Behavioral/Cognitive

Pattern Separation Underpins Expectation-Modulated Memory

Darya Frank, Marcelo A. Montemurro, and Daniela Montaldi
Division of Neuroscience and Experimental Psychology, School of Biological Sciences, University of Manchester, Manchester M13 9PL, United Kingdom

Pattern separation and completion are fundamental hippocampal computations supporting memory encoding and retrieval. However, despite extensive exploration of these processes, it remains unclear whether and how top-down processes adaptively modulate the dynamics between these computations. Here we examine the role of expectation in shifting the hippocampus to perform pattern separation. In a behavioral task, 29 participants (7 males) learned a cue-object category contingency. Then, at encoding, one-third of the cues preceding the to-be-memorized objects, violated the studied rule. At test, participants performed a recognition task with old objects (targets) and a set of parametrically manipulated (very similar to dissimilar) foils for each object. Accuracy was found to be better for foils of high similarity to targets that were contextually unexpected at encoding compared with expected ones. Critically, there were no expectation-driven differences for targets and low similarity foils. To further explore these effects, we implemented a computational model of the hippocampus, performing the same task as the human participants. We used representational similarity analysis to examine how top-down expectation interacts with bottom-up perceptual input, in each layer. All subfields showed more dissimilar representations for unexpected items, with dentate gyrus (DG) and CA3 being more sensitive to expectation violation than CA1. Again, representational differences between expected and unexpected inputs were prominent for moderate to high levels of input similarity. This effect diminished when inputs from DG and CA3 into CA1 were lesioned. Overall, these novel findings strongly suggest that pattern separation in DG/CA3 underlies the effect that violation of expectation exerts on memory.

Key words: **episodic memory; expectation; pattern separation; representational similarity**

Significance Statement

What makes some events more memorable than others is a key question in cognitive neuroscience. Violation of expectation often leads to better memory performance, but the neural mechanism underlying this benefit remains elusive. In a behavioral study, we found that memory accuracy is enhanced selectively for unexpected highly similar foils, suggesting expectation violation does not enhance memory indiscriminately, but specifically aids the disambiguation of overlapping inputs. This is further supported by our subsequent investigation using a hippocampal computational model, revealing increased representational dissimilarity for unexpected highly similar foils in DG and CA3. These convergent results provide the first evidence that pattern separation plays an explicit role in supporting memory for unexpected information.

Introduction

The hippocampus supports memory by storing each experience as a unique memory representation in a process known as *pattern separation* (PS), and by later reinstating that representation from a partial cue using *pattern completion* (PC; McClelland et al., 1995; Norman and O'Reilly, 2003). Sparse activity in dentate gyrus (DG) and mossy fiber projections to CA3 are believed to underlie pattern separation, whereas the recurrent collaterals in CA3 and projection via Schaffer collaterals to CA1 support pattern completion (Leutgeb et al., 2007; Yassa and Stark, 2011). PS and PC are postulated to synchronize along the theta cycle, with PS occurring at the trough and PC at the peak of the cycle (Hasselmo et al., 2002; Kunec et al., 2005). Although these computations have been assessed in neuropsychological, computational, and neuroimaging studies, the role of potential cognitive modulators remains unclear.

One such process is contextual expectation, which guides adaptive behavior (Bar, 2009) and modulates memory performance (Long et al., 2016; Frank et al., 2018; Kafkas and Montaldi, 2018a). A mismatch between predicted and received information prepares cellular mechanisms toward encoding, including a shift

Received Aug. 23, 2019; revised Jan. 17, 2020; accepted Feb. 1, 2020.
Author contributions: D.F., M.A.M., and D.M. designed research; D.F. performed research; D.F. analyzed data; D.F., M.A.M., and D.M. wrote the paper.
This work was supported by a PDS award from The University of Manchester to D.F.
The authors declare no competing financial interests.
Correspondence should be addressed to Darya Frank at darya.frank@manchester.ac.uk.
https://doi.org/10.1523/JNEUROSCI.2047-19.2020

in the theta rhythm (O'Reilly et al., 1994; Meeter et al., 2004; Axmacher et al., 2010; Douchamps et al., 2013). The hippocampus, perhaps through interaction with the dopaminergic system, is believed to support this effect (Lisman and Grace, 2005; Shohamy and Wagner, 2008; Kafkas and Montaldi, 2018b). Findings from animal models provide some support for such adaptive encoding, showing context-sensitive shifts between PS and PC (Colgin et al., 2008). However, evidence for the relationship between expectation-violation and PS in humans is lacking. To shed light on this, comparison is needed between expected and unexpected items in tasks where performance is dependent on successful PS.

Given that PS entails disambiguating similar inputs, a common approach to probe it, is using perceptually similar foils (Leal and Yassa, 2018). In such tasks, correct rejections can capture PS by reflecting differentiation of the similar foils from the target. As similarity increases between target and foils, discrimination becomes more difficult, therefore requiring PS. If a violation of expectation triggers a shift toward a PS-driven encoding state, foils similar to an unexpected target would be more easily discriminated. Alternatively, if expectation does not modulate memory, performance should reflect solely the degree of item similarity. Using similar foils also allows us to directly test whether the beneficial effect of surprise selectively targets a PS mechanism or provides a more general memory boost. If the latter is true, we should observe more hits as well as more correct rejections of all unexpected objects (a global enhancement effect). On the other hand, an effect that is selective to correct rejections of unexpected similar foils (i.e., an interaction between expectation and similarity) would suggest expectation-violation triggers a shift toward PS. This mechanism would only be engaged when input differentiation is essential to task performance (i.e., high similarity).

Here we tested whether the beneficial influence of expectation-violation on memory is driven by a shift toward PS, and explored the specificity of such mechanism. We first tested this in a behavioral task (Experiment 1) and subsequently used a neural network model to examine the contribution of different hippocampal subfields (Experiment 2). To elicit contextual expectation, we used a rule-learning task where participants (and the network) learned a cue-category contingency. This contingency was later violated on some unexpected trials. We examined recognition performance and representational dissimilarity for versus unexpected targets and foils of parametrically-manipulated similarity, presented at retrieval. This approach allows us to bridge the gap between different levels of analysis, combining a cognitive task adapted from human research and a neural mechanism characterized by computational and animal models. We predicted that highly similar unexpected items would benefit from enhanced discrimination performance and more dissimilar representations, compared with expected ones, driven by a shift toward PS on expectation-violation.

Methods

Experiment 1

Participants

Twenty-nine participants (mean age = 19.5, 7 males) gave informed consent and took part in the experiment. Participants had normal or corrected-to-normal vision and no history of neurologic or psychiatric disorders. Five participants were excluded from analysis due to memory performance below chance (1 participant) or failure to reach criterion during the rule-learning task (4 participants). All procedures were approved by the University of Manchester Research Ethics Committee.

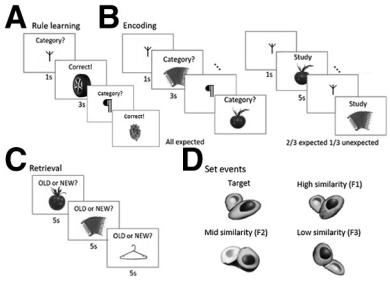

Figure 1. Experimental design. ***A***, In the rule-learning task participants learned a contingency between a cue and an object's category, manmade or natural. ***B***, During the first round of encoding, participants were presented with the same cues (all rule-abiding) and had to indicate whether the object was manmade or natural. In the second round, participants are asked to study the object carefully and 1/3 of the cues are misleading (unexpected). ***C***, In the final retrieval task targets and new similar foils are presented, and participants are asked to respond whether the object is old or new. ***D***, Illustration of all set events presented during retrieval (target, F1, F2, and F3).

Materials

Seventy-eight images, natural (39) and manmade (39), were selected using the Dissimilarity Index from the Similar Object-Lures Image Database (SOLID; Frank et al., 2019). These images were used as the target objects, presented during encoding. Using a custom MATLAB (MathWorks) function (https://github.com/frdarya/SOLID/blob/master/ChooseFoils.m), three foils of decreasing levels of similarity were selected (DIs 1300, 2000, and 2700) for each target image. Similarity was parametrically manipulated by keeping the average distance between the levels constant (average DI between foils 2100). For the rule-learning task 56 more images (28 from each manmade/natural category) were taken from SOLID.

Procedure

The experiment was controlled using PsychoPy2 v1.82 (Peirce, 2007) and consisted of three main parts (Fig. 1), similar to the design used by Kafkas and Montaldi (2018a).

Rule learning task. Participants learned an association between an arbitrary symbol cue and a category (manmade or natural), to generate contextual expectation. There was a total of four cues, two for manmade and two for natural objects. Following a fixation cross, a cue appeared on the screen for 1 s, during which participants were asked to predict the category of the next object. They were instructed to guess during the first few trials, but to learn the contingency as the task progressed. Subsequently, a manmade or natural object (not tested) appeared and participants received feedback about their prediction (3 s). Each cue was repeated 14 times and cues were counterbalanced across participants.

Encoding task. In this task, participants encoded 78 object images (targets), presented twice, in two rounds. Before the task began, participants were told their memory for these items would be tested using similar foils and truly old objects. During the first round, a previously-learned cue (1 s) was followed by an object (3 s), and participants were asked to indicate whether the object is manmade or natural. In this round, all cues were consistent with the rule (no expectation violation). In the second round, participants were asked to study the perceptual details of the image carefully (5 s), to correctly recognize them and reject similar foils as new items. Importantly, two-thirds of the cues in this round were consistent with the rule (expected stimuli) while the other one-

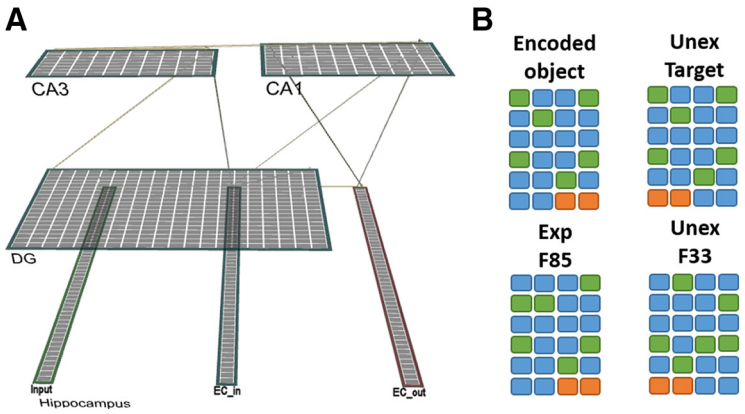

Figure 2. ***A***, Network architecture. Illustration of the hippocampal network in emergent. ***B***, Simplified representations of stimuli. Green slots represent active stimulus units (clamped to 1) and orange slots represent the category cue (clamped at 0.9). Top, Left, The object the network was exposed to during training. In the other corners there are examples of items presented during testing, varying in expectation condition and level of overlap.

Table 1. Projection parameters

Projection	Weights	Scale	Connectivity	Learning rate
Input → EC_{in}	0.25–0.75	1/1	1–1	0
EC_{in} → DG	0.25–0.75	1/1	25%	0.2
EC_{in} → CA3	0.25–0.75	1/1	25%	0.2
DG → CA3	0.89–0.91	1/8	5%	0
CA3 → CA3	0.25–0.75	1/1	100%	0.2
CA3 → CA1	0.25–0.75	1/1	100%	0.05
EC_{in} → CA1	0.25–0.75	3/1	100%	0.02
CA1 → EC_{out}	0.25–0.75	1/1	100%	0.02
EC_{out} → CA1	0.25–0.75	1/1	100%	0.02
EC_{out} → EC_{in}	0.49–0.51	2/0.5	1–1	0

Weight range is specified by a mean and SD of activations. Scale takes into account the absolute multiplier on weights divided by the weighting relative to other projections. Connectivity is the percentage of units projecting from the sending layer to the receiving one; 1–1 indicates each unit in the sending layer projects to one unit in the receiving layer. Learning rate captures how fast the weights in each layer change per presentation.

third of cues violated the rule (unexpected stimuli). In both encoding rounds participants were instructed to ignore the cue and focus on the main task. Stimulus presentation order was random, and allocation to contextual expectation condition pseudorandom, maintaining equal number of expected/unexpected targets for the two categories. Before the retrieval task, an arithmetic distractor task was used for 5 min.

Retrieval task. The final task was a continuous recognition memory paradigm. Targets and their associated foils formed sets, and each occurrence of an object constitutes a unique set event, capturing the object's set identity (e.g., apple), mnemonic status (target or foil), and position within set during retrieval (e.g., the second appearance of an apple). A set event (target, F1, F2, or F3) appeared on the screen for 5 s during which participants had to decide if it was old (target) or new (foil).

Statistical analyses

Given the nature of our recognition task, we collated object sets (target + 3 foils) and ran mixed-effects binary logistic regression models on these ungrouped data. Models were computed using the lme4 package (Bates et al., 2015) in the R environment (R Development Core Team, 2008). The parameters of such models can be used to assess the probability of giving a correct response ("old" for targets, "new" for foils) while accounting for each participant's unique intercept. To assess the independent effect on hit rate, of contextual expectation established during encoding, we used a simple model using expectation as the only predictor. For foils, all correct rejections (CRs) were collated and a model including main effects (similarity level: F1, F2, F3, and expectation), as well as the interaction, was devised. To examine whether this effect persisted over presentation of multiple set events, we also created models for predicting each event, presented in all other positions within a set as a function of the contextual expectation during encoding. Extraction and plotting of the effects reported in the results section below was conducted using the effects package (Fox, 2003) and ggplot2 (Wickham, 2009). As the mixed-effects approach incorporates trial-by-trial decisions, in Fig. 4-2) we also calculated the average hit and false alarm (1 − CR) rates for each participant, to calculate a corrected-recognition score (hit – FA). The data and analysis code are available at https://github.com/frdarya/PS_expectation.

Experiment 2

Model architecture

We used a neural network model of the hippocampus implemented in the Emergent simulation software (v7.0.1; Aisa et al., 2008; for illustration, see Fig. 2*A*). The model's architecture and projections are based on the hippocampal component of the CLS framework (McClelland et al., 1995; Norman and O'Reilly, 2003), which has been used to demonstrate neural memory mechanisms, emulating findings from human and animal research (Meeter et al., 2004; Elfman et al., 2014; Schapiro et al., 2017; Pilly et al., 2018).

The model includes entorhinal cortex input (ECin) and output (ECout), DG, CA3, and CA1 layers. Inputs are presented to the model via an input layer with one-to-one connections to ECin, which then projects to DG, CA3, and CA1 via the trisynaptic pathway (TSP). TSP is believed to support encoding of new memories and conjunctions with pattern-separated representations generated in DG, and transferred to CA3 via the sparse mossy fibers (each CA3 unit receives input from 5% of DG). Recurrent collaterals in CA3 (modelled as a fully-connected projection) then allow pattern completion from partial cues to occur. The pattern-completed representation is then projected onto CA1 via the fully-connected Schaffer collateral pathway. EC is also connected directly to CA1 via the fully-connected monosynaptic pathway (MSP), which supports memory retrieval by associating direct input with diffuse inputs from the Schaffer collaterals. The network parameters of these projections are outlined in Table 1.

Activity levels of units in the network ranged between 0 and 1. Each unit's activity level was modulated by local inhibition from other units within the same layer, modeling inhibitory interneurons. Following previous work (O'Reilly and Munakata, 2000; Schapiro et al., 2017) this inhibition was implemented using a *k*-winner-takes-all equation (for specific kWTA values used in each layer, see Table 2). In EC_{in} and EC_{out} $k = 12$, which is the number of active units in each stimulus presented to the network (excluding the previous cue, see Stimulus presentation). Similar results were obtained using a lower inhibition at $k = 15$ (total number of stimulus + cue units).

Table 2. kWTA parameters

Layer	No. of units	Proportion of activity	kWTA pt
EC (in and out)	54	$k = 12$ (0.22)	0.5
DG	400	0.01	0.9
CA3	80	0.06	0.7
CA1	100	0.25	0.7

The proportion of activity in each layer, k is determined by the layer size and desired proportion of activity. kWTA pt represents the global inhibition in each layer.

The network utilizes a combination of Hebbian and error-driven learning (Ketz et al., 2013; Schapiro et al., 2017) to adjust its connection weights, such that an input presented to EC_{in} can be reproduced in EC_{out}. To achieve this, the network goes through minus and plus phases during the learning period, akin to theta oscillations (Ketz et al., 2013). In the minus phase, EC_{in} projects to CA1 while CA3 input is inhibited, followed by a reversed effect whereby CA3 input to CA1 resumes while EC_{in} inputs are weakened. Theta troughs and peaks have been suggested to reflect encoding (driven by external inputs) and retrieval states (driven by internal inputs), respectively (Hasselmo et al., 2002; Kunec et al., 2005). In the plus phase, the network is exposed to the "ground truth" via a loop between EC_{in}, EC_{out} and CA1 during which CA1 is forced to represent the correct output (given the symmetry between EC_{in} and EC_{out}). The goal of the network is to adjust its weights such that activity in the minus phase resembles activity during the plus phase, thus constantly reducing the error between minus and plus phases.

Stimulus presentation

The network was trained on four inputs, presented in a random order, each simulating an object. Inputs were generated using a custom Python code (available at https://github.com/frdarya/PS_expectation, together with all necessary code to reproduce the data reported below). Of 54 feature dimensions, every object had 12 active units (clamped to 1) and three units clamped to 0.9, representing a previously presented cue (Schapiro et al., 2017). Two cues were randomly set, one for "manmade" objects and one for "natural" objects, simulating the conditions used in Experiment 1. Therefore, two stimuli were associated with a manmade category and the other two with a natural category. During a training trial (100 processing cycles), the network was presented with a single object and the associated cue for two minus phases and a single plus phase. A full set of trials including all four objects completed an epoch. The network was trained for 10 epochs and tested following the last one.

At test, no changes were applied to connection weights, keeping the network at a retrieval-like neutral state. In every test trial, we presented the network an input and recorded the activation levels of each hidden layer. We used cue units to simulate the "expected" and "unexpected" conditions used in Experiment 1. Expected trials had the same cue–object association the network was trained on, whereas in unexpected trials the cue was flipped (i.e., a manmade cue for a natural object, and vice versa). Critically, because the network has learned the contingency of both cues, flipping them at test does not create a novel input, but rather an unexpected pairing of learned inputs (for items tested without a cue at all, see Fig. 6-1). In addition to testing each encoded object, we also created parametrically manipulated similar foils (expected and unexpected). These foils were created pseudorandomly by varying the percentage of overlap between the target (original item) and the foils, ensuring flipped units between foils were also independent (e.g., units changed in F50 were not the same as those changed in F67). The very high similarity foil (F85) had an 85% overlap with the target, F67 had 67% overlap with the target, F50 had 50% overlap, whereas the lowest similarity foil (F33) had only 33% overlap with the encoded target. Therefore, for each object, 10 test trials were used (1 target and 4 foils, each tested once as expected and once as unexpected), resulting in 40 test trials altogether (for illustration of the stimuli used, see Fig. 2*B*).

Lesions

To examine the contribution of pattern separation to differences in representational dissimilarity between expected and unexpected inputs, we simulated lesions in TSP. The following projection strengths were set to 0: $EC_{in} \rightarrow DG$, $EC_{in} \rightarrow CA3$, $DG \rightarrow CA3$, $CA3 \rightarrow CA3$, and $CA3 \rightarrow CA1$. These lesions reveal the independent function of MSP to the representation created in CA1 without any pattern separation from DG and CA3 (as DG and CA3 are "turned off", they cannot be examined). As noted by Schapiro et al. (2017) MSP lesions do not reveal the independent contribution of TSP, as MSP serves communications between EC and TSP, therefore they were not examined.

Statistical analyses

To assess representational similarity between targets and foils in both expectation conditions, we recorded unit activity in each hidden layer, per test trial, and calculated distances as 1–Pearson correlations. Therefore, each batch had one 40×40 representational dissimilarity matrix (RDM), capturing all trials across conditions and objects. To examine differences between expected and unexpected trials, we averaged across objects and computed two 5×5 RDMs; one correlating expected-expected trials (EE), another unexpected-unexpected (UU; symmetrical matrices, meaningless diagonals). It is important to note this analysis overcomes the inherent difference in number of slots changed between expected and unexpected inputs. For example, the correlation between unexpected target and unexpected F85 is only driven by their perceptual similarity, as both of them had the same number of slots changed. Therefore, by comparing correlations from EE RDMs to ones from UU RDMs, we could examine how the perceptual similarity between inputs was modulated by the expectation manipulation. For comparisons between theoretical RDMs and simulated data we used the RSAtoolbox (Nili et al., 2014) implemented in MATLAB 2018a (MathWorks). Kendall's Tau A was computed to assess the second-order correlation between categorical models and data RDMs; Pearson's r was used for comparison between models of input similarity (i.e., multiple computational models) and data RDMs (Nili et al., 2014). All tests were FDR-corrected and analyses were done per batch and then averaged across them, using each randomly reinitialized network batch as a random effect.

Results

Experiment 1

Rule-learning performance

To ensure contextual expectations were established, we examined participants' performance during the rule-learning task (Fig. 3). Average performance, excluding four participants who did not reach criterion (above chance in the first half of task, and >75% accuracy in the second half of task), was 84.7% (SD = 8%), significantly above chance ($t_{(23)} = 21.02$, $p < 0.001$, Cohen's $d = 4.29$). Next, we examined reaction times (RTs) for the predictions, within the 1 s decision window. There was no significant

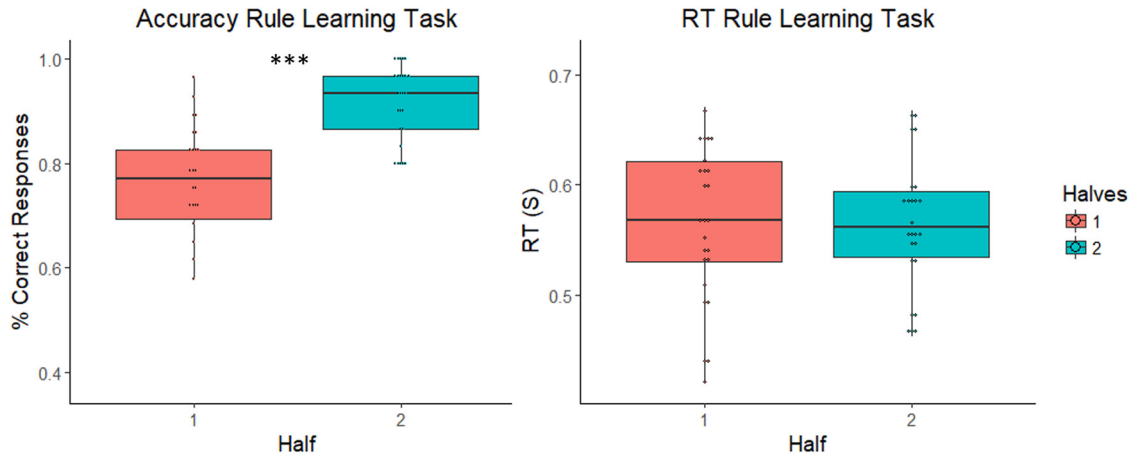

Figure 3. Rule learning task results. Accuracy in rule learning task as a factor of task progression. Participants learned the cue-category contingency well and their overall accuracy was above chance (left). RT was restricted to 1 s, so no differences in mean RT were observed, but there was a reduction in variance as the task progressed and participants learned the contingency (right). Unless otherwise stated, error bars reflect SE. ***$p < 0.001$.

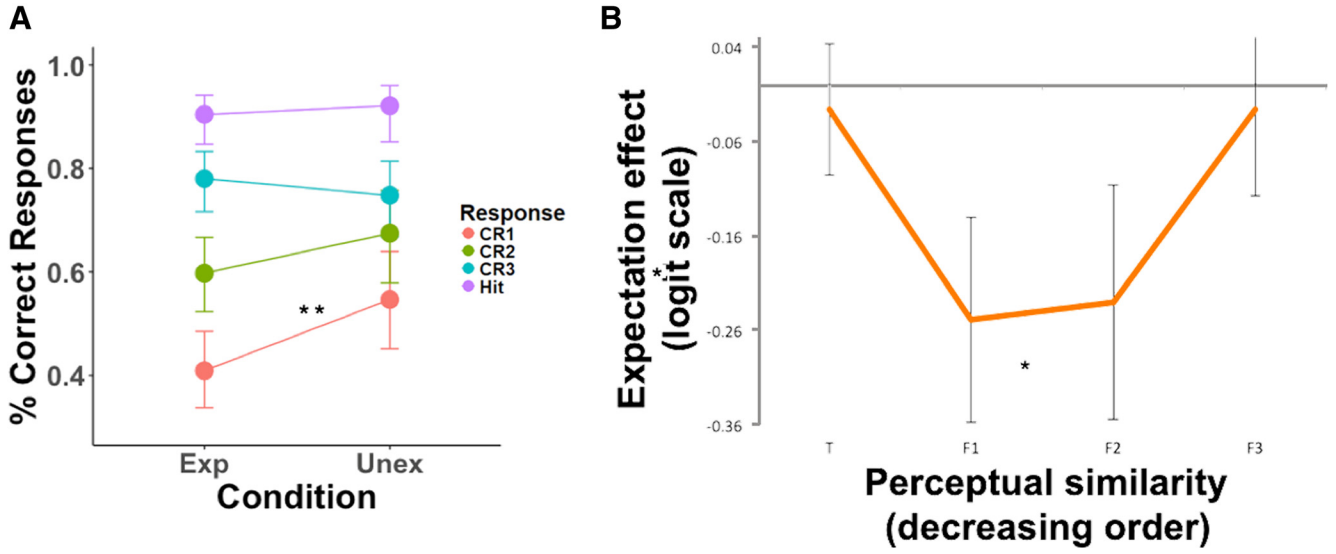

Figure 4. Recognition performance. ***A***, Raw recognition decisions (mixed-effects logistic regression). Contextually unexpected high (F1) similarity foils were correctly rejected (CR1) more than expected ones. No differences were observed for low similarity foils (F3) or targets. Error bars reflect 95% confidence interval (see Figure 4-1 for recognition performance in last trials). ***B***, Expectation effect as a function of input similarity. A quadratic effect of level of similarity on the unexpected-expected recognition difference (negative logit-transformed values correspond to positive percentages; see Figure 4-2 for raw percentages). *$p < 0.05$, **$p < 0.01$.

difference in mean RT between the first and second halves of the task ($t_{(23)} = 0.434$, $p = 0.669$).

Recognition memory performance

As contextual expectation was manipulated at encoding, we looked at the first set event at retrieval (Fig. 4), to examine contextual expectation effects without any interference from other similar foils (for analysis of subsequent set events, see Fig. 4-1). Using a mixed-effects logistic regression to predict hit rate for versus unexpected targets [in lme4: hit ~ expectation + (1|participant)], we did not observe a significant difference between expectation conditions ($\beta = 0.213$, $\chi^2_{(1)} = 0.371$, $p = 0.542$). As performance for first target approached ceiling (average at 90%), we also examined RT to test whether the lack of contextual expectation effect was masked by the high level of accuracy. If hits also benefit from the violation of expectation, we should see faster response times for these trials. Running a similar mixed-effect linear model [in lme4: Hit_RT ~ expectation + (1|participant)], we did not find an effect of expectation ($\beta = 0.023$, $\chi^2_{(1)} = 0.253$, $p = 0.614$).

To examine recognition performance for foils, we devised a mixed-effects binary logistic regression with participant as a random intercept [in lme4: CR ~ foil similarity × expectation + (1|participant)], which allowed us to account for each participant's unique bias (for grouped corrected recognition results, see Fig. 4-2). This mixed-effects analysis revealed main effects of item ($\chi^2_{(2)} = 77.2$, $p < 0.001$) and expectation ($\chi^2_{(1)} = 7.22$, $p = 0.007$), and a significant item by expectation interaction $\chi^2_{(2)} = 6.05$, $p = 0.048$). *post hoc* contrasts revealed that F1 foils that were similar to unexpected targets were better recognised than ones similar to expected targets ($\beta = 0.55$, $z = 2.68$, $p = 0.007$). Responses to F2 items were not significantly modulated by expectation at encoding ($\beta = 0.33$, $z = 1.53$, $p = 0.12$), as were responses for F3 items ($\beta = -0.18$, $z = 0.807$, $p = 0.42$).

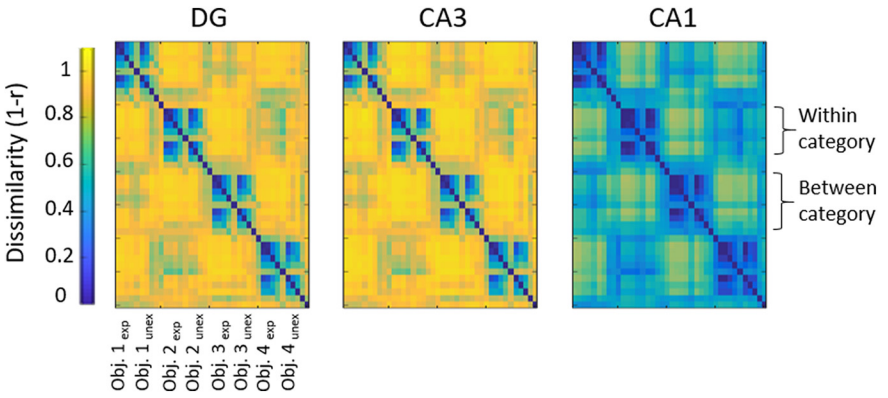

Figure 5. Overall 40 × 40 RDMs. These represent all of the test trials presented to the network and their corresponding dissimilarity (1- Pearson's *r*), split by hidden layer. The four objects used are represented along the diagonal, with the split in the middle of each one representing the expected and unexpected conditions tested. Objects 1 and 3 shared the same category, as did Objects 2 and 4. Between-category dissimilarity is higher than within-category. Warmer colors indicate more dissimilarity (see Figures 5-1 and 5-2, for the second-order correlation matrices between the data RDM and the theoretical RDMs).

To further explore the difference between contextual expectation conditions as a function of similarity, we calculated the mean correct response rate for each first set event per participant and logit-transformed these values. In cases where there were values of 1 or 0, these were replaced by calculating ε, the smallest value observed in the sample divided by 2, and replacing 0 with ε, and 1 with $1 - \varepsilon$. This method ensure the original form of the transformation is kept, but allows conversion of 1 and 0 s to values that match the overall shape of the logit function. We then subtracted the expected logit-transformed values from the unexpected ones. This calculation represents the magnitude of the boost (or decline) in recognition performance, along the continuum of input similarity (with the target being 100%, and similarity parametrically decreasing from target to F1 and across F1, F2, and F3), and is not the memory decision per se The quadratic contrast for the effect was significant ($t_{(23)} = 2.03$, $p = 0.046$). These results therefore show a selective increase in the correct rejections of highly similar foils similar to contextually unexpected targets.

These results offer the first direct evidence for the role of expectation in engaging a pattern separation mechanism. We found the contextual manipulation exerted an effect only on the first set event, when there was no interference from other similar events. High similarity foils from sets whose target was unexpected at encoding, produced more correct rejections at retrieval, compared with foils whose target was expected at encoding. Targets, as well as low similarity foils, were unaffected by this manipulation. As the retrieval task progressed, and participants were exposed to other set events, the effect of contextual expectation diminished, and we did not observe differences between expectation conditions for these later events. Exposure to other set events leads to interference due to multiple memory traces ("was it this apple or the apple I saw a few trials ago?") and mnemonic attributions ("I said old to the previous apple, but now I see this one is the target"). Therefore, it might be the case that the PS mechanism involved in the expectation effects is now needed to resolve this interference to make a correct recognition decision.

Experiment 2

To establish that the network learned to represent the different categories used (manmade and natural), we first tested whether between-category dissimilarity was greater than within-category. A 40 × 40 RDM was computed representing all of the trials tested, sorted by object and expectation condition. As can be seen in Figure 5, within-category dissimilarity is lower than the between-category one. To formally demonstrate that the network's representations capture the input's category, as well as the presence of different objects and expectation conditions, we created four theoretical models and computed the second-order similarity between each of them and the data RDM. The theoretical models represented main effects of category (manmade vs natural), object (A vs B vs C vs D), expectation (vs Unex) and a random model. We found the second-level similarity between the data RDM and the category and object RDMs was significant in all layers (all p values < 0.01), whereas the expectation and random models could not explain the data (for second-order correlations, see Fig. 5-1). This pattern indicates that the overall structure of the data was most affected by the object's own identity and the category to which it belonged.

We also tested whether the network successfully captured the amount of overlap between the trained inputs and tested foils. This test assesses whether the network tracks the continuous gradient of input similarity, or whether it responds in a binary way (vs dissimilar) by setting a threshold of similarity. Sets (target and foils) from both expectation conditions were averaged across objects and compared with three theoretical models. The first model represented a scaled response, mirroring the percentage of overlap between inputs. The second model represented a thresholded response, with high overlap foils (F85 and F67) being more similar to target than F50 and F33. Finally, the third model represented a random distribution. While both theoretical models outperformed the random model, using second-order Pearson's correlations, the scaled model showed the best performance (Fig. 5-2). This can also be seen intuitively in the RDMs (Fig. 6), with decreasing levels of overlap associated with more representational dissimilarity in all hippocampal layers. This effect was most prominent in DG and CA3 layers, consistent with their role in pattern separation.

Next, we sought to examine whether patterns of activations elicited by unexpected items differ from those elicited by expected ones. In a univariate analysis, each cell (i.e., pairwise distance) in the UU RDM was subtracted from its EE counterpart. We then subjected the data to a two-tailed t test against a value of 0 (no difference between conditions). Across the three layers, all differences were significantly smaller than 0 (all p values < 0.001) indicating larger dissimilarity in UU compared with EE RDMs. However, as can be seen in Figure 6, these differences varied in magnitude, both between layers and overlap levels. To assess which RDM cells had the most variability between EE and UU, we turned to a multivariate approach. Singular value decomposition (SVD) analysis is a dimensionality-reduction technique used to expose the substructure of a given dataset. In this analysis, a matrix or an image can be compressed while preserving its most informative features. Using this method we were

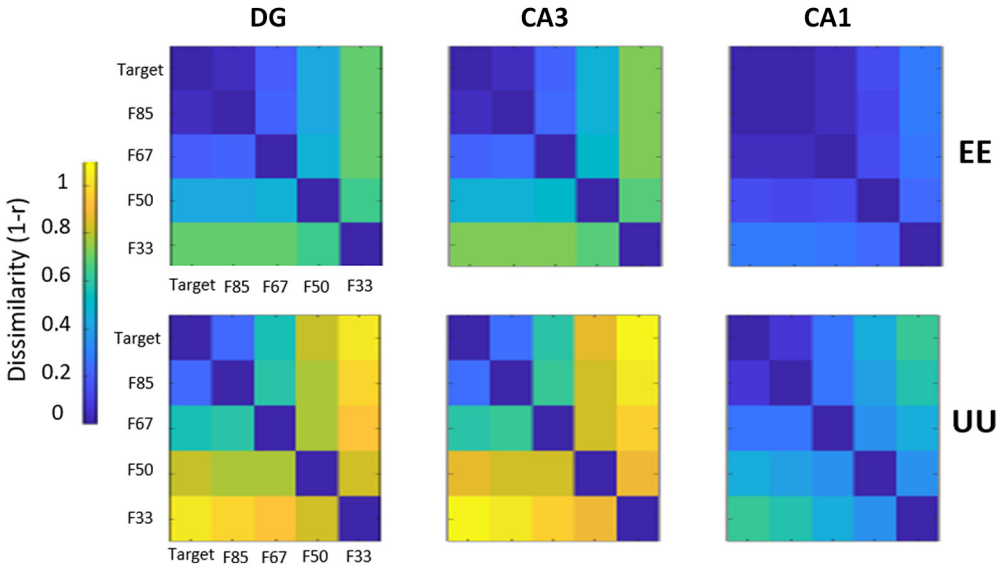

Figure 6. 5 × 5 Object RDMs split by condition. Top row, RDMs for EE, whereas the bottom row shows UU ones. There is greater representational dissimilarity in UU compared with EE in all hidden layers. Warmer colors indicate more dissimilarity (see Figure 6-1 for the RDMs for trials presented without a cue).

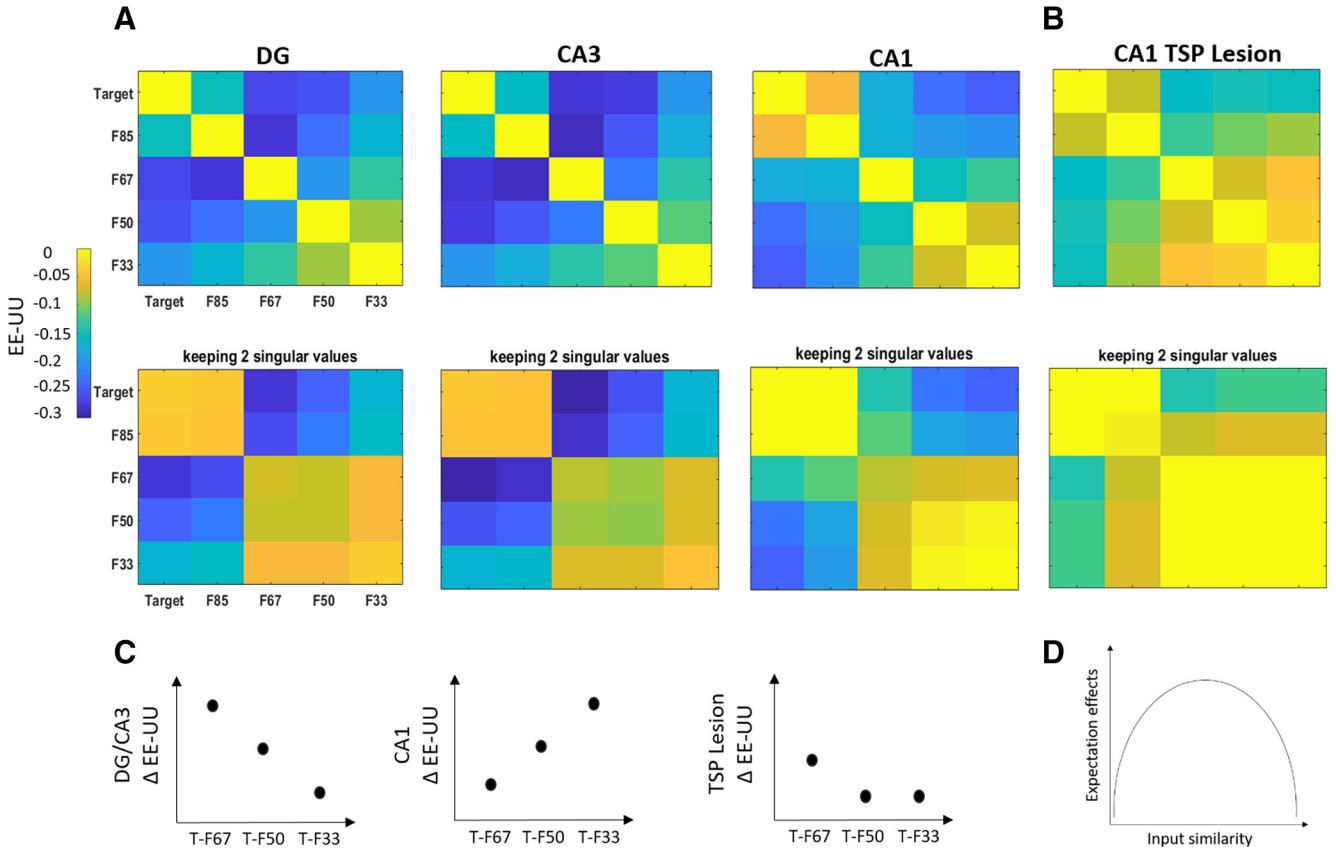

Figure 7. ***A***, SVD analysis for EE-UU RDMs. Top row, The raw EE-UU RDMs for each layer; bottom row, the SVD results, keeping the most informative two singular values. Minimal differences were observed for very high and low overlap, whereas moderate-high levels showed a graded difference between conditions. ***B***, SVD analysis for TSP lesion. When inputs from DG and CA3 are muted, the observed differences between EE and UU RDMs diminish, and expand to higher levels of overlap. Warmer colors indicate less EE–UU difference. ***C***, Illustration of DG/CA3 and CA1 moderate effects. Whereas DG and CA3 show a negative linear relationship between moderate to high levels of overlap and EE-UU dissimilarity, CA1 shows the opposite effect. ***D***, Conceptual illustration of top-down and bottom-up interaction. Expectation effects peak at moderate to high levels of input similarity.

able to capture higher-order structural differences between expected and unexpected RDMs (i.e., identifying the most informative cells in the matrix). For each layer, we used the EE-UU subtracted RDM. This RDM was then decomposed using SVD, keeping two singular values. Figure 7*A* depicts the reconstructed matrix, keeping two singular values, reflecting a compression ratio of 0.88 (warmer colors indicate no difference between conditions).

In DG and CA3, when target-foil similarity is very high (F85), the differences between EE and UU matrices were minimal. However, when comparing target and F67 to F33 foils, a graded difference emerged; as target-foil overlap decreases, the representational dissimilarity between expected and unexpected stimuli diminished (EE and UU are represented similarly). When mid and low similarity foils were compared with one another (as foils share little overlap among themselves, these represent low similarity), there were again minimal differences between the expected and unexpected conditions. This analysis suggests the representational difference between expected and unexpected RDMs in DG and CA3 was characterized by an interaction between level of overlap and expectation condition. In CA1, the SVD analysis revealed even smaller differences between EE and UU for target F85 and between mid-low similarity foils. However, when comparing target and F67 to F33 foils, the pattern in CA1 mirrored that of DG and CA3. As overlap decreased, the difference between EE and UU increased. To ensure these higher-order structures were significantly different from arbitrary noise, we randomly shuffled each layer's RDM 10,000 times. In each permutation we computed Spearman's rank correlation between the shuffled and the original matrices. The aggregated correlations were compared with 0 (no correlation between shuffled and data matrices) using a two-sided t test. None of these tests was significant (all p values > 0.1), indicating the subtracted EE-UU RDMs were not significantly correlated with random noise.

Finally, we examined whether the difference between EE and UU is reduced when TSP is muted. To do so, we ran the same SVD procedure on the lesion data (only in CA1, as projections from DG and CA3 are set to 0) and compared the result to the data presented above. As can be seen in Figure 7*B*, whereas the overall structure in CA1 remained similar, the differences between expected and unexpected diminished (warmer colors in the lesion data). Furthermore, fewer differences were observed for F85 correlations, compared with data without the lesion. To quantify the changes introduced by the TSP lesion, we compared the raw EE-UU RDM in CA1 from both datasets using a two-tailed t test. For the T-F85 correlation difference, lesioned CA1 showed larger differences between EE and UU than the spared network $t_{(199)} = 3.251$, $p = 0.001$. The T-F67 difference was significantly smaller in the lesion data $t_{(199)} = -2.02$, $p = .0044$. All other differences were in the same direction, with EE-UU differences being smaller in the lesioned TSP data (all p values < 0.001).

In Experiment 2, we tested how interactions between top-down expectation and bottom-up perceptual inputs are represented in different layers of a hippocampal neural network model. We found the network learned to represent the perceptual similarity of inputs and that this representation was modulated by a violation of the learned cue-item contingency. Unexpected inputs showed greater representational dissimilarity compared with expected items. This effect was modulated by the degree of overlap between the originally encoded item and the current input, as well as the layer tested. In very high and low levels of item similarity, minimal differences were observed between expected and unexpected inputs. However, in moderate-high levels, the magnitude and direction of the effects differed between DG/CA3 and CA1. First, in DG and CA3 differences between expected and unexpected inputs were greater than in CA1. Furthermore, the overall structure of this interaction differed between layers, with DG and CA3 showing a positive linear effect, whereas CA1 showed a negative one. Finally, a lesion to TSP resulted in smaller differences between EE and UU RDMs in CA1, suggesting pattern-separated inputs from DG and CA3 drive this interaction. Together, these results offer a mechanistic neural account for the behavioral results obtained in Experiment 1.

Discussion

In two experiments, we examined whether expectation-modulated memory is driven by a pattern separation mechanism, supporting improved memory for unexpected information. In Experiment 1, we found an interaction effect between contextual expectation at encoding and item similarity. Specifically, high similarity foils, from sets whose target was unexpected at encoding, produced more correct rejections at retrieval, compared with foils whose target was expected at encoding. In Experiment 2, we used a neural network model emulating the behavioral task to examine the neural mechanism. We found the network learned to represent the perceptual similarity of inputs and that this representation was modulated by a violation of the learned cue-item contingency. Again, we found an expectation by item similarity interaction. Unexpected inputs showed greater representational dissimilarity than expected inputs; this effect was modulated by the degree of overlap between the originally encoded item and the current input, as well as the layer tested. In moderate to high levels of input similarity, maximal differences were observed between expected and unexpected items. Finally, a lesion to TSP resulted in smaller differences between EE and UU RDMs in CA1, suggesting pattern-separated inputs from DG and CA3 drive this interaction. Overall, our results demonstrate that violation of expectation elicits an adaptive mechanism that is sensitive to the level of similarity between bottom-up inputs and existing representations.

Previous studies have shown that contextual expectation plays an important role in modulating hippocampal involvement and behavioral memory responses (Kumaran and Maguire, 2007; Kafkas and Montaldi, 2015; Frank et al., 2018). However, the extent of these effects, and the underlying mechanism supporting them, remain unclear. Here we show contextual surprise does not enhance memory indiscriminately, but specifically aids the disambiguation of overlapping inputs. This suggests that the beneficial effect of expectation-violation is selective and stems from pattern separation engagement rather than a more general memory boost, which could be mediated by an extra-hippocampal circuit. We found that violations of contextual expectation at encoding support the ability to correctly identify similar foils as such. Additionally, we did not observe any beneficial effect of contextual expectation on hits or correct rejections of lower similarity foils. Converging results were found in Experiment 2. Again, for very high (target F85) or low (foil-foil) overlap, EE and UU differences were minimal in all layers, but the expectation manipulation elicited distinct patterns across layers in moderate-high levels of overlap, with unexpected items showing increased dissimilarity compared with expected ones. The increased dissimilarity points to the formation of a more distinct representation, complementary to subsequent memory effects, where targets that are successfully remembered show greater encoding-retrieval pattern similarity (Xue, 2018). Together, these results strongly suggest that a modulating mechanism is used when unexpected events occur, leading to a shift toward pattern separation of the surprising information.

Whereas very high or low levels of overlap offer too much or too little perceptual interference, respectively, moderate-high overlap between encoded and current inputs require engagement

of pattern separation (Norman and O'Reilly, 2003; Yassa and Stark, 2011). Indeed, more distinct representations and better recognition performance were observed for unexpected F1. This suggests that at peak perceptual interference levels, unexpected items engage pattern separation more than expected ones. This finding dovetails with the suggestion that unexpected information can bias hippocampal computations toward an encoding state (Lisman and Grace, 2005; Shohamy and Wagner, 2008; Axmacher et al., 2010; Gruber et al., 2018; Kafkas and Montaldi, 2018b). Our findings are the first to show the consequences of such a shift, creating more distinct memory representations for unexpected items. In Experiment 1, the diminished effect of contextual expectation at encoding on performance for later events from a set, potentially points to the sensitivity of this boost to performance; the more immediate interference from other, perceptually similar, set events during retrieval masks the beneficial effect of the enhanced encoding of unexpected targets. As more exemplars are presented during retrieval, the task of comparing each one to the originally encoded target, relying on successful PS, becomes more demanding, comprising multiple comparisons with other set events presented at retrieval.

In Experiment 2, we also found dissociations in responses between the different hippocampal subfields. In CA1, as target-foil overlap decreased, the representational dissimilarity between expected and unexpected items increased. In DG and CA3, on the other hand, a mirrored effect was observed; as target-foil overlap decreased, the representational dissimilarity between expected and unexpected stimuli diminished. This suggests that in CA1 expectation-violation had a more prominent effect in the lower levels of the similarity scale (T-F33), whereas in DG/CA3 expectation-violation exerted the largest effect in T-F67. CA1 receives pattern-separated inputs from CA3 and projections from EC, reflecting retrieval of existing representations (Norman and O'Reilly, 2003). Given these connections, CA1 has been postulated to act as a match/mismatch detector (Chen et al., 2011; Elfman et al., 2014; Valenti et al., 2018). CA1 expectation effects are most pronounced at lower levels of perceptual interference (T-F33 overlap) for which PS is not critical. This result indicates CA1 representations are indeed sensitive to mismatches, but more so for the coarse perceptual differences (i.e., stimulus novelty of low-similarity foils in relation to the target; for example, a new breed of dog) compared with memory-based mismatches, for high-similarity foils that engage PS (for example, 2 dogs of the same breed, but pictured from a slightly different angle). Together, the contrasting results from DG/CA3 and CA1 offer an interesting view on the division of labor between these hippocampal subfields when it comes to mismatches originating both from bottom-up and top-down sources. Our findings suggest top-down unexpected information is represented more distinctly in DG/CA3 when bottom-up interference is also high, whereas CA1 is more responsive to top-down manipulations when interference from bottom-up inputs is lower.

Finally, we examined how CA1 representations change when a TSP lesion is introduced, muting inputs from DG and CA3. Despite this lesion, CA1 representations still managed to capture the higher-order structure of the data, in accordance with previous models (Schapiro et al., 2017). Nevertheless, without the sparse inputs from TSP, differences between expected and unexpected items diminished considerably in CA1. This suggests that improved memory performance for unexpected items is driven by pattern separation in DG/CA3. It is also important to note how these findings fit with the dynamic nature of hippocampal processing. Previous research suggests pattern separation and completion occur at different stages of the theta cycle (Hasselmo et al., 2002; Kunec et al., 2005). Although the learning algorithm of the hippocampal model used here reflects these oscillatory properties (Ketz et al., 2013; Schapiro et al., 2017), because of the nature of the task simulated, our manipulation and tests were conducted at retrieval, where weights are not adjusted. Therefore, although our layer-by-layer analysis strongly indicates that pattern separation underlies the effects reported here, future electrophysiological research could examine the online dynamics of these effects (Axmacher et al., 2006; Hanslmayr et al., 2016) and how they relate to pattern completion (e.g., failure to pattern complete to target). Based on the differential pattern of responses across layers and the lesion data (most prominent effects in DG/CA3), as well as previous findings (Meeter et al., 2004; Axmacher et al., 2010; Douchamps et al., 2013; Gruber et al., 2018), our model predicts that violation of expectation would modulate the hippocampal theta cycle toward encoding. Future studies should examine whether similar results are observed when expectation is manipulated during retrieval, and the directionality of such effects.

In conclusion, our highly novel results offer a hippocampus-driven mechanism for the interaction between top-down expectation and bottom-up perceptual inputs and its effect on memory representations and performance. When the level of overlap between existing and current input is moderate to high, violation of expectation helps disambiguate these representations by engaging pattern separation, resulting in improved memory performance. However, at the two extremes, very high and low levels of overlap, this mechanism is not engaged, for different reasons. When overlap is very high, this mechanism could be turned on, and, alas, fail to exert an effect (failure of pattern separation). Conversely, when the level of overlap is low, disambiguation is redundant, and therefore further engagement of pattern separation is unnecessary. The findings reported here have important implications for our understanding of the neural bases of top-down and bottom-up interactions in memory.

References

Aisa B, Mingus B, O'Reilly R (2008) The Emergent neural modeling system. Neural Netw 21:1146–1152.

Axmacher N, Cohen MX, Fell J, Haupt S, Dümpelmann M, Elger CE, Schlaepfer TE, Lenartz D, Sturm V, Ranganath C (2010) Intracranial EEG correlates of expectancy and memory formation in the human hippocampus and nucleus accumbens. Neuron 65:541–549.

Axmacher N, Mormann F, Fernández G, Elger CE, Fell J (2006) Memory formation by neuronal synchronization. Brain Res Rev 52:170–182.

Bar M (2009) The proactive brain: memory for predictions. Philos Trans R Soc Lond B Biol Sci 364:1235–1243.

Bates D, Mächler M, Bolker B, Walker S (2015) Fitting linear mixed-effects models using lme4. J Stat Soft 67:1–48.

Chen J, Olsen RK, Preston AR, Glover GH, Wagner AD (2011) Associative retrieval processes in the human medial temporal lobe: hippocampal retrieval success and CA1 mismatch detection. Learn Mem 18:523–528.

Colgin LL, Moser EI, Moser MB (2008) Understanding memory through hippocampal remapping. Trends Neurosci 31:469–477.

Douchamps V, Jeewajee A, Blundell P, Burgess N, Lever C (2013) Evidence for encoding versus retrieval scheduling in the hippocampus by theta phase and acetylcholine. J Neurosci 33:8689–8704.

Elfman KW, Aly M, Yonelinas AP (2014) Neurocomputational account of memory and perception: thresholded and graded signals in the hippocampus. Hippocampus 24:1672–1686.

Fox J (2003) Effect displays in R for generalised linear models. J Stat Soft 8:15.

Frank D, Montaldi D, Wittmann B, Talmi D (2018) Beneficial and detrimental effects of schema incongruence on memory for contextual events. Learn Mem 25:352–360.

Frank D, Gray O, Montaldi D (2019) SOLID-Similar object and lure image database. Behav Res Methods 52:151–161.

Gruber MJ, Hsieh LT, Staresina BP, Elger CE, Fell J, Axmacher N, Ranganath C (2018) Theta phase synchronization between the human hippocampus and prefrontal cortex increases during encoding of unexpected information: a case study. J Cogn Neurosci 30:1646–1656.

Hanslmayr S, Staresina BP, Bowman H (2016) Oscillations and episodic memory: addressing the synchronization/desynchronization conundrum. Trends Neurosci 39:16–25.

Hasselmo ME, Bodelón C, Wyble BP (2002) A proposed function for hippocampal theta rhythm : separate phases of encoding and retrieval enhance reversal of prior learning. Neural Comput 14:793–817.

Kafkas A, Montaldi D (2015) Striatal and midbrain connectivity with the hippocampus selectively boosts memory for contextual novelty. Hippocampus 25:1262–1273.

Kafkas A, Montaldi D (2018a) Expectation affects learning and modulates memory experience at retrieval. Cognition 180:123–134.

Kafkas A, Montaldi D (2018b) How do memory systems detect and respond to novelty? Neurosci Lett 680:60–68.

Ketz N, Morkonda SG, O'Reilly RC (2013) Theta coordinated error-driven learning in the hippocampus. PLoS Comput Biol 9:e1003067.

Kumaran D, Maguire EA (2007) Match mismatch processes underlie human hippocampal responses to associative novelty. J Neurosci 27:8517–8524.

Kunec S, Hasselmo ME, Kopell N (2005) Encoding and retrieval in the CA3 region of the hippocampus: a model of theta-phase separation. J Neurophysiol 94:70–82.

Leal SL, Yassa MA (2018) Integrating new findings and examining clinical applications of pattern separation. Nat Neurosci 21:163–173.

Leutgeb JK, Leutgeb S, Moser MB, Moser EI (2007) Pattern separation in the dentate gyrus and CA3 of the hippocampus. Science 315:961–966.

Lisman JE, Grace AA (2005) The hippocampal-VTA loop: controlling the entry of information into long-term memory. Neuron 46:703–713.

Long NM, Lee H, Kuhl BA (2016) Hippocampal mismatch signals are modulated by the strength of neural predictions and their similarity to outcomes. J Neurosci 36:12677–12687.

McClelland JL, McNaughton BL, O'Reilly RC (1995) Why there are complementary learning systems in the hippocampus and neocortex. Psychol Rev 102:419–457.

Meeter M, Murre JMJ, Talamini LM (2004) Mode shifting between storage and recall based on novelty detection in oscillating hippocampal circuits. Hippocampus 14:722–741.

Nili H, Wingfield C, Walther A, Su L, Marslen-Wilson W, Kriegeskorte N (2014) A toolbox for representational similarity analysis. PLoS Comput Biol 10:e1003553.

Norman KA, O'Reilly RC (2003) Modeling hippocampal and neocortical contributions to recognition memory: a complementary-learning-systems approach. Psychol Rev 110:611–646.

O'Reilly RC, Munakata Y (2000) Computational explorations in cognitive neuroscience: Understanding the mind by simulating the brain. Cambridge, MA: MIT.

O'Reilly RC, McClelland JL, Reilly RCO (1994) Hippocampal conjunctive encoding, storage, and recall: avoiding a trade-off. Hippocampus 4:661–682.

Peirce JW (2007) PsychoPy-Psychophysics software in Python. J Neurosci Methods 162:8–13.

Pilly PK, Howard MD, Bhattacharyya R (2018) Modeling contextual modulation of memory associations in the hippocampus. Front Hum Neurosci 12:442.

R Development Core Team (2008) R: a language and environment for statistical computing. Available at http://www.r-project.org.

Schapiro AC, Turk-Browne NB, Botvinick MM, Norman KA (2017) Complementary learning systems within the hippocampus: a neural network modelling approach to reconciling episodic memory with statistical learning. Philos Trans R Soc Lond B Biol Sci 372:20160049.

Shohamy D, Wagner AD (2008) Integrating memories in the human brain: hippocampal-midbrain encoding of overlapping events. Neuron 60:378–389.

Valenti O, Mikus N, Klausberger T (2018) The cognitive nuances of surprising events: exposure to unexpected stimuli elicits firing variations in neurons of the dorsal CA1 hippocampus. Brain Struct Funct 223:3183–3211.

Wickham H (2009) ggplot2. New York: Springer.

Xue G (2018) The neural representations underlying human episodic memory. Trends Cogn Sci 22:544–561.

Yassa MA, Stark CEL (2011) Pattern separation in the hippocampus. Trends Neurosci 34:515–525.