33. Olshausen BA & Field DJ Sparse coding of sensory inputs. Curr. Opin. Neurobiol 14, 481–487 (2004). [PubMed: 15321069]
34. Rust NC & DiCarlo JJ Balanced Increases in Selectivity and Tolerance Produce Constant Sparseness along the Ventral Visual Stream. J. Neurosci 32, 10170–10182 (2012). [PubMed: 22836252]
35. Bassett DS & Sporns O Network neuroscience. Nat. Neurosci 20, 353–364 (2017). [PubMed: 28230844]

36. Murray JD et al. Stable population coding for working memory coexists with heterogeneous neural dynamics in prefrontal cortex. Proc. Natl. Acad. Sci 114, 394–399 (2017). [PubMed: 28028221]
37. Stokes MG et al. Dynamic Coding for Cognitive Control in Prefrontal Cortex. Neuron 78, 364–375 (2013). [PubMed: 23562541]
38. Freedman DJ, Riesenhuber M, Poggio T & Miller EK Visual Categorization and the Primate Prefrontal Cortex: Neurophysiology and Behavior. J. Neurophysiol 88, 929–941 (2002). [PubMed: 12163542]
39. Fuster JM & Alexander GE Neuron Activity Related to Short-Term Memory. Science 173, 652–654 (1971). [PubMed: 4998337]
40. Warden MR & Miller EK The Representation of Multiple Objects in Prefrontal Neuronal Delay Activity. Cereb. Cortex 17, i41–i50 (2007). [PubMed: 17726003]
41. Spaak E, Watanabe K, Funahashi S & Stokes MG Stable and Dynamic Coding for Working Memory in Primate Prefrontal Cortex. J. Neurosci 37, 6503–6516 (2017). [PubMed: 28559375]
42. Miller P & Wang X-J Inhibitory control by an integral feedback signal in prefrontal cortex: A model of discrimination between sequential stimuli. Proc. Natl. Acad. Sci 103, 201–206 (2006). [PubMed: 16371469]
43. Postle BR The cognitive neuroscience of visual short-term memory. Curr. Opin. Behav. Sci 1, 40–46 (2015). [PubMed: 26516631]
44. Chaudhuri R & Fiete I Computational principles of memory. Nat. Neurosci 19, 394–403 (2016). [PubMed: 26906506]
45. Meyers EM Dynamic population coding and its relationship to working memory. J. Neurophysiol 120, 2260–2268 (2018). [PubMed: 30207866]
46. Riley MR & Constantinidis C Role of Prefrontal Persistent Activity in Working Memory. Front. Syst. Neurosci 9, (2016).
47. Perez F & Granger BE IPython: A System for Interactive Scientific Computing. Comput. Sci. Eng 9, 21–29 (2007).
48. Millman KJ & Aivazis M Python for Scientists and Engineers. Comput. Sci. Eng 13, 9–12 (2011).
49. Oliphant TE Python for Scientific Computing. Comput. Sci. Eng 9, 10–20 (2007).
50. Pedregosa F et al. Scikit-learn: Machine Learning in Python. Mach. Learn. PYTHON 6.
51. Walt S van der Colbert, S. C. & Varoquaux G The NumPy Array: A Structure for Efficient Numerical Computation. Comput. Sci. Eng 13, 22–30 (2011).
52. McKinney W Data Structures for Statistical Computing in Python. 6 (2010).
53. Hunter JD Matplotlib: A 2D Graphics Environment. Comput. Sci. Eng 9, 90–95 (2007).
54. Manly B Randomization, Bootstrap and Monte Carlo Methods in Biology (Chapman & Hall/CRC, 1997).
55. Nicosia V, Mangioni G, Carchiolo V & Malgeri M Extending the definition of modularity to directed graphs with overlapping communities. J. Stat. Mech. Theory Exp 2009, P03024 (2009).
56. Blondel VD, Guillaume J-L, Lambiotte R & Lefebvre E Fast unfolding of communities in large networks. J. Stat. Mech. Theory Exp 2008, P10008 (2008).
57. Rousseeuw PJ Silhouettes: A graphical aid to the interpretation and validation of cluster analysis. J. Comput. Appl. Math 20, 53–65 (1987).
58. McInnes L, Healy J & Melville J UMAP: Uniform Manifold Approximation and Projection for Dimension Reduction. ArXiv180203426 Cs Stat (2018).
59. Beyer K, Goldstein J, Ramakrishnan R & Shaft U When Is “Nearest Neighbor” Meaningful? in Database Theory — ICDT’99 (eds. Beeri C & Buneman P) 217–235 (Springer, 1999). doi:10.1007/3-540-49257-7_15.
60. Wasmuht DF, Spaak E, Buschman TJ, Miller EK & Stokes MG Intrinsic neuronal dynamics predict distinct functional roles during working memory. Nat. Commun 9, 3499 (2018). [PubMed: 30158572]

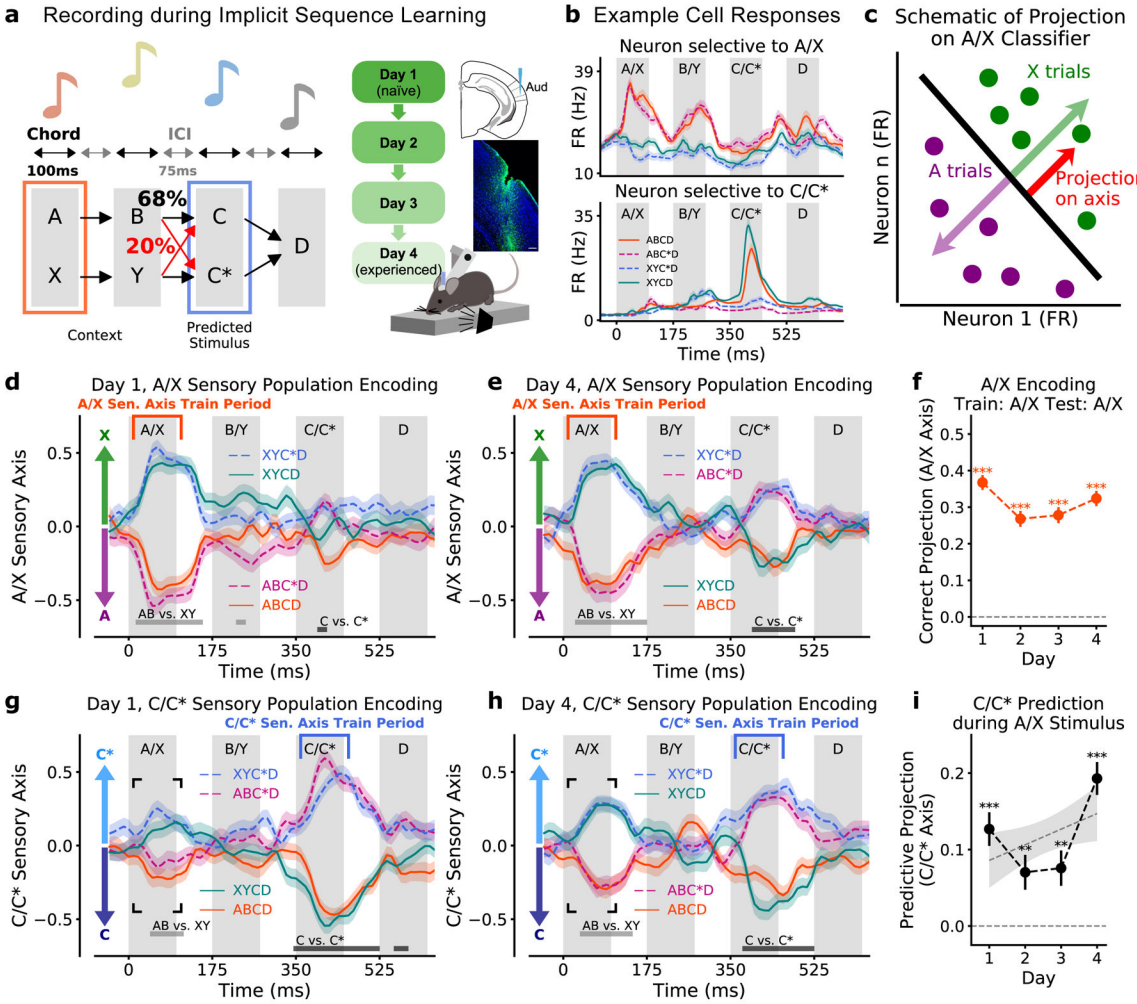


Fig. 1: Associative Learning of Sequences Facilitates Prediction.

(a) Schematic of implicit sequence learning paradigm. Beginning naïve, animals heard 1500 sequences of auditory chords every day over 4 days. Sequences had statistical regularities between chords: 68% expected trials (ABCD, XYC*D), 20% unexpected (ABC*D, XYCD), 12% mixed stimuli (see methods). Inset shows schematic of silicon probe placement in right auditory cortex. Bottom panel: example histology of electrode location (scale bar is 150 μm). Green is immunolabel for astrocytes (highlighting electrode track); blue is a Hoechst stain for cells. **(b)** Example sequence responses from two neurons preferring (top) AB context stimulus (Mouse #496, Day 2) and (bottom) C stimulus (Mouse #537, Day 2). Lines and bands show mean±SEM of firing rate. Gray patches indicate stimulus periods. Legend shows four types of sequences experienced; colors are maintained throughout the manuscript. **(c)** Schematic of classifier trained to discriminate neural responses to A/X stimulus; shown with projection of withheld response onto encoding axis. **(d and e)** The neural population encoding of A/X shown on **(d)** Day 1 and **(e)** Day 4. Lines show mean ±SEM of the population projection onto the A/X sensory axis for all four conditions. Positive and negative projections indicate X (green) and A (purple) encoding, respectively. Light and dark grey horizontal bars mark significant differences for AB vs. XY and C vs.

C*, respectively (two-sided t-test, $p \leq 0.001$, Bonferroni corrected). Orange outlines A/X training period. For panels d-i, n=1064 withheld trials, combined across animals per day. **(f)** Points show mean±SEM of encoding of A/X stimulus during stimulus presentation: Day 1 = 0.37±0.02, Day 2 = 0.27±0.022, Day 3 = 0.28±0.022, Day 4 = 0.32±0.021, all greater than zero, p<1/5000, two-sided bootstrap tests. Negatively labeled conditions (i.e., A) were inverted, such that positive values on y-axis indicate A and X trials are 'correctly' encoded as A and X, respectively. Slope mean±SEM over days = −0.012±0.009, p=0.094, one-sided bootstrap test. **(g and h)** Lines show mean±SEM of population encoding of C/C* information across the sequence timecourse on **(g)** Day 1 and **(h)** Day 4. Plots as in panels d-e. Blue outlines C/C* training period. Positive and negative projections indicate C* (light blue) and C (dark blue) encoding, respectively. **(i)** Predictive encoding of the upcoming C/C* stimulus during A/X increased with experience. Points show mean±SEM encoding of the predicted C/C* stimulus, measured as projection onto C/C* sensory axis during the A/X stimulus (black box in panels g-h). Day 1 = 0.13±0.022, p<1/5000, Day 2 = 0.07±0.023, p=0.0036, Day 3 = 0.076±0.023, p<1/5000, Day 4 = 0.19±0.022, p<1/5000, all two-sided bootstrap tests against zero. Line and shaded region show mean and 95% CI of bootstrapped linear regressions. Slope mean±SEM over days = 0.02±0.01, p=0.022, one-sided bootstrap test; see Supplemental Fig. 2b for predictive encoding during blocks of trials within days. For all panels, p-values: * ≤ 0.05, ** ≤ 0.01, *** ≤ 0.001.

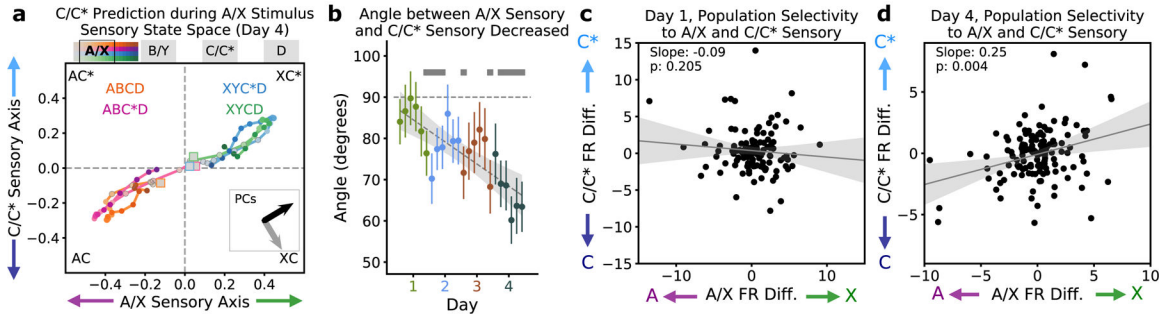


Fig. 2: Sensory Representations Align with Experience.

(a) Mean projection of neural activity onto the A/X sensory axis (x-axis) and C/C* sensory axis (y-axis) during the presentation of the A/X stimulus (−10–170 ms) on Day 4 for all four conditions (n=266 trials each, n=1064 total). The oblique response reflects prediction of the C/C* stimulus. Marker saturation increases with time (key shown along top); squares indicate timepoints before stimulus onset. Inset shows PCs of neural trajectories in grey, black arrow size matches percentage of explained variance per PC. The angle of the first PC increased with experience (Extended Data Fig. 4a–b). **(b)** The angle between A/X and C/C* sensory axes decreased across days. Points show mean±SEM of angles calculated per block of trials across 4 days (marker color indicates day; 500 trials per block, stepped by 200 trials). Top grey squares indicate significant difference from 90 degrees ($p \leq 0.01$, one-sided bootstrap test, n=5000 resamples of neurons). Line and shaded region show mean and 95% CI of linear regression of change in angle over blocks: slope mean±SEM = −0.89±0.18, $p < 1/5000$, one-sided bootstrap test. **(c and d)** Correlation between A/X selectivity (x-axis) and C/C* selectivity (y-axis) of individual neurons for **(c)** Day 1 and **(d)** Day 4. Selectivity is the z-scored firing rate difference calculated during A/X (10–175 ms) and C/C* (360–525 ms). Dots show individual neurons. Line and shaded region show mean and 95% CI of linear regression: Day 1 slope mean±SEM = −0.09±0.1, p=0.20, n=121, bootstrap test; Day 4 slope=0.25±0.087, p=0.002, n=143. Consistent with axis alignment shown in panel b, the slope relating A/X and C/C* selectivity increased over days; change in slope across days mean±SEM = 0.08±0.04, p=0.028. All one-sided bootstrap tests.

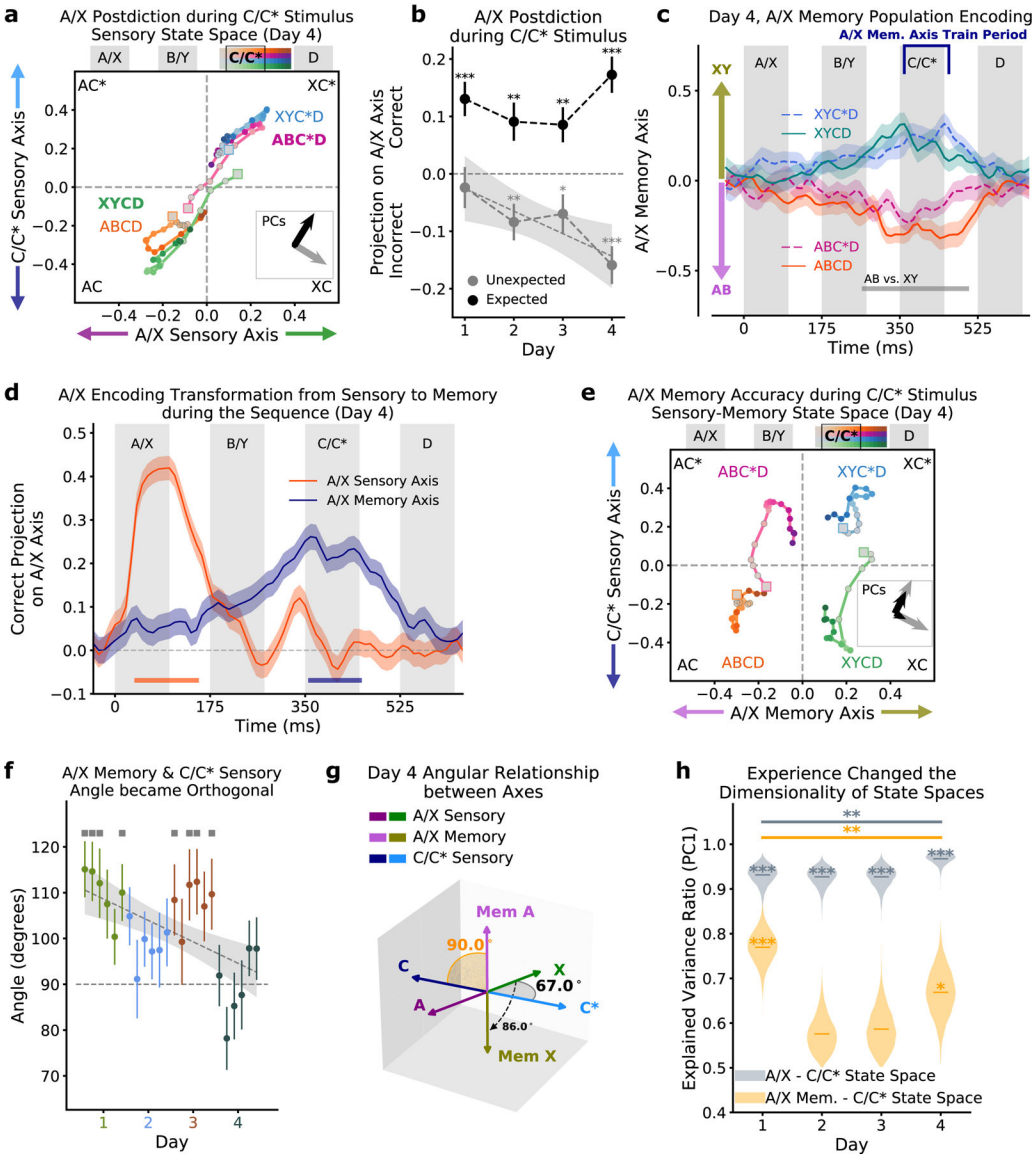


Fig. 3: Orthogonal Memory Representation Avoids Interference.

(a) Neural activity is projected into the A/X-C/C* sensory state space on Day 4 during the C/C* stimulus (340–520 ms, n=1064 trials; see Extended Data Fig. 4c for Day 1). The A/X sensory encoding (x-axis) of unexpected trials (ABC*D, pink, and XYCD, green) was incorrect after C/C* stimulus onset. Format follows Fig. 2a. **(b)** Experience increased the postdiction of A/X encoding during the C/C* stimulus (360–460 ms). Points show mean ±SEM of A/X encoding; positive and negative values indicate correct and incorrect A/X encoding, respectively. Interference on unexpected trials (grey) increased across days (n=532 trials, Day 1 mean±SEM = −0.024±0.036, p=0.48, Day 2 = −0.084±0.032, p=0.0056, Day 3 = −0.07±0.034, p=0.034, Day 4 = −0.16±0.033, p<1/5000, two-sided bootstrap tests; change over days slope mean±SEM = −0.039±0.015, p=0.0068, one-sided bootstrap test). A/X encoding on expected trials (black) remained correct over days (n=532 trials, Day 1 = 0.13±0.03, p<1/5000, Day 2 = 0.091±0.033, p=0.0052, Day 3 = 0.086±0.03, p=0.004, Day 4

= 0.17±0.031, p<1/5000, two-sided bootstrap tests; change over days slope = 0.012±0.014, p=0.19, one-sided bootstrap test). **(c)** The memory of A/X was maintained during the C/C* stimulus along an A/X memory axis on Day 4 (and Day 1, see Extended Data Fig. 5). Lines show mean±SEM of neural activity projections onto the A/X memory axis. A/X memory classifier was trained during the C/C* stimulus (360–460 ms, blue range) on preceding A/X stimulus (i.e., ABCD, ABC*D vs. XYCD, XYC*D; see methods). Positive and negative projections indicate XY (green) and AB (purple) memory encoding, respectively. Significant differences between AB and XY trials shown by horizontal bar (n=1064, p≤0.001, two-sided t-tests, Bonferroni corrected for multiple comparisons). Format as in Fig. 1d–e. Panels c-d, grey patches indicate timing of chords. **(d)** Projection of neural response onto A/X sensory axis (orange) and A/X memory axis (blue) over time. Orange/blue horizontal bars indicate stronger A/X encoding along sensory/memory axis, respectively (n=1064, p≤0.001, two-sided t-tests, Bonferroni corrected). Extended Data Fig. 6c shows when encoding switched from sensory to memory across days. **(e)** Plot shows neural activity projected into the A/X memory (x-axis) - C/C* sensory (y-axis) state space on Day 4. All four conditions are correctly encoded. Format follows panel a. **(f)** The angle between A/X memory and C/C* sensory axes became orthogonal with experience. Points show mean±SEM of angles, as in Fig. 2b. Change of angle across blocks: slope mean±SEM = −0.78±0.2, p<1/5000, one-sided bootstrap test (n=5000 resamples). **(g)** Schematic shows angles between the three axes of interest – A/X sensory, C/C* sensory and A/X memory on Day 4. **(h)** Dimensionality of state spaces during C/C* (340–520 ms) was estimated by the explained variance ratio (EVR) of the first principle component (PC1) of the neural trajectories within a given state space. Violin plots show distribution of EVR, bootstrapped across trials (n=5000 resamples). In the A/X - C/C* sensory state space (grey, shown in panel a), the bootstrapped EVR was significantly greater than chance on all four days (Day 1 = 0.93±0.02, Day 2 = 0.93±0.02, Day 3 = 0.93±0.02, and Day 4 = 0.97±0.011, all p≤1/5000 by permutation tests, see methods). EVR increased from Day 1 to 4 (horizontal bar), D4-D1 = 0.036, p=0.0054, permutation test; regression across days is trending: slope mean±SEM = 0.01±0.01, p=0.066, bootstrap test. In the A/X memory - C/C* state space (orange, shown in panel d), the bootstrapped EVR was greater than chance on Day 1 (0.77±0.034, p≤1/5000) but decreased with experience (Day 2 = 0.58±0.04, p=0.43; Day 3 = 0.59±0.04, p=0.15; Day 4 = 0.67±0.05, p=0.015; all permutation tests; D4-D1 = −0.11, p=0.011, permutation test; regression is trending: slope mean±SEM = −0.03±0.02, p=0.059, bootstrap test). The EVR of the A/X - C/C* sensory state space was significantly higher than the EVR of the A/X memory - C/C* sensory state space: difference on days 1–4: 17%, 40%, 40%, and 32%, all p≤1/5000 by permutation test. All tests in panel h are one-sided. For all panels, p-values: * ≤ 0.05, ** ≤ 0.01, *** ≤ 0.001.

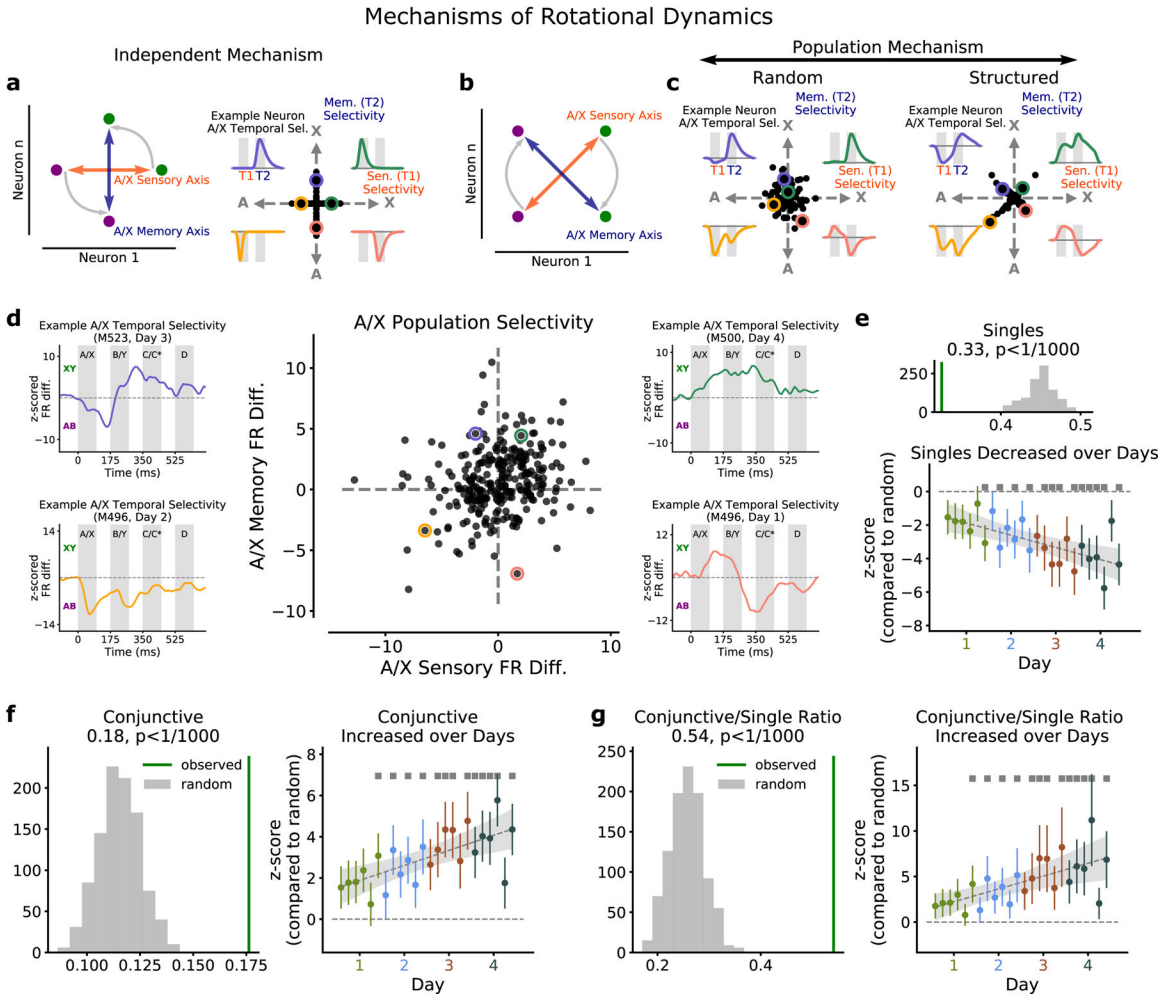


Fig. 4: Rotation Creates an Orthogonal Memory Representation.

(a-c) Schematics of mechanisms for creating orthogonal representations. **(a)** The independent mechanism predicts separate populations of neurons are selective during sensory (e.g., green and yellow) and memory (e.g., purple and pink) periods. Left plot schematizes neural responses rotating the representation; right plot shows associated neuronal selectivity distributions during the A/X stimulus presentation (x-axis, 'Sen. (T1) Selectivity') and the memory period (y-axis, 'Mem. (T2) Selectivity'), circles and insets show example timecourses of selectivity predicted by each mechanism. **(b)** Population mechanisms represent sensory and memory in a shared population of neurons. Orthogonalization occurs through a rotation induced by either **(c)** random (left) or structured (right) changes in selectivity. In contrast to random, a structured rotation predicts both positive (e.g., green and yellow) and negative (e.g., purple and pink) correlations in selectivity across time. **(d)** Center plot shows distribution of sensory and memory selectivity of individual neurons (z-scored A-X firing rate difference during A/X sensory period, x-axis, and A/X memory period, y-axis; data from all days, non-selective neurons not shown). Subplots show A/X selectivity traces from example neurons (marked in distribution). Left top: A-X switching, left bottom: A-A stable; right top: X-X stable, right bottom: X-A switching. **(e-g)** Observed proportion of **(e)** single neurons, **(f)** conjunctive neurons, and **(g)**

the ratio of conjunctive/single neurons. Conjunctive neurons are selective during both sensory and memory ($p \leq 0.025$); single neurons are selective during one time period ($p \leq 0.025$) but not the other ($p > 0.025$) All selectivity tests were by permutation (n=1000 shuffles) and Bonferroni corrected. Histograms compare the observed proportion of neuron types across all days (green line; n=522) to null distribution (grey histogram; estimated by permuting selectivity across neurons within each time period, n=1000 shuffles). Scatter plots show mean±SEM of proportion of neurons (z-scored by chance) across blocks and days. Grey squares indicate significant difference from zero ($p \leq 0.01$, bootstrap test, n=5000 resamples) **(e)** The neural population contained a lower proportion of single neurons than expected by chance (top; 0.33, $p < 1/1000$, permutation test) and decreased across blocks (bottom; slope mean±SEM = −0.12±0.03, $p < 1/5000$, bootstrap test). **(f)** The proportion of conjunctive neurons was higher than expected by chance (left; 0.18, $p < 1/1000$, permutation test) and increased across blocks (right; slope = 0.12±0.03, $p < 1/5000$, bootstrap test). **(g)** The conjunctive/single ratio was higher than expected by random chance (left; 0.54, $p < 1/1000$, permutation test) and increased across blocks (right; slope = 0.24±0.08, $p < 1/5000$, bootstrap test). All one-sided tests.

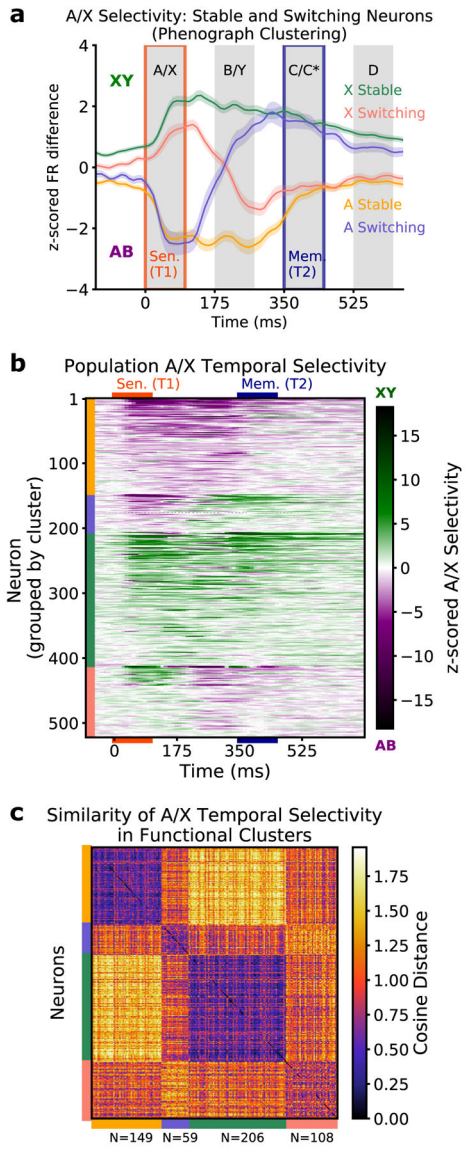


Fig. 5: Rotation Creates an Orthogonal Memory Representation.

(a) Unsupervised Phenograph clustering of the A/X temporal selectivity profiles revealed four clusters in neural population (n=522). Lines show mean±SEM A/X selectivity over time per cluster (counts per group indicated in panel c). Purple and pink are 'switching' neurons (32% of neurons, of which 35% prefer AB then XY and 65% prefer XY then AB). Yellow and green are 'stable' neurons (68% of neurons; 42% prefer AB and 58% prefer XY). **(b)** A/X selectivity over time shown across the total population of neurons (n=522), grouped by cluster (color bars along y-axis match panel a) and sorted by selectivity within group. The color axis indicates the z-scored difference in firing rate (XY-AB). 36% and 31% of neurons were significantly selective during the A/X sensory and memory periods, respectively; 50% were selective during either period. **(c)** Matrix of cosine distances between neurons' A/X temporal selectivity profiles sorted by Phenograph labels (as in panel a, colored on both axes; counts per group indicated on x-axis).

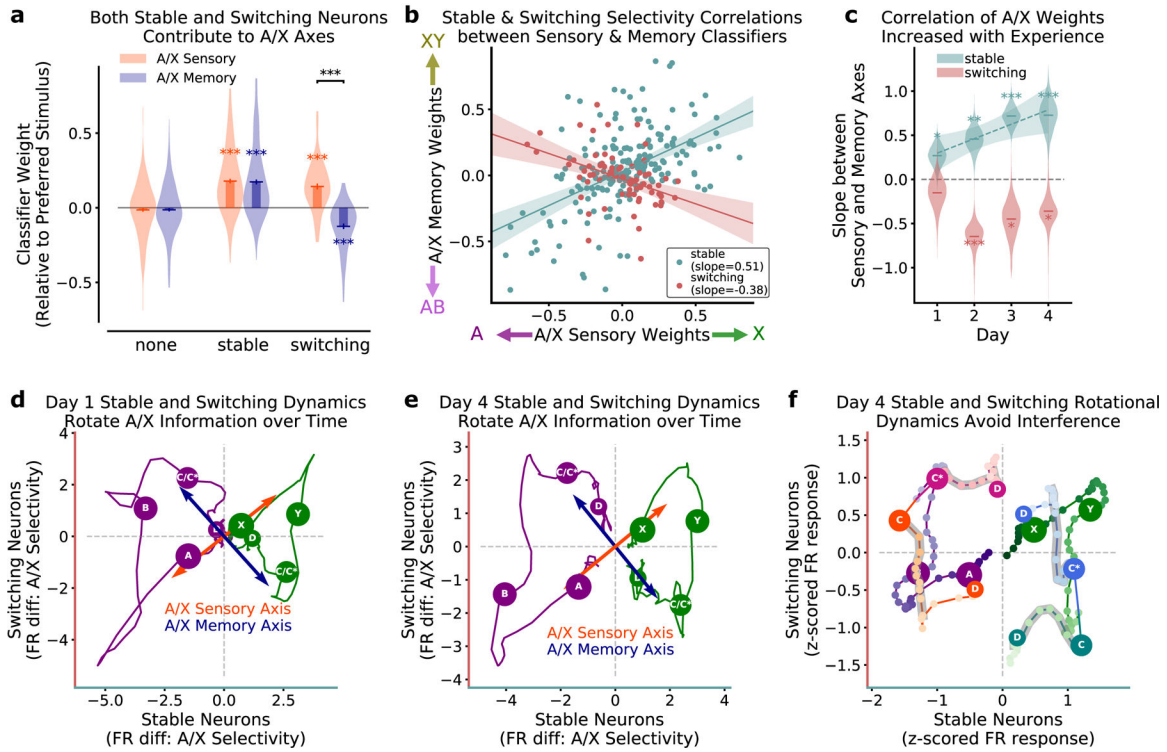


Fig. 6: Neurons with Stable and Switching Selectivity Rotate the Sensory Representation into a Memory Representation.

(a) Stable and switching neurons contributed to both A/X sensory and A/X memory axes. Violin plots show full distribution of classifier weights (mean shown by bars). Weights were re-oriented such that positive weights reflect a match to the neuron's A/X sensory preference. Stable neurons (n=209) have positive weights in both A/X classifiers (sensory axis: 0.18±0.015; memory axis: 0.17±0.16, both p<1/5000 by two-sided bootstrap test). Switching neurons (n=70) are positively weighted in the A/X sensory axis (0.14±0.02) but negatively weighted in the A/X memory axis (−0.12±0.018; both p<1/5000 by two-sided bootstrap tests). Switching neurons invert their selectivity between axes (difference in weights is 0.27, p≤1/5000, one-sided permutation test). 'None' neurons do not have significant selectivity at any time. **(b)** The selectivity (classifier weights) of stable (green) and switching (red) neurons was correlated between the A/X sensory (x-axis) and A/X memory (y-axis) classifiers. Lines show mean and 95% CI of bootstrapped linear regressions for each neuron type. Correlation was positive for stable neurons (slope mean±SEM = 0.51±0.069, p<1/5000) and negative for switching neurons (slope = −0.38±0.089, p=0.0004). Non-selective neurons (not shown) had no correlation (slope = 0.01±0.063, p=0.44). All one-sided bootstrap tests. **(c)** Experience increased the correlation between A/X sensory and A/X memory classifier weights. Violin plots show bootstrapped distribution of slope of linear regression for each neuron type, on each day (horizontal lines indicate mean). Stable slope mean±SEM on Day 1 = 0.27±0.12, p=0.013, n=60; Day 2 = 0.46±0.13, p=0.0008, n=56; Day 3 = 0.72±0.11, p<1/5000, n=43; Day 4 = 0.73±0.18, p<1/5000, n=50. Switching slope on Day 1 = −0.15±0.17, p=0.18, n=22; Day 2 = −0.64±0.16, p<1/5000, n=19; Day 3 = −0.45±0.2, p=0.015, n=10; Day 4 = −0.36±0.2, p=0.036, n=19. All one-sided bootstrap tests. Differences between stable and switching neurons' regression slopes were

significant (Day 1 = 0.41, p=0.046; Day 2 = 1.08, p=0.0008; Day 3 = 1.18, p=0.0018; Day 4 = 1.1, p=0.0004, all one-sided permutation tests). Experience increased the correlation of weights for stable neurons (slope across days = 0.16±0.07, p=0.009, one-sided bootstrap test), but not switching neurons (slope across days = −0.04±0.08, p=0.27, one-sided bootstrap test). **(d-e)** The combined activity of stable and switching neurons rotates the A/X sensory axis into an orthogonal A/X memory axis, shown for **(d)** Day 1 and **(e)** Day 4. Neurons are grouped by Phenograph labels (Fig. 5a) and initial sensory period preference (purple: A preferring neurons, green: X preferring neurons). Neurons without A/X selectivity were removed. Average z-scored firing rate differences are plotted for both stable neurons (x-axis) and switching neurons (y-axis). Circle size indicate time during sequence (larger radius: earlier). Labels indicate time period and preference (i.e., preferring A or X at 0 ms). A/X sensory (orange) and the A/X memory (blue) arrows are the average stable/switching selectivity taken during the sensory (0–100 ms) and memory (350–450 ms) periods, respectively. **(f)** Average stable and switching neuron z-scored responses to the four conditions (ABCD - orange, ABC*D - pink, XYCD - green, XYC*D - blue). Note, the response to C/C* is aligned to the sensory response to A/X, but not A/X memory, thereby avoiding interference. Only neurons with C/C* selectivity are included. Responses to A (purple) and X (green) conditions were merged prior to C/C* onset for clarity. For all panels, permutation tests used 5000 shuffles and bootstrap tests used 5000 resamples across neurons. p-values: * ≤ 0.05, ** ≤ 0.01, *** ≤ 0.001.

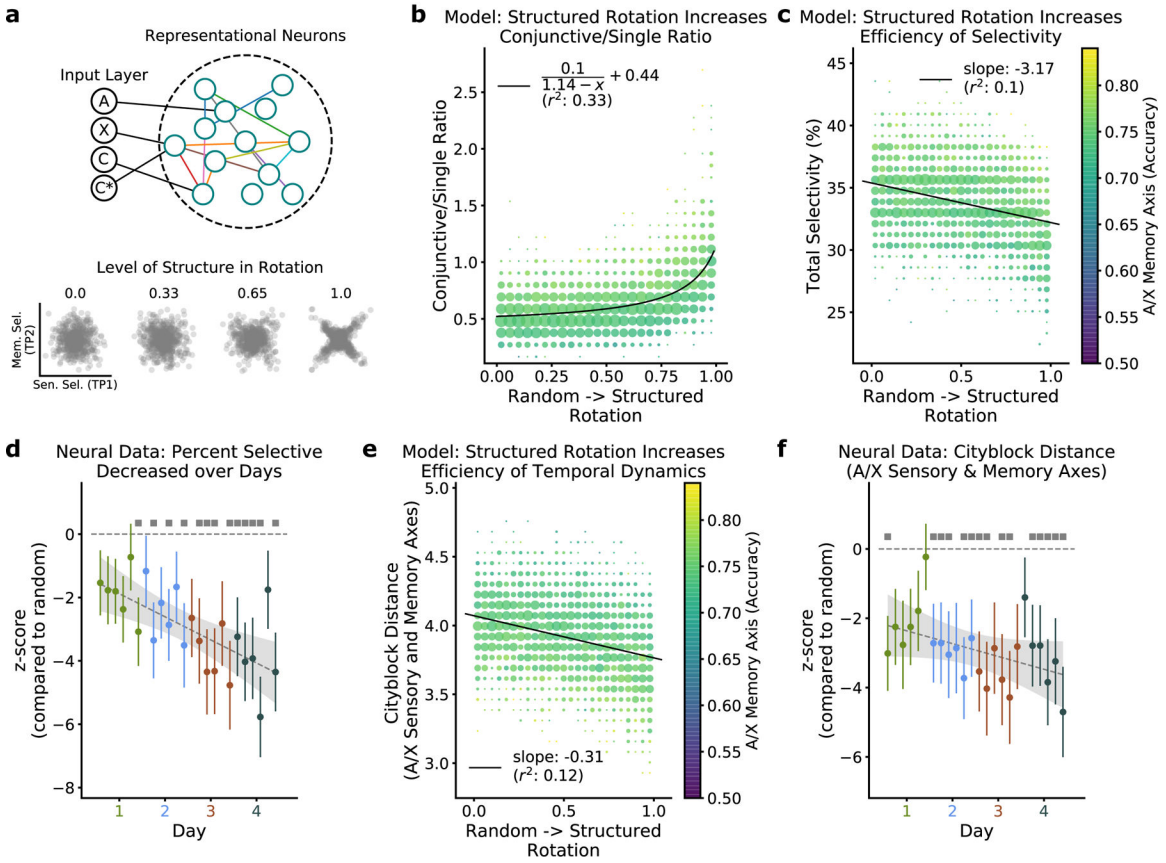


Fig. 7: Rotational Dynamics is an Efficient Mechanism for Generating Orthogonality.

(a) Schematic of neural network model. Upper panel shows network diagram. 4 input nodes (A, X, C, C*) were connected to a recurrent representational layer (n=150 neurons; see methods). Trials consisted of binary activation of stimulus input nodes over 2 time periods (e.g., AC, XC, AC* and XC* over TP1 and TP2). Responses from representational layer were used to train classifiers and calculate selectivity. Changing recurrent weights parametrically controlled the structure in the rotation. Bottom plots show four example networks with increasing levels of structure, reflected in correlation of selectivity between A/X sensory (TP1) and memory (TP2; see Supplementary Fig. 4c for full range). The number of selective neurons per time period was fixed across all networks (n=50 selective neurons). Classifier accuracy and mean angular relationships were similar across all levels of structure (Supplementary Fig. 4d–e). **(b)** In the model, increasing structure in the rotation (x-axis) increased the ratio of conjunctive/single neurons (y-axis; conjunctive neurons are selective during both TP1 and TP2, single neurons are selective during just one period; see methods). Data shown from 5000 model runs. Each dot represents a local density statistic (larger area = more model runs); dot color indicates average A/X memory accuracy (scale in c). Line shows functional fit, matching analytical predictions (equation shown in plot; see methods). **(c)** Increasing structure in the model's rotation increased the efficiency of representations: neural network models with increased structured (x-axis) required fewer selective neurons (y-axis) to achieve the same representation accuracy. Format follows panel b. **(d)** In the neural data, the percent of neurons selective during one or both time periods

was less than expected by random mechanism (0.5, p<1/1000, n=522, one-sided permutation test). Points show mean±SEM of the percent selectivity decreasing over blocks of trials (slope mean±SEM = −0.12±0.03, p<1/5000, one-sided bootstrap test). Format follows Fig. 4e–g. **(e)** Increasing structure in the model's rotation (x-axis) increased the efficiency of the transformation, measured as a decrease in the cityblock distance between sensory and memory axes (y-axis). Format follows panels b-c. **(f)** In the neural data, the A/X rotation was more efficient than expected by chance. Points show mean±SEM of the cityblock distance calculated between the A/X sensory and A/X memory and divided by total neuron count (0.21, p<1/1000, n=522, one-sided permutation test against random mechanism). Over blocks of trials, the cityblock distance decreased slightly (slope mean±SEM = −0.06±0.04, p=0.04, one-sided bootstrap test). Format follows Fig. 4e–g.