Statistical learning drives predictive shifts in object memory

Dhairyya Singh*[1], Cody V. Dong*[2], Marlie C. Tandoc[1], and Anna C. Schapiro[1]

[1] Department of Psychology, University of Pennsylvania, Philadelphia, PA, USA
[2] Department of Psychology, Princeton University, Princeton, NJ, USA
*** These authors contributed equally**
*** Authors for correspondence:** dsin@sas.upenn.edu, codydong@princeton.edu

Abstract

We are keenly attuned to the temporal relationships among the objects in our environment, allowing us to anticipate their likely futures. Do these relationships also change our memory for the objects themselves? We investigated this question in a series of statistical learning experiments in which participants viewed sequences of colored shape pairs, where the first shape predicted the second with varying strength. Despite the task being unrelated to color prediction, we found that color memory for strongly predictive shapes systematically distorted towards the color of their successors. Our biologically constrained neural network model of the hippocampus recapitulated human behavior, with predictive object representations asymmetrically distorting toward their successors and the magnitude of the shift increasing in proportion to transition strength. Together, this work shows how our memory for objects shifts toward the predictable future.

Key Words: Prediction, Statistical learning, Memory distortion, Color memory, Neural network model

Introduction

The world around us is highly regular, with objects clustering predictably in time and space (Girvan & Newman, 2002). Learning these statistics is a ubiquitous ability across the lifespan and across species, often occurring incidentally and within minutes (Frost et al., 2019; Isbilen & Christiansen, 2022; Santolin & Saffran, 2018). Investigations of statistical learning have typically involved exposing participants to repeating sequences of objects, presented across a range of sensory modalities, and assessing sequence knowledge during and after exposure. This literature has demonstrated recognition of reliable sequences, facilitation of responses to predictable items, and has delineated the conditions under which regularities are extracted, revealing much about how we learn the temporal relationships between objects in our environment. Despite this extensive literature, however, little is known about how this learning influences the way we think about the constituent objects themselves.

There is evidence for temporal regularities impacting neural object representations in neuropsychological, physiology, and neuroimaging work, with the medial temporal lobe (MTL) often implicated. Patients with hippocampal and broader MTL damage are impaired across a range of statistical learning tasks (Covington et al., 2018; Schapiro et al., 2014). Neuroimaging studies have found increased hippocampal activity in response to structured vs. random sequences (Ellis et al., 2021; Turk-Browne et al., 2009) and hippocampal anticipation of predictable objects (Sherman & Turk-Browne, 2020; Turk-Browne et al., 2010). And, critically, single neuron, local field potential, and fMRI multivariate pattern analyses have indicated that statistical learning shapes individual object representations themselves, with co-occurring objects represented more similarly to one another (Henin et al., 2021; Schapiro et al., 2012, 2016; Tacikowski et al., 2024; Wammes et al., 2022). There is evidence for a forward-looking bias in these shifts, with objects moving closer to their reliable successors in the hippocampus (Schapiro et al., 2012; Tacikowski et al., 2024). There is also evidence for predictive shifts in the responses of single cells in the rodent hippocampus, where after repeated runs along a linear track, place cells corresponding to later parts of the track fire progressively earlier (e.g. Mehta et al., 1997). It is unknown whether these neural changes correspond to changes in object memory, including whether anticipatory shifts in object memory might arise.

Several models have been proposed to account for the role of the hippocampus in learning predictive relationships. The successor representation posits that the hippocampus represents a state in terms of expected future state occupancies, weighted by the temporal proximity and likelihood of transition to the future states (Brunec & Momennejad, 2022; Dayan, 1993; Gershman, 2018; Stachenfeld et al., 2017). Experiencing predictive relationships thus leads to asymmetric representational morphing, such as that observed in rodent place cells. The Temporal Context Model of the hippocampus (Howard & Kahana, 2002) encodes the associations between sensory stimuli and a drifting temporal context, with learned context-item associations developing a predictive asymmetry and producing anticipatory recall behavior in a manner that is consistent with the successor representation (Gershman et al., 2012). Predictive coding proposes that higher regions in a sensory processing hierarchy generate predictions for the activity in lower regions (de Lange et al., 2018; Friston, 2005; Rao & Ballard, 1999), with the hippocampus's position at the apex of these hierarchies making it especially important for expressing predictive models (Barron et al., 2020; Mehta et al., 2000). Under this account, the hippocampus constantly generates predictions of upcoming sensory data for the purpose of minimizing prediction error, with these predictions unfolding in time (as in "theta sweeps") or with representational morphing as in the asymmetric expansion of place fields (Barron et al., 2020). These theories, among others, have in common a crucial role for the hippocampus in prediction, with predictive responses

associated with representational shifts, though they have not tended to map these functions to the more detailed subfield circuitry of the hippocampus, which we will undertake here.

To explore whether and how anticipatory representational shifts arise in behavior, we employed a color memory paradigm where changes in object color memory allow continuous assessment of memory distortions (Brady et al., 2018; Chanales et al., 2021; Tandoc et al., 2023; Zhao et al., 2021). We conducted three behavioral experiments in which we probed color memory over the course of statistical learning to investigate how statistical structure impacts object memory, consistently finding that object memories distort toward strongly anticipated future objects. We applied our neural network model of the hippocampus (Schapiro et al., 2017; Zhou et al., 2023) to the task in order to ascertain whether simple auto-encoding with a model incorporating biological properties of the hippocampus can support the predictive feature memory shifts found in behavior. We found that indeed our model was able to recapitulate these shifts, exhibiting anticipatory activation of representational units associated with successor objects, providing a candidate mechanism for how anticipatory shifts are learned in the subfield circuitry of the hippocampus. Together, the work shows how reliable predictive transitions in our environment distort object memory and how simple hippocampal learning mechanisms might underlie these changes.

Experiment 1

Methods

***Participants*.** We recruited 81 participants (28 female, 52 male, 1 chose not to respond; mean age = 39.68, SD = 11.11) from Amazon Mechanical Turk to participate in an online experiment. Participants who failed to meet one of the following criteria were excluded from analysis: average accuracy on the cover task lower than 90%, or cover task accuracy on any individual block lower than 70%; evidence of stepping away from the computer (taking longer than 5 minutes to respond on a color memory trial); indicating color-blindness in the post-experiment survey; and reporting taking pictures of the shapes. 64 participants remained post-exclusion. Out of these 64, 46 participants completed an additional post-test assessing explicit knowledge of regularities. Participants were provided with monetary compensation, with bonuses for speed and accuracy on the cover task. The study protocol was approved by the Institutional Review Board of the University of Pennsylvania.

***Stimuli*.** Stimuli consisted of twelve simple shapes (Fiser & Aslin, 2002) sized 256 x 256 pixels, viewed one at a time in sequences at the center of a bounding box sized 280 x 280 pixels. Embedded in the sequences were pairs of shapes in which one member (A) was predictive of the other (B) at three transition probability levels: 33% (weak), 67% (medium), and 100% (strong). Each shape was assigned a color drawn from a circular spoke arrangement in a two dimensional LAB color space slice (L = 65, centered around a=0, b=5), where paired items were 14 LAB units apart (**Fig. 2A**). Shape assignment to transition probability level as well as to pair position (A or B) was randomized for each participant. The position of the A and B pairmate shapes on the inner versus outer radius of the spokes was counterbalanced across the two pairs within each transition probability condition such that one A and one B shape from the two pairs for a given probability level was on the inner radius while one A and one B shape was on the outer radius.

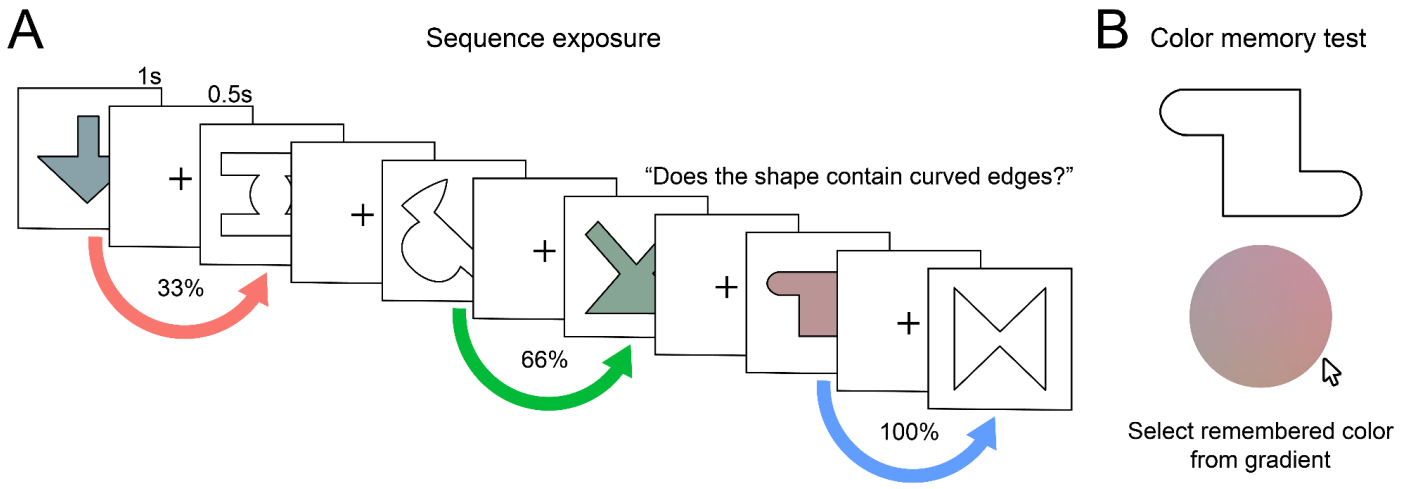


Fig. 1. Experimental design. **(A)** Participants were exposed to sequences of colored shapes constructed from temporal pairs of varying transition probability strength. Shapes were displayed with their color on half of the exposure trials. **(B)** Participants were tested on their memory for each shape's color at the end of every exposure block (10 blocks in Experiment 1, 8 in Experiments 2 and 3). They selected a color from a continuous 2D color gradient.

Procedure. Participants were exposed to 10 blocks of shape sequences. Each block had 72 trials, where each of the 12 shapes was shown six times. Stimulus duration was 1 s, with a 0.5 s interstimulus interval during which a black fixation cross was shown (**Fig. 1A**). One block lasted approximately 1.8 minutes. For half of the shape presentations, the shape was shown with its black border but no color. During paired presentations (i.e. when B appeared after A), one pairmate was always shown without color. Color was removed in this manner in order to reduce color-based segmentation of the stream of shapes, given that pair colors were nearby in color space, and to avoid any perceptual interactions for similar colors presented back to back. Shape coloring was counterbalanced such that each shape appeared in color three times per block.

Six shapes contained at least one curved edge and six contained only straight edges. A cover task required participants to quickly identify whether each presented shape contained any curved edges. Participants were informed that speed and accuracy on this task would factor into a monetary bonus they would receive at the end of the experiment. At the end of each exposure block, participants were told the bonus they had earned in that round. Speed and accuracy on this task also served as an online measure of statistical learning. The keys F and J were randomly assigned for each participant to either map onto a curved edge or straight edge response. Participants were required to respond within the 1s window that the shape was on screen. In order to familiarize participants with the task, they first completed a short 16-trial practice phase in which they were presented with colored shapes different from those encountered in the exposure blocks. Feedback was provided on the accuracy of the curved/straight response during the practice phase but not the exposure blocks.

After each exposure block, participants were tested on their color memory for each of the 12 shapes in random order. In each color memory test trial, participants were presented with an uncolored shape and were required to select their remembered color from a circular slice of continuous color space of radius 8 LAB units (**Fig. 1B**). This was a self-paced task, and participants were allowed to click around the slice and change their decision as many times as they wanted before indicating a final color choice. The circular slice was sampled around the true color of each shape. On every test trial, the slice was randomly jittered (2 lab units in either axis

of the a*b* plane) and rotated around its center in order to avoid the center of the slice being the true color of the shape. Participants were informed that performance on the color testing blocks would not affect their bonus.

After the 10 exposure and color testing blocks, 46 out of 64 participants completed a post-test in which we explicitly assessed knowledge of the pairs. For this test, participants were informed that there were regularities in the exposure stream and then tested on whether they could identify the pairs in an alternative-forced-choice task. On each trial, they were prompted with a cue shape on screen and asked to select the shape that appeared most often with the cued shape. If the cued shape was an A shape (first member of a pair), all six possible B shapes were shown as choices. Similarly, if the cued shape was a B shape, all six possible A shapes were shown as choices. Shapes in the post-test were shown without color, and trial order was random.

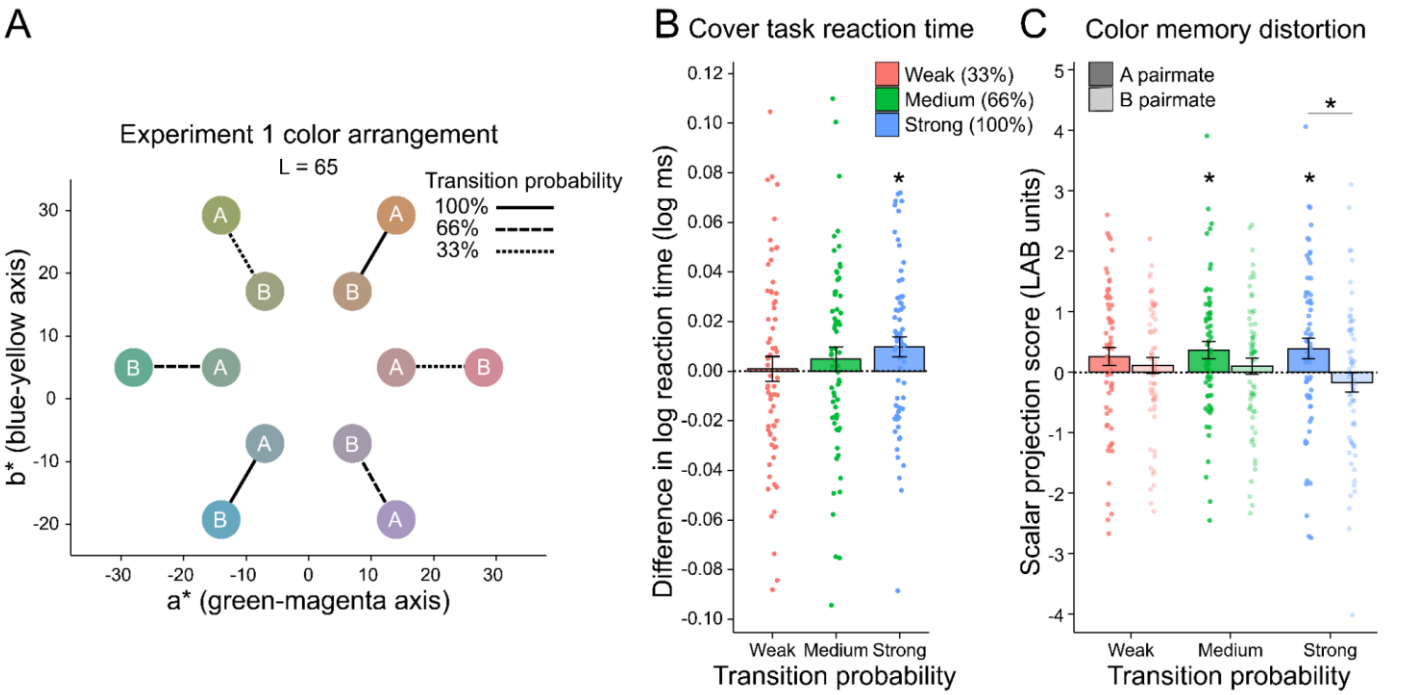


Fig. 2. Experiment 1 results. **(A)** Pair color arrangement. **(B)** Cover task reaction time difference between the predictable B shapes and predictive A shapes, with positive differences indicating faster responses on the more predictable shape. **(C)** Color memory distortions, with positive scalar projection scores indicating attraction toward the pairmate. Dots correspond to individual participants, and error bars represent +/- 1 SEM. *$p < .05$.

Results and discussion

Sensitivity to statistical structure. To examine whether participants showed sensitivity to the regularities in the exposure sequence, we compared reaction times in the cover task for the predictable (B) shapes with the predictive (A) shapes. We expected anticipation of B items given corresponding A presentations to lead to faster responses on the curved/straight classification. We restricted our analysis to paired occurrences of the shapes and only included correct trials (mean proportion trials correct = 94.8%). We log-transformed reaction times to normalize their distributions. We found that participants were significantly faster to respond to B items compared to A items (paired A - B; **Fig 2B**) in the strong condition ($t(63) = 2.38$, $p = .020$) but not in the medium ($t(63) = 1.01$, $p = .317$) or weak conditions ($t(63) = 0.180$, $p = .858$), suggesting that responses were facilitated for the most predictable objects.

For participants who completed the AFC post-test (n = 46), we assessed accuracy in identifying the correct pairmates **(Supplementary Fig. 1A)**. Because we expected skewed distributions for proportion correct measures, we used a two-tailed one-sample bootstrap test to identify whether participants had above chance (16.67%) accuracy in identifying the true pairmate. Participants reliably identified the pairmate for A shapes from strong pairs (bootstrap 10,000 replicates; $M = 0.283$, $p = .009$) but were at chance for medium and weak pairs (medium A: $M = 0.217$; weak A: $M = 0.152$; ps > .284). Participants were at chance in identifying B shape pairmates for all transition strengths (weak B: $M = 0.174$; medium B: $M = 0.185$; strong B: $M = 0.163$; ps > .699). These results suggest a potential asymmetry in explicit identification of pairmates, with significantly higher accuracy for A than B cues in the strong condition (bootstrap 10,000 replicates; $p = .038$; other conditions: ps > .609).

***Color memory*.** To evaluate whether color memory was distorted by the statistics, we calculated scalar projection scores, defined as the magnitude of the shift in color memory in the direction of a shape's pairmate along the axis in color space connecting the two shapes (**Fig 2A**). Positive scalar projection scores would indicate attraction towards the shape's pairmate, while negative scalar projection scores would signify repulsion. We excluded color memory trials in which participants took more than 15 seconds to select a color. We found that memory for the color of strong A shapes shifted toward their associated B shapes significantly more than B shapes shifted toward A shapes ($t(63) = 2.261$, $p = .027$; **Fig 2C**). This difference in memory distortion for the strong probability pairs was not due to a difference in overall error between pairmates (mean error for strong A color memory = 4.88 LAB units, mean error for strong B color memory = 4.85 LAB units; $t(63) = 0.412$, $p = .682$); rather, this difference demonstrates a bias for memory representations to be shifted in the forward temporal direction (i.e., movement from A towards B). We also assessed the directionality of the shifts for each shape individually, finding that the strong A shapes showed significant attraction towards B pairmates ($t(63) = 2.304$, $p = .024$) as did the medium A shape towards B pairmates ($t(63) = 2.587$, $p = .012$). The weak (TP = 33%) A shape attraction towards its pairmate was marginally significant ($t(63) = 1.721$, $p = .090$). However, interpreting the directionality of change for individual objects may be complicated by the arrangement of pairs in the color space, as we found that participants showed an overall bias to recall the colors of shapes as more saturated (closer to the periphery of the color space; **Supplementary Fig. 2**) and the AB pairmate axis lies parallel to the saturation axis of the colorspace. In subsequent experiments we thus decided to use a circular color arrangement (**Fig 3A, 4A**), where the within-pair axis is orthogonal to the saturation axis, allowing us to more easily interpret the directionality of the memory shifts for individual shapes.

Relationship between measures. To examine whether online sensitivity to statistical structure was associated with color memory distortions, we computed across-participant correlations between (i) response time facilitation for pairs with (ii) the degree of movement from A pairmates towards B for all three levels of transition probability. We found that none of these correlations were significant (ps > .169). The null effects could either indicate a lack of relationship between these measures or could be due to measurement noise. We discuss the relationship between memory behavior and representational change in the context of our model simulations below, where we do find a reliable relationship.

Experiment 2

In Experiment 1, the pair color arrangement was parallel to the saturation axis of the colorspace, and we found that saturation itself impacted color memory, complicating our interpretation of the directionality of the color shifts. Therefore, for Experiment 2, we placed pairs perpendicular to the saturation axis. We first ran a pre-registered experiment which included 12 colors arranged equidistantly on a circle, with two pairs in each of the strong, medium, and weak conditions (https://osf.io/xdr5e?revisionId=61a7a33cd283381bc4d13a5e; **Supplementary Fig. 3**). We did not find the basic RT facilitation effect in this experiment, which could have been due to the change in color arrangement and/or a switch in compensation to fixed course credit. Thus, for the current experiment, in order to maximize our ability to detect the statistical learning and color memory effects found in Experiment 1, we maintained the perpendicular arrangement but reverted to monetary bonus and also focused on the strong pair condition with fewer pairs, allowing more separation in color space. We pre-registered this experiment and a subset of analyses (https://osf.io/xdr5e?revisionId=61dc94998ee4ee000e9d843c).

Methods

***Design*.** Experiment 2 had the same design as Experiment 1 except that four pairs were used and all were strong (100% transition probability). The pairs were randomly arranged along a circle (**Fig 3A**) with within-pair distances of 14.5 LAB units and adjacent pairs spaced further apart, 29 LAB units away. The directionality within pairmates from A to B (clockwise or counterclockwise) was counterbalanced across pairs within each participant. The exposure stream consisted of 8 blocks of 64 shape presentations (8 presentations per shape per block), with color memory tests between each block. The explicit post-test had four alternative choices, as there were only four pairs.

***Participants*.** In Experiment 2, we recruited 158 participants from Amazon Mechanical Turk (65 female, 92 male, 1 nonbinary; mean age = 40.16, SD = 12.11). The same exclusion criteria as in Experiment 1 were applied in Experiment 2, with the addition of participants who reported not turning off color-shift programs (ex: f.lux, Redshift, blue light filters, night light setting, etc.) after 6 PM local time. 103 participants remained post-exclusion.

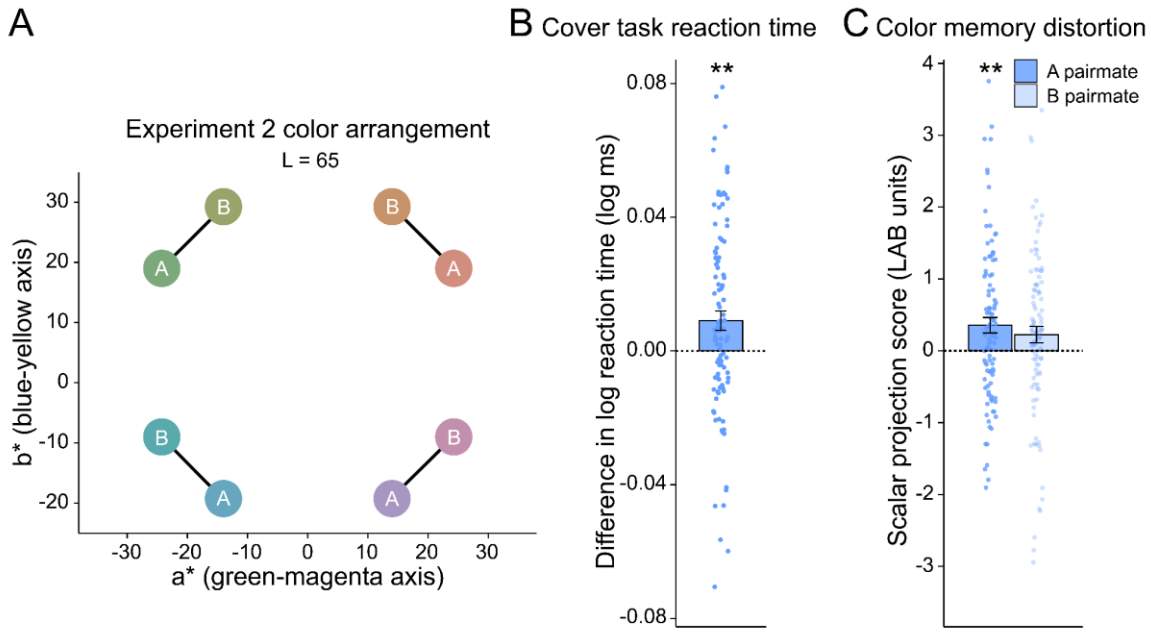


Fig. 3. Experiment 2 results. **(A)** Pair color arrangement. **(B)** Cover task reaction time difference between the predictable B shapes and predictive A shapes, with positive differences indicating faster responses for B than A. **(C)** Color memory distortions on the axis connecting pairmates, with positive scores indicating attraction toward the pairmate. **$p < .01$.

Results and discussion

Sensitivity to statistical structure. We again found that participants were faster to respond to the more predictable B shapes of the strong pairs (pre-registered; $t(102) = 3.14$, $p = .002$; **Fig 3B**). On the explicit post-test **(Supplementary Fig. 1C)**, participants were significantly above chance (25%) at identifying the true pairmates for both the A (pre-registered; bootstrap, 10,000 replicates; $M = 0.369$; $p < .001$) and B shapes (bootstrap, 10,000 replicates; $M = 0.330$; $p = .002$). Contrary to our pre-registered hypothesis, we found no significant difference between pairmates (A vs. B; bootstrap 10,000 replicates: $p = .233$).

Color memory. We found that A pairmates showed significant attraction towards the B pairmates (pre-registered; $t(102) = 3.326$, $p = .001$; **Fig 3C**). B pairmates showed marginal attraction towards A pairmates ($t(102) = 1.971$, $p = .051$), and there was no reliable difference between A and B distortion (pre-registered; $t(102) = 0.742$, $p = .46$). There was also no significant difference in the overall error magnitude between pairmates (strong A: $M = 4.795$ LAB units; strong B: $M = 4.803$ LAB units; $t(102) = -0.14$, $p = .89$). We therefore replicated the predictive memory shift we found in Experiment 1 for A shapes towards their predictable successor in the context of this new color pair arrangement.

Relationship between measures. As in Experiment 1, we did not observe a significant correlation between the degree that A pairmates were attracted towards B pairmates in color memory and faster reaction times in responding to the predictable B pairmates compared to A ($p = .909$).

Experiment 3

In Experiment 2 we replicated the predictive color memory shift observed in Experiment 1 in the context of color pairs that lay perpendicular to the saturation access. In Experiment 3 we aimed to additionally replicate the transition strength results from Experiment 1.

Methods

Design. In Experiment 3 we used the same color arrangement as Experiment 2 but replaced two strong pairs with weak pairs (**Fig 4A**).

Participants. We recruited 150 participants from Amazon Mechanical Turk (69 female, 81 male; mean age = 39.18, SD = 10.69). The same exclusion criteria from Experiment 2 were applied in Experiment 3. 103 participants remained post-exclusion.

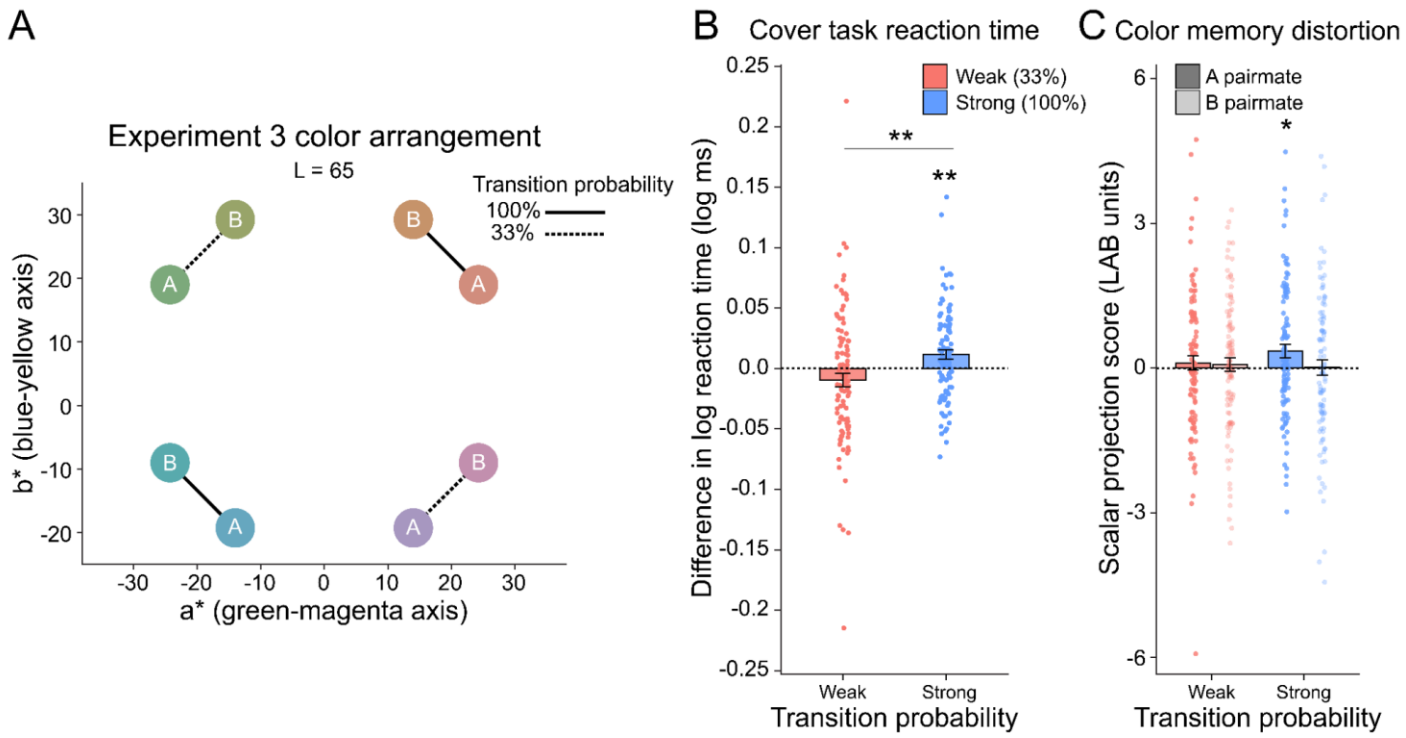


Fig. 4. Experiment 3 results. **(A)** Pair color arrangement. **(B)** Cover task reaction time difference between the predictable B shapes and predictive A shapes. **(C)** Color memory distortions on the axis connecting pairmates. *p < .05; **p < .01.

Results and discussion

Sensitivity to statistical structure. The cover task results replicated Experiments 1 and 2, with participants responding more quickly to B than A shapes (t(102) = 3.084, p = .003; **Fig 4B**). Participants were marginally slower to respond to the B shape compared to the A shape for the weak pairs (t(102) = -1.724, p = .088). The difference between strong and weak pairs was significant (t(102) = 3.230, p = .002).

For the explicit post-test **(Supplementary Fig. 1D)**, we found that participants were above chance (25%) in identifying the true pairmate for both A and B shapes for the strong pairs (bootstrap, 10,000 replicates; strong A: $M = 0.320$, $p = .025$ and strong B: $M = 0.350$, $p = .005$, respectively) and also the B shape of the weak pairs (bootstrap, 10,000 replicates; weak B: $M = 0.340$, $p = .009$), but not above chance for the A shape of the weak pairs (bootstrap, 10,000 replicates; weak A: $M = 0.301$, $p = .112$). The difference in identification accuracy (A vs. B) between pairmates was not significant in either condition (bootstrap, 10,000 replicates; ps > .378). Taken together, Experiments 2 and 3 did not replicate Experiment 1's A shape advantage in this task, suggesting that there is not a reliable asymmetry in the explicit pairmate judgments.

***Color memory*.** Once again, we found significant evidence for A shape attraction towards B in the strong transition probability condition ($t(102) = 2.564$, $p = .012$; **Fig 4C**). We did not observe this memory distortion effect for the A pairmate for the weak pairs ($t(102) = 0.731$, $p = .467$). B shapes did not show significant levels of attraction towards A pairmates in the strong or weak conditions ($t(102) = 0.103$, $p = .918$; $t(102) = 0.502$, $p = .617$ respectively). The difference in attraction between pairmates (A vs. B) was not significant in either condition (ps > .136). There were no significant differences in overall error magnitude between pairmates in either condition (ps > .184). Overall, we again replicated the predictive shift in color memory, while also finding that this shift is driven by stronger temporal relationships.

Relationship between measures. We again did not observe significant correlations between color memory distortion and reaction time (ps > .417).

Model Simulations

Across our three experiments we consistently observed that exposure to temporal statistics drives predictive shifts in memory. What neural learning mechanisms might underlie this phenomenon? There is evidence that the hippocampus is crucial both for statistical learning (Covington et al., 2018; Schapiro et al., 2014) and for predictive processing (Barron et al., 2020; Stachenfeld et al., 2017). We thus explored whether and how a biologically-grounded neural network of the hippocampus could account for our findings.

Methods

Architecture. Our model of the hippocampus, C-HORSE (Complementary Hippocampal Operations for Representing Statistics and Episodes), is implemented in the Emergent neural network simulation framework (Aisa et al., 2008). The model's architecture (**Fig 5A**) is biologically constrained by known properties of hippocampal subfield circuitry and has been previously employed to model an array of tasks that engage the hippocampus, including statistical learning, associative inference, and category learning (Heffernan et al., 2021; Schapiro et al., 2017; Singh et al., 2022; Sučević & Schapiro, 2022; Zhou et al., 2023). The model consists of input and output layers EC_in and EC_out corresponding to superficial and deep layers of the Entorhinal Cortex (EC) respectively, as well as hidden layers representing Dentate Gyrus (DG), Cornu Ammonis 3 (CA3) and Cornu Ammonis 1 (CA1). We used the version of the model from Zhou et al., (2023), which incorporates topographic organization of projections from CA3 and superficial EC to CA1 along its proximodistal axis (Sun et al., 2014; Witter et al., 2006). The full input to output path involving the canonical trisynaptic pathway (TSP) was: EC_in → DG → CA3 → proximal(p) CA1 ↔ EC_out, while the monosynaptic pathway (MSP) was EC_in

→ distal(d) CA1 ↔ EC_out. This modification allows each pathway's contributions to be separately expressed to EC_out. EC_out was connected back into EC_in, in correspondence with "big-loop" recurrence in the hippocampal circuit (Kumaran & McClelland, 2012).

EC_in and EC_out had 12 units each (one for each shape), DG 400 units, CA3 80 units, and p/dCA1 each had 50 units. Each unit could have activation values ranging from 0-1, representing a rate code. Activation values were calculated as a non-linear transformation of the sum of all inputs to each unit weighted by their connection weights (O'Reilly et al., 2024). Layer-wise feedforward feedback inhibition modulated overall activity levels in each layer, mimicking inhibitory interneuron dynamics. The TSP had sparse projections from EC_in to DG and CA3 and from DG to CA3 and high levels of inhibition (Piatti et al., 2013; Vazdarjanova & Guzowski, 2004). EC_in to DG and CA3 also had a high learning rate, corresponding to the rapid learning capabilities of the TSP (Nakashiba et al., 2008; Nakazawa et al., 2003). The MSP had full connectivity from EC_in to dCA1, and between dCA1 and EC_out. These connections had a lower learning relative to the TSP and lower inhibition levels, inspired by the more incremental learning and dense code of the MSP (Leutgeb et al., 2004; Nakashiba et al., 2008; Skaggs & McNaughton, 1998). See **Supplementary Tables 1-2** for the full list of parameters.

Sequence presentation. Sequences were generated for model training with the same structure as Experiment 1. Each of the 12 shapes was converted into a unique one-hot vector, and inputs were presented to the model as a moving window (following Schapiro et al., 2017) which included the current item at full activation strength, and the previous item at 60% of full activation strength. We used distortions to the internal representations of the model as our analog to color memory distortion, so there was no explicit simulation of shape color.

Training & Testing. The model was set up as an autoencoder: Weights were adjusted at the end of each trial via a mix of Hebbian and error-driven learning to recreate the inputs provided to the input layer (EC_in) on the output layer (EC_out). Each training trial implemented theta oscillations as described in Ketz et al. (2013) which involved two separate "minus" phases and a "plus" phase. In the first minus phase, the MSP drove responses, mimicking the trough of a theta phase, while in the second minus phase, the TSP drove responses, mimicking the peak of the theta phase. In the plus phase, the input pattern was directly clamped onto EC_out. The difference in activation states for each pathway's minus phase and the plus phase was used for local error signals.

The model was trained for 142 trials per epoch for 10 epochs (a total of twice the number of trials as human participants). To assess model representations, after each training epoch the model was tested in trials where each individual shape's one-hot vector was clamped onto the input layer with no clamping of an output vector and no weight changes allowed. The whole network's activity was recorded twice, first once activity had spread to EC_out but not yet back to EC_in ("initial response") and again after big-loop recurrence had occurred and activity in the network was stable ("settled response"). The initial response allows for cleaner measurement of each pathway's contribution, as big-loop recurrence allows activity to spread between pathways, creating a more mixed response.

Analysis. All analyses were averaged across 500 models initialized with different weights and sparse projections. For the representational similarity analysis (**Fig 5B**), at the end of model training we calculated the pairwise Pearson correlation between patterns of activity evoked by the 12 different items. This was separately computed for initial and settled responses in each hidden layer. We present DG and dCA1 representation

similarity, which are representative of the TSP and MSP representations, respectively. As an analog to the reaction time measure of statistical learning, we subtracted the activation of the A shape on EC_out given B input on EC_in from the activation B on EC_out given A on EC_in. We then subtracted the mean nonpairmate activity from this measure. The resulting measure provides an index of the asymmetry in model prediction / expectation for A-B pairs.

To isolate the role of the MSP in statistical learning, we also performed analyses in an MSP-only network. The representational asymmetry analysis (**Fig 5D**) was performed as an analogue to the color memory scalar projection analysis for the behavioral experiments. In order to characterize representational movement toward or away from the pairmate, we calculated Pearson correlations between the post-training initial response (Post) for each shape and its pairmate's pre-training (i.e. baseline) initial response (Pre) in dCA1: Corr(A_Post, B_Pre) and Corr(B_Post, A_Pre). This allowed us to measure the extent to which each shape's post-training representation had moved towards the pre-training baseline representation of the pairmate (Schapiro et al., 2012). We subtracted from this quantity the average correlation values for unpaired shapes. As another window into the development of a shared representation between pairmates, we calculated the proportion of units selective to one shape that also became selective to its pairmate over the course of training within dCA1 (**Fig 5E, 5F**). We first computed the proportion of units that each shape and its pairmate shared at baseline (i.e. units shared between A_Pre and B_Pre). A unit was classified as participating in a shape representation if its activity level surpassed 0.5 in the initial response in dCA1. We then computed the same proportional overlap measure between A_Pre and B_Post and between B_Pre and A_Post. This allowed us to assess whether there was asymmetry in the extent to which B-selective units became integrated with A's representation versus A-selective units becoming integrated with B's representation.

Results and discussion

Representation similarity. We conducted a representational similarity analysis to assess the structure of hidden layer activities for the two pathways in the model after 10 epochs of training. Figure 5B shows DG and dCA1 activity correlations, which were representative of TSP and MSP region representations, respectively. DG's initial response (before big-loop recurrence) showed weak sensitivity to the statistical relationships between shapes in the sequence, representing strong items as modestly more similar compared to medium and weak pairs, which were represented as no more similar to one another than non-pairs. In contrast, dCA1's initial response showed robust sensitivity to statistical structure, representing items more similarly as an increasing function of their co-occurrence strength. In the settled response, representations were more similar across the network, as big-loop recurrence allows dCA1's impact on EC_out to be propagated back into the network. These results show that the MSP is driving sensitivity to statistical structure in the network, which mirrors our previous simulations of statistical learning with this model (Schapiro et al., 2017), extending to a case with varying transition probability strength. Given the dominance of the MSP in representing the statistical structure, we turned off the TSP for remaining analyses in order to isolate the mechanisms underlying statistical learning in the MSP. (Including the TSP does not change any of the qualitative patterns of results).

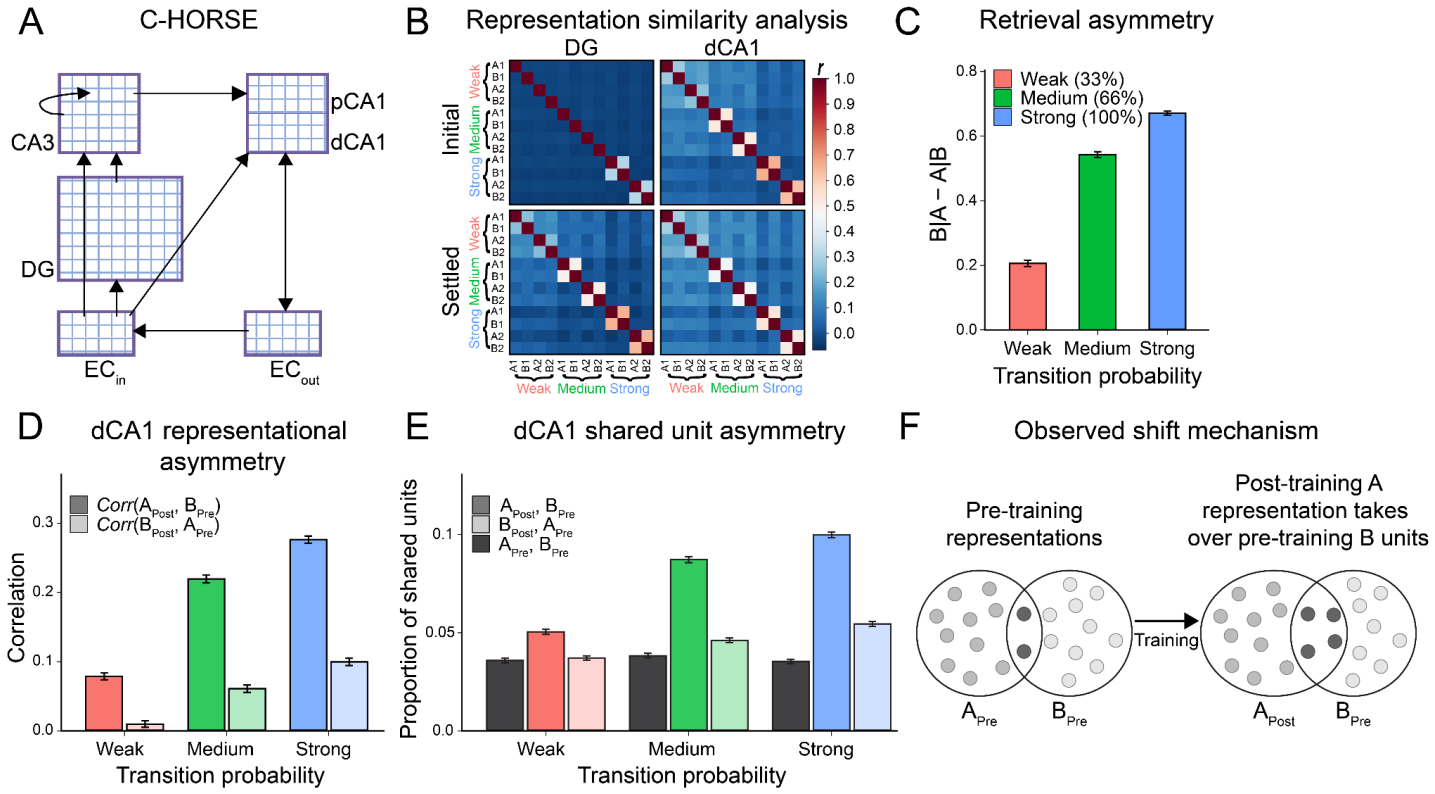


Fig. 5. Model simulations. (**A**) The architecture for C-HORSE, which mirrors the anatomy of the hippocampus. (**B**) Representation similarity analysis of initial and settled responses for DG and dCA1. (**C**) Behavioral asymmetry indicates the difference in strength of retrieval for B shapes given A inputs vs. A shapes given B inputs. (**D**) The correlation of the post-training shape representations with their pairmates' pre-training representations. (**E**) The mean proportion of units shared by the trained A and B item representations with the baseline representations of their pairmates, compared with the initial proportion of shared units. (**F**) Schematic for the asymmetric shifting of representations in our model. Units from baseline B representations become part of trained A representations over the course of learning.

Retrieval asymmetry. As an analogue to the reaction time analysis that showed that participants were anticipating predictable items, we examined the extent to which the model exhibited forward predictions in its output behavior, as measured by its tendency to retrieve B items given A inputs, relative to retrieving A items given B inputs. We found that the model's behavior indeed exhibited this forward asymmetry, with stronger effects for higher transition probabilities (**Fig 5C**).

Representational asymmetry. We assessed whether the model could account for the predictive morphing of color memory that we found in humans. We examined the extent to which the dCA1 representations of A and B items shifted over the course of training toward their pairmate's initial untrained representation. We found that there were greater shifts for A items toward their B pairmates as compared to B items toward their A pairmates across all three transition probability conditions in dCA1 (**Fig 5D**). This asymmetry suggests that over the course of training, A's representations morphed more toward B than B's did towards A, demonstrating a predictive shift akin to the human color memory results. Though lower than A, we did find small amounts of attraction for B shapes, unlike in the human data. This may be an aspect of the model that does not align with

human behavior, or it is possible that more sensitive assessments would uncover some movement for B shapes in humans.

Relationship between retrieval and representational asymmetries. We conducted a Spearman correlation between retrieval asymmetry and representational asymmetry values averaged across transition probability conditions for each model initialization and found a weak but statistically significant positive relationship ($r(498) = 0.16$, $p < .001$). While the representational asymmetry is determined mainly by weights feeding into CA1, the model's retrieval is additionally impacted by weights from CA1 to EC_out, creating the opportunity for some independence in these asymmetry measures. We did not find a relationship between memory distortion and forward behavioral facilitation in human behavior, which could be consistent with the very weak relationship predicted by the model.

Understanding representational change. We ran a discrete version of the representational shift asymmetry analysis to provide intuition for the nature of the observed representation shifts. We calculated the proportion of units each shape shares after training with the representation of its pairmate prior to training and found that A integrates a greater proportion of B's original units than vice versa (**Fig. 5E, 5F**). In other words, A comes to share some of the representational real estate initially assigned only to B.

Post-training feature transfer. Yu & Zhao (2018a) demonstrated an asymmetric feature transfer after statistical learning which on the surface seems opposite to our empirical and simulation results. Participants were exposed to sequences with embedded shape (Experiments 1-3) or face pairs (Experiments 4 & 5) and then briefly presented with predictive A objects with one feature updated, and immediately queried on their memory for the associated B object's value on the same feature. For example, an A shape was briefly presented as larger than its original size, and its pairmate's size was then recalled as also being larger. This transfer did not occur when changed B objects were presented and A memory was queried. In other words, *predictable* object memory was distorted in these experiments, whereas in our experiments *predictive* object memory was distorted. We propose that the apparent discrepancy is because the Yu & Zhao (2018a) findings are due to distortions in short term memory, as opposed to long term memory. We ran a simple proof-of-concept simulation to demonstrate this. We appended an extra unit onto EC_in and EC_out representing the feature that is updated during the post-training test. This unit's activity was set at a consistent value of 0.5 (medium size) for every training trial across all of training, such that the model could learn the baseline value of the feature as consistently associated with every shape. Then during the post-training test, we presented each shape with either a higher (0.8) or lower (0.2) value on the feature unit (to simulate increased or decreased size, for example) and replicated the behavioral asymmetries (**Fig. 5C**) in these contexts. This means that object A input in concert with a feature indicating large size strongly evoked pairmate object B, and thus the large size was still "in mind" when object B was activated (whereas pairmate activation occurred less when object B is presented). The feature-transfer asymmetry resulting in greater distortion to B could thus be due to predictive activation dynamics post-learning, whereas we suggest that the color memory distortions reflect weight-based changes.

General Discussion

Across three experiments, we found that temporal statistical learning shaped memory for the individual constituent objects, with distortions exhibiting a future-oriented bias. In Experiment 1, participants viewed sequences of colored objects with embedded pairs with strong, medium, or weak transition probabilities. After verifying successful statistical learning through faster response times to predictable objects, we found asymmetric predictive color memory distortions for stronger pairs, with participants recalling the predictive object's color as being more similar to the successor's color than vice versa. Experiments 2 and 3 replicated these effects with different pair strength conditions and arrangements of objects in color space. These findings show for the first time that reliable statistical relationships between objects directly induce an asymmetric predictive skew in object feature memory. We employed our neural network model of the hippocampus, C-HORSE, which we have previously used to account for the role of the hippocampus in statistical learning (Schapiro et al., 2017) to investigate potential mechanisms of representational change. Training the model on a simple autoencoding task with the same input structure as Experiment 1, we found behavioral asymmetries analogous to the response time effects in humans, and representational asymmetries in subfield CA1 that mirrored the color memory distortions. Examining the representational change, we found that the predictive shifts were explainable by units that initially represented predictable objects becoming part of predictive object representations with experience. The model demonstrates how simple hippocampal learning mechanisms give rise to representational shifts that could underlie the observed behavioral distortions.

The predictive shifts we see in behavior and in our model are highly consistent with the predictive skew observed in place cells in the hippocampal subfield CA1 in rodents, where neurons that are associated with a certain location on a linear track become part of the representation of earlier locations with experience (Battaglia et al., 2004; Chaudhuri-Vayalambrone et al., 2023; Mehta et al., 1997; Muller & Kubie, 1989). While CA1 place cells have also been shown to fire postdictively, as is also seen to some extent in our model, the modal cell bias as well as the population average is predictive (Chaudhuri-Vayalambrone et al., 2023; Dong et al., 2021). A predictive bias has also been observed in subfield CA3 but tends to develop to a greater extent in CA1, indicating its greater sensitivity to predictive temporal and spatial relationships (Dong et al., 2021; Madar et al., 2023; Roth et al., 2012). The hippocampus, and CA1 in particular, has also been strongly implicated in the learning of sequential temporal information (Allen et al., 2016; Davachi & DuBrow, 2015; Eichenbaum, 2014; Howard & Eichenbaum, 2013; Manns et al., 2007; Mau et al., 2018; Shimbo et al., 2021), with lesions to CA1 but not CA3 disrupting temporal order memory in rats (Hoge & Kesner, 2007), inhibition of EC inputs to CA1 disrupting temporal association memory in mice (Suh et al., 2011), and evidence for CA1 in particular representing temporal context in humans (F. Wang & Diana, 2016) and rats (Mankin et al., 2012).

We found monotonic impacts of co-occurrence strength in our study: Stronger statistics led to stronger learning and distortions. Neuroimaging studies of statistical learning in humans have sometimes reported non-monotonic shifts (Ritvo et al., 2019, 2023) in the hippocampus as a function of transition probability and visual similarity (Schapiro et al., 2012; Wammes et al., 2022), with higher levels leading to representational integration but moderate levels to repulsion. It is possible that the stimulus and transition properties in our paradigm differed from the prior paradigms in a manner resulting in only monotonic changes, or that there is a disconnect between the movement of underlying neural representations and behavioral shifts. Future work could more closely replicate the prior paradigms to investigate whether non-monotonic behavioral memory distortion patterns would emerge.

Schapiro et al. (2012) also found evidence for asymmetric integration of strong pairs, with the neural representations for predictive items shifting more toward the predictable item than vice versa. We found that behavioral memory distortions also exhibit this predictive bias. Our simulations with C-HORSE accounted for both the overall shifts and predictive asymmetries, implicating the representations learned in CA1. While the predictive asymmetry in neural representations was found in DG/CA3 in Schapiro et al. (2012), the mapping of representational change to CA1 is consistent with other studies that have implicated CA1 or anterior hippocampus, an area that overrepresents CA1, in statistical learning (Ellis et al., 2021; Schapiro et al., 2016; Turk-Browne et al., 2010; H. S. Wang et al., 2023).

Two classes of hippocampal circuit mechanisms have been proposed to underlie learned predictive shifts. The first involves "static" shifts, where A representations come to directly include parts of B representations with learning, a mechanism which we argue is executed by the MSP (Kumaran & McClelland, 2012; Schapiro et al., 2017). The second class involves retrieval-based mixing of representations driven by the iterative reinstatement of related representations, argued to be executed by the TSP via "big-loop" recurrence (Kumaran & McClelland, 2012) or within the TSP itself through recurrent connections in CA3 (Lisman, 1999). The successor representation framework proposes that hippocampal population activity represents a state in terms of the states that are expected to follow it, manifesting as static representational shifts analogous to those in the MSP of C-HORSE (Gershman, 2018; Stachenfeld et al., 2017). The Temporal Context Model of the hippocampus (Howard & Kahana, 2002) encodes associations between sensory stimuli and a drifting temporal context, with learned context-item associations similarly developing a static asymmetry (Gershman et al., 2012). Predictive coding, the proposal that higher regions in a processing hierarchy, such as the hippocampus, constantly generate predictions to explain away activity in lower regions, posits that the learning of predictive relationships can involve static shifts as well as recurrent retrieval-based processing within the hippocampus (Barron et al., 2020; Mehta et al., 2000). Given the grounding of our model in the biological circuitry of the hippocampus, we are able to map static predictive representational skewing to the MSP and account for the strong predictive skew seen in region CA1 in particular (Dong et al., 2021; Madar et al., 2023; Roth et al., 2012). Our account is also compatible with recurrence-based mechanisms executed by the TSP, but suggests that the representational changes observed in statistical learning are likely outcomes of static shifts within the MSP.

There have been a few prior behavioral findings relevant to how statistical learning interacts with object feature memory. One study found that memory for the orientation of an object became more similar to the orientation of a predictable successor, consistent with our results, but did not assess the reverse direction (Yu & Zhao, 2018b). Another probed how changing an object's size after learning impacted immediate memory for the pairmate's size (Yu & Zhao, 2018a), finding that a size shift in a predictive object led to immediate distortion in size memory in the same direction for the predicted object, but not vice versa. This distortion for predicted but not predictive objects appears at first glance in tension with our findings that predictive objects undergo more distortion, but our simulation of this paradigm showed that the discrepancy can be explained by the difference between post-learning working memory dynamics and long term representational shifts: Forward behavioral predictions can support immediate feature transfer in activity dynamics that would distort predicted objects, while weight based representational shifts distort predictive objects over the course of learning. A third study exposed participants to statistical regularities at the category level while maintaining trial-unique exemplars and found that predictive exemplars were remembered less well in a recognition memory paradigm (Sherman & Turk-Browne, 2020). We did not find analogous differences in overall error between predictive and predictable

objects across our experiments, but this could be because our study involved many learning opportunities for each specific colored shape (as opposed to just one).

In summary, we provide evidence that reliable temporal statistical relationships induce asymmetric predictive shifts in memory. Our model of the hippocampus simulates these shifts and provides a hypothesis for their implementation in the subfield circuitry of the hippocampus. Predictable statistical structure is ubiquitous in our environment and the work begins to reveal how this structure shapes the content of our memories.

Acknowledgements

We thank Cybelle Smith for helpful discussions and Aditya Pratap Singh for model implementation advice.

References

Aisa, B., Mingus, B., & O'Reilly, R. (2008). The emergent neural modeling system. *Neural Networks*, *21*(8), 1146–1152. https://doi.org/10.1016/j.neunet.2008.06.016

Allen, T. A., Salz, D. M., McKenzie, S., & Fortin, N. J. (2016). Nonspatial sequence coding in CA1 neurons. *The Journal of Neuroscience*, *36*(5), 1547–1563. https://doi.org/10.1523/JNEUROSCI.2874-15.2016

Barron, H. C., Auksztulewicz, R., & Friston, K. (2020). Prediction and memory: A predictive coding account. *Progress in Neurobiology*, *192*, 101821. https://doi.org/10.1016/j.pneurobio.2020.101821

Battaglia, F. P., Sutherland, G. R., & McNaughton, B. L. (2004). Local sensory cues and place cell directionality: Additional evidence of prospective coding in the hippocampus. *Journal of Neuroscience*, *24*(19), 4541–4550. https://doi.org/10.1523/JNEUROSCI.4896-03.2004

Brady, T., Schacter, D. L., & Alvarez, G. (2018). *The adaptive nature of false memories is revealed by gist-based distortion of true memories* [Preprint]. OSF. https://doi.org/10.31234/osf.io/zeg95

Brunec, I. K., & Momennejad, I. (2022). Predictive representations in hippocampal and prefrontal hierarchies. *The Journal of Neuroscience*, *42*(2), 299–312. https://doi.org/10.1523/JNEUROSCI.1327-21.2021

Chanales, A. J. H., Tremblay-McGaw, A. G., Drascher, M. L., & Kuhl, B. A. (2021). Adaptive repulsion of long-term memory representations is triggered by event similarity. *Psychological Science*, *32*(5), 705–720. https://doi.org/10.1177/0956797620972490

Chaudhuri-Vayalambrone, P., Rule, M. E., Bauza, M., Krstulovic, M., Kerekes, P., Burton, S., O'Leary, T., & Krupic, J. (2023). Simultaneous representation of multiple time horizons by entorhinal grid cells and CA1 place cells. *Cell Reports*, *42*(7), 112716. https://doi.org/10.1016/j.celrep.2023.112716

Covington, N. V., Brown-Schmidt, S., & Duff, M. C. (2018). The necessity of the hippocampus for statistical learning. *Journal of Cognitive Neuroscience*, *30*(5), 680–697. https://doi.org/10.1162/jocn_a_01228

Davachi, L., & DuBrow, S. (2015). How the hippocampus preserves order: The role of prediction and context. *Trends in Cognitive Sciences*, *19*(2), 92–99. https://doi.org/10.1016/j.tics.2014.12.004

Dayan, P. (1993). Improving generalization for temporal difference learning: The successor representation. *Neural Computation*, *5*(4), 613–624. https://doi.org/10.1162/neco.1993.5.4.613

de Lange, F. P., Heilbron, M., & Kok, P. (2018). How do expectations shape perception? *Trends in Cognitive*

Sciences, *22*(9), 764–779. https://doi.org/10.1016/j.tics.2018.06.002

Dong, C., Madar, A. D., & Sheffield, M. E. J. (2021). Distinct place cell dynamics in CA1 and CA3 encode experience in new environments. *Nature Communications*, *12*(1), Article 1. https://doi.org/10.1038/s41467-021-23260-3

Eichenbaum, H. (2014). Time cells in the hippocampus: A new dimension for mapping memories. *Nature Reviews Neuroscience*, *15*(11), Article 11. https://doi.org/10.1038/nrn3827

Ellis, C. T., Skalaban, L. J., Yates, T. S., Bejjanki, V. R., Córdova, N. I., & Turk-Browne, N. B. (2021). Evidence of hippocampal learning in human infants. *Current Biology*, *31*(15), 3358-3364.e4. https://doi.org/10.1016/j.cub.2021.04.072

Fiser, J., & Aslin, R. N. (2002). Statistical learning of higher-order temporal structure from visual shape sequences. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *28*(3), 458–467. https://doi.org/10.1037/0278-7393.28.3.458

Friston, K. (2005). A theory of cortical responses. *Philosophical Transactions of the Royal Society B: Biological Sciences*, *360*(1456), 815–836. https://doi.org/10.1098/rstb.2005.1622

Frost, R., Armstrong, B. C., & Christiansen, M. H. (2019). Statistical learning research: A critical review and possible new directions. *Psychological Bulletin*, *145*(12), 1128–1153. https://doi.org/10.1037/bul0000210

Gershman, S. J. (2018). The successor representation: Its computational logic and neural substrates. *The Journal of Neuroscience*, *38*(33), 7193–7200. https://doi.org/10.1523/JNEUROSCI.0151-18.2018

Gershman, S. J., Moore, C. D., Todd, M. T., Norman, K. A., & Sederberg, P. B. (2012). The successor representation and temporal context. *Neural Computation*, *24*(6), 1553–1568. https://doi.org/10.1162/NECO_a_00282

Girvan, M., & Newman, M. E. J. (2002). Community structure in social and biological networks. *Proceedings of the National Academy of Sciences*, *99*(12), 7821–7826. https://doi.org/10.1073/pnas.122653799

Heffernan, E. M., Schlichting, M. L., & Mack, M. L. (2021). Learning exceptions to the rule in human and model via hippocampal encoding. *Scientific Reports*, *11*(1), 21429. https://doi.org/10.1038/s41598-021-00864-9

Henin, S., Turk-Browne, N. B., Friedman, D., Liu, A., Dugan, P., Flinker, A., Doyle, W., Devinsky, O., & Melloni, L. (2021). Learning hierarchical sequence representations across human cortex and hippocampus. *Science Advances*, *7*(8), eabc4530. https://doi.org/10.1126/sciadv.abc4530

Hoge, J., & Kesner, R. P. (2007). Role of CA3 and CA1 subregions of the dorsal hippocampus on temporal processing of objects. *Neurobiology of Learning and Memory*, *88*(2), 225–231. https://doi.org/10.1016/j.nlm.2007.04.013

Howard, M. W., & Eichenbaum, H. (2013). The hippocampus, time, and memory across scales. *Journal of Experimental Psychology: General*, *142*(4), 1211–1230. https://doi.org/10.1037/a0033621

Howard, M. W., & Kahana, M. J. (2002). A distributed representation of temporal context. *Journal of Mathematical Psychology*, *46*(3), 269–299. https://doi.org/10.1006/jmps.2001.1388

Isbilen, E. S., & Christiansen, M. H. (2022). Statistical learning of language: A meta-analysis into 25 years of research. *Cognitive Science*, *46*(9), e13198. https://doi.org/10.1111/cogs.13198

Ketz, N., Morkonda, S. G., & O'Reilly, R. C. (2013). Theta coordinated error-driven learning in the hippocampus. *PLoS Computational Biology*, *9*(6), e1003067. https://doi.org/10.1371/journal.pcbi.1003067

Kumaran, D., & McClelland, J. L. (2012). Generalization through the recurrent interaction of episodic memories: A model of the hippocampal system. *Psychological Review*, *119*(3), 573–616. https://doi.org/10.1037/a0028681

Leutgeb, S., Leutgeb, J. K., Treves, A., Moser, M.-B., & Moser, E. I. (2004). Distinct ensemble codes in hippocampal areas CA3 and CA1. *Science*, *305*(5688), 1295–1298. https://doi.org/10.1126/science.1100265

Lisman, J. E. (1999). Relating hippocampal circuitry to function: Recall of memory sequences by reciprocal Dentate–CA3 interactions. *Neuron*, *22*(2), 233–242. https://doi.org/10.1016/S0896-6273(00)81085-5

Madar, A. D., Dong, C., & Sheffield, M. E. J. (2023). *BTSP, not STDP, drives shifts in hippocampal representations during familiarization* [Preprint]. bioRxiv. https://doi.org/10.1101/2023.10.17.562791

Mankin, E. A., Sparks, F. T., Slayyeh, B., Sutherland, R. J., Leutgeb, S., & Leutgeb, J. K. (2012). Neuronal code for extended time in the hippocampus. *Proceedings of the National Academy of Sciences*,

109(47), 19462–19467. https://doi.org/10.1073/pnas.1214107109

Manns, J. R., Howard, M. W., & Eichenbaum, H. (2007). Gradual changes in hippocampal activity support remembering the order of events. *Neuron*, *56*(3), 530–540. https://doi.org/10.1016/j.neuron.2007.08.017

Mau, W., Sullivan, D. W., Kinsky, N. R., Hasselmo, M. E., Howard, M. W., & Eichenbaum, H. (2018). The same hippocampal CA1 population simultaneously codes temporal information over multiple timescales. *Current Biology*, *28*(10), 1499-1508.e4. https://doi.org/10.1016/j.cub.2018.03.051

Mehta, M. R., Barnes, C. A., & McNaughton, B. L. (1997). Experience-dependent, asymmetric expansion of hippocampal place fields. *Proceedings of the National Academy of Sciences*, *94*(16), 8918–8921. https://doi.org/10.1073/pnas.94.16.8918

Mehta, M. R., Quirk, M. C., & Wilson, M. A. (2000). Experience-dependent asymmetric shape of hippocampal receptive fields. *Neuron*, *25*(3), 707–715. https://doi.org/10.1016/S0896-6273(00)81072-7

Muller, R. U., & Kubie, J. L. (1989). The firing of hippocampal place cells predicts the future position of freely moving rats. *Journal of Neuroscience*, *9*(12), 4101–4110. https://doi.org/10.1523/JNEUROSCI.09-12-04101.1989

Nakashiba, T., Young, J. Z., McHugh, T. J., Buhl, D. L., & Tonegawa, S. (2008). Transgenic inhibition of synaptic transmission reveals role of ca3 output in hippocampal learning. *Science*, *319*(5867), 1260–1264. https://doi.org/10.1126/science.1151120

Nakazawa, K., Sun, L. D., Quirk, M. C., Rondi-Reig, L., Wilson, M. A., & Tonegawa, S. (2003). Hippocampal CA3 NMDA receptors are crucial for memory acquisition of one-time experience. *Neuron*, *38*(2), 305–315. https://doi.org/10.1016/S0896-6273(03)00165-X

O'Reilly, R. C., Munakata, Y., Frank, M. J., Hazy, T. E., & Contributors. (2024). *Computational cognitive neuroscience*. Online Book, 5th Edition. https://compcogneuro.org/book

Piatti, V., Ewell, L., & Leutgeb, J. (2013). Neurogenesis in the dentate gyrus: Carrying the message or dictating the tone. *Frontiers in Neuroscience*, *7*. https://doi.org/10.3389/fnins.2013.00050

Rao, R. P. N., & Ballard, D. H. (1999). Predictive coding in the visual cortex: A functional interpretation of some extra-classical receptive-field effects. *Nature Neuroscience*, *2*(1), Article 1. https://doi.org/10.1038/4580

Ritvo, V. J. H., Nguyen, A., Turk-Browne, N. B., & Norman, K. A. (2024). Differentiation and integration of competing memories: A neural network model. *eLife*, *12*, e88608. https://doi.org/10.7554/eLife.88608.3

Ritvo, V. J. H., Turk-Browne, N. B., & Norman, K. A. (2019). Nonmonotonic plasticity: How memory retrieval drives learning. *Trends in Cognitive Sciences*, *23*(9), 726–742. https://doi.org/10.1016/j.tics.2019.06.007

Roth, E. D., Yu, X., Rao, G., & Knierim, J. J. (2012). Functional differences in the backward shifts of CA1 and CA3 place fields in novel and familiar environments. *PLOS ONE*, *7*(4), e36035. https://doi.org/10.1371/journal.pone.0036035

Santolin, C., & Saffran, J. R. (2018). Constraints on statistical learning across species. *Trends in Cognitive Sciences*, *22*(1), 52–63. https://doi.org/10.1016/j.tics.2017.10.003

Schapiro, A. C., Gregory, E., Landau, B., McCloskey, M., & Turk-Browne, N. B. (2014). The necessity of the medial temporal lobe for statistical learning. *Journal of Cognitive Neuroscience*, *26*(8), 1736–1747. https://doi.org/10.1162/jocn_a_00578

Schapiro, A. C., Kustner, L. V., & Turk-Browne, N. B. (2012). Shaping of object representations in the human medial temporal lobe based on temporal regularities. *Current Biology*, *22*(17), 1622–1627. https://doi.org/10.1016/j.cub.2012.06.056

Schapiro, A. C., Turk-Browne, N. B., Botvinick, M. M., & Norman, K. A. (2017). Complementary learning systems within the hippocampus: A neural network modelling approach to reconciling episodic memory with statistical learning. *Philosophical Transactions of the Royal Society B: Biological Sciences*, *372*(1711), 20160049. https://doi.org/10.1098/rstb.2016.0049

Schapiro, A. C., Turk-Browne, N. B., Norman, K. A., & Botvinick, M. M. (2016). Statistical learning of temporal community structure in the hippocampus. *Hippocampus*, *26*(1), 3–8. https://doi.org/10.1002/hipo.22523

Sherman, B. E., & Turk-Browne, N. B. (2020). Statistical prediction of the future impairs episodic encoding of the present. *Proceedings of the National Academy of Sciences*, *117*(37), 22760–22770. https://doi.org/10.1073/pnas.2013291117

Shimbo, A., Izawa, E.-I., & Fujisawa, S. (2021). Scalable representation of time in the hippocampus. *Science Advances*, *7*(6), eabd7013. https://doi.org/10.1126/sciadv.abd7013

Singh, D., Norman, K. A., & Schapiro, A. C. (2022). A model of autonomous interactions between hippocampus and neocortex driving sleep-dependent memory consolidation. *Proceedings of the National Academy of Sciences*, *119*(44), e2123432119. https://doi.org/10.1073/pnas.2123432119

Skaggs, W. E., & McNaughton, B. L. (1998). Spatial firing properties of hippocampal ca1 populations in an environment containing two visually identical regions. *Journal of Neuroscience*, *18*(20), 8455–8466. https://doi.org/10.1523/JNEUROSCI.18-20-08455.1998

Stachenfeld, K. L., Botvinick, M. M., & Gershman, S. J. (2017). The hippocampus as a predictive map. *Nature Neuroscience*, *20*(11), 1643–1653. https://doi.org/10.1038/nn.4650

Sučević, J., & Schapiro, A. C. (2022). A neural network model of hippocampal contributions to category learning. *eLife*, *11*, e77185. https://doi.org/10.7554/eLife.77185

Suh, J., Rivest, A. J., Nakashiba, T., Tominaga, T., & Tonegawa, S. (2011). Entorhinal cortex layer III input to the hippocampus is crucial for temporal association memory. *Science*, *334*(6061), 1415–1420. https://doi.org/10.1126/science.1210125

Sun, Y., Nguyen, A. Q., Nguyen, J. P., Le, L., Saur, D., Choi, J., Callaway, E. M., & Xu, X. (2014). Cell-type-specific circuit connectivity of hippocampal CA1 revealed through cre-dependent rabies tracing. *Cell Reports*, *7*(1), 269–280. https://doi.org/10.1016/j.celrep.2014.02.030

Tacikowski, P., Kalender, G., Ciliberti, D., & Fried, I. (2024). Human hippocampal and entorhinal neurons encode the temporal structure of experience. *Nature*, *635*, 160–167. https://doi.org/10.1038/s41586-024-07973-1

Tandoc, M. C., Dong, C., & Schapiro, A. (2023). *Object feature memory is distorted by category structure* [Preprint]. PsyArXiv. https://doi.org/10.31234/osf.io/9a24e

Turk-Browne, N. B., Scholl, B. J., Chun, M. M., & Johnson, M. K. (2009). Neural evidence of statistical learning: Efficient detection of visual regularities without awareness. *Journal of Cognitive Neuroscience*, *21*(10), 1934–1945. https://doi.org/10.1162/jocn.2009.21131

Turk-Browne, N. B., Scholl, B. J., Johnson, M. K., & Chun, M. M. (2010). Implicit perceptual anticipation triggered by statistical learning. *Journal of Neuroscience*, *30*(33), 11177–11187. https://doi.org/10.1523/JNEUROSCI.0858-10.2010

Vazdarjanova, A., & Guzowski, J. F. (2004). Differences in hippocampal neuronal population responses to modifications of an environmental context: Evidence for distinct, yet complementary, functions of ca3 and ca1 ensembles. *Journal of Neuroscience*, *24*(29), 6489–6496. https://doi.org/10.1523/JNEUROSCI.0350-04.2004

Wammes, J., Norman, K. A., & Turk-Browne, N. (2022). Increasing stimulus similarity drives nonmonotonic representational change in hippocampus. *eLife*, *11*, e68344. https://doi.org/10.7554/eLife.68344

Wang, F., & Diana, R. A. (2016). Temporal context processing within hippocampal subfields. *NeuroImage*, *134*, 261–269. https://doi.org/10.1016/j.neuroimage.2016.03.048

Wang, H. S., Rosenbaum, R. S., Baker, S., Lauzon, C., Batterink, L. J., & Köhler, S. (2023). Dentate gyrus integrity is necessary for behavioral pattern separation but not statistical learning. *Journal of Cognitive Neuroscience*, *35*(5), 900–917. https://doi.org/10.1162/jocn_a_01981

Witter, M. P., Wouterlood, F. G., Naber, P. A., & Van Haeften, T. (2006). Anatomical organization of the parahippocampal-hippocampal network. *Annals of the New York Academy of Sciences*, *911*(1), 1–24. https://doi.org/10.1111/j.1749-6632.2000.tb06716.x

Yu, R. Q., & Zhao, J. (2018a). Implicit updating of object representation via temporal associations. *Cognition*, *181*, 127–134. https://doi.org/10.1016/j.cognition.2018.08.015

Yu, R. Q., & Zhao, J. (2018b). Object representations are biased toward each other through statistical learning. *Visual Cognition*, *26*(4), 253–267. https://doi.org/10.1080/13506285.2018.1435596

Zhao, Y., Chanales, A. J. H., & Kuhl, B. A. (2021). Adaptive memory distortions are predicted by feature representations in parietal cortex. *The Journal of Neuroscience*, *41*(13), 3014–3024. https://doi.org/10.1523/JNEUROSCI.2875-20.2021

Zhou, Z., Singh, D., Tandoc, M. C., & Schapiro, A. C. (2023). Building integrated representations through interleaved learning. Journal of Experimental Psychology: General, 152(9), 2666–2684. https://doi.org/10.1037/xge0001415

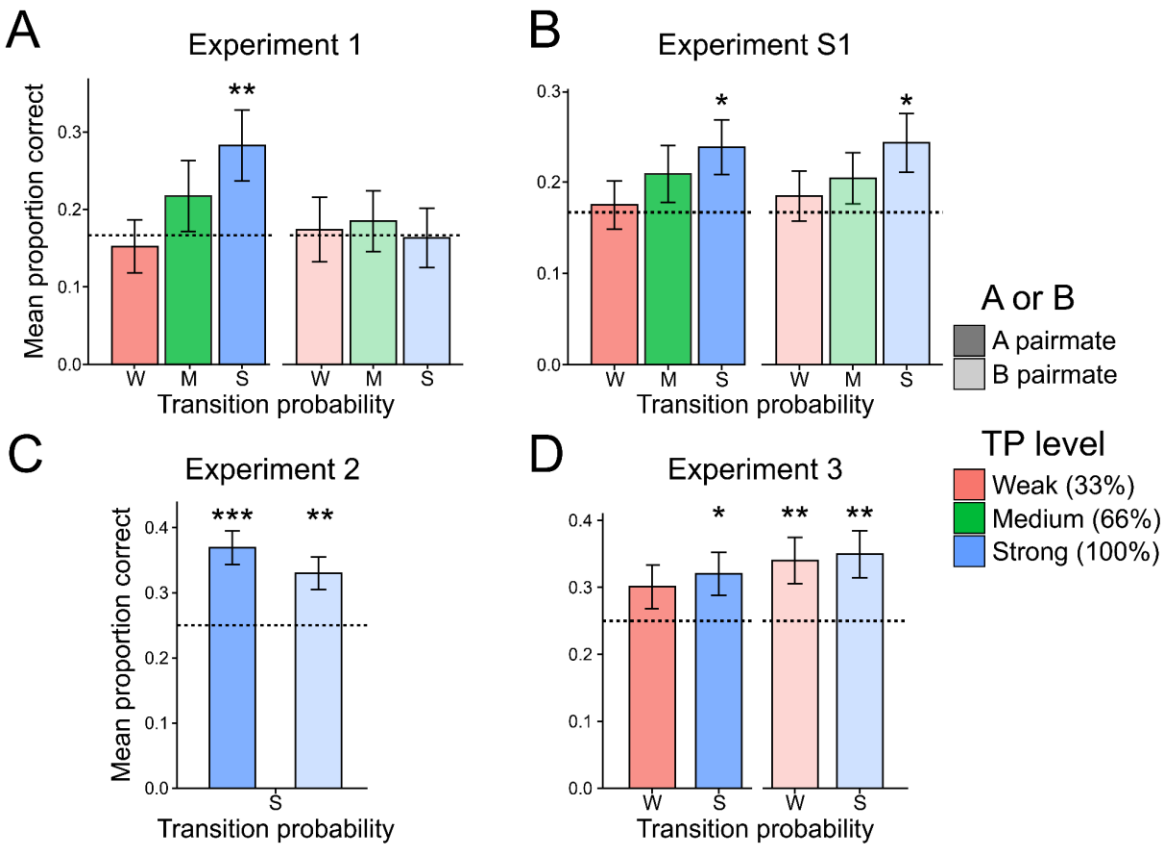
Supplementary Information for:

Statistical learning drives predictive shifts in object memory

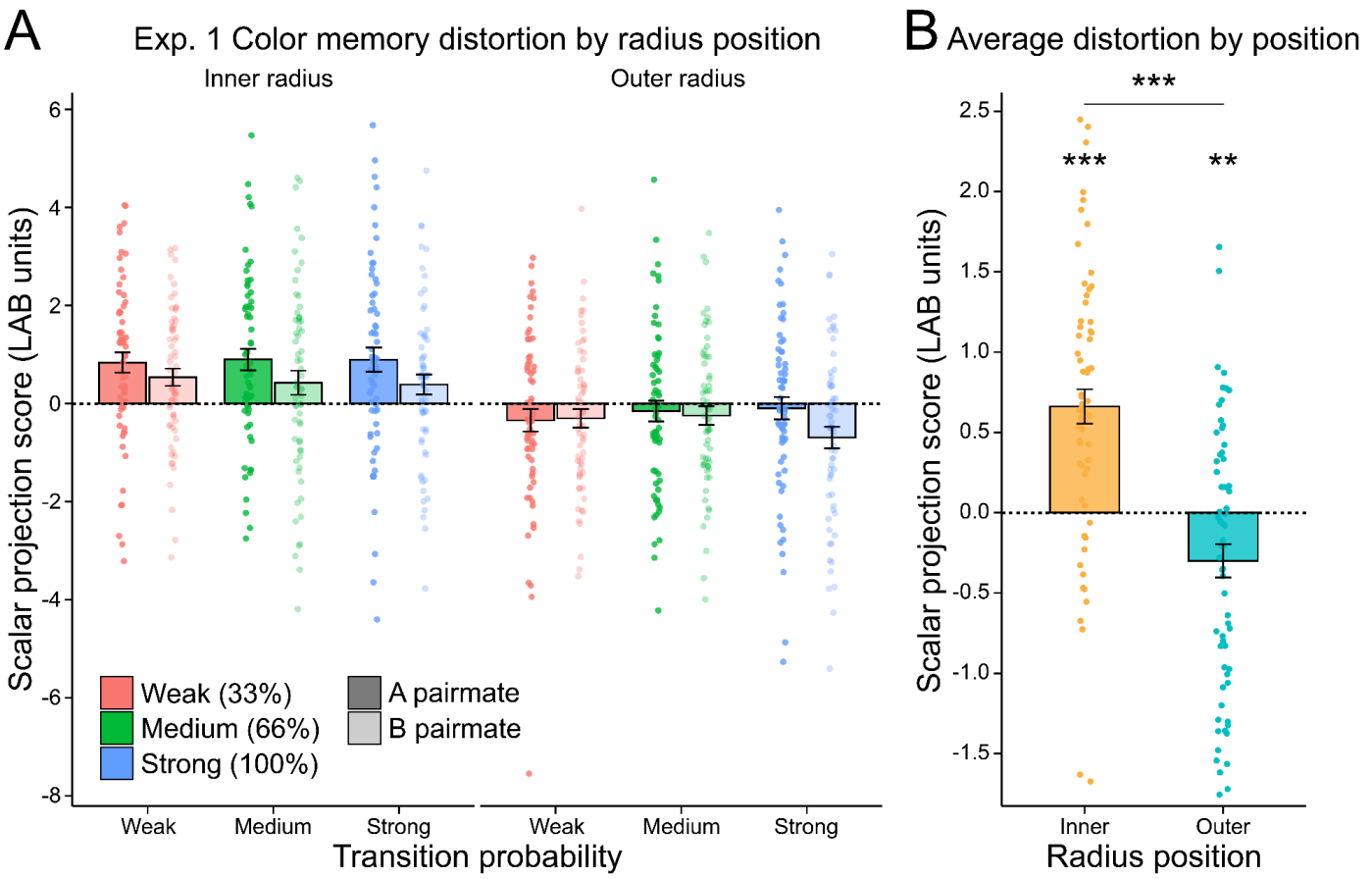
Dhairyya Singh, Cody V. Dong, Marlie C. Tandoc, and Anna C. Schapiro

This document includes:

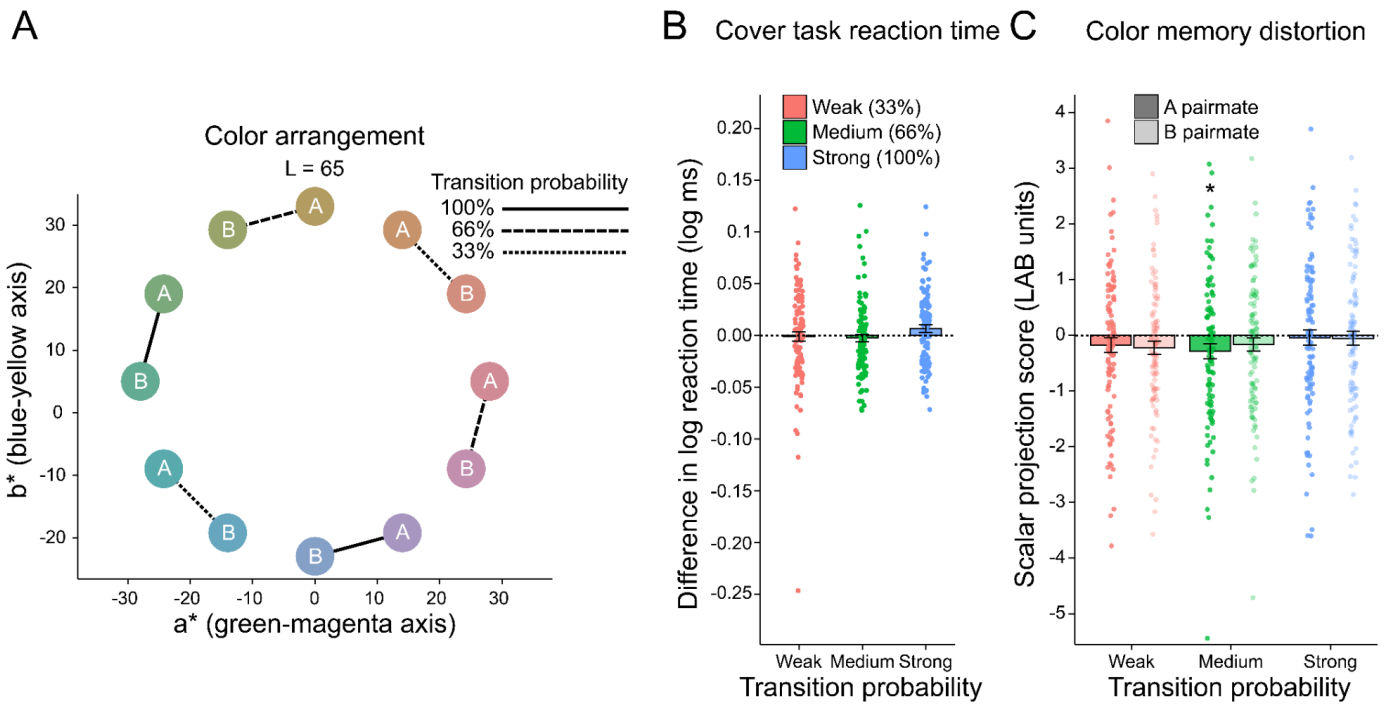
Supplementary Figures 1-3
Supplementary Tables 1-2



Supplementary Fig. 1. Explicit AFC post-test results. **(A)** Experiment 1. **(B)** Experiment S1. Participants were significantly above chance at selecting the true pairmate for only Strong A and B shapes (bootstrap 10,000 replicates, A: $p = .015$; bootstrap 10,000 replicates, B: $p = .013$; other shapes ps > .17). The difference in identification accuracy (A vs. B) between pairmates was not significant in any condition (bootstrap, 10,000 replicates; strong: ps > .821). **(C)** Experiment 2. **(D)** Experiment 3. See main text for description of results for Experiments 1-3. *p < .05; **p < .01, ***p < .001. Dashed lines denote chance levels.



Supplementary Fig. 2. Saturation bias in Experiment 1 color memory distortion. **(A)** Scalar projection score values from **Fig. 2C** split by placement on the inner or outer radius of the color arrangement **(Fig. 2A)**. Memory for inner radius colors is biased toward greater levels of attraction and memory for outer radius colors is biased toward greater repulsion. Despite differences in baseline levels, there is an overall trend of A pairmates showing greater attraction towards B than vice versa, conserved across inner and outer radius colors, with the strongest effect in the strong transition probability condition (Inner: $t(63) = 1.58$, $p = .118$; Outer: $t(63) = 1.91$, $p = .061$). **(B)** Distortion averaged across pairmate and transition probabilities shows a significant attraction for inner positions towards their pairmates ($M = 0.661$; $t(63) = 6.17$, $p < .001$) and repulsion for outer positions away from their pairmates ($M = -0.300$; $t(63) = -2.875$, $p = .005$), indicating a strong overall repulsion away from the unsaturated center of the color space and towards the saturated periphery. The difference between scalar projection values between inner and outer radius colors was also significant ($t(63) = 5.867$, $p < .001$). Dots correspond to individual participants, and error bars represent +/- 1 SEM. ***$p < .001$.



Supplementary Fig. 3. Experiment S1 results. **(A)** Pair color arrangement. **(B)** Cover task reaction time difference between the predictable B shapes and predictive A shapes. There was only marginal evidence of faster responses for predictable shapes in paired trials (*p*s > .067). **(C)** Color memory distortions on the axis connecting pairmates. Only the medium A pairmate showed a significant shift in the negative direction ($t(102) = -2.18$, $p = .032$, other *p*s > .06). Differences between A and B scalar projection scores were not significant (*p*s > .475). *$p < .05$.

Supplementary Table 1. Model layer parameters

Layer	Parameter	Value
EC_{in}	# Units	12
	Inhibition (Gi)	1.8
EC_{out}	# Units	12
	Inhibition (Gi)	1.8
DG	# Units	400
	Inhibition (Gi)	24
CA3	# Units	80
	Inhibition (Gi)	4.5
pCA1	# Units	50
	Inhibition (Gi)	3
dCA1	# Units	50
	Inhibition (Gi)	2.1

Supplementary Table 2. Model projection parameters

Projection	**Parameter**	**Value**
Default	Weight init distribution	Uniform 0.25 - 0.75
	Connectivity	Full
	WtScale.Abs / WtScale.Rel	1/1
EC_{in} → DG/CA3	Connectivity	Sparse - 0.25
	Learning rate	0.4
DG → CA3	Weight init distribution	Uniform 0.89 - 0.91
	Connectivity	Sparse - 0.05
	Learning rate	0
	WtScale.Rel	8
CA3 → CA3	WtScale.Rel	2
CA3 → pCA1	Connectivity	Sparse - 0.25
	Learning rate	0.04
pCA1 ↔ EC_{out}	Learning rate	0.04
EC_{in} → dCA1	Learning rate	0.04
dCA1 ↔ EC_{out}	Learning rate	0.04
EC_{out} → EC_{in}	Learning rate	0
	WtScale.Abs	2