Hippocampal Repulsion as a Function of Memory Interference and Subjective Beliefs

by

Wanjia Guo

A dissertation accepted and approved in partial fulfillment of the

requirements for the degree of

Doctor of Philosophy

in Psychology

Dissertation Committee:

Brice A. Kuhl, Chair

Sarah DuBrow, In Memoriam

Ben Hutchinson, Core Member

Ulrich Mayr, Core Member

Luca Mazzucato, Institutional Representative

University of Oregon
Spring 2024

Dissertation Abstract

Wanjia Guo

Doctor of Philosophy in Psychology

Title: Hippocampal Repulsion as a Function of Memory Interference and Subjective Beliefs.

Resolving memory interference is critical for performing essential tasks in our daily life. The hippocampus is believed to play a critical role in distinguishing similar memories. This dissertation focused on understanding the mechanisms of hippocampal repulsion, which stands for when the representations of two overlapping memories are actively pushed away to be represented less similarly to each other than non-overlapping memories. The first chapter draws direct connections between repulsion and behavioral expression of memory interference resolution. In particular, we show that the timing when repulsion happens is exactly when memory inference is resolved. The second chapter focuses on why repulsion occurs. It provides evidence that repulsion can occur with distinct internal states, even when external stimuli are identical. The third chapter focuses on how the intensity of repulsion changes with different levels of experience and shows that repulsion is not simply a linear process that accumulates with learning. Instead, it is transient and subsides after memory interference is resolved. Across all 3 chapters, the hippocampus was also segmented into subfields, and we consistently found CA3/DG to be the region that showed the repulsion effect, but not CA1.

Acknowledgements

None of the study in this dissertation would have been possible without the generous guidance and help from **Dr. Brice Kuhl**. With the deepest influence in my scientific career, I am grateful to have the experience learning from his rigorous way of conducting research, creativity in designing experiment and analysis, and articulate writing. Above all, I want to thank him for his compassion and understanding through various events, to support my growth not only as a trainee, but also as an individual. I will always be grateful for his offer from six years ago, and for his immense support during these six years.

During the time in the lab, I want to thank **Alex** for being a close friend and ally. I often think back to the time when we had frequent dog dates, and the endless conversations we had. I am grateful for everyone whom I met in the Kuhl Lab – **Max, Zhifang, Futing, Lindsay, Anisha, and Subin** – for your help and companion since we met. I also want to thank all the fellow graduate students and research assistants I met while at the University of Oregon, for all the interesting conversations and fun time together.

In the department, I received valuable support from **Dr. Sarah DuBrow**, with whom I had the courage to share my fear and emotions; **Dr. Ben Hutchinson**, who always gives me the most unexpected questions after my presentations; and **Dr. Ulrich Mayr**, who generously agreed to be in my dissertation committee and provided extremely useful feedback. I also want to send a special thank you to **Lori**, as she is always there when I had last minute questions, and always made sure I am on time for the countless logistic requirements.

Dr. **Anthony Wagner** and Dr. **Beth Mormino** had the biggest influence in my scientific career before I entered graduate school. Thank you for taking an undergrad into your lab, and patiently teaching her to become a qualified graduate student. I want to especially thank Beth, for all the hours you spent seating with me to write Python script, and the empathetic personal interactions in our daily life.

I also want to thank **my parents**, who had the courage to support my wild dream of studying abroad at the age of 18, even though I had never been to the US before. I now realized that underlying this decision, it shines the strong belief you had in me and the unspoken sacrifices you had to make. Thank you.

Lastly, I want to thank **Eric**, who is my husband, my closest friend, and my strongest advocator. Thank you for always taking a first look at all my presentations and writings, for the endless discussions regarding hippocampus repulsions – to a level that you started talking about it in your dreams as well. Mostly, thank you for always being there, to support me and to listen to me with the kindest heart. I would not be the person I am today without you. Thank you for being the constant in my life.

Table of Content

LIST OF FIGURES

Introduction

The hippocampus is essential for forming **episodic memories** (long-term memories for experiences and the context in which they occurred). However, many experiences contain overlapping elements – common places, people, emotions, etc. – and this overlap is the primary source of **memory interference,** the tendency for one memory to 'get in the way' of another memory. Thus, a fundamental computational priority for the hippocampus is to minimize interference between overlapping memories[1,2].

While this topic has been central to theoretical models of the hippocampus and its role in episodic memory[3–5], these models draw heavily from rodent studies that have focused on how the hippocampus represents spatial information, as opposed to episodic memories. The rationale for these cross-species comparisons is that *distinguishing similar episodic memories may rely on the same hippocampal computations that allow for similar spatial environments to be distinguished.* In rodent studies, it is well established that when spatial environments change in subtle ways, hippocampal representations of the environment **remap** in such a way that individual cells (place cells) change their spatial firing preferences. It has been speculated that remapping may also occur for human episodic memories and play a critical role in resolving memory interference[6]. While several recent human fMRI studies have provided initial evidence of remapping-like phenomena in the human hippocampus[7–9], these studies have not directly related remapping to episodic memory interference. Thus, there remains a critical need to establish whether and how hippocampal remapping relates to episodic memory interference.

‘Remapping’ usually refers to a populational change in place cells’ firing patterns or frequency in rodent studies[10–13]. However, using human MRI, we can only collect neural activation evoked by various stimuli in the brain at a voxel level. We call the array of activation level for all the voxels within a given brain region an activation pattern. With distinct task manipulation and stimuli, we can then repeatedly measure the different levels of activity patterns in different brain regions. For the current dissertation, to connect remapping with human episodic memory, we will refer ‘remapping’ more broadly as changes in hippocampal activation patterns.

An appealing reason for remapping to occur is to decorrelate initial overlapped signals and therefore resolve memory interference[6]. In human fMRI studies and computational models, pattern separation is a well-studied hippocampal coding scheme that is believed to facilitate memory interference resolution[1,2]. As a result of pattern separation, two initially overlapped memories became orthogonalized to each other[2,14]. In here, orthogonalization indicates a complete decorrelation. In other words, two overlapping memories were represented as distinct from each other as two non-overlapping memories.

However, another appealing way to consider the result of remapping is to not just decorrelate initially overlapping signals, but *actively push* them as far away from each other as possible to systematically decrease overlap between overlapping memories. Indeed, recent human fMRI studies identified such a hippocampal coding mechanism, which is called hippocampal repulsion[15–17]. Repulsion is distinct from pattern separation because it is characterized by overlapping memories being repulsed away from each other, even further than non-overlapping memories. In other words, the result of pattern

separation is orthogonalization, same level of distance between overlapping memories or between non-overlapping memories, but the result of repulsion showed a more exaggerated difference between overlapping memories than non-overlapping memories. Studies found that repulsion happened with learning[15], and that increased levels of repulsion are related to better behavioral performance in memory interference resolution[15,16].

A few questions remained to be answered to better understand the mechanisms of hippocampal repulsion and to draw the link between remapping and repulsion.

Firstly, even though previous literature has linked repulsion with better behavioral performance in resolving memory interference[16], there still exists a gap to directly measure when hippocampus representations become repulse while participants learn to resolve episodic memory interference. Testing whether, and if so when, hippocampal repulsion happens during the learning process is critical to understand the relationship between repulsion and behavior.

Secondly, the question of what exactly causes repulsion in hippocampus representational patterns is still unsolved. Previous literature has shown that similar memories that suffered from memory interference could lead to hippocampal repulsion[15,16]. Moreover, recent computational models and rodent studies have suggested that remapping might depend on changes in internal states[18]. Borrowing from this line of thinking it certainly stands to reason that repulsion might also be influenced by internal states, but direct evidence supporting this idea is still absent, especially with human fMRI. Understanding whether repulsion can happen in human hippocampus with

changes in internal states will hugely advance our understanding regarding the mechanisms underlying this phenomenon.

Lastly, research investigating how repulsion changes as a function of experience is still limited. Even though previous studies have shown that exposure and learning are related to hippocampal repulsion[15], no study to our knowledge has controlled for how different levels of experiences influence the intensity of the repulsion effect. A direct test aimed at quantifying how the strength of repulsion changes with various levels of experience would be critical in helping us fill this particular gap in our understand of the repulsion effect.

In my dissertation projects, we aim to answer the above 3 questions to shed lights towards the underlying mechanisms of the repulsion effect. We firstly establish the relationship between episodic memory interference and hippocampal repulsion (Chapter 1). Specifically, we link rodent remapping with the “repulsion effect” in the human hippocampus. We showed that the timing of repulsion is temporally coupled with the exact timing when memory interference is being resolved, and that the effect is limited to CA2/3/Dentate Gyrus (DG), a subfield within the hippocampus[19], but not in CA1, or other visual-attention regions such as early visual cortex or parahippocampal place area.

Then, in Chapter 2, we will focus on the underlying mechanisms of the repulsion effect, or, what causes the repulsion between overlapping events. We have already noted that recent literatures with rodents suggests that a change in internal belief can lead to remapping, even if the external environments were hold the same[20,21]. We designed a spatial task that uses ideas from rodents’ T-maze to test whether repulsion

occurs when internal beliefs changes with human fMRI. This chapter shows that repulsion is not driven by external environments or distinct visual inputs. Instead, repulsion can happen if participants hold the internal beliefs that two similar (competing) events are distinct.

Lastly, in Chapter 3, we investigate how different levels of experience influence both the level and the timing of repulsion. Specifically, we separated stimuli into high and low training groups and found that the high training group showed repulsion in the early half of the experiment but faded away afterwards. On the other hand, low training group didn't show repulsion until the latter half of the experiment. This result suggests that even though repulsion is depended to experience, but repulsion is also not simply a linear accumulation as experience increased.

Chapter 1. Abrupt hippocampal remapping signals resolution of memory interference

1.1 Abstract

Remapping refers to a decorrelation of hippocampal representations of similar spatial environments. While it has been speculated that remapping may contribute to the resolution of episodic memory interference in humans, direct evidence is surprisingly limited. We tested this idea using high-resolution, pattern-based fMRI analyses. Here we show that activity patterns in human CA3/dentate gyrus exhibit an abrupt, temporally-specific decorrelation of highly similar memory representations that is precisely coupled with behavioral expressions of successful learning. The magnitude of this learning-related decorrelation was predicted by the amount of pattern overlap during initial stages of learning, with greater initial overlap leading to stronger decorrelation. Finally, we show that remapped activity patterns carry relatively more information about learned episodic associations compared to competing associations, further validating the learning-related significance of remapping. Collectively, these findings establish a critical link between hippocampal remapping and episodic memory interference and provide insight into why remapping occurs.

1.2 Introduction

The hippocampus is critical for forming long-term, episodic memories[1–3]. However, one of the fundamental challenges that the hippocampus faces is that many experiences are similar, creating the potential for memory interference[4,5]. In rodents, it is well established that minor alterations to the environment can trigger sudden changes in hippocampal activity patterns—a phenomenon termed remapping[6,7]. An appealing

possibility is that hippocampal remapping also occurs in human episodic memory, allowing for similar memories to be encoded in distinct activity patterns that prevent interference[8]. At present, however, there remains an important gap between evidence of place cell remapping in the rodent hippocampus and episodic memory interference in humans. To bridge this gap, it is informative to consider how properties of place cell remapping, as demonstrated in the rodent hippocampus, might translate to episodic memory interference in humans.

One of the most important properties of remapping in the rodent hippocampus is that it is characterized by abrupt transitions between representations[9–12]. These abrupt transitions, evidenced by decorrelations in patterns of neural activity, have most typically been observed as a function of the degree of environmental change[9,11]. However, abrupt remapping can also occur as a function of experience with a new environment[10,12]. Evidence of experience-dependent remapping[6,13,14] suggests an important point: that remapping fundamentally reflects changes in internal representations, as opposed to changes in environmental states[15,16]. An emphasis on internal representations lends itself well to human episodic memory in that it suggests that hippocampal remapping should occur as memories change. More specifically, this perspective makes the critical prediction that when two events are highly similar, hippocampal remapping will occur if, and when, corresponding memories become distinct. To date, a number of human fMRI studies have observed experience-dependent decorrelations in hippocampal representations of similar memories[17–22] and/or have linked hippocampal pattern overlap to memory interference[20,23–25]. However, to test the prediction that hippocampal activity patterns abruptly remap when

memory interference is resolved it is necessary to precisely track changes in memories as a function of temporally-specific changes in hippocampal representations. Critically, standard approaches of averaging fMRI data across different stimuli (memories), stimulus repetitions, and/or participants can easily obscure or wash out abrupt changes in hippocampal representations if the timing of those changes varies across memories or participants.

Evidence of place cell remapping in rodents also motivates specific predictions regarding the relative contributions of hippocampal subfields, with a major distinction being between CA3/dentate gyrus and CA1[8,26,27]. In general, CA3 and dentate gyrus are thought to be more important than CA1 for discriminating between similar stimuli[16,28,29,27,30,31] and remapping has been shown to occur more abruptly in CA3 than in CA1[10,12,32]. High-resolution fMRI studies in humans have also tested for and confirmed distinctions between these subfields. For example, fMRI studies have found that, relative to CA1, activity patterns/responses in CA3 and dentate gyrus are more sensitive to subtle differences between similar memories[17,19,33,34] or spatial environments[23,24,33]. Moreover, responses in human CA3/dentate gyrus have specifically been linked to behavioral discrimination of similar memories[23,24,35]. However, these studies have not directly established a link between temporally abrupt remapping in CA3/dentate gyrus and changes in corresponding episodic memories.

Here, we tested whether the resolution of interference between highly similar episodic memories is associated with an abrupt remapping of activity patterns in human CA3/dentate gyrus. We used an associative memory paradigm in which participants learned and were repeatedly tested on associations between scene images and object

images[20]. The critical design feature was that the set of scene images included pairs of highly similar scenes (**Fig. 1a**). These scene pairmates were intended to elicit associative memory interference. Across six rounds of learning, we tracked improvement in associative memory for each set of pairmates while also continuously tracking representational changes indexed by fMRI. Specifically, after each associative memory test round, participants were shown each scene image one at a time (exposure phase) which allowed us to measure the activity pattern evoked by each scene and, critically, the representational distance between scene pairmates. To preview, we find that behavioral expressions of memory interference resolution are temporally-coupled to abrupt, stimulus-specific remapping of human CA3/dentate gyrus activity patterns. This remapping specifically exaggerated the representational distance between similar memories. In additional analyses, we show that the magnitude of remapping that individual memories experienced was predicted by the degree of initial pattern overlap among CA3/dentate gyrus representations and that remapped CA3/dentate gyrus representations carried increased and highly specific information about learned episodic associations.

1.3 Results

Participants completed six rounds of the experimental paradigm while inside an fMRI scanner. Each round included a study phase, an associative memory test phase, and a scene exposure phase (**Fig. 1b)**. fMRI scanning was only conducted during the exposure phases. During the study phases, participants viewed scene-object associations one at a time. During the associative memory test phases, participants were shown scenes, one at a time, along with two very similar object choices (e.g., two

guitars); one object was the target (i.e., the object that had been paired with the current scene) and the other object was the competitor (i.e., the object that had been paired with the scene pairmate). After selecting an object, participants indicated their confidence (high or low). During exposure phases, participants were shown each scene, along with novel scenes, and made a simple old/new judgment (mean ± 95% CI: d' = 5.40 ± 0.88; one-sample t-test vs. 0: t_{30} = 12.58, p < 0.001, Cohen's d = 2.26).

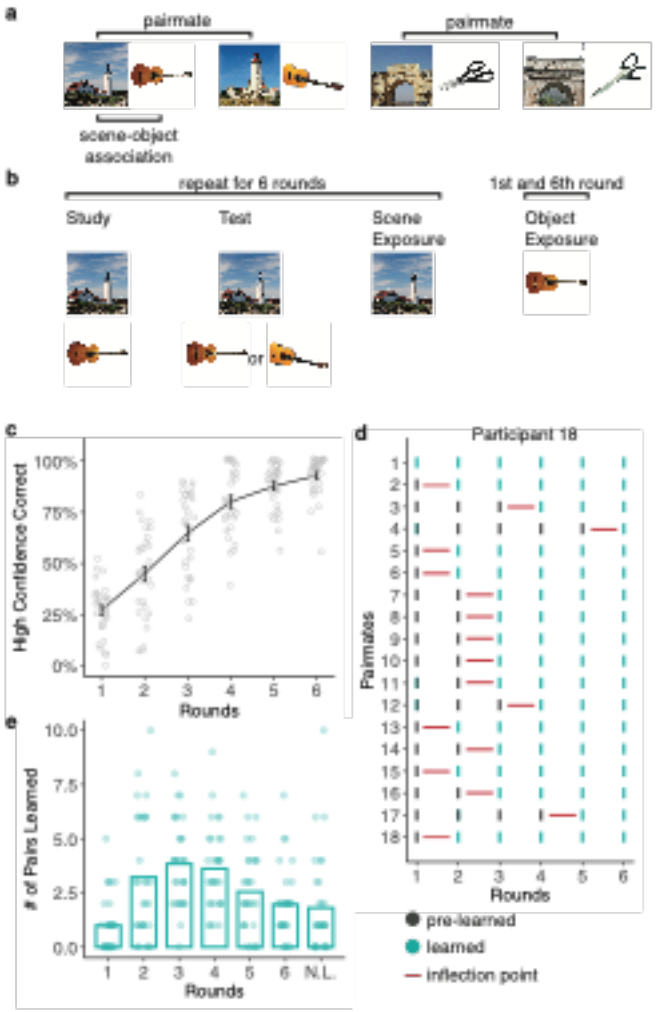


Figure 1. Experimental Design and Behavior. **a.** Participants learned 36 scene-object associations. The 36 scenes comprised 18 scene pairmates which consisted of highly similar image pairs (e.g., 'lighthouse 1' and 'lighthouse 2'). Scene pairmates were also associated with similar objects (e.g., 'guitar 1' and 'guitar 2'). **b.** Participants completed 6 rounds of study, test, and exposure phases. During the study, participants viewed scenes and associated objects. During the test, participants were presented with scenes and had to select the associated object from a set of two choices, followed by a confidence rating (high or low confidence; not shown). During exposure, scenes (rounds 1-6) or objects (round 1 and 6) were presented and participants made an old/new judgment. fMRI data were only collected during the scene and object exposure phases. **c.** Mean percentage of high confidence correct responses for each test round. **d.** Data from a representative participant showing the 'inflection point' in learning (red horizontal line), for each pairmate. The inflection point was defined as the point at which participants transitioned to high-confidence correct retrieval for both scenes within a pairmate—a transition from 'pre-learned' (black) to 'learned' (aqua). **e.** Mean number of scene pairmates that transition to a learned state at each round. N.L. indicates pairmates that were never learned. Notes: Data are presented as mean values ± S.E.M., n = 31 independent participants. Source data are provided as a Source Data file.

Behavior.

During the associative memory test phases, participants chose the correct object with above-chance accuracy in each of the 6 rounds (round 1: t_{30} = 2.65, p = 0.013, d = 0.48, CI = [0.56 ± 0.05]; round 2: t_{30} = 7.77, p < 0.001, d = 1.40, CI = [0.69 ± 0.05];

round 3: t_{30} = 10.78, p < 0.001, d = 1.94, CI = [0.79 ± 0.05]; round 4: t_{30} = 19.39, p < 0.001, d = 3.48, CI = [0.87 ± 0.04]; round 5: t_{30} = 29.71, p < 0.001, d = 5.34, CI = [0.92 ± 0.03]; round 6: t_{30} = 41.38, p < 0.001, d = 7.43, CI = [0.95 ± 0.02]; one-sample t-tests vs. 50%). Accuracy markedly improved across rounds (main effect of round: $F_{1,30}$ = 318.86, p < 0.001, η^2 = 0.91). The rate of choosing the correct object with high-confidence also robustly increased across rounds, from a mean of 27.15% ± 4.71% in round 1 to 92.83% ± 3.58% in round 6 (main effect of round: $F_{1,30}$ = 574.44, p < 0.001, η^2 = 0.95; **Fig. 1c**). See **Supplementary Figure 1** for test accuracy for each set of scene pairmates.

To test whether hippocampal remapping was temporally coupled with the resolution of memory interference, we identified, for each participant and for each set of pairmates, the learning round in which scene-object associations were recalled with high confidence (for both scenes in a pairmate). We refer to this timepoint as the 'learned round' (LR; see Methods). Of critical interest for our remapping analyses was the correlation of activity patterns evoked by scene images during the learned round (LR) with activity patterns evoked immediately prior to the learned round (LR-1). We refer to this transition (from pre-learned to learned) as the 'inflection point' (IP) in learning (**Fig. 1d**). For example, if the learned run for a particular set of pairmates was round 4, then the inflection point was the transition from round 3 to 4. Our rationale for correlating activity patterns from the learned round with activity patterns from the preceding round (LR-1) was that this correlation would capture the critical change in hippocampal representations (remapping) that putatively supports learning.

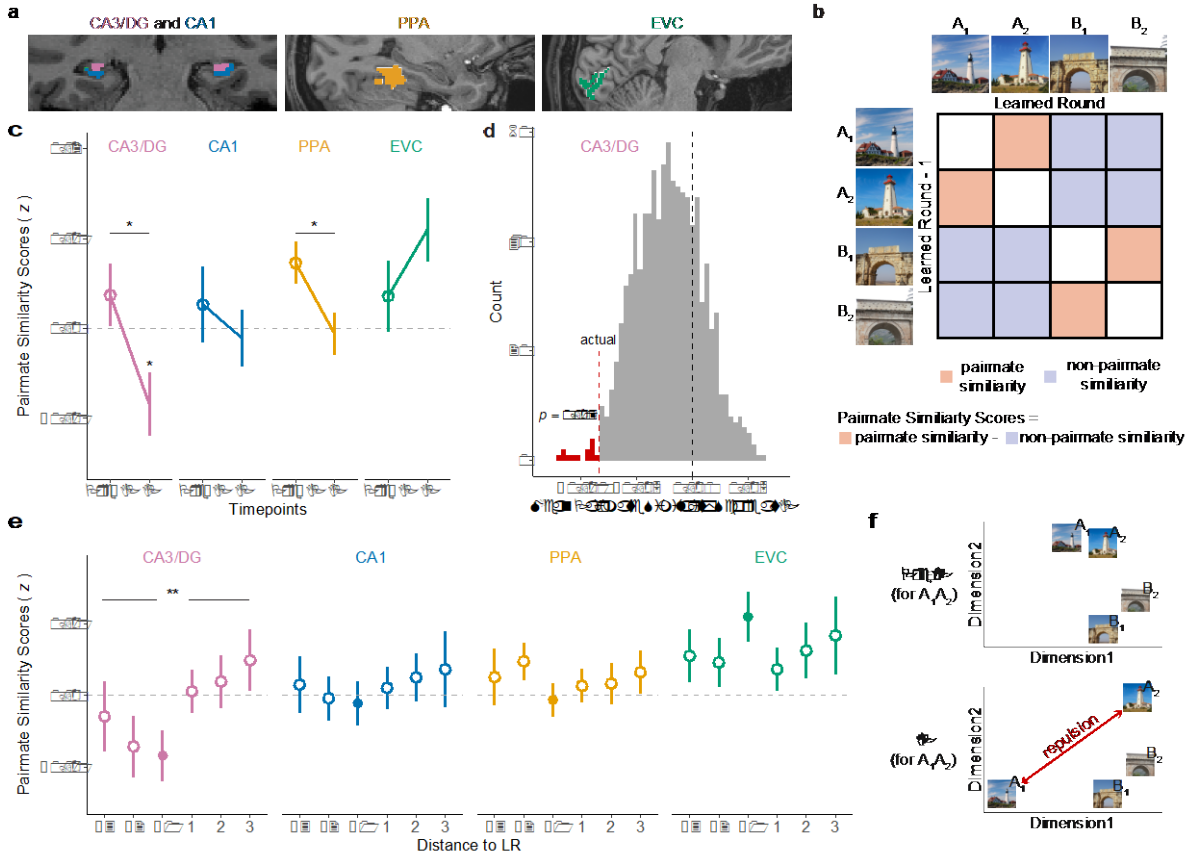


Figure 2. Pairmate similarity scores change at the behavioral inflection point.

a. Regions of interest included CA3/dentate gyrus (CA3/DG, pink) and CA1 (blue) in the hippocampus, the parahippocampal place area (PPA, yellow), and early visual cortex (EVC, green). **b.** Correlation matrix illustrating how pairmate similarity scores were computed at the behavioral inflection point. See Methods for details. **c**. Pairmate similarity scores at the behavioral inflection point (IP) and just prior to the inflection point (pre-IP) across different regions of interest (ROIs). Pairmate similarity scores significantly varied by ROI (p = 0.009, repeated measures ANOVA) and there was a significant interaction between ROIs and behavioral state (p = 0.037, repeated measures ANOVA). In CA3/DG, pairmate similarity scores at the IP were significantly lower than 0 (p = 0.025, two-tailed one sample t-test) and significantly lower than the pre-IP state (p = 0.033, two-tailed paired samples t-test). In PPA, pairmate similarity scores decreased from pre-IP to IP (p = 0.030, two-tailed paired samples t-test). **d.** A permutation test (1,000 iterations) was performed by shuffling, within participants, the mapping between the behavioral inflection point and scene pairmates. In CA3/dentate gyrus the actual mean group-level pairmate similarity score at the IP was lower than 98.70% of the permuted mean similarity scores (p = 0.013, one-tailed permutation test). **e.** Pairmate similarity scores calculated by correlating the learned round (LR) with each of the three preceding rounds (– distance to LR) and each of the three succeeding rounds (+ distance to LR). [Note: the inflection point was defined as the correlation between the LR and the immediately preceding round (LR - 1); the inflection points are depicted by filled circles and are the same values as in **c**]. In CA3/dentate gyrus, pairmate similarity scores were significantly lower when the LR was correlated with preceding rounds compared to succeeding rounds (p = 0.006, two-tailed paired samples t-test). The difference was not significant for any other ROIs (CA1: p = 0.435; PPA: p = 0.955; EVC: p = 0.760; two-tailed paired sample t-tests). **f.** Conceptual illustration of a decrease in pairmate similarity scores from pre-IP to IP. In the pre-IP state (top panel), A_1 and A_2 are nearby in representational space. In the IP state (bottom panel), the representational distance between A_1 and A_2 has been exaggerated. When pairmates (e.g., A_1 and A_2) are farther apart in representational space than non-pairmates (e.g., A_1 and B_2) the pairmate similarity score will be negative (i.e., pairmate similarity < non-pairmate similarity), consistent with a repulsion of competing representations. Notes: * p < .05, ** p < .01. No correction for multiple comparisons was applied given the a priori predictions for CA3/dentate gyrus. Data are presented as mean ± S.E.M. and all data reflect n = 31 independent participants. Source data are provided as a Source Data file.

Remapping in CA3/dentate gyrus is time-locked to the inflection point in learning.

For our fMRI analyses, our primary focus was on pattern similarity between scene pairmates. Pattern similarity was measured by correlating patterns of fMRI activity evoked by each scene during the scene exposure phases. Pairmate similarity was defined as the correlation between activity patterns evoked by scene pairmates (e.g., 'lighthouse 1' and 'lighthouse 2'; **Fig. 2b**). Correlations between scenes that were not pairmates (e.g., 'lighthouse 1' and 'arch 2'; **Fig. 2b**) provided an important baseline measure of non-pairmate similarity. We refer to the difference between these two measures (pairmate – non-pairmate similarity) as the *pairmate similarity score*[20]. A positive pairmate similarity score would indicate that visually similar scenes (e.g., two lighthouses) are associated with more similar neural representations than two unrelated scenes. Critically, because pairmate similarity scores are a relative measure, they can be directly compared across different brain regions[36]—something that would be inadvisable with raw correlation values. For all pattern similarity analyses, correlations were always performed across learning rounds—for example, correlating 'lighthouse 1' at the learned round (LR) with 'lighthouse 2' at LR-1. This ensured independence of fMRI data[37], but was also intended to capture transitions in hippocampal representations (remapping).

Following a prior study that used similar stimuli and analyses[20], fMRI analyses targeted the following regions of interest (ROIs): hippocampus, parahippocampal place area (PPA), and early visual cortex (EVC). PPA and EVC served as important control regions indexing high-level (PPA) and low-level (EVC) visual representations. We did not anticipate that these regions would demonstrate learning-related remapping. Within

the hippocampus, we leveraged our high-resolution fMRI protocol to segment the hippocampus body into subfields comprising CA1 and a combined CA3/dentate gyrus (see Methods). Motivated by past empirical findings[33,38] and theoretical models[8], we predicted that remapping would occur in CA3/dentate gyrus. More specifically, we predicted that CA3/dentate gyrus remapping would occur at the inflection point (IP) in learning. To test this prediction, we compared pairmate similarity scores at the inflection point to pairmate similarity scores at a timepoint just prior to the inflection point (pre-IP). Whereas pairmate similarity scores at the inflection point were based on correlations between activity patterns from the learned round (LR) and the preceding round (LR-1), pairmate similarity scores at the pre-IP were based on correlations shifted back one step in time: i.e., between LR-1 and LR-2. Thus, whereas the inflection point captured the transition from pre-learned to learned, the pre-IP was an important reference point that corresponded to a 'non-transition' (pre-learned to pre-learned).

An ANOVA with factors of behavioral state (pre-IP, IP) and ROI (CA3/dentate gyrus, CA1, PPA, and EVC) revealed a significant main effect of ROI ($F_{3,90} = 4.08$, $p = 0.009$, $\eta^2 = 0.04$), reflecting overall differences in pairmate similarity scores across ROIs. Scores were numerically lowest in CA3/dentate gyrus and numerically highest in EVC. There was no main effect of behavioral state ($F_{1,30} = 2.71$, $p = 0.110$, $\eta^2 = 0.01$), indicating that learning did not have a global effect on representational structure across ROIs. Critically, however, the interaction between behavioral state and ROI was significant ($F_{3,90} = 2.95$, $p = 0.037$, $\eta^2 = 0.04$), indicating that learning differentially influenced pairmate similarity scores across ROIs.

Within CA3/dentate gyrus, pairmate similarity scores were significantly lower at the inflection point than the pre-IP (t_{30} = -2.24, p = 0.033, d = 0.40, CI = [-0.012 ± 0.011]), consistent with our prediction that remapping would specifically occur at the behavioral inflection point. Importantly, we also confirmed via permutation test (see Methods) that CA3/dentate gyrus pairmate similarity scores at the inflection point were lower than would be expected if the mapping between pairmates and inflection points was shuffled within participants (p = 0.013, one-tailed; **Fig. 2d**).

Notably, CA3/dentate gyrus pairmate similarity scores not only decreased at the inflection point, but they were significantly below 0 at the inflection point (t_{30} = -2.36, p = 0.025, d = 0.19, CI = [-0.008 ± 0.007]). In other words, pairs of scenes with high visual similarity were represented as less similar than completely unrelated scenes in CA3/dentate gyrus. While seemingly counterintuitive, several recent fMRI studies have also found that, in certain situations, hippocampal pattern similarity is lower for similar than dissimilar events[18,20,33]. This has led to the proposal that similarity triggers a repulsion of hippocampal representations. That is, just as physical proximity triggers repulsion of like magnetic poles, representational proximity triggers repulsion of similar memories (**Fig. 2f**). The present results, however, provide critical evidence that this repulsion is time-locked to—and may, in fact, underlie—the resolution of interference between competing memories.

In CA1, pairmate similarity scores did not significantly differ by learning state (t_{30} = -0.72, p = 0.474, d = 0.13, CI = [0.004 ± 0.01]) or differ from 0 either at the pre-IP (t_{30} = -0.63, p = 0.531, d = 0.11, CI = [0.003 ± 0.009]) or inflection point (t_{30} = -0.34, p = 0.735, d = 0.06, CI = [-0.001 ± 0.006]). In PPA, pairmate similarity scores decreased

from pre-IP to the inflection point (t_{30} = -2.28, p = 0.030, d = 0.41, CI = [0.008 ± 0.007]), with scores significantly greater than 0 at the pre-IP (t_{30} = 3.14, p = 0.004, d = 0.56, CI = [0.007 ± 0.005]) but not different from 0 at the inflection point (t_{30} = -0.26, p = 0.798, d = 0.05, CI = [-0.0006 ± 0.005]). In EVC, pairmate similarity scores did not significantly vary by learning state (t_{30} = -1.39, p = 0.175, d = 0.25, CI = [-0.007 ± 0.01]); but there was a numerical increase from pre-IP to the inflection point, with scores significantly above 0 at the inflection point (t_{30} = 3.13, p = 0.004, d = 0.56, CI = [0.01 ± 0.007]) but not at the pre-IP (t_{30} = 0.92, p = 0.366, d = 0.16, CI = [0.004 ± 0.008]).

The qualitative difference between CA3/dentate gyrus and EVC is notable in that, at the inflection point, these regions exhibited fully opposite representational structures: scene pairmates were more similar than non-pairmates in EVC, but less similar than non-pairmates in CA3/dentate gyrus. This finding parallels prior evidence of opposite representational structures in the hippocampus and EVC[18,20] and argues against the possibility that CA3/dentate gyrus 'inherited' representational structure from early visual regions. More generally, pairmate similarity scores markedly varied across the four ROIs at the inflection point ($F_{3,90}$ = 8.73, p < 0.001, η^2 = 0.14), but not at the pre-IP ($F_{3,90}$ = 0.33, p = 0.804, η^2 = 0.008), underscoring the influence of learning on representational structure.

For the preceding fMRI analyses, the inflection point was defined as the correlation between the learned round (LR) and the immediately preceding round (LR-1). To more fully characterize how the representational state at the learned round compared to other rounds, we additionally correlated representations at the learned round to representations at LR-2 and LR-3 (i.e., other rounds that preceded the learned

round) and also correlated the learned round with LR+1, LR+2, and LR+3 (rounds that followed the learned round). Within CA3/dentate gyrus, pairmate similarity scores were significantly lower when correlating the learned round with rounds that preceded learning compared to rounds that followed learning (t_{30} = -2.98, p = 0.006, d = 0.54, CI = [-0.009 ± 0.006]; **Fig. 2e**; see **Supplementary Figure 2** for related analyses). This asymmetry indicates that CA3/dentate gyrus representations expressed at the learned round were systematically biased away from the initial representational position of competing memories. More generally, these data support the idea of an abrupt representational change (remapping) in CA3/dentate gyrus that was time-locked to the specific round at which learning occurred for individual pairmates. For CA1, PPA, and EVC, there were no significant differences in pairmate similarity scores when correlating the learned round to rounds that preceded learning vs. followed learning (CA1: t_{30} = -0.79, p = 0.435, d = 0.14, CI = [-0.002 ± 0.006]; PPA: t_{30} = 0.06, p = 0.955, d = 0.01, CI = [-0.0002 ± 0.005]; EVC: t_{30} = 0.31, p = 0.760, d = 0.06, CI = [-0.001 ± 0.006]; **Fig. 2e**).

Overlap of CA3/dentate gyrus representations triggers remapping.

The fact that pairmate similarity scores in CA3/dentate gyrus were negative at the inflection point (**Fig. 2c**) suggests that learning-related remapping involved an active repulsion of competing hippocampal representations (**Fig. 2f**). Conceptually, the key feature of a repulsion account is that separation of hippocampal representations is a reaction to initial overlap among memories[25]. Here, because we measured representational states throughout the course of learning, we were able to test this hypothesis directly. Specifically, we tested the prediction that relatively greater pairmate similarity scores (i.e., higher overlap between memories) at a given timepoint is

associated with relatively lower pairmate similarity scores (i.e., lower overlap between memories) at a successive timepoint.

To test this hypothesis, we first translated the 6 learning rounds into 5 'timepoints' (see Methods). Each timepoint corresponded to the set of scene pair similarity scores obtained by correlating activity patterns across consecutive learning rounds [e.g., timepoint 1 = *r*(round 1, round 2)]. These scores reflected the representational structure at each timepoint (i.e., which pairmates were relatively similar, which pairmates were relatively dissimilar). We then rank correlated the pairmate similarity scores across successive timepoints [*r*(timepoint 1, timepoint 2)]. Whereas a positive rank correlation would indicate that representational structure is preserved across time points, a negative rank correlation would indicate that representational structure is inverted across time points. Critically, an inversion of representational structure is precisely what would be predicted if initial overlap among activity patterns (i.e., high pairmate similarity scores) triggers a repulsion of activity patterns (i.e., low pairmate similarity scores).

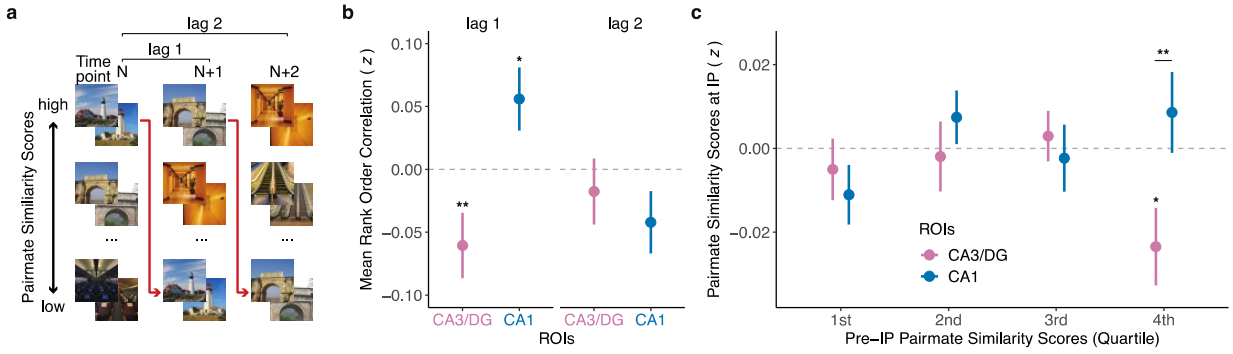


Figure 3. Representational structure across timepoints.

a. Schematic illustration showing the rank order of scene pairmates based on pairmate similarity scores at various time points (N, N+1, N+2). If scene pairmates with relatively high pairmate similarity scores at a given timepoint are systematically associated with relatively low pairmate similarity scores at a succeeding time point (red arrows), this will produce a negative rank correlation. **b.** Mean rank order correlations of pairmate similarity scores across timepoints for CA3/dentate gyrus (CA3/DG, pink) and CA1 (blue). Lag 1 correlations reflect correlations between a given timepoint and an immediate succeeding timepoint (e.g., timepoints 2 and 3). Lag 2 correlations reflect correlations between a given timepoint and a timepoint two steps away (e.g., timepoints 2 and 4). At lag 1, there was a negative correlation in CA3/dentate gyrus (p = 0.006, two-tailed one sample t-test), but a positive correlation in CA1 (p = 0.043, two-tailed one

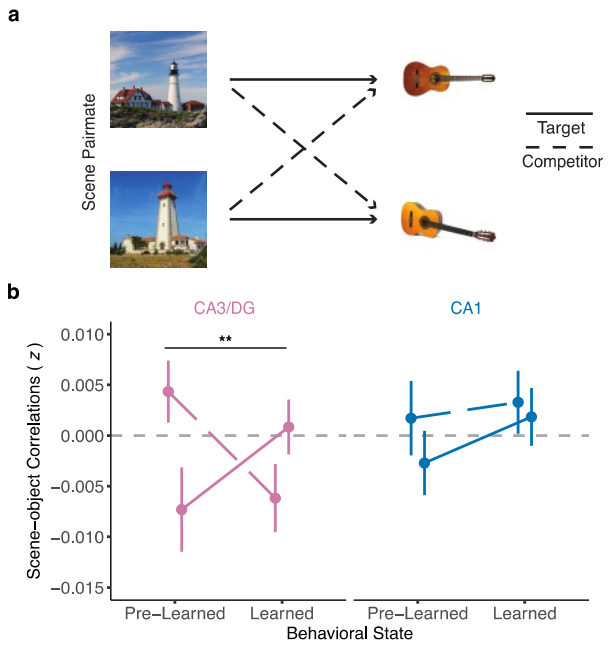
sample *t*-test). At lag 2, correlations were not significant in either CA3/dentate gyrus (p = 0.485, two-tailed one sample *t*-test) or CA1 (p = 0.120, two-tailed one sample *t*-test) indicating that correlations in representational structure were specific to temporally adjacent rounds. **c.** Pairmate similarity scores at the inflection point (IP) as a function of relative pairmate similarity scores in the pre-IP state (1st quartile = lowest similarity, 4th quartile = highest similarity). Pairmate similarity scores in CA3/dentate gyrus were significantly lower than CA1 (p = 0.008, two-tailed paired samples *t*-test) and significantly below 0 (p = 0.017, two-tailed one sample *t*-test) for pairmates with the highest pre-IP similarity (4th quartile). See **Supplementary Figure 4** for the distributions of pre-IP pairmate similarity scores. Notes: * $p < .05$, ** $p < .01$. No correction for multiple comparisons was applied given the a priori predictions for CA3/dentate gyrus. Data are presented as mean ± S.E.M. and all data reflect n = 31 independent participants. Source data are provided as a Source Data file.

Notably, the rank correlation in CA3/dentate gyrus was significantly negative (t_{30} = -2.99, p = 0.006, d = 0.54, CI = [-0.06 ± 0.04], **Fig. 3b**). In contrast, the rank correlation in CA1 was significantly positive (t_{30} = 2.11, p = 0.043, d = 0.38, CI = [0.06 ± 0.05]). The difference between CA3/dentate gyrus and CA1 was also significant (t_{30} = 3.73, $p < 0.001$, d = 0.67, CI = [0.12 ± 0.06]). Importantly, the negative correlation in CA3/dentate gyrus cannot be explained by regression to the mean (see Methods). As a control, we also tested correlations at a lag of 2 [r(timepoint N, timepoint N+2)]; however lag 2 correlations did not significantly differ from 0 for either CA3/dentate gyrus (t_{30} = -0.71, p = 0.485, d = 0.13, CI = [-0.02 ± 0.05]) or CA1(t_{30} = -1.60, p = 0.120, d = 0.29, CI = [-0.04 ± 0.05]). The interaction between lag (1, 2) and ROI (CA3/dentate gyrus, CA1) was also significant ($F_{1,30}$ = 7.09, p = 0.012, η^2 = 0.06). Thus, for CA3/dentate gyrus and CA1, representational structure at a given time point specifically predicted representational structure at a successive timepoint. Rank correlations did not differ from 0 in either PPA or EVC, either for lag 1 or lag 2 (PPA lag 1: t_{30} = 0.83, p = 0.412, d = 0.15, CI = [0.02 ± 0.05]; PPA lag 2: t_{30} = -0.80, p = 0.433, d = 0.14, CI = [-0.02 ± 0.05]; EVC lag 1: t_{30} = 1.12, p = 0.272, d = 0.20, CI = [0.03 ± 0.06]; EVC lag 2: t_{30} = 0.69, p = 0.493, d = 0.12, CI = [0.02 ± 0.06]). Additionally, rank order correlations did not differ from 0 when representational structure at timepoint N was defined from EVC and representational structure at timepoint N+1 (lag1) or N+2 (lag 2) was defined from

CA3/dentate gyrus (lag 1: t_{30} = -0.12, p = 0.902, d = 0.02, CI = [-0.003 ± 0.05]; lag 2: t_{30} = -0.22, p = 0.825, d = 0.04, CI = [-0.005 ± 0.05]).

To better visualize the relationship in representational structure across successive timepoints—and to specifically connect this relationship to learning (as in **Fig. 2c**)—we computed pairmate similarity scores at the inflection point as a function of pre-IP pairmate similarity scores. Specifically, we binned all pairmates, by quartiles, according to pre-IP pairmate similarity scores, with the 4th quartile representing pairmates with the highest pre-IP pairmate similarity scores (see Methods for additional rationale; see **Supplementary Figure 3** for alternative binning procedures). We then computed the mean pairmate similarity scores at the inflection point for each of the pre-IP quartiles. Again, this analysis was separately performed for CA3/dentate gyrus and CA1. An ANOVA with factors of ROI (CA3/dentate gyrus, CA1) and pairmate similarity scores at the pre-IP (4 quartiles) revealed a significant interaction ($F_{3,90}$ = 3.19, p = 0.027, η^2 = 0.03), indicating that pre-IP representational overlap was differentially related to representational overlap at the inflection point for CA3/dentate gyrus versus CA1. Critically, this interaction was driven by a marked difference between CA3/dentate gyrus and CA1 when considering the bin with the highest overlap at the pre-IP (i.e., 4th quartile: t_{30} = -2.87, p = 0.008, d = 0.51, CI = [-0.03 ± 0.02], **Fig. 3c**). For CA3/dentate gyrus, pairmate similarity scores at the inflection point were significantly below 0 and numerically lowest for pairmates with the highest pre-IP similarity (4th quartile comparison to 0: t_{30} = -2.54, p = 0.017, d = 0.46, CI = [-0.023 ± 0.019]); the pattern in CA1 was qualitatively opposite. Collectively, these results provide theory-consistent

evidence that remapping of competing representations in CA3/dentate gyrus is actively triggered by initial representational overlap.



a. Example associations between scene pairmates and objects. The scene-object similarity was calculated by correlating activity patterns evoked during the scene exposure phases (at different behavioral states) and the object exposure phases. Target similarity refers to correlations between a given scene and the object with which it was studied. Competitor similarity refers to correlations between a given scene and the object with which its pairmate was studied. **b.** Scene-object similarity as a function of object relevance (target, competitor), ROI (CA3/dentate gyrus, pink; CA1, blue), and behavioral state (pre-learned round, learned round). Mean correlations between unrelated scenes and objects (across pairmate similarity; not shown) were subtracted from target and competitor similarity values. For CA3/dentate gyrus (CA3/DG), there was a significant interaction between behavioral state and object relevance (p = 0.002, repeated measures ANOVA). Note: ** p < .01. No correction for multiple comparisons was applied given the a priori predictions for CA3/dentate gyrus. Data are presented as mean ± S.E.M. and all data reflect n = 31 independent participants. Source data are provided as a Source Data file.

Figure 4 Scene-object similarity as a function of behavioral state.

CA3/dentate gyrus scene representations differentiate between competing object associations.

Thus far, we have focused on similarity among neural representations evoked while viewing the scene images (scene exposure phase). However, our paradigm also included two fMRI runs during which participants viewed each of the objects associated with the scene images (object exposure phase; see Methods). This allowed us to test whether hippocampal activity patterns evoked while viewing the scenes resembled—or came to resemble—activity patterns evoked while viewing corresponding object images.

Whereas pairmate similarity scores were computed by correlating activity patterns across different rounds of the scene exposure phase, here we computed correlations between a single round of the scene exposure phase and the average of

the two object rounds (see Methods; see **Supplementary Figure 5** for data separated by object round). For this analysis, there were three important factors that we considered. First, we considered whether scene representations were in a 'learned' state (i.e., scene representations from the learned round) or a 'pre-learned' state (i.e., scene representations from LR-1). Second, we separately tested correlations between each scene and (a) the target object (e.g., 'guitar 1') vs. (b) the competing object (e.g., 'guitar 2') (**Fig. 4a**). Third, we again compared results in CA3/dentate gyrus vs. CA1.

A repeated measures ANOVA with factors of ROI (CA3/dentate gyrus, CA1), behavioral state (pre-learned, learned), and object relevance (target, competitor) revealed a significant interaction between behavioral state and object relevance ($F_{1,30}$ = 12.42, p = 0.001, η^2 = 0.02). Qualitatively, this interaction reflected a learning-related change wherein hippocampal representations of scene images became relatively more similar to target objects and less similar to competitor objects. However, this 2-way interaction between behavioral state and object relevance was qualified by a trend toward a 3-way interaction between behavioral state, object relevance, and ROI ($F_{1,30}$ = 4.07, p = 0.053, η^2 = 0.01). Specifically, the interaction between behavioral state (pre-learned, learned) and object relevance (target, competitor) was significant in CA3/dentate gyrus ($F_{1,30}$ = 11.98, p = 0.002, η^2 = 0.06) but not in CA1 ($F_{1,30}$ = 0.44, p = 0.510, η^2 = 0.002) (**Fig. 4b**). For CA3/dentate gyrus, there was a qualitative increase, from the pre-learned to learned state, in the similarity between scenes and target objects and a qualitative decrease, from the pre-learned to learned state, in similarity between scenes and competing objects. In other words, the remapping of CA3/dentate gyrus scene representations that occurred at the learned round yielded a relative

strengthening of information related to target object associations and a relative weakening of information related to competing object associations. This dissociation in CA3/dentate gyrus is notable when considering that target and competitor objects were highly similar (see **Fig.1a, Fig. 4a**) and even more so when considering that during the scene and object exposure phases participants were not instructed or required in any way to recall the corresponding images. The 2-way interaction between behavioral state and object relevance was not significant for PPA or EVC (PPA: $F_{1,30} = 1.97$, $p = 0.170$, $\eta^2 = 0.01$; EVC: $F_{1,30} = 3.23$, $p = 0.082$, $\eta^2 = 0.02$, see **Supplementary Figure 6**). Interestingly, for CA3/dentate gyrus, scene representations in the pre-learned state were significantly more similar to competitor objects than to target objects ($t_{30} = 2.70$, $p = 0.011$, $d = 0.48$, CI = [0.012 ± 0.009]). While this result was not anticipated, we consider potential interpretations in the Discussion.

1.4 Discussion

Here, we show that learning to discriminate competing episodic memories is associated with an abrupt remapping of activity patterns in CA3/dentate gyrus. Specifically, fMRI pattern similarity in CA3/dentate gyrus decreased precisely when behavioral expressions of learning emerged. Additionally, the degree to which remapping occurred in CA3/dentate gyrus was predicted by the degree of initial pattern overlap among competing memories. Finally, remapped CA3/dentate gyrus representations contained relatively stronger information about relevant episodic associations and relatively weaker information about competing episodic associations, confirming the learning-related significance of the remapping effect.

Our experimental paradigm and analyses were inspired by—and our findings are consistent with—evidence of abrupt remapping in the rodent hippocampus[9–12]. Our findings also complement recent evidence of remapping-like phenomena in the human hippocampus[23,39,40]. However, the current findings provide unique and direct support for the proposal that hippocampal remapping is associated with the resolution of human episodic memory interference[8]. Specifically, we demonstrate an abrupt transition in hippocampal representations that occurred at an important inflection point in learning—the point at which participants were able to correctly discriminate similar memories and retrieve associations with high confidence. Notably, this finding was only possible because (a) we repeatedly probed episodic memory and hippocampal representations over the course of learning and (b) we identified inflection points in a participant- and pairmate-specific manner. Indeed, inflection points varied considerably across and within participants (**Fig. 1d, Supplementary table 1, and Supplementary Figure 1**) and the observed hippocampal remapping effect was significantly weaker when the specific mapping between behavior and fMRI data was shuffled within participants (**Fig. 2d**).

The fact that CA3/dentate gyrus remapping occurred precisely at the inflection point in learning strongly suggests that remapping was related to learning. This argument is also reinforced by our independent finding that remapped CA3/dentate gyrus activity patterns, evoked while participants viewed individual scene images, carried more information (compared to the pre-learning state) about target versus competing object associations. In other words, the inflection point defined from behavioral expressions of associative memory also captured a critical change in

associative representations encoded in CA3/dentate gyrus activity patterns. The fact that CA3/dentate gyrus exaggerated the representational distance between competing scenes (remapping) while simultaneously reflecting learned associations (scene-object similarity) is consistent with the idea that CA3 balances both pattern separation and pattern completion mechanisms[4,27,28,41]. The fact that remapped activity patterns contained information about learned associations is also consistent with the argument that hippocampal remapping does not simply reflect changes in the external environment—which did not change over the course of the experiment—but instead fundamentally reflects changes in internal models of the environment[15,16].

One aspect of our findings which does not, to our knowledge, have a direct analog in rodent studies of remapping is the negative pairmate similarity score we observed at the inflection point in CA3/dentate gyrus. The negative score indicates that scene pairmates—which were highly similar images—were associated with less overlapping CA3/dentate gyrus representations than completely unrelated scenes. In rodents, the most extreme version of remapping occurs when two similar environments are associated with fully independent place codes[8]. In our study, however, if each scene was associated with an independent representation, then the similarity between pairmates would be equal to, but not lower than, the similarity between non-pairmates. Instead, the negative pairmate similarity score requires a dependence between competing hippocampal representations wherein a given memory representation systematically moves away from the representational position of a competing memory (**Fig. 2f**). We refer to this dependence as ‘repulsion’ in order to emphasize the oppositional influence that competing memories exerted. Several recent human fMRI

studies have reported conceptually similar effects in the hippocampus[18,20,22,42]—and in CA3/dentate gyrus, specifically[17,33,19,24]. However, the current findings directly establish that the repulsion of competing hippocampal representations is temporally coupled to the resolution of memory interference.

Based on computational models[25,43,44], our prediction was that the repulsion effect in CA3/dentate gyrus was a direct consequence of initial overlap among activity patterns. Indeed, a recent study found that hippocampal repulsion was more likely to occur for behaviorally-confusable memories[18], potentially because confusable memories are associated with greater pattern overlap during initial learning. In the current study, we tested—and confirmed—this account directly. Specifically, we found that the representational structure (relative pairmate similarity) in CA3/dentate gyrus at a given timepoint was negatively correlated with representational structure at an immediately following timepoint. This negative relationship is highly consistent with the idea that overlap, itself, triggers plasticity that 'punishes' those features which are shared across memories[19,25,43,44]. While our study does not afford inferences about the causal relationship between repulsion and learning, the idea that repulsion (or remapping more generally) is triggered by representational overlap, combined with the fact that remapping was associated with learning, is consistent with the possibility that repulsion of CA3/dentate gyrus representations is a causal factor in learning. This account also offers a potential explanation for an otherwise surprising finding: that CA3/dentate gyrus scene representations from the round that immediately preceded learning (LR-1) were significantly more similar to competitor objects than to target objects. Although speculative, it is possible that when a given scene activated the 'wrong' object

association (at LR-1), this actively triggered a correction in favor of the target object association that supported learning. This account is consistent with evidence that prediction errors can powerfully drive episodic memory[45,46] as well as differentiation of hippocampal activity patterns[19]. More broadly—and consistent with our findings, in general—prediction errors may induce abrupt state changes in the hippocampus that facilitate the separation of episodic memories[47].

Across multiple analyses, we observed dissociations between CA3/dentate gyrus and CA1. The fact that the remapping effects were selective to CA3/dentate gyrus is consistent with evidence from rodent studies of remapping and pattern separation[8,26,28] and with several human fMRI studies[17,19,24,28,33]. Perhaps the most notable dissociation between CA3/dentate gyrus and CA1 comes from our analysis of representational structure across time points. Whereas CA3/dentate gyrus exhibited a negative rank correlation across successive timepoints, CA1 exhibited a positive rank correlation (**Fig. 3b**). Thus, in contrast to CA3/dentate gyrus, CA1 was characterized by stability (though only modest stability) of representational structure across timepoints[4]. This dissociation between CA3/dentate gyrus and CA1 is consistent with the idea that CA3, in particular, supports rapid plasticity that allows for changes in memory representations on short time scales[48] and is also consistent with evidence of faster remapping in CA3/dentate gyrus than in CA1[10,12,32]. It is also notable that the remapping effect we observed in CA3/dentate gyrus at the inflection point in learning strongly contrasted with the pattern of data in early visual cortex. Whereas CA3/dentate gyrus exhibited a negative pairmate similarity score at the inflection point, EVC exhibited a significant, positive pairmate similarity score at the inflection point. This finding makes the important point that

CA3/dentate gyrus was not inheriting representational structure from early sensory regions (e.g., due to visual attention)—rather, CA3/dentate gyrus fully inverted the representational structure that was expressed in early visual cortex[20].

Taken together, our findings reveal remapping of human CA3/dentate gyrus representations that is temporally-coupled to the resolution of episodic memory interference. These findings were motivated by—and complement—existing evidence of remapping in the rodent hippocampus. Yet, our findings also go beyond existing rodent or human studies by establishing a direct link between remapping and changes in internal memory states[15,16]. Additionally, our conclusion that overlap among CA3/dentate gyrus representations actively triggers a repulsion of memory representations has important implications for theoretical accounts of how the hippocampus resolves memory interference[5,8,28,43] and will hopefully inspire targeted new analyses that test for similar mechanisms in rodent models.

1.5 Methods

Participants.

Thirty-six participants (21 female; mean age = 23.69 yrs, range = 18 – 34 yrs) were enrolled in the experiment following procedures approved by the University of Oregon Institutional Review Board. Written informed consent was collected for each participant prior to the experiment. All participants were right-handed native-English speakers with normal or corrected-to-normal vision, with no self-reported psychiatric or neurological disease. One participant was excluded due to excess motion in the scanner (max FD > 3.5 mm); another 4 participants were excluded due to low behavioral performance (see

Results for more details). The final analysis included 31 participants. All participants received monetary compensation for participating.

Stimuli.

Thirty-six images of scenes and 36 images of everyday objects were used in the experiment. The set of 36 scenes and the set of 36 objects were each comprised of 18 'pairmates' of visually and semantically similar images (**Fig. 1a**). An additional 36 scenes and 12 objects were used as lures for the scene and object exposure phases of the study, respectively. Separately for each participant, scene pairmates were randomly assigned to object pairmates (**Fig. 1a**). For example, if 'lighthouse 1' was assigned to 'guitar 1', then 'lighthouse 2' would be assigned to 'guitar 2.' Note: the scene and object images shown in the figures are not the actual stimuli used in the experiment, but are public domain images representative of the stimuli that were used. See Data Availability for access to the actual stimuli.

Experimental procedure.

After providing consent and reviewing the instructions, participants entered the MRI scanner. Inside the scanner, participants completed 6 rounds of the experimental paradigm (**Fig. 1b**). The first round and the last round included 4 phases: study, test, scene exposure (scanned), and object exposure (scanned). Rounds 2–5 were the same, except they did not include the object exposure phase. Across all phases, stimuli were displayed on a grey background, projected from the back of the scanner. After exiting the scanner, participants completed a separate memory task that involved learning new scene-object associations (not reported here). The experiment was

implemented in PsychoPy[49] and lasted approximately 3 hrs, with about 2 hrs 15 min inside the scanner.

Study Phase. During the study phases, participants learned 36 scene-object associations, one association at a time. Each trial began with the presentation of a scene image (1000 ms), followed by a white fixation cross (200 ms), the associated object image (1000 ms) and then another white fixation cross (1200 ms) until the start of the next trial. The order in which the 36 scene-object associations were studied was randomized for each round and for each participant.

Test Phase. During the test phases, participants attempted to retrieve the object associated with each of the 36 scenes. Each trial began with the presentation of a scene (1000 ms), followed by a white fixation cross (200 ms), and then the presentation of two object pairmates (e.g., 'Guitar 1' and 'Guitar 2'). One of the object images was the 'target' (i.e., the object associated with the cued scene) and the other object image was the 'competitor' (i.e., the object associated with the cued scene's pairmate). Participants had a maximum of 4000 ms to select the correct object image (target) via a button box in their right hand. If no response was made, the next trial began after a white fixation cross was displayed for 1200 ms. If a response was made, a confidence rating then appeared beneath the objects and participants had a maximum of 3000 ms to indicate whether their response was a "Guess" or "Sure." After indicating their confidence (or after time ran out), a white fixation cross appeared (1200 ms) until the start of the next trial. The location of the correct object (left or right) and the

order in which each of the 36 scene-object associations were tested were randomized for each round and for each participant.

Scene Exposure Phase. During the scene exposure phases, which were conducted during fMRI scanning, participants saw 39 scene images in each of two blocks (78 scenes per round). Each block included the 36 studied scenes and 3 novel lure scenes. Participants made an old/new judgment for each scene. Each trial began with the presentation of a scene image (500 ms), followed by a red fixation cross (1500 ms) which represented the response window. Participants again responded using the button box. After the red fixation cross, a white fixation cross (2000 ms) was presented until the start of the next trial. The order of the 39 scene trials within each block was randomized for each block, round, and participant. Between the two blocks of 39 trials, participants performed a short odd/even judgment task (4 trials). Each odd/even trial consisted of a single-digit number displayed on the screen (500 ms), followed by a red fixation cross (1000 ms) which represented the response window, and then a white fixation cross (1000 ms) until the start of the next trial.

Object Exposure Phase. The object exposure phase (conducted during fMRI scanning) was only included in the first and sixth rounds and followed an identical structure and procedure as the scene exposure phase. The only difference was that the 39 trials in each block corresponded to the 36 studied objects and 3 novel lure objects.

MRI acquisition.

All images were acquired on a Siemens 3T Skyra MRI system in the Lewis Center for Neuroimaging at the University of Oregon. Functional data were acquired with a T2*-weighted echo-planar imaging sequence with partial-brain coverage that prioritized full coverage of the hippocampus and early visual cortex (repetition time = 2000 ms, echo time = 36 ms, flip angle = 90°, 72 slices, 1.7x1.7x1.7mm voxels). A total of 8 functional scans were acquired. Each functional scan comprised 177 volumes and included 10 s of lead-in time and 10 s of lead-out time at the beginning and end of each scan, respectively. The 8 functional scans corresponded to 6 scans of the scene exposure phase (scans 1 and 3–7) and 2 scans of the object exposure phase (scans 2 and 8). Anatomical scans included a whole-brain high-resolution T1-weighted magnetization prepared rapid acquisition gradient echo anatomical volume (1x1x1mm voxels) and a high-resolution (coronal direction) T2-weighted scan (0.43x0.43x2mm voxels) to facilitate segmentation of hippocampal subfields.

Anatomical data preprocessing.

Preprocessing was performed in Python 3.7 using *fMRIPrep* 1.5.0[50,51] (RRID:SCR_016216), which is based on *Nipype* 1.2.2[52,53] (RRID:SCR_002502). The T1-weighted (T1w) image was corrected for intensity non-uniformity (INU) with N4BiasFieldCorrection[54] (ANTs 2.2.0[55], RRID:SCR_004757), and used as the T1w-reference throughout the workflow. The T1w-reference was skull-stripped with the antsBrainExtraction.sh workflow (ANTs) in *Nipype*, using OASIS30ANTs as target template. Brain tissue segmentation of cerebrospinal fluid (CSF), white-matter (WM) and gray-matter (GM) was performed on the brain-extracted T1w using fast[56] (FSL 5.0.9, RRID:SCR_002823). Volume-based spatial normalization to one standard space

(MNI152NLin2009cAsym) was performed through nonlinear registration with antsRegistration (ANTs 2.2.0), using brain-extracted versions of both T1w reference and the T1w template. ICBM 152 Nonlinear Asymmetrical template version 2009c[57] (RRID:SCR_008796; TemplateFlow ID: MNI152NLin2009cAsym) was used for spatial normalization.

Functional data preprocessing.

For each of the 8 BOLD scans per participant, the following preprocessing was performed. First, a reference volume and its skull-stripped version were generated using *fMRIPrep*. A deformation field to correct for susceptibility distortions was estimated based on two echo-planar imaging (EPI) references with opposing phase-encoding directions, using 3dQwarp, AFNI[58]. Based on the estimated susceptibility distortion, an unwarped BOLD reference was calculated for a more accurate co-registration with the anatomical reference. The BOLD reference was then co-registered to the T1w reference using bbregister (FreeSurfer) which implements boundary-based registration[59]. Co-registration was configured with six degrees of freedom. Head-motion parameters with respect to the BOLD reference (transformation matrices, and six corresponding rotation and translation parameters) were estimated before any spatiotemporal filtering using mcflirt FSL 5.0.9[60]. BOLD scans were slice-time corrected using 3dTshift AFNI[58](RRID:SCR_005927). The BOLD time-series (including slice-timing correction when applied) were resampled onto their original, native space by applying a single, composite transform to correct for head-motion and susceptibility distortions. Framewise displacement (FD) confounding time-series were calculated based on the resampled BOLD time-series for each functional scan[61].

fMRI first-level general linear model (GLM) analyses.

After *fMRIPrep* preprocessing, the first 5 volumes (10 s) of each functional scan were discarded. Then, the brain mask generated by *fMRIPrep* from the T1 anatomical image was used to perform brain extraction for each of the 8 functional scans. Each functional scan was then median centered. For the 6 scans of the scene exposure phase and 2 scans of the object exposure phase, all first level GLMs were performed in participants' native space with *FSL* using a Double-Gamma HRF with temporal derivatives, implemented with *Nipype*. GLMs were calculated using a variation of the Least Squares – Separate method[62]: a separate GLM was calculated for each of the 36 scenes (for scene exposure phases) or objects (for object exposure phases) across both repeats within a scan. For each GLM, there was one regressor of interest (representing a single scene or object image across its two repetitions per scan). All other trials (including lure images), framewise displacement, xyz translation and xyz rotation were represented with nuisance regressors. Additionally, a high pass filter (128 Hz) was applied for each GLM. This model resulted in 36 beta-maps per scan (one map per scene/object) which were converted to *t*-maps that represented the pattern of activity elicited by each scene/object for each scan.

Regions of interest.

A region of interest (ROI) for early visual cortex (EVC) was created from the probabilistic maps of Visual Topography[63] in the MNI space with a 0.5 threshold. This ROI was transformed into each participant's native space using inverse T1w-to-MNI non-linear transformation. For each participant, the top 300 EVC voxels were then selected by averaging the *t*-maps of all scenes and objects and then choosing the

voxels with the highest *t*-statistics (i.e., the voxels most responsive to visual stimuli). An ROI for the parahippocampal place area (PPA) was created by first using an automated meta-analysis in Neurosynth with the key term “place”. Then, clusters were created using voxels with a z-score > 2 based on the Neurosynth associative tests. Since these clusters were generated through an automated meta-analysis and were not anatomically exclusive to PPA, we visually inspected the results and manually selected the two largest clusters that were spatially consistent with PPA. One cluster was in the right hemisphere (voxel size = 247) and one cluster was in the left hemisphere (voxel size = 163). These clusters were combined into a single PPA mask. This mask was then transformed into each participant's native space using the inverse T1w-to-MNI transformation. For each participant, a final PPA ROI was generated by averaging the *t*-maps of all scene exposure phase scans and then selecting the 300 voxels with the highest average *t*-statistics (i.e., the most scene-responsive voxels). To create hippocampal ROIs, we used the Automatic Segmentation of Hippocampal Subfields (ASHS)[64] toolbox with the upenn2017 atlas to generate subfield ROIs in each participant's hippocampal body, including CA3/dentate gyrus (which included CA2, CA3 and dentate gyrus) and CA1. The most anterior and posterior slices of the hippocampal body were manually determined for each participant based on the T2-weighted anatomical structure (see **Supplementary Figure 7** for a sample demarcation). Each participant's subfield segmentations were also manually inspected to ensure accuracy of the segmentation protocol. Then, each subfield ROI was transformed into each participant’s native space using the T2-to-T1w transformation, calculated with FLIRT (fsl) with 6 degrees of freedom, implemented with *Nipype*. All ROIs were again visually

inspected following the transformation to native space to ensure the ROIs were anatomically correct.

fMRI pattern similarity analyses.

Pairmate Similarity Scores. Pattern similarity was calculated as the Fisher *z*-transformed Pearson correlation between *t*-maps within each ROI. All pattern similarity analyses were performed by correlating the *t*-maps for stimuli across scans (i.e., correlations were never performed within the same scan). For our primary analyses related to pattern similarity between scene images, of critical interest was similarity between pairmate scenes (*pairmate similarity*) relative to similarity between non-pairmate scenes (*non-pairmate similarity*). Specifically, for each set of pairmates, mean non-pairmate similarity was subtracted from mean pairmate similarity to yield a *pairmate similarity score* for each set of pairmates. As an example, to calculate pairmate similarity scores for 'lighthouse 1' and 'lighthouse 2' across scans 3 and 4, pairmate similarity would be defined as the mean of the following two *z*-transformed correlations: *r*(lighthouse $1_{\text{scan 3}}$, lighthouse $2_{\text{scan 4}}$) and *r*(lighthouse $2_{\text{scan 3}}$, lighthouse $1_{\text{scan 4}}$). Corresponding non-pairmate similarity scores would be defined as the mean of all *z*-transformed correlations, across the same scans (scans 3 and 4), between either pairmate (lighthouse 1 or lighthouse 2) and each non-pairmate stimulus [e.g., *r*(lighthouse $1_{\text{scan 3}}$, arch $1_{\text{scan 4}}$), *r*(arch $2_{\text{scan 3}}$, lighthouse $1_{\text{scan 4}}$), …].

Learned Round. To relate pairmate similarity scores to behavioral measures of learning, we identified the *learned round* (LR) for each pairmate, separately for each participant. The learned round was based on performance in the

associative memory test. Specifically, the learned round was defined as the first round in which the target object was selected with high confidence for both scenes in a pairmate, with the additional requirement that performance remained stable in all subsequent rounds. It was, therefore, possible that both scenes in a pairmate were associated with high confidence correct responses in round N, not in round N+1, and then (again) in round N+2 and thereafter; in this case, the learned round would be round N+2.

Inflection Point. The *inflection point* (IP) was defined as the transition from LR-1 (the round that immediately preceded the learned round) to LR (the learned round). Thus, pairmate similarity scores at the inflection point were based on correlations of *t*-maps from LR-1 with *t*-maps from LR. We hypothesized that the behavioral state change from LR-1 to LR would correspond to a reduction in pairmate similarity scores. Pairmate similarity scores at the inflection point were contrasted against the 'pre-IP' state, which was based on the correlation of *t*-maps from LR-2 and LR-1 (i.e., a non-transition from 'pre-learned' to 'pre-learned') (**Fig. 2c**). Pairmates for which participants never reached and sustained high-confidence correct responses (mean ± s.d., 1.81 ± 2.27 per participant) and pairmates that were learned in the 1st round (LR = 1; mean ± s.d., 1.00 ± 1.26) were excluded from the inflection point analyses because neither the pre-IP nor inflection point states could be measured. For pairmates that were learned in the 2nd round (LR = 2; mean ± s.d., 3.23 ± 2.80), pattern similarity at the inflection point was calculated and included in the analyses, but pattern similarity at the pre-IP state could not be calculated because an LR-2 did not exist. For the rest of the

pairmates (LR = 3, 4, 5, or 6), we calculated pattern similarity for both the pre-IP and inflection point states (**Fig. 1e**). Similar restrictions applied to correlations between LR and LR-3, LR + 1, LR + 2, and LR + 3 (**Fig. 2e**). The number of pairmates included in each comparison and for each participant are reported in **Supplementary Table 1**.

Representational Structure Across Time Points. To test whether representational overlap triggered remapping (related to **Fig. 3**), the 6 rounds were translated into 5 timepoints. Each timepoint corresponded to a pair of consecutive rounds ([1,2], [2,3], [3,4], [4,5], [5,6]). For each timepoint, pairmate similarity scores were calculated, as described above, by correlating activity patterns from consecutive rounds (e.g., pairmate similarity scores at timepoint 1 were based on correlations between round 1 and round 2). This yielded a set of pairmate similarity scores at each of the 5 timepoints. These sets of similarity scores reflected the representational structure at each timepoint (i.e., which pairmates were relatively similar and which pairmates were relatively dissimilar). Pairmate similarity scores were then correlated across timepoints using Spearman's rank correlation (Fisher *z* transformed). Lag 1 correlations refer to rank correlations between successive timepoints whereas lag 2 correlations refer to correlations between timepoints two steps apart. To facilitate a direct comparison between lag 1 vs. lag 2 correlations, correlations were computed for the following timepoints: Lag 1 = *r*(timepoint 1, 2), *r*(timepoint 2, 3), *r*(timepoint 3, 4); Lag 2 = *r*(timepoint 1, 3), *r*(timepoint 2, 4), *r*(timepoint 3, 5). It is important to emphasize that we did not correlate initial pairmate similarity scores with the change in pairmate similarity

as this would produce an artifactual correlation (via regression to the mean). In contrast, a negative rank correlation (as we observed in CA3/dentate gyrus) cannot be explained by regression to the mean. Mathematically, if all values at timepoint N partially regressed toward the mean at timepoint N+1, this would yield a positive rank correlation (i.e., representational structure would be partially preserved). If all values fully regressed toward the mean (i.e., variance at timepoint N+1 = 0), this would yield a null correlation ($r = 0$; representational structure fully abolished).

To specifically consider the relationship between representational structure at pre-IP and representational structure at the inflection point (IP) (related to **Fig. 3c**), we binned pairmates, by quartile (using the *cut* function in base R), according to pairmate similarity scores at pre-IP and then computed pairmate similarity scores, for each quartile, at the inflection point. The quartile analysis was performed within subject and separately for CA3/dentate gyrus and CA1. The mean number of pairmates included in each pre-IP bin were 3.42, 2.90, 3.10, and 2.55 for quartiles 1-4, respectively. The decision to divide pre-IP pairmate similarity scores into four bins was motivated by evidence, from conceptually-related studies, of non-monotonic relationships between initial memory activation/competition and experience-dependent plasticity[43,65]. While formally testing for non-monotonic relationships was beyond the scope of the current study, the goal was to allow for qualitative inspection of the relationship. Notably, similar results were obtained when pre-IP pairmate similarity scores were binned by terciles or quintiles (**Supplementary Figure 3**).

Scene-Object Similarity. To calculate pattern similarity between scenes and objects (related to **Fig. 4**), activation patterns for objects were first generated by averaging *t*-maps across the two object exposure phases, resulting in a single, mean activity pattern for each object. These object-specific activity patterns were then correlated with activity patterns from the scene exposure phases at LR – 1 (i.e., the pre-learned state) and LR (i.e., the learned state). Correlations were separated into three groups: (1) target correlations refer to the correlation between a scene and the object it was associated with during the study phase (e.g., 'lighthouse 1' and 'guitar 1'), (2) competitor correlations refer to the correlation between a scene and the object that was associated with that scene's pairmate during the study phase (e.g., 'lighthouse 1' and 'guitar 2'), and (3) across pairmate correlations refer to correlations between a scene and an object that was not associated with that scene or its pairmate during the study phase (e.g., 'lighthouse 1' and 'scissors 1'). Target and competitor correlations were expressed relative to across pairmate correlations.

Statistics and Reproducibility.

To compare pairmate similarity scores and other measures across ROIs and learning states, repeated measures ANOVAs and paired-samples *t*-tests were used. To test whether pairmate similarity scores and other measures were significantly positive or negative (i.e., above/below 0), one-sample *t*-tests were used. To test whether the negative pairmate similarity score observed in CA3/dentate gyrus at the inflection point depending on the specific mapping between behavioral and fMRI measures, we randomly shuffled the mapping between the behavioral inflection point and scene

pairmate, within each participant (see **Fig. 1d**), and then computed the group-level mean pairmate similarity score at the permuted inflection point. This was repeated 1,000 times, producing a distribution of 1,000 permuted means. The observed pairmate similarity score at the inflection point was then compared against this distribution of permuted means. Data analysis was performed in R 3.5.0 and its associated libraries. All of the data and results reported here reflect a single experiment; an independent replication was not conducted.

Data Availability.

The MRI data generated in this study have been deposited on Openneuro.org (DOI: 10.18112/openneuro.ds003707.v1.0.0)[66]. The stimuli used and the behavioral data generated in this study have been deposited on osf.io (DOI: 10.17605/OSF.IO/VPQ2X)[67]. The source data underlying all Figures and Supplementary Figures are provided as a source data file with this paper.

Code Availability.

Analysis scripts are available at [https://github.com/wanjiag/NEUDIF_analysis].

Chapter 2. Hippocampal repulsion can be driven by internal beliefs

2.1 Abstract

The hippocampus plays an important role in distinguishing similar events. Recent literature showed that hippocampal repulsion happened when competing memories representations were being actively pushed away from each other. It was shown that the repulsion effect in the hippocampus is beneficial for the resolution of memory interference. However, the reason underlying why repulsion occurs remains undetermined. In the current study, we directly tested the hypothesis that hippocampal repulsion could be triggered by distinct internal beliefs, while holding the external stimuli constant. Utilizing a route-learning behavioral task with high-resolution human fMRI data, we showed that hippocampal repulsion happened if participants held distinct beliefs between routes, but the repulsion became absent if the beliefs were the same. Moreover, we showed that repulsion was the strongest while participants were actively resolving the memory interference, showing a temporal lock between the hippocampal repulsion effect and the behavioral resolution of memory interference. We also showed that hippocampus patterns were highly dynamic and could change abruptly within a trial. Critically, hippocampus showed lower pattern similarities between competing routes when the visual similarities between the routes were high, which was opposite to the pattern similarity patterns for the early visual cortex. Lastly, we showed that the hippocampal repulsion occurred only within the CA2/3/DG subfields within the hippocampus, and this effect was absent for CA1 in the hippocampus, the parahippocampal place area, or the early visual cortex.

2.2 Introduction

Many of our daily experiences share overlapping features, leading to confusability among similar memories, termed memory interference. The hippocampus has been shown to play a critical role in distinguishing similar memories. Previous rodent studies have established that the place cells in the hippocampus remap – a rapid change in cell firing activities – when there were small changes in the environment. More recent rodent studies have further shown that remapping can be caused not only by subtle change of the external stimuli, but also by changes of internal models[1], latent information[2], or reward locations[3]. These recent literatures suggested that remapping depends on the animal's internal model of a particular environment. Building on these findings, recent theoretical and computational models also propose a framework in which the hippocampus remaps as a function of changes in internal beliefs about the environment, instead of changes in the objective, external environment itself[4].

Similar to rodent literature, the human hippocampus also demonstrates remapping-like phenomena in the context of episodic memory[5]. To resolve memory interference, the activity patterns in human hippocampus not only orthogonalize similar memories, but actively *repulse* two similar memories apart from each other. This repulsion effect is well-established in the literature and has shown to be experience-dependent and facilitate the resolution of memory interference[6,7]. However, it remains a question in human hippocampus repulsion regarding whether the repulsion of hippocampus activity patterns can be caused by changing internal belief alone while holding the external environment constant.

One critical question that remains unanswered is why repulsion occurs. One of the reasons could be that repulson is triggered by participants noticing diagnostic features between competitive memories, which then leads to the repulsion in hippocampus activity patterns between those memories. This hypothesis suggests that hippocampus repulsion stems from and depends on a neural signal from the visual cortex. However, we do not think this hypothesis accounts for the exaggerated differences in the repulsion effect: the visual differences between two overlapping events should not be more distinct than that of two non-overlapping events. In other words, diagnostic features should at most lead to orthogonalization: when overlapping events are represented as dissimilarly as non-overlapping events. However, the repulsion effects that have been observed previously—and described in Chapter 1—demonstrate that overlapping events were represented as *more dissimilar* than non-overlapping events. Inspired by recent rodent studies[1–3] and computational models[4] in hippocampus remapping, the alternative hypothesis articulates that repulsion can be triggered even if external stimuli are identical, if participants hold the believes that they were distinct events. The alternative hypothesis emphasizes that subtly distinct visual inputs are not necessary to trigger hippocampus repulsion and repulsion does not originate from visual cortex.

Here, we used a spatial learning paradigm inspired from a T-maze in rodent studies[8,9] that includes pairs of real-world overlapping routes[6]. Participants studied continuous images of each route and associated each route to a specific destination. For each route, we separated it into 3 segments that differed in the degree of similarity to the overlapping route. Two overlapping routes started at a segment with identical

images (‘same’ segment), followed by a segment with extremely similar but not identical images (‘similar’ segment), and ended with a segment of different images (‘different’ segment). The change of similarity levels across segments within a route allowed us to understand how representational patterns in the hippocampus change with different levels of similarities. More critically, to understand how internal belief influences representational patterns, we also manipulated participants’ internal beliefs of which route they were viewing by adding a possibility cue before the start of the route. Therefore, we can test whether hippocampus repulsion depends on the memories of highly similar yet observable external stimulus, or internal believes that two memories are different from each other. This is because at the same segment, when two overlapping routes shared identical images and path, participants’ belief of the current route was only influenced by the cue.

Our study design allowed us to test two critical hypotheses regarding hippocampal repulsion. The first of these hypotheses is that hippocampus repulsion does not depend on external stimuli and can be triggered by distinct internal belief. Previous literature proposed computational models[4] supporting this hypothesis for hippocampus remapping, but to our knowledge, there was no previous rodent or human literature that tested the influence of internal beliefs on hippocampus representations, while holding external environment constant. The second of our critical hypotheses is that hippocampal repulsion occurs abruptly at the moment of interference resolution. In our recent publications, we showed that the repulsion effect occurred at the exact time when memory interference was successfully resolved, and quickly subsided afterwards. However, the previous literature measured the changes in hippocampus representation

at a coarse temporal sampling (round by round) in the context of associative learning. It remains unknown whether repulsion happened abruptly at a more fine-grained temporal measurement (second-to-second).

To preview the current work, we show that the moment of interference resolution showed repulsion effect in CA2/3/DG, a subfield within the hippocampus, and the repulsion effect disappeared afterwards. Moreover, our results suggest that even with identical external stimuli, distinct internal beliefs can lead to repulsion effect in CA23DG, however, strikingly, identical internal beliefs did not lead to repulsion effect with the same external stimuli.

2.3 Results

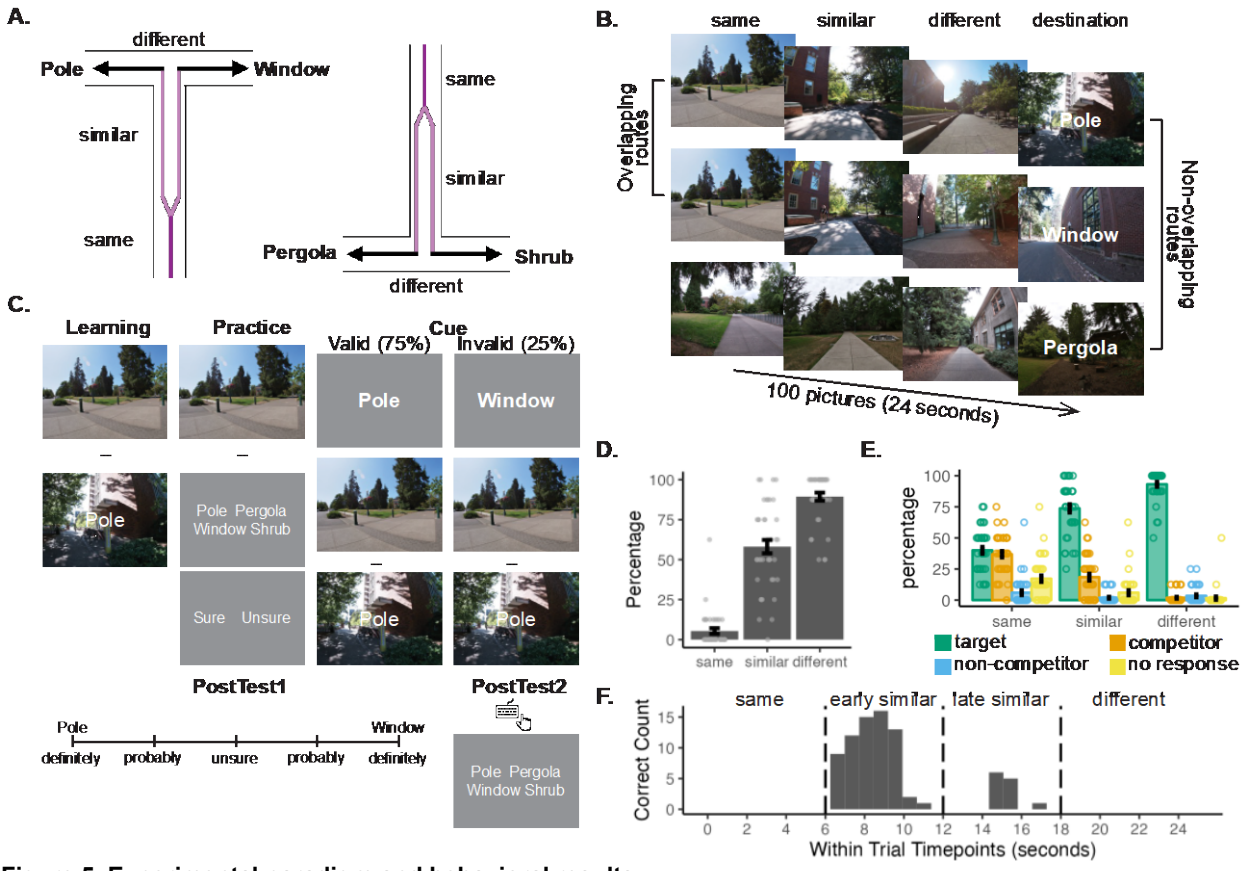


Figure 5. Experimental paradigm and behavioral results.

(A) Stimuli schema: each participant studied four routes, that could be grouped into two pairs of overlapping routes. Overlapping routes were initially identical before becoming subtly different and ultimately diverging to terminate at

unique destinations. In this example, pole and window, shrub and pergola are overlapping pairs, respectively. **(B)** Each route consisted of a stream of 100 continuous images that were displayed to the participants rapidly. During the 'same' segment (picture index 1-25, 6 seconds), overlapping pair (e.g., pole and window) contained identical images and traveled identical paths. During the 'similar' segment (picture index 26-75, 12 seconds), images were similar but non-identical, and the paths remained identical for overlapping routes. During the different segment (picture index 76-100, 6 seconds), the images and paths became both distinct. For non-overlapping pairs (e.g., pole and pergola; window and pergola), both images and paths were always distinct. **(C)** Participants finished Learning, Practice, Cue, and 2 Posttests. Throughout the whole experiment, each route appeared four times per round. During the Learning phase (1 round), participants passively viewed each route. During the Practice phase (2 rounds), for one of the four displays of each route, the display would pause once at each segment, and the participants were instructed to choose the correct destination for the current route, followed by indicating their confidence level. During the Cue phase (10 rounds), each trial was preceded by a cue indicating the likely destination. These cues could be **valid**, indicating the correct destination, or **invalid**, indicating the competitor's destination. For each scan round for each route, there were 2 valid trials, 1 invalid trial, and 1 catch trial (Catch trials were excluded from all fMRI analysis; see Methods for more details). Lastly, participants performed two posttests. In the first posttest (1 round), each route was displayed from start to the end of the similar segment (picture index 1 - 75). The route paused 4 time within the similar segment at picture index 30, 45, 60, and 75. Participants then used a slider to indicate the correct destination, as well as their confidence. In the second Posttest (1 round), participants were instructed to press a button when they reached 90% certainty about the destination, followed by choosing the destination. **(D)** As the routes became more distinct, the percentage of confident correct responses also increased for the associative tests (Same vs. Similar: $t(39) = 12.76$, $p < 0.001$; Similar vs. Different: $t(39) = 7.78$, $p < 0.001$). **(E)** As the routes became more distinct, the percentage of accuracy also increased for the associative tests (Save vs. Similar: $t(39) = 10.24$, $p < 0.001$; Similar vs. Different: $t(39) = 9.41$, $p < 0.001$). Within the same segment, participants chose the correct destination and the competitor destination at a similar rate ($t(39) = 0.77$, $p = 0.447$), but correct destination was chosen significantly higher than both non-overlapping destinations ($t(39) = 8.61$, $p < 0.001$) and no response ($t(39) = 5.13$, $p < 0.001$). Within the similar segment, participants selected the correct destination at a higher rate than both the competitor destination ($t(39) = 9.93$, $p < 0.001$) and non-overlapping destination ($t(39) = 20.60$, $p < 0.001$), with the competitor destination being selected more than non-overlapping destinations ($t(39) = 5.90$, $p < 0.001$). **(F)** Mean timepoints for indicating 90% certainty of the destination for each overlapping pair for each participant.

Each participant learned and repeatedly viewed a slideshow of images depicting four routes (2 overlapping route pairs) within the University of Oregon campus. Within each overlapping pair, the two routes started on identical paths with identical images ('same' segment, 6 seconds, 25 pictures), followed by identical paths with highly similar images and overlapping paths ('similar segment', 12 seconds, 50 pictures), then different images with diverged paths ('different' segment, 6 seconds, 25 pictures), and ultimately ended at distinct destinations (Fig 1A). Therefore, each route (e.g., 'pole') would be the competitor route for the other route (e.g., 'window') within an overlapping pair, and the non-competitor route for the rest of the routes (e.g., 'shrub' and 'pergola') (Fig 1B). Throughout the experiment, participants were instructed to learn the specific route to each destination as early as possible.

Behavioral results.

Prior to fMRI scanning, participants completed two rounds of a pre-scan test in which they viewed routes which occasionally paused. At the pauses, participants were asked to select the route destination from a set of 4 choices. After making a selection, participants indicated whether their confidence was high or low. The pauses occurred either during the same, similar, or different segments, allowing for consideration of performance at each segment. During the same segment, (where the images were identical for the overlapping routes), participants were more likely to select the correct destination or the competitor destination than the destination for one of the non-overlapping routes (correct vs. non-overlapping: $t(39) = 8.62$, $p < 0.001$; competitor vs. non-overlapping: $t(39) = 8.85$, $p < 0.001$) (Fig. 1E). However, given that the images were identical for the overlapping routes during the same segment, participants were no more likely to select the correct destination than the destination for the competitor ($t(39) = 0.77$, $p = 0.447$). During the similar segment, participants were more likely to select the correct destination than one of the destinations for the non-overlapping routes ($t(39) = 20.60$, $p < 0.001$) or the destination for the competitor ($t(39) = 9.93$, $p < 0.001$). However, the competitor destination was still more likely to be selected than one of the destinations for the non-overlapping routes ($t(39) = 5.90$, $p < 0.001$). Finally, during the different segment, participants were more likely to select the correct destination than one of the destinations for the non-overlapping routes ($t(39) = 20.60$, $p < 0.001$) or the destination for the competitor ($t(39) = 9.93$, $p < 0.001$). Importantly, during the different segment, participants were no more likely to select the competitor destination than one of the destinations for the non-overlapping routes ($t(39) = 1.30$, $p = 0.201$), indicating that competition between the overlapping routes was fully resolved once the images

were distinct. The percentage of high-confidence, correct responses also robustly increased from the same segment to the similar segment (5.31% to 58.13%, $t(39) = 12.76$, $p < 0.001$) and from the similar segment to the different segment (58.13% to89.38%, $t(39) = 9.41$, $p < 0.001$) (Fig. 1D).

During fMRI scanning, routes occasionally paused (25% of trials) and participants were instructed to select the correct destination. However, these test trials during scanning were only intended to encourage vigilance; performance on these trials is not easily interpreted given that the test trials were always preceded by a valid cue (see Methods).

After fMRI scanning, participants completed two post-tests. The first posttest was similar to the pre-scan test except that the routes only paused during the similar segment. Each route was tested 4 times, with pauses occurring at picture index 30, 45, 60, and 75 (note: the similar segment spanned picture index 26 to 75). Of interest was the probability that participants selected the correct destination and the confidence of these responses. A one-way repeated measures ANOVA revealed a significant main effect of picture index on accuracy ($F(3, 117) = 88.975$, $p < 0.001$) and on high-confidence correct responses ($F(3, 117) = 164.07$, $p < 0.001$). Splitting the same segment into an early part (picture index 30 and 45) and late part (picture index 60 and 75) revealed a significant increase in accuracy from early to late ($t(39) = 8.5217$, $p < 0.001$) and an increase in high-confidence correct responses ($t(39) = 12.25$, $p < 0.001$).

In the second posttest, participants viewed each route and were instructed to press a button, during the route viewing, as soon as they were are at least ‘90% sure’ of the destination. Subjects were then prompted to select the destination from a set of 4

options. The average accuracy was 94.84% with a standard error of 1.26%. For correct trials, the mean response time was 9.25 s, with a range across subjects from 6.30 to 16.63 s (Fig. 1F). Thus, all participants were able to anticipate the correct destination at some point during the similar segment, which spanned 6.0 to 18.0 s, but with variability across participants and routes. Moreover, when separating the route into the early-similar (picture index 25-50) and late-similar (picture index 51-75), participants responded more frequently in the early-similar (mean=12.68, se=0.56) than the late-similar (mean=2.43, se=0.54), which demonstrated the behavioral performance differences within the similar segment of the route.

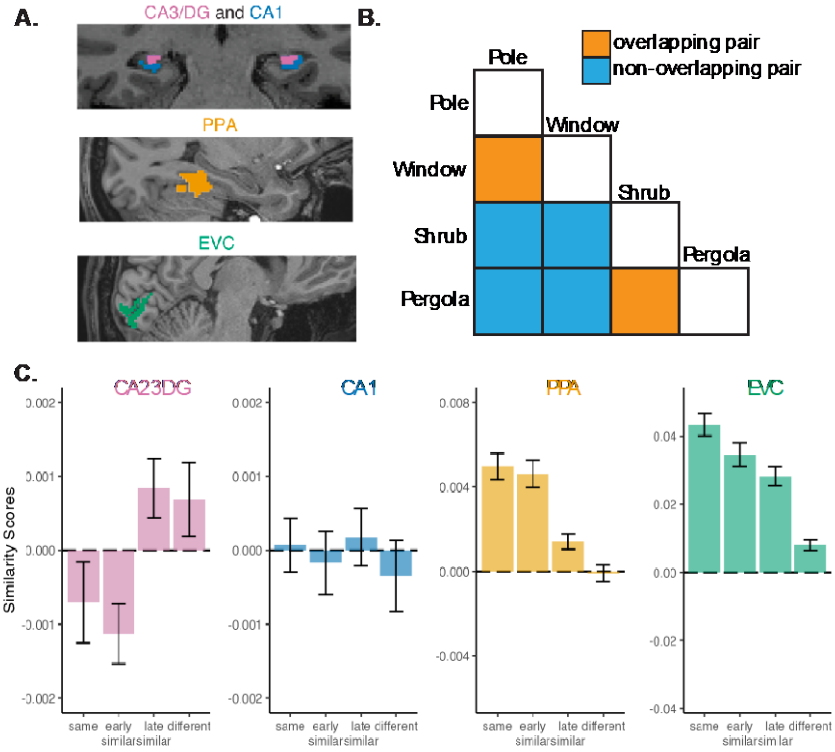


Figure 6. Similarity scores decreased as the competing routes became increasingly distinct.

(A) Regions of interest included CA2/CA3/dentate gyrus (CA2/3/DG, pink) and CA1 (blue) in the hippocampus, the parahippocampal place area (PPA, yellow), and early visual cortex (EVC, green). (B) Correlation matrix illustrating how Similarity Scores were defined. Here, Pole and Window, as well as Shrub and Pergola, were overlapping routes with each other (orange). Pole and Shrub, Pole and Pergola, Window and Shrub, as well as Window and Pergola, were non-overlapping routes with each other (blue). Similarity scores were calculated using the average correlations of overlapping routes subtracted by the average correlations of non-overlapping routes, timepoint-by-timepoint within a trial. (C) Similarity scores across each segment of the route for different Regions of Interests (ROIs). Two-way

ANOVA showed that similarity scores changed significantly with different segments ($F(3, 585) = 34.35$, $p < 0.001$) and ROIs ($F(3, 585) = 379.40$, $p < 0.001$), as well as a significant interaction between segments and ROIs ($F(9, 585) = 26.58$, $p < 0.001$). Within CA23DG, the early-similar segment was significantly lower than 0 ($t(39) = -2.80$, $p = 0.008$), whereas the late-similar segment was significantly higher than 0 ($t(39) = 2.09$, $p = 0.043$). The same ($t(39) = -1.29$, $p = 0.203$) and different ($t(39) = 1.38$, $p = 0.176$) segments were not significantly different from 0.

CA23DG repulsion occurred when ambiguity was highest.

For each subject, trial, and region of interest (ROI), we extracted each voxel's activation at each TR—from the cue period through the destination (27 timepoints in total). As in prior, related studies, we focused analyses on the hippocampus, Parahippocampal Place Area (PPA) and Early Visual Cortex (EVC)[7]. We split the hippocampus into two ROIs: CA1 and CA2/CA3/dentate gyrus (CA23DG) (Fig 2A, see Methods for more details).

Pattern similarity values were computed by correlating activity patterns across two different routes from independent two routes within an overlapping pair (e.g., pole and window) and not within the same scanning round are correlated with each other TR by TR, resulting in a correlation score for each timepoint. The non-overlapping pair (e.g., pole and shrub) from different scanning rounds are also correlated with each other, resulting in another correlation score for each timepoint. Similarity scores are calculated as the difference between the correlations from overlapping pair and non-overlapping pair. A positive similarity score would indicate that a pair of overlapping routes (e.g., pole and window) with visual similarities are associated with more similar neural representations than two non-overlapping routes (e.g., pole and shrub; pole and pergola). Based on prior studies, we specifically predicted repulsion within CA23DG and that CA23DG would 'invert' the representational structure in PPA and EVC.

Of central interest, we investigated how the similarity scores changed with various degrees of similarities. Based on behavioral performance differences within the

similar segment, we divided the route into 4 equal-length segments: same, early similar (first half of the similar segment), late similar (second half of the similar segment), and different, with each segment being 6 seconds in length. A 2-way ANOVA with factors of segments (same, early similar, late similar, different) and ROIs (CA23DG, CA1, PPA, EVC) showed significant main effect of segments ($F(3, 585) = 34.35$, $p < 0.001$) and ROIs ($F(3, 585) = 379.40$, $p < 0.001$), as well as a significant interaction ($F(9, 585) = 26.58$, $p < 0.001$), indicating that similarity scores exhibited different patterns across segments for different ROIs. With the exception of CA1 ($F(3, 177) = 0.36$, $p = 0.782$), all other three ROIs demonstrated a significant main effect of segments (CA23DG: $F(3, 177) = 5.22$, $p = 0.002$; PPA: $F(3, 117) = 28.73$, $p < 0.001$; EVC: $F(3, 117) = 41.27$, $p < 0.001$).

Among the 3 ROIs with a significant main effect of segment, we wanted to identify whether there is a linear trend across time. To do so, we separated each route into the first half (first 12 seconds, picture index 1-50) and the second half (last 12 seconds, picture index 51-100), and conducted paired t-tests within each ROI. In visual attention regions including PPA ($t(39) = -7.21$, $p < 0.001$) and EVC ($t(39) = -3.47$), there is a decreasing linear trend, meaning the similarity scores decreased as the overlapping routes become visually more distinct. However, in CA23DG, there is an opposite increasing linear trend ($t(39) = 3.47$, $p = 0.001$), meaning the similarity scores increased as the overlapping routes become more visually distinct (Fig 2C). Moreover, one-sample t-tests in CA23DG showed that in the first half of the route, when the overlapping routes are more similar and participants displayed lower performance in the memory

association test, the similarity scores were significantly lower than 0 ($t(39) = -2.31$, $p = 0.026$).

However, in the second half of the route, when the overlapping routes become more distinct and memory competitions are resolved, the similarity scores were significantly higher than 0 ($t(39) = 2.26$, $p = 0.030$). Overall, CA23DG not only displayed opposite linear trend comparing to EVC and PPA, but also displayed the repulsion effect only in the first half of the route, when memory competitions were still yet to be resolved.

To more precisely locate the timing of the repulsion, we also conducted one-sample t-tests for similarity scores within each segment of the route for CA23DG and found the early-similar segment was significantly lower than 0 ($t(39) =-2.80$, $p = 0.008$), whereas the late-similar segment was significantly higher than 0 ($t(39) =2.09$, $p = 0.043$). The same ($t(39) = -1.29$, $p = 0.203$) and different ($t(39) = 1.38$, $p = 0.176$) segments were not significantly different from 0, even though the same segment was numerically negative, whereas the different segment was numerically positive (Fig 2C). The increase of similarity scores from same and early-similar segment to late-similar and late segment was consistent in timing with participants behavioral performance increase: the similarity scores were lower in CA23DG when memory competition has not been resolved, and increased after the competition was resolved.

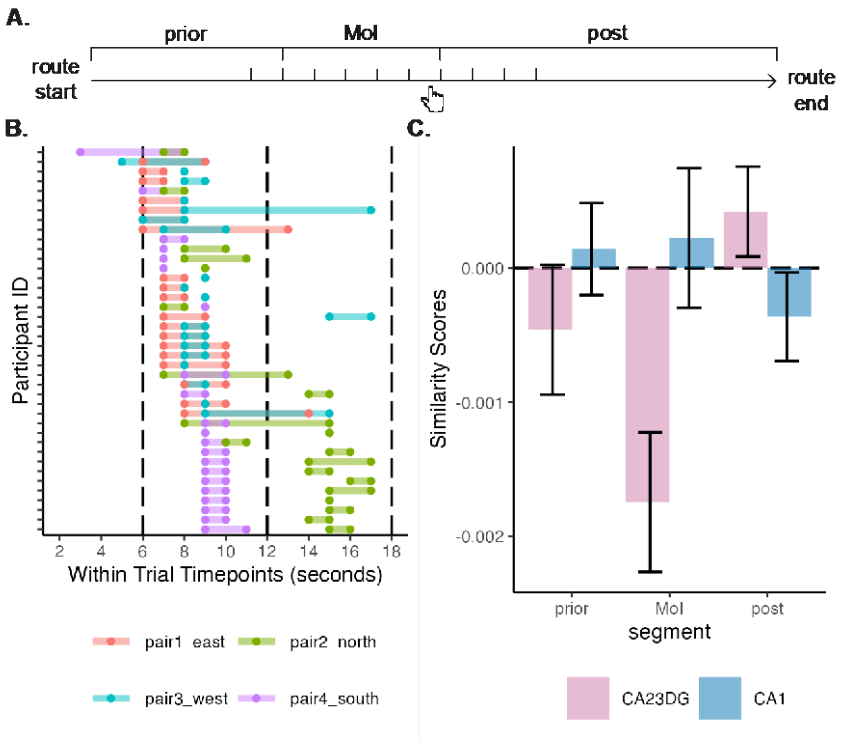


Figure 7. Similarity scores change with behavioral states.

(A) Schematic illustration showing how MoI was defined for each pair of overlapping routes. The earliest timepoint of MoI was the earliest moment when participants pressed a button to indicate they were at least 90% sure of the destination and chose the destination correctly across two routes within an overlapping pair. The end of MoI was the latest time the participants did so for a pair of overlapping routes. All timepoints before the earliest MOI were considered as Prior, and all timepoints after the latest MOI were considered as Post. **(B)** Behavioral distribution of the MoI for each participant for each pair of overlapping routes. Each participant studied two of the four possible routes pairs (east and west, or south and north). **(C)** Similarity scores grouped on each participants' behavioral definition of MoI for CA23DG and CA1. A one-way ANOVA showed significant main effect for segment in CA23DG ($F(2, 78) = 5.84$, $p = 0.004$) for similarity scores, but not in CA1 ($F(2, 78) = 0.69$, $p = 0.506$). The repulsion effect was significant in CA23DG with negative similarity scores at MOI ($t(39) = -3.35$, $p = 0.002$), and increased significantly from MoI to Post ($t(39) = 3.31$, $p = 0.002$).

Repulsion in CA23DG is time-locked to memory interference resolution moment (moment of insights).

To better understand the relationship between participants' behavioral indication of memory interference resolution and the repulsion effect in CA23DG, we used second posttest responses to pinpoint the exact timing as Moment of Insights (MoI) when participants indicated they were certain of the correct destination for each route. We defined the start of MoI as the earliest when participants pressed the button to indicate

they were certain of the destination for a pair of overlapping routes and the end of MoI as the latest time when participants pressed the button. Then, we used all timepoints prior to MoI as Prior, and all timepoints post to MoI as Post (Fig 3A). On average, participants responded within a narrow time window for each overlapping pair, with a length of only 1.7 seconds, indicating most participants used a consistent cue to distinguish the correct destination of an overlapping pair. The Prior segment had an average length of 8.7 seconds, and the post segment had an average length of 13.6 seconds (Fig 3B).

After grouping the similarity scores based on each participants' MoI for each route pair, a two-way ANOVA with factors of behavioral state (Prior, MoI, Post) and ROIs (CA23DG, CA1, PPA, EVC) showed a significant main effect of segment ($F(2, 429) = 22.614$, $p < 0.001$), significant main effect of ROIs ($F(3, 429) = 323.611$, $p < 0.001$), and a significant interaction ($F(6, 429) = 17.006$, $p < 0.001$) (not sure if I should change the major figure into 4 ROIs or adding the 4 ROI figure as a supplementary figure), indicating the similarity scores change differently within each ROIs across the behavioral resolution of memory interference. A one-way ANOVA within each ROI showed significant main effect in CA23DG ($F(2, 78) = 5.84$, $p = 0.004$), PPA($F(2, 78) = 11.36$, $p < 0.001$), and EVC ($F(2, 78) = 31.21$, $p < 0.001$), but not in CA1 ($F(2, 78) = 0.69$, $p = 0.506$). Both EVC and PPA showed no significant difference from prior to MoI segment (PPA: $t(39) = -0.71$, $p = 0.478$; EVC: $t(39) = 0.21$, $p = 0.834$), but the patter similarity decreased significantly from MoI to Post (PPA: $t(39) = - 3.44$, $p = 0.001$; EVC: $t(39) = -7.66$, $p < 0.001$), showing pattern similarity in visual attention region decreased as visual input became more distinct. However, CA23DG displayed an opposite pattern,

with a significant increase in pattern similarities from MoI to Post (t(39) = 3.31, p = 0.002), and no significant difference from Prior to MoI (t(39) = -1.77, p = 0.085), which supported that pattern similarities was depended on participants resolving memory interference and the similarities increased once memory interference was resolved. Lastly, one-sample t-tests showed a significant negative patter similarities within the CA23DG at MoI (t(39) = -3.35, p = 0.002), showing the repulsion effect in CA23DG was time-locked to the moment when participants were actively resolving memory interference.

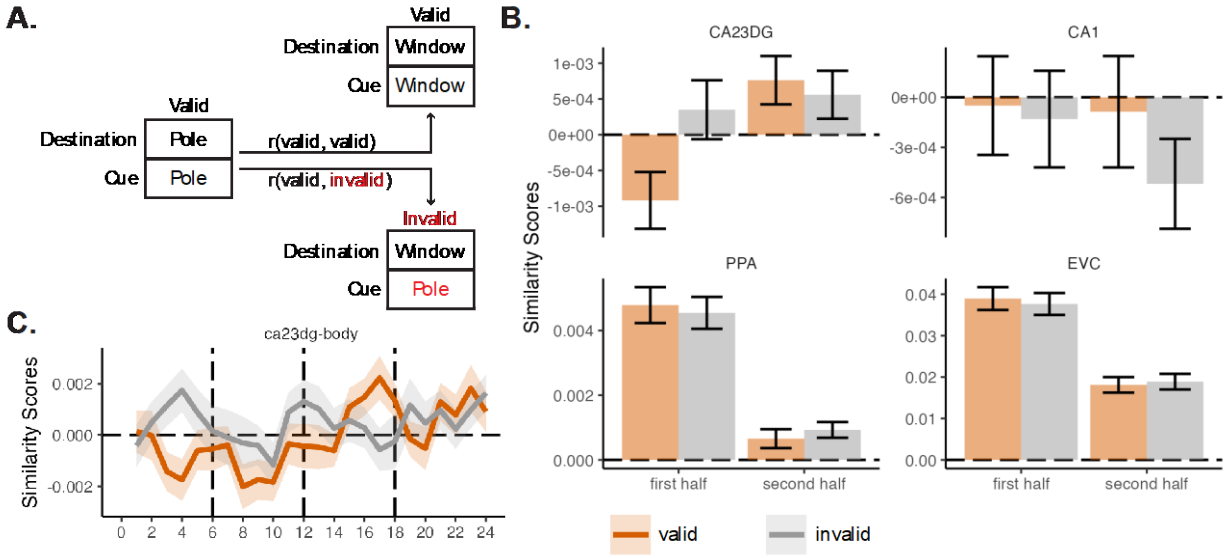


Figure 8. Similarity scores as a function of Cues (Valid vs. Invalid).

(A) Example schematic for the definition of valid-valid correlations ('Valid' condition) and valid-invalid correlations ('Invalid' condition). Correlations were always calculated between two routes that were overlapping routes (e.g., Pole and Window). We either correlated two overlapping routes both with valid cues (valid-valid, 'valid' condition), or one valid and one invalid cues (valid-invalid, 'invalid' condition). When cues were valid, both cues and destinations from an overlapping pair were distinct from each other. However, when one cues was valid and the other cue was invalid, two overlapping routes shared the same cue, but distinct destinations. Through manipulating whether participants saw identical or distinct cues, participants were led to a biased belief of which destination they were heading towards. **(B)** Similarity Scores separated into the first (picture index 1-50) and second (picture index 51-100) half of the route, separated by different internal states and different ROIs. Only CA23DG showed significant interaction between segment and internal states (F(1,117) = 4.48, p = 0.036). In CA23DG, for the early half, different states showed significant negative similarities scores (repulsion effect) (t(39) = -2.31, p = 0.026), which was significantly lower than same states from the early half (t(39) = -2.75, p = 0.009). The similarity scores in CA23DG also increased from the first to the second half of the route in CA23DG (t(39) = 3.47, p = 0.001). **(C)** Timepoint-by-timepoint similarity scores in CA23DG, separated by valid or invalid condition. Visually, there are two negative dips in CA23DG for valid condition, when participants saw distinct cues and destinations. One dip was at the same segment and the other one was at the early-similar segment.

Different cues led to repulsion between competing memories in CA23DG with identical visual inputs.

To test whether the repulsion effect was related to subjective internal states, one key manipulation we added to the task was the cue before each route started. With the cue, we can manipulate participants' internal belief of destination. To calculate similarity scores, we are always correlating two overlapping routes with different destination (e.g., Pole and Window). When we correlation two overlapping routes both with valid cue before the routes started (referred to as 'valid' condition), participants were more inclined to believe these two routes led to different destination, thus led to subjectively different internal states. However, when correlating two overlapping routes with one valid cue and one invalid cue (referred to as 'invalid' condition), participants were more inclined to believe these two routes led to the same destination (Fig 4A). Critically, in both valid and invalid conditions, the only difference in visual input is the cue that showed up prior to the start of the route. Once the route start, the visual inputs are identical since we were always correlation two overlapping routes. Therefore, any neural differences were caused by the cue manipulation, or subjective internal states, instead of external stimuli inputs.

We ran a two-way ANOVA with two halves of the route: first (picture index 1-50), second (picture index 51-100), and Cue (valid vs. invalid) for each ROI separately (Fig 4B). CA23DG showed a main effect between first and second half of the route ($F(1, 117) = 7.40$, $p = 0.008$) and a significant interaction between the halves and Cue ($F(1,117) = 4.48$, $p = 0.036$). CA1 didn't show any significant main effect or interaction ($F(1,117) < 0.77$, $p > 0.382$). PPA and EVC only showed a significant main effect of the

halves (PPA: $F(1, 117) = 131.87$, $p < 0.001$; EVC: $F(1, 117) = 214.05$, $p < 0.001$), but not significant main effect of Cue nor interactions. Thus, CA23DG was the only region that the Cue prior to the start of a route influenced pattern similarity.

To follow up, we ran a two-way ANOVA between the first two segments of the route ('same', 'early-similar') and different Cues (valid vs. invalid) and found a significant main effect of Cue ($F(1,117) = 7.69$, $p = 0.006$) but no main effect of segment ($F(1,117) = 1.34$, $p = 0.250$) or interactions between segment and Cue ($F(1,117)= 0.052$, $p = 0.820$). When running a two-way ANOVA between the two segments in the second half of the route ('late-similar', 'different') and Cue (valid vs. invalid), no significant main effect of Cue or segments was shown. Therefore, the effect of Cue within CA23DG was only present in the first half of the route, consistent with the timing that participants exhibited memory interference through behavioral performance. After participants used the external stimuli to identify the correct destination for the second half of the route, the effect of initial cue disappeared.

Within CA23DG, paired t-tests showed that in the first half of the route, the pattern similarity for Valid Cue was significantly lower than Invalid Cue ($t(39) = -2.75$, $p = 0.009$). The first half of the route with Valid Cue also showed pattern similarities significantly lower than 0 ($t(39) = -2.31$, $p = 0.026$), but the similarity scores from Invalid Cue was not significantly different from 0 ($t(39) = 0.85$, $p = 0.402$). These results support the idea that repulsion effect is dependent to participants' internal belief that two memories were distinct from each other, that was driven by different cues. When participants did not hold such belief (same cue), the repulsion effect was absent.

For a more precisely understanding in how the pattern similarities changed across the time of route, we also plotted the pattern similarities in CA23DG from timepoint to timepoint (Fig 4C). As we can visually inspect, there were two “dips” in CA23DG with different states. One sample t-tests showed that timepoint 3, 4, 8, 9, 10 were significantly below 0 ($0.007 <= p < 0.039$). Importantly, all these timepoints with the repulsion effect were at the first half of the route, while participants suffered from memory interference. Moreover, paired t-tests between valid and invalid cues showed significant differences only at time point 3 ($p = 0.009$) and 4 ($p = 0.002$), both of which located at the same segment, while the visual inputs were identical between two overlapping routes. Thus, participants had to use the cue at the start of the route to estimate the potential destinations.

2.4 Discussion

Here, we showed that high similarity between routes led to hippocampal pattern representations of the routes being pushed away from each other. These activity patterns were pushed so far from each other, that the representations of overlapping routes were further from each other than non-overlapping routes. We also found that this repulsion effect between overlapping routes was limited to the timepoints when routes were highly similar and when memory interference still exists. We identified that hippocampal similarity was the lowest at the moment when participants resolved memory interference. Lastly, we showed that, consistent to rodent research, hippocampus pattern repulsion can happen even when visual stimuli were identical, if the participants were led to believe they were distinct.

The current study design and analysis were inspired by both computational modeling and rodent empirical work related to hippocampus remapping[4]. We were interested in testing the idea that hippocampal representational patterns can change with different internal, subjective beliefs, independent to changes in the external, objective environments. Our finding provided a novel perspective showing that hippocampus representational change can happen solely with changes in participants' internal beliefs. Several details in our paradigm were critical to interpret our findings. Firstly, we controlled the levels of similarities between overlapping routes. Each pair of overlapping routes started at a 'same segment', during which participants viewed identical images and made it impossible to distinguish between overlapping routes. Second, by adding a probabilistic cue prior to the start of the route, we manipulated participants' internal belief regarding which route they were traversing for each trial. Lastly, the similarity scores were always calculated between overlapping routes, keeping the visual inputs identical, despite the distinct cues. Overall, between two overlapping routes, at the 'same segment', participants had no external information to decide the destination and could only utilize the cues we provided at the start. Then, during the 'similar' and 'different' segments, participants could use visual differences to identify the correct destination.

This study design provides an important baseline to test the effects of distinct internal beliefs, relative to identical internal beliefs, while holding the external visual inputs identical. This baseline comparison is absent in previous rodent and human literatures to our knowledge. In the current study, we showed that during the first half of the route, while participants were still uncertain of the destination, the pattern similarity

between distinct cues were lower than same cues, controlling for the visual inputs. However, this difference was absent in the second half of the route, when participants behaviorally indicated they were certain of the destination. It is possible that while the external information was not sufficient to resolve memory interference, participants had no choice but to trust the cue, leading to lower pattern similarity between overlapping routes with distinct cues, relative to same cues. Later, after more external information was gathered, participants relied less on the initial cues and updated their internal belief regarding the destination. Overall, this result provides novel evidence that hippocampus representations depended on what participants believed internally, instead of what they saw externally.

Different from rodent remapping studies, to calculate similarities between overlapping routes, we used non-overlapping routes similarity as the baseline. The baseline here is meaningful: the representational patterns between two non-overlapping routes should be orthogonal to each other or share no overlap. Pattern separation is believed to be the process that the hippocampus disambiguates similar events by decorrelate the overlapping signal from each other[10,11]. Pattern separation's defining characteristic is the orthogonal signal between two events. In other words, pattern separation should result in the same similarities as non-overlapping routes. Therefore, using pattern similarities between non-overlapping routes as baseline, we can measure how the relationship between overlapping routes change. When the similarity score, which is the similarities between overlapping routes subtracted by non-overlapping routes, is above 0, it indicates that overlapping routes were represented more similarly than non-overlapping routes. When the score is equal to 0, it indicates the same level of

distinction between overlapping and non-overlapping routes, which is pattern separation. Lastly, also what we saw in our results from CA23DG, if similarity scores were lower than 0, it indicates that the pattern similarities between overlapping routes were lower than non-overlapping routes, which is even more distinct than being orthogonal to each other. We call this the repulsion effect in the hippocampus, since the two overlapping events were systematically pushed away from each other.

Critically, after aligned the pattern similarity with behavioral indication of memory interference resolution for each participant, the similarity scores were negative at the Moment of Insight (MoI), the moment right when participants resolved memory interference. This result suggested that similar memories, instead of being orthogonalized, was *abruptly* pushed away from each other. Numerous previous human fMRI studies have shown the repulsion effect and demonstrated its relationship with learning and memory interference. Consistent with Chapter 1, the fact that repulsion happened abruptly at the MoI suggested that it was strongly associated with learning and can be generalized into spatial memory domain. Furthermore, our current result provided novel evidence that the timing of the abrupt repulsion can happen in seconds consistently within a trial, while participants were actively resolving memory interference.

In our study, the change in representational patterns happened consistently in CA23DG but not CA1. This is in line with our prediction that CA1, with its role more emphasized on integrating existing memories with incoming ones[11], didn't show the repulsion effect. The abruptness of the repulsion effect supports the idea that repulsion was rooted from the attractor dynamic in CA3[12–14]. This is also consistent with the

findings that CA3/DG showed faster remapping than CA1[15–17]. The representational patterns in CA3/DG also provided striking contrast with visual regions, both EVC and PPA[7]. EVC and PPA's pattern similarity was highly consistent to stimuli similarity, in a sense that the pattern similarity decreased as the overlapping routes became more distinct. Moreover, EVC and PPA didn't show cue-related changes in representational patterns, indicating they were not influenced by internal state like CA3/DG.

Overall, our current study reveals that hippocampal repulsion can be triggered with distinct internal belief, without different external inputs. These findings provided novel insights to understand the computation in the hippocampus while resolving memory interference. These results are also consistent with rodent remapping literature and computational modeling theory, focusing on the subjective model within the animal instead of the objective environments. The findings that related hippocampus repulsion with behavioral resolution also draw a direct link between repulsion and memory interference resolution, suggesting repulsion being a general mechanism in the hippocampus across different types of memories.

2.5 Methods

Participants.

Forty-eight participants (27 female; mean age = 20.40 years, range = 18–32 years) were enrolled in the experiment following procedures approved by the University of Oregon Institutional Review Board. Written informed consent was collected for each participant prior to the experiment. All participants were right-handed native-English speakers with normal or corrected-to-normal vision, with no self-reported psychiatric or

neurological disease. 3 participants were excluded because they didn't finish the experiment (n = 2) or exited the scanner in the middle of the experiment (n = 1). 5 participants were excluded due to low behavioral performance (see Results for more details). The final analysis included 40 participants. All participants received monetary compensation for participating.

Stimuli.

The stimuli consisted of eight routes, with a stream of 100 images each, that traversed the University of Oregon campus. Images were screenshots taken from videos at a constant time interval. The videos were taken from an egocentric perspective while a researcher walking along the route. All routes started at the same 4-way intersection on the campus and ended at eight distinct destinations named after a visual object at the ending location. The 8 routes can be grouped into 4 pairs of overlapping pairs (Fig 1A). Within an overlapping pair (e.g., Pole and Window), the first 25 images of the route not only shared the same path, but the images were also identical to each other, making it impossible to distinguish one route from another within a pair ('same' segment). The next 50 images of the overlapping routes were traversing at the same path, but at slightly different time, led to images with subtle differences from one another ('similar' segment). For last 25 images, the overlapping routes diverged after a turn to their respective destinations ('different' segment) (Fig 1B). Each overlapping pair left the starting intersection at a different cardinal direction (north, south, east, west). Each participant was assigned to either the north/south routes or the east/west routes alternatively. Therefore, each participant was only exposed to 4 out of the 8 total routes.

Experimental procedure.

At a preparation room, participants were provided with consent and given instructions for the whole experiment. Then, participants finished the Learning phase at the preparation room before entering the MRI scanner. Inside the scanner, participants finished 2 rounds of the Practice phase (not scanned) and 10 rounds of the Cue phase (scanned). After exiting the scanner, participants completed two additional tasks measuring their memorization of each route. All stimuli were presented on a gray background, projected from the back of the scanner. Lights were turned off during the scan to ensure better contract for the display. The experiment was implemented in PsychoPy2021.2.3 and lasted for ~3 hours, with 2 h 30 min inside the scanner.

Learning. During the Learning phase, participants viewed each route for 4 times at random order. Each image in the route was shown for 240 ms. After the route was shown from start to finish (24 seconds), the destination for the route appeared (2000 ms), followed by a white fixation cross (3000 ms). Then the next route start.

Practice. During each round of the Practice phase, participants also viewed each route for 4 times at random order. For 3 of the 4 times, the route is identical to the previous Learning phase. However, for once per route, the route will pause once at each segment at random picture index, followed by an associative test. The associative test showed all 4 possible destinations on the screen and the participants had a maximum of 4000 ms to choose destination of the current route. If they answered within the given time, a confidence choice would appear on the screen (‘sure’ vs. ‘unsure’), and the participants had another 3000 ms max

to answer. The possible picture indices are 10-25 for the same segment, 26-75 for the similar segment, and 76-90 for the different segment.

Cue. In each round of the Cue phase, participants viewed each route for 4 times. Importantly, each trial was preceded by a cue indicating the likely destination. These cues could be **valid**, indicating the correct destination (e.g., Pole), or **invalid**, indicating the overlapping route's destination (e.g., Window). Within each scan round, there were 2 valid trials, 1 invalid trial, and 1 catch trial for each route. The catch trial was always proceeded with a valid cue and would stop at random picture index during the similar segment, followed by the same associative test as the Practice phase to ensure participants' attention during the scan. Different from the Practice phase which the route continued and ended at the destination after the associative test, the route ended after the associative test for the catch trials in the Cue phase. Otherwise, each trial used the same timing as the Learning phase.

PostTest1. Each route appeared 4 times at a random order for the first posttest. The route started from the beginning (without cue prior to the route). Then an associative test appeared for every trial at picture index 30, 45, 60, and 75, all in the similar segment. In this associative test, a slider is used to measure participants' memory and confidence at the same time. On the slider, the correct destination is at one side, and the overlapping route's destination at the other. Critically, in the middle of the slider, participants can also indicate 'unsure' of the destination. Participants can also choose "probably" or "definitely" as the confidence rating for each destination. There is no time limit for response. After

the associative test at picture index 75, next trial starts after a white fixation cross (3000 ms).

PostTest2. Each route appeared 4 times at a random order for the second posttest. The route started from the beginning (without cue prior to the route). Participants were instructed to press '6' when they reached 90% certainty towards the destination, then all 4 choices of the possible destinations appeared on the screen, and the participants had to choose the destination for the route. The next trial started right after the choice after a white fixation cross (3000 ms).

MRI acquisition.

All images were acquired on a Siemens 3T Prisma MRI system in the Lewis Center for Neuroimaging at the University of Oregon. Functional data were acquired with a T2*-weighted echo-planar imaging sequence with partial brain coverage that prioritized full coverage of the hippocampus and EVC (repetition time = 1000 ms, echo time = 33 ms, flip angle = 55°, 66 slices, 1.7 × 1.7 × 1.7 mm voxels). A total of 10 functional scans were acquired. Each functional scan comprised 458 volumes and included 6 s of lead-in time and 6 s of lead-out time at the beginning and end of each scan, respectively. Anatomical scans included a whole-brain high-resolution T1-weighted magnetization prepared rapid acquisition gradient-echo anatomical volume (1 × 1 × 1 mm voxels) and a high-resolution (coronal direction) T2-weighted scan (0.43 × 0.43 × 1.8 mm voxels) to facilitate segmentation of hippocampal subfields.

Anatomical data preprocessing.

Preprocessing was performed *using fMRIPrep* 21.0.1 (RRID:SCR_016216), which is based on *Nipype* 1.6.1 (RRID:SCR_002502). The T1-weighted (T1w) image

was corrected for intensity nonuniformity (INU) with N4BiasFieldCorrection54 (ANTs 2.3.3, RRID: SCR_004757), and used as the T1w reference throughout the workflow. The T1w reference was skull-stripped with the antsBrainExtraction.sh workflow (ANTs) in Nipype, using OASIS30ANTs as the target template. Brain tissue segmentation of cerebrospinal fluid (CSF), white-matter (WM), and gray-matter (GM) was performed on the brain-extracted T1w using fast (FSL 6.0.5.1:57b01774, RRID:SCR_002823). Brain surfaces were reconstructed using recon-all (FreeSurfer 6.0.1, RRID:SCR_001847). Volume-based spatial normalization to one standard space (MNI152NLin2009cAsym) was performed through nonlinear registration with antsRegistration (ANTs 2.3.3), using brain-extracted versions of both T1w reference and the T1w template. *ICBM 152 Nonlinear Asymmetrical template version 2009c* was selected for spatial normalization (RRID:SCR_008796; TemplateFlow ID: MNI152NLin2009cAsym).

Functional data preprocessing.

fMRIPrep.

For each of the 10 BOLD scans per participant, the following preprocessing was performed. First, a reference volume and its skullstripped version were generated by aligning and averaging 1 single-band references (SBRefs). Fieldmap was collected and estimated based on two (or more) echo-planar imaging (EPI) references with topup (FSL 6.0.5.1:57b01774). The estimated *fieldmap* was then aligned with rigid-registration to the target EPI (echo-planar imaging) reference run. The field coefficients were mapped on to the reference EPI using the transform. BOLD runs were slice-time corrected to 0.445s (0.5 of slice acquisition range 0s-0.89s) using 3dTshift from AFNI (RRID:SCR_005927). The BOLD reference was then co-registered to the T1w

reference using bbregister(FreeSurfer). Several confounding time-series were calculated based on the *preprocessed BOLD*: framewise displacement (FD), DVARS and three region-wise global signals. Additionally, a set of physiological regressors were extracted to allow for component-based noise correction (*CompCor*). Principal components are estimated after high-pass filtering the *preprocessed BOLD* time-series (using a discrete cosine filter with 128s cut-off) for the two *CompCor* variants: temporal (tCompCor) and anatomical (aCompCor). Frames that exceeded a threshold of 0.5 mm FD or 1.5 standardized DVARS were annotated as motion outliers. The BOLD time-series were resampled into standard space, generating a *preprocessed BOLD run in MNI152NLin2009cAsym space*.

fMRI first-level general linear model (GLM) analyses.

After fMRIPrep preprocessing, the first 6 frames of each *preprocessed BOLD* were discarded. Then, eight brain masks were generated by fMRIPrep for each of the eight functional scans. The intersection of all eight masks was used to perform brain extraction. Each *processed BOLD* was then scaled at a mean equal 100, with an upper bound of 200 and a lower bound of 0. Each voxel in *preprocessed BOLD* A high-pass filter of 128 was applied to the *preprocessed BOLD*, and any volumes with spike higher than 3 standard deviation or FD higher than 0.5mm was discarded from the analysis. A GLM with only nuisance regressors was applied for each functional scan with FD, xyz translation, xyz rotation, aCompCor00-05, and csf. Each functional scan was then smoothed with a sliding window of 3 seconds.

Regions of interest.

A region of interest (ROI) for EVC was created from the probabilistic maps of Visual Topography63 in the MNI space with a 0.5 threshold. This ROI was transformed into each participant's native space using inverse T1w-to-MNI nonlinear transformation. For each participant. An ROI for the PPA was created by first using an automated metanalysis in Neurosynth with the key term "place". Then, clusters were created using voxels with a z-score >2 based on the Neurosynth associative tests. Since these clusters were generated through an automated meta-analysis and were not anatomically exclusive to PPA, we visually inspected the results and manually selected the two largest clusters that were spatially consistent with PPA. One cluster was in the right hemisphere (voxel size = 247) and one cluster was in the left hemisphere (voxel size = 163). These clusters were combined into a single PPA mask. This mask was then transformed into each participant's native space using the inverse T1w-to-MNI transformation. For each participant, a final PPA ROI was generated by averaging the t-maps of all scene exposure phase scans. To create hippocampal ROIs, we used the Automatic Segmentation of Hippocampal Subfields (ASHS)64 toolbox with the upenn2017 atlas to generate subfield ROIs in each participant's hippocampal body, including CA3/dentate gyrus (which included CA2, CA3, and dentate gyrus) and CA1. The most anterior and posterior slices of the hippocampal body were manually determined for each participant based on the T2-weighted anatomical structure (see Supplementary Fig. 7 for a sample demarcation). Each participant's subfield segmentations were also manually inspected to ensure the accuracy of the segmentation protocol. Then, each subfield ROI was transformed into each participant's native space using the T2-to-T1w transformation, calculated with FLIRT (fsl) with six

degrees of freedom, implemented with Nipype. All ROIs were again visually inspected following the transformation to native space to ensure the ROIs were anatomically correct.

fMRI pattern similarity analyses.

Similarity Scores.

Pattern similarity was calculated as the Pearson correlation coefficient between BOLD signal with 6 seconds lag relative to visual input to account for the hemodynamic response. All pattern similarity analyses were performed by correlating the BOLD signal across scans (e.g., run 1 correlating with run 2-10) after removing all catch trials within each ROI. To measure how the pattern similarity change within a route, correlations were performed on a frame-to-frame fashion relative to the timing of each trial: for example, the first second of one trial was always correlated the first second of other trials from a different run. Of central interests was the similarity between overlapping routes relative to the similarity between non-overlapping routes. Specifically, for each set of overlapping routes, the mean non-overlapping similarity was subtracted from mean overlapping similarity to yield a Similarity Score for each set of overlapping routes.

As an example, to calculate the **valid** Similarity Scores between Pole and Window at the first second of the trial, we took the BOLD signal at the first second of valid trials, and calculated the mean of r(Pole, Window) from different runs, subtracted by the mean of r($Pole_{Valid}$, $Shrub_{Valid}$), r($Pole_{Valid}$, $Pergola_{Valid}$), r($Window_{Valid}$, $Shrub_{Valid}$), and r($Window_{Valid}$, $Pergola_{Valid}$) from different runs.

To calculate invalid Similarity Scores, we would still take differences between overlapping routes and non-overlapping routes. However, one of the routes in the correlations would have an invalid cue at the start of the route. For example, to calculate the **invalid** Similarity Scores between Pole and Window for each second within the trial, we calculated the mean of $r(Pole_{Valid}, Window_{Invalid})$ and $r(Pole_{Invalid}, Window_{Valid})$ from all different runs, subtracted by the mean of $r(Pole_{Valid}, Shrub_{Invalid})$, $r(Pole_{Invalid}, Shrub_{Valid})$ … $r(Window_{Valid}, Pergola_{Invalid})$, and $r(Window_{Invalid}, Pergola_{Valid})$ from different runs at the same second within the trial.

Moment of Insights (MoI).

To relate similarity scores to behavioral measurements of resolving memory interference, we identified the MoI separately for each overlapping routes, for each participant. The MoI was based on the performance in the second Posttest. For each route, participants indicated for 4 times the exact timing when they are at least 90% certain of the destination. Therefore, for each overlapping routes, participants indicated 8 times the timing when they were able to distinguish between the pair. We took the earliest timepoint of the 8 responses as the start of MoI and the latest timepoint of the 8 responses as the end of MoI. Everything prior to the earliest timepoint was considered as Prior and everything after the latest timepoint was considered as Post.

Chapter 3. Hippocampal repulsion as a function of exposure

3.1 Abstract

When events are similar to each other, memories of the events can be easily confused with one another, which is termed ‘Memory interference’. Memory interference can be resolved through repeated exposure and recall. In human fMRI studies, hippocampal repulsion was directly linked to the resolution of memory interference, even though the direct link between levels of experience and repulsion has yet to be established. The current study uses high resolution fMRI to measure how repulsion changes as a function of experience. We show that with an accumulation of experience, repulsion does not simply increase linearly. Instead, repulsion is shown to be a transient effect that is the strongest while memory competition is the greatest, and subsides after interference is resolved.

3.2 Introduction

The hippocampus plays a critical role in encoding and retrieving memories. However, a lot of daily memories overlap with each other, causing memory interference. Evidence from recent human fMRI studies indicates that one of the critical hippocampal mechanisms that facilitates the resolution of memory interference is *hippocampal repulsion*[1–3]. Distinct from hippocampus pattern separation, which refers to the orthogonalization of memory events[4], hippocampal repulsion refers to a more targeted differentiation process in which similar memories are actively ‘pushed away’ from each other (in representational space). Clear evidence for repulsion occurs when two initially overlapping memories become associated with activity patterns (representations) that are less similar to each other than to other, non-overlapping memories. In contrast,

orthogonalization of memory representations would result in overlapping memories becoming as distinct as non-overlapping memories, but not more distinct[5].

Several previous human fMRI studies have established connections between hippocampal repulsion and differing levels of experience with similar events. Both Favila et. and Chanales et al. showed hippocampal repulsion happened after training. Also, in the first chapter of the current dissertation, our results indicate that hippocampal repulsion is an abrupt change that is time locked to the moment when interference is resolved, providing evidence to support the idea that repulsion depends highly on participants' experience. However, empirical evidence also shows that more experience does not necessarily lead to more repulsion. For example, data presented in the first chapter shows that repulsion only occurred during the transition of behavioral performance from not being able to resolve memory interference to being able to. However, for the runs afterwards, when participants could consistently resolve memory competitions between pairmates, repulsion effect was absent. This result indicates that the similar memories didn't stay being repulsed from each other with increased experience. Moreover, Chanales et al and Chapter 2 provide evidence that hippocampal representational patterns can change dramatically even within a trial. Specifically, timepoint-by-timepoint data shows that repulsion can happen when routes share overlap but disappears when the routes diverge. These results further suggest that hippocampal representations are highly dynamic and change in a non-linear pattern, even at the level of seconds. However, a direct test regarding how repulsion changes as a function of experience remains missing.

Within the hippocampus, different subfields make distinct contributions regarding memory formation and memory interference resolution, with major comparisons between CA1 and CA3/ dentate gyrus (DG). In previous literatures, CA1 was shown to play an important role in forming integrated memory, by combining incoming information with existing memories[6–9]. On the other hand, due to the 'sparse coding' nature of the DG, it was shown in multiple studies that DG represents incoming information from ERC less similar to each other[10]. However, one mechanism that we think plays a critical role in hippocampal repulsion, is the attractor dynamics in CA3[11]. The attractor dynamics in CA3 can result in a non-linear relationship between input similarities and the representations in CA3[12,13], which is a striking contrast from CA1 that showed a more linear correspondence with perceptual changes[7,14]. The nonlinearity of attractor dynamics makes it a reasonable underlying mechanism for hippocampal repulsion, that also showed a nonlinear relationship between stimuli similarities and hippocampal representations[15,16]. Even though the resolution of human fMRI made it hard to separate CA3 and DG as separate subfields in the hippocampus, previous chapters of the dissertation showed that CA3/DG showed repulsion between similar stimuli, whereas the effect was absent in CA1.

In the current chapter, we utilized an associative learning paradigm to investigate how different levels of experience will influence the levels of repulsion in CA3/DG in the hippocampus. Participants learned 48 associations between scene and object images. The stimuli in the current study include 24 scenes of beaches and 24 scenes of gazebos. Critically, two categories of scenes (beaches and gazebos) were assigned as either high or low training group, counter-balanced across participants. With different

levels of training, we are interested in whether levels of experience will influence the level of repulsion.

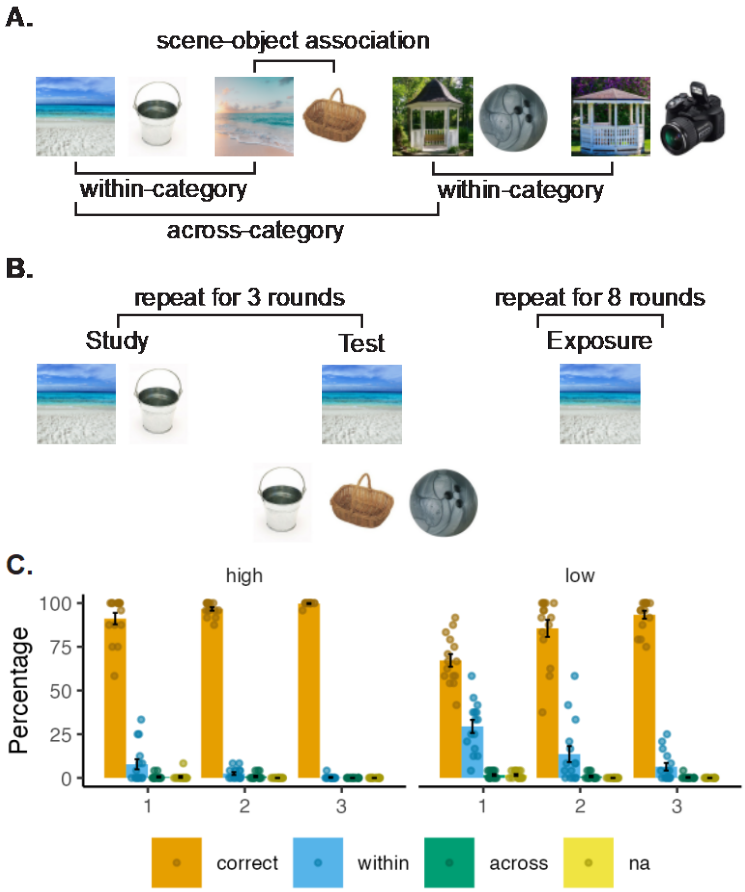


Figure 9. Experimental procedure and behavioral results.

(A). Participants learned 48 scene-object associations. The 48 scenes comprised 2 categories of scenes: 24 beaches and 24 gazebos. For each participant, one scene category was assigned as the 'high training' group, and the other scene category was assigned as the 'low training' group. Objects were randomly assigned to each scene for each participant. **(B).** Participant finished 3 rounds of learning before entering the fMRI scanner. Each round of learning included a study section and an associative memory test phase. During the Study sections, participants viewed scenes and associated object, with high training group been shown at higher frequency and longer length. During the Testing phase, participants were presented with scenes and had to select the associated object from a set of three choices: one correct object, one object that was associated with the within-category scene, and one object that was associated with the across-category scene. Only high-training group scenes were followed by feedback after participants' selection. Then, inside the scanner, participants finished 8 rounds of Exposure, during which they performed an old/new judgement for scenes. **(C).** Participants' performance during the associative memory test, separated by two levels of training and the number of rounds. A two-way ANOVA showed accuracy significantly varied with different level of training ($F(1, 70) = 53.80$, $p < 0.001$), number of rounds ($F(2, 70) = 29.73$, $p < 0.001$), as well as the interaction of the two ($F(2, 70) = 7.72$, $p < 0.001$).

3.3 Results

Participants were instructed to memorize 48 scene-object associations, with 24 scenes of beaches and 24 scenes of gazebos. Participants finished three rounds of Study and Test before entering the fMRI scanner. Importantly, for one of the categories (beaches or gazebos, counterbalanced) participants received extensive training on the associations (high training) whereas for the other category they received more limited training (low training) (See Methods for more detail). Then, in the fMRI scanner, participants completed eight rounds of Exposure phase. Within each round of the Exposure phase, each scene was shown once, intermixed with some novel scenes (also beaches or gazebos). Participants were instructed to make a button press whenever a novel scene appeared.

Behavior

During the associative memory test, participants chose the correct object with above-chance accuracy in each of the three rounds for both the high training group (round1: $t(14) = 17.51$, $p < 0.001$, CI = [0.91 ± 0.06] ; round2: $t(14) = 62.55$, $p < 0.001$, CI = [0.97 ± 0.02] ; round3: $t(14) = 239$, $p < 0.001$, CI = [0.99 ± 0.01] ; one sample t-tests vs. 1/3) and the low training group (round1: $t(14) = 9.38$, $p < 0.001$, CI = [0.67 ± 0.07] ; round2: $t(14) = 10.75$, $p < 0.001$, CI = [0.86 ± 0.10] ; round3: $t(14) = 26.17$, $p < 0.001$, CI = [0.93 ± 0.04]; one sample t-tests vs. 1/3). An ANOVA with factors of training level (high vs. low) and Rounds revealed a significant main effect of both training level ($F(1, 70) = 53.80$, $p < 0.001$) and rounds ($F(2, 70) = 29.73$, $p < 0.001$), as well as a significant interaction ($F(2, 70) = 7.72$, $p < 0.001$) in the rate of choosing the correct object. Overall, high training scenes showed higher accuracy in associative test than the

low training scenes ($t(14) = 5.79$, $p < 0.001$). The older adult participant performed below the mean for the younger adults across rounds and training groups (z-scores for high training rounds 1 to 3: -1.26, -2.34, -3.61; low training: -0.64, -1.45, -1.60). During the Exposure phase, participants showed higher d' for the high training scenes than the low training scenes ($t(14) = 2.22$, $p = 0.043$; high training CI = [0.78 ± 0.06]; low training CI = [0.69 ± 0.09]). The older adult's z-score for d' was -0.23 for the high training scenes and 1.13 for the low training scenes.

Extensive training accelerated hippocampal repulsion effect in CA23DG.

For our fMRI analyses, our primary focus was on pattern similarity among similar scenes—that is, scenes from the same category. Pattern similarity was measured by correlating fMRI BOLD activity evoked while viewing each scene during the exposure phase. All correlations were calculated across rounds. Within category similarity was calculated as the pairwise correlations between every possible pair of scenes from the same category (e.g., 'beach 1' was correlated with 'beach 2', with 'beach 3' … and with 'beach 24'). Correlations were also computed between every possible pair of scenes across categories (e.g., for 'beach 1' - 'gazebo 1', 'beach 1' - 'gazebo 2', etc.). These across-category correlations were used as a baseline. Specifically, for each participant and region of interest, the mean across-category correlation was subtracted from the within-category correlation, yielding a similarity score. If the similarity score was above 0, it indicated that scenes from the same category were more similar to each other than scenes from different categories. If the similarity score was negative, it indicated that scenes from the same category were *less* similar to each other than scenes from

different categories. Negative similarity scores were taken as evidence for a repulsion effect.

Following the previous chapters, as well as other literature in the field with similar stimuli and analyses, fMRI analyses targeted the following regions of interest (ROIs): hippocampus, parahippocampal place area (PPA), and early visual cortex (EVC). PPA and EVC served as important control regions indexing high-level (PPA) and low-level (EVC) visual representations. Within the hippocampus, we leveraged our high-resolution fMRI protocol to segment the hippocampus body into subfields comprising CA1 and a combined CA2/3/dentate gyrus (CA23DG) as the previous chapters (see Methods). Motivated by past empirical findings and theoretical models, we predicted that repulsion would occur in CA23DG and different levels of training would influence the timing at which repulsion happens. To understand the timing when repulsion happened based on different level of training, we separated eight rounds of exposure into two halves: early (1-4 rounds) and late (5-8 rounds). For the early exposure, we only took correlations between trials that were both during the first 4 rounds of the scan, and for the late exposure, we only took correlations between trials that both appeared in the last 4 rounds of the exposure.

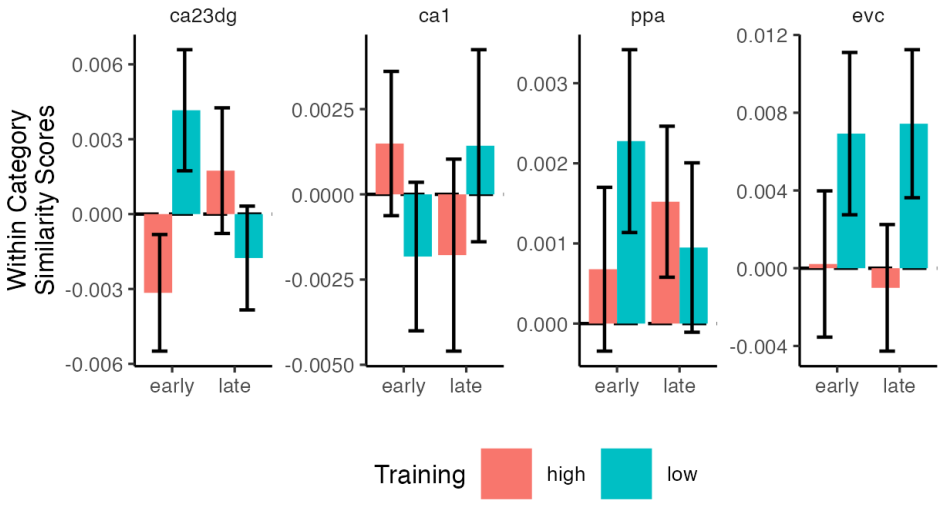


Figure 10. Similarity Scores separated by levels of training (high vs. low) and Exposure rounds (early vs. late).

We ran two-way ANOVAs for each ROI to test the effect of training (high vs. low) and Rounds (early vs. late) on similarity scores. Of our central interests, in CA23DG, there was no main effect of training ($F(1,56) = 0.66$, $p = 0.42$) or exposure ($F(1,56) = 0.05$, $p = 0.828$) alone, but there was a significant interaction between training and exposure ($F(1,56) = 5.31$, $p = 0.025$), indicating that the change in similarity scores changed from the early to the late half of rounds depended on the level of training. As can be seen in Figure 10, similarity scores tended to increase across rounds (early to late rounds) in the high training condition but to decrease across rounds in the low training condition. Similarity scores were numerically negative in the high training group during early exposure rounds (consistent with a repulsion effect), whereas the low training group exhibited numerically negative similarity scores during the late half of the exposure rounds. High training scenes showed accelerated repulsion effect comparing to low training scenes might be caused by the extensive training prior to the exposure

rounds. With more extensive training, memory interference resolution happened earlier for the high training scenes than the low training scenes.

For other ROIs, EVC showed a significant main effect of training for similarity scores ($F(1, 56) = 4.05$, $p = 0.049$). Specifically, high training group showed lower pattern similarities than the low training group, indicating differing levels of training influence representational structures in the visual area. Otherwise, CA1, PPA, and EVC didn't show other additional significant main effect of training (CA1: $F(1, 56) = 0.00$, $p = 0.983$; PPA: $F(1, 56) = 0.24$, $p = 0.624$) or exposure (CA1: $F(1, 56) = 0.00$, $p = 0.996$; PPA: $F(1, 56) = 0.054$, $p = 0.817$; EVC: $F(1, 56) = 0.009$, $p = 0.925$), nor a significant interaction between the two (CA1: $F(1, 56) = 1.70$, $p = 0.198$; PPA: $F(1, 56) = 1.08$, $p = 0.303$; EVC: $F(1, 56) = 0.05$, $p = 0.819$) for similarity scores.

3.4 Discussion

Here, the initial results showed that both the level of training and the number of exposure rounds had an influence on repulsion effects in CA23DG. Despite the relatively small number of participants (n = 15), we still found that the number of exposure rounds showed different influences for images with different levels of training on pattern similarities in CA23DG. We found hints of hippocampal repulsion only in CA23DG, not CA1, PPA, or EVC, which is similar to previous literatures[2,17]. This result support the idea that hippocampal repulsion might be related to the attractor dynamics in CA3 subfield within the hippocampus.

In addition to the repulsion effect, the interaction results showed that the influence of exposure and training were not simply additive. Specifically, the trend of repulsion appeared in the high training group during the first half of the exposure but

was absent afterwards. On the other hand, the low training group didn't show signs of repulsion during the first half of the exposure, but repulsion emerged afterwards during the second half of the exposure. This suggested that the extensive training prior to the exposure phase made the repulsion for the high-training category happen at an earlier timepoint. Based on behavioral performance, the high-training category also had significantly better performance than the low training category. Therefore, this result is also consistent with the idea that repulsion happens at the exact same time as the resolution of memory interference: high training with earlier memory interference resolution showed earlier repulsion. At the same time, the low training category still demonstrated hints of a repulsion effect, though only during the late half of the exposure. It is possible that during the first half of the exposure, low training scenes had a higher to be seen by the participants after limited training, and the increased familiarity led to more interference resolution among images. Therefore, low training scenes only displayed negative similarity values in the latter half of the route. Lastly, consistent with previous literature, CA1, PPA, and EVC didn't show training related changes in pattern similarity.

So far, the results for the current study are still preliminary. A few planned analyses will be conducted after more data are collected, and we will focus the future investigation in the following two perspectives.

First, we are interested in how images similarities within a category influences the pattern similarity in CA3/DG across the length of exposure. To be more specific, image similarities can be measured in multiple angles. In here, I will focus on two potential measurements: visual similarities and memory similarities. To measure how

visual similarities influence pattern similarities in CA3/DG, we can relate the pattern similarity in each participants' visual cortex with the pattern similarity in CA3/DG. In this way, we can test the hypothesis that the relationship between visual input information and hippocampus representations is nonlinear at a subject-level. To measure the relationship between memory similarity and pattern similarity in the hippocampus, in addition to the coarse separation of high and low training category, we can also identify different level of memory performance within each category. It makes sense that some images were easier to remember the associated objects than others. This measurement in behavioral performance is related, but not identical to, the visual similarity. We are interested in testing whether the memorability within each category can lead to distinct representational similarity patterns over the course of exposure. For example, are the images that are more easily memorized being repulsed from all other images during early exposure? If two images are especially similar relative to all other images, how will these two be repulsed from each other, and what will repulsion look like between either one and any other image within the same category?

Second, we plan to include older adult in the study, though the current analysis only included 1 older adult data. One of the goals of the current study is to provide a novel perspective in how the repulsion effect changes with age. Even though older adults usually perform worse in memory tasks with similar objects, providing a high and low training category gives us a chance to compare younger and older adults at a similar behavioral performance level (e.g., comparing older adults in the high training category against younger adults in the low training category). There are two interesting potential hypotheses: 1) older adults will show lower levels of hippocampal repulsion

than younger adults. In this case, the results should suggest that even though older adults' hippocampal mechanisms were hindered by aging, displayed by a weaker repulsion effect in the hippocampus, with more training, there might be some other neural mechanisms in the cortex that can facilitate the resolution of memory interference. 2) older adults will show a stronger level of hippocampal repulsion than younger adults. This might mean that the repulsion effect, instead of representing the resolution of memory interference, represents the effort towards that resolution. In other words, with a longer process of attempting to resolve memory interference, older adults might require stronger effort to push similar memories away from each other, thereby amplifying the eventual repulsion effect. Of course, it will be hard to come to a decisive conclusion regarding these two hypotheses with a single study, but we believe the current study will shed light in future research towards how aging influence hippocampal mechanisms and its relationship to the rest of the brain.

3.1 Methods

Participants

Fifteen younger participants (11 female, mean age = 23.33, range = 19 – 31 years) and one older participant (1 female, age = 71) finished the experiment procedure. All participants' enrollment followed procedures approved by the University of Oregon Institutional Review Board. Written informed consent was collected for each participant prior to the experiment. All participants were right-handed native-English speakers with normal or corrected-to-normal vision, with no self-reported psychiatric or neurological disease. One younger participant's data was excluded due to mistakes in MRI prescription during data collection. The final analysis was comprised of 15 participants,

including one older adult. All participants received monetary compensation for participating.

Stimuli

24 images of gazebos, 24 images of beaches, and 48 images of everyday objects were used in the experiment (Fig. 1A). An additional 24 images of gazebos and 24 images of beaches were used as lures for the exposure phase of the study. Separately for each participant, the set of 48 scenes and the set of 48 objects were randomly assigned with each other to form 48 associative pairs. For each participant, one category of the scenes (e.g., beaches) was assigned as high-training category, and the other category of the scenes (e.g., gazebo) was assigned as the low-training category. The high- and low-training category assignment alternate between participants.

Experimental procedure

After providing consent and reviewing the instructions, participants finished the Study and Test Phase in the preparation room prior to entering the MRI scanner. Inside the scanner, participants completed eight rounds of the Exposure phase. Participants then left the scanner and finished an additional memory task in the preparation room. Across all phases, stimuli were displayed on a gray background, either directly on a laptop screen or projected from the back of the scanner. The experiment was implemented in PsychoPy2022.1.1 and lasted for about 2 h, with about 1 h 15 min inside the scanner.

Learning There were 3 rounds of the Study and Test phase. Within each round, participants performed one study section followed by a test section. During the

study, a scene and its associated object was displayed on the screen at the same time, with a white fixation cross in between. Then, two images disappeared with only a white fixation cross on the screen (1000ms), followed by the next trial. The display time for the scene and object decreased across rounds and were lower for the low-training group than the high-training group. For the high-training group, the display time were 3000ms (round 1), 2500ms (round 2), and 2000ms (round 3); for the low-training group, the display time were 2500ms (round 1), 2000ms (round 2), and 2000ms (round 3).

For each test round, participants viewed the scene at the top half of the screen, with 3 objects at the bottom of the screen: the correct object (e.g., water bucket), an random object that was associated with the scene from the same category ('within-category object'; e.g., basket), and an random object that was associated with the other category ('across-category object'; e.g., bowling ball). Participants has a max of 10 seconds to make their choice. immediately after participants made their choice - or after 10 seconds had passed without any choice being recorded - high-training group received feedback by removing the two incorrect objects and displaying only the scene and the correct object (1000ms). There was no feedback for low-training group. The next trial started after a 1000ms white fixation cross.

Exposure The exposure phase was conducted during fMRI scanning and consisted of eight rounds. Participants were shown 54 scenes in a random order for each round. The 54 scenes consisted of 48 scenes from the associative pairs in the previous Learning phase (24 gazebos, 24 beaches), as well as 6 new lure images (3 gazebos, 3 beaches). The lure images never repeat across the whole

exposure phase. Participants made an old/new judgment for each scene. After 6000ms of lead time, each trial started with the scene being displayed for 2000ms, followed by a white fixation cross of 4000ms. Participants were instructed to respond during the latter 6000ms period. Then the next trial started with the next image.

Post-test The post-test was conducted in the preparation room after participants finished the fMRI scanning. Each of the 48 scenes from the associative pair appear once on the screen for each trial. Within a trial, the scene was at the top half of the screen, with all 24 objects that was associated with the same category at the bottom. Participants were instructed to use their mouse to choose the correct object that was associated with the scene. The maximum permitted response time was 99 seconds. The next trial started after a 500ms white fixation.

MRI acquisition

All images were acquired on a Siemens 3T Skyra MRI system in the Lewis Center for Neuroimaging at the University of Oregon. Functional data were acquired with a T2*-weighted echo-planar imaging sequence with partial brain coverage that prioritized full coverage of the hippocampus and EVC (repetition time = 2000 ms, echo time = 34 ms, flip angle = 90°, 66 slices, 1.7 × 1.7 × 1.7 mm voxels). A total of 8 functional scans were acquired. Each functional scan comprised 168 volumes and included 6 s of lead-in time and 6 s of lead-out time at the beginning and end of each scan, respectively. Anatomical scans included a whole-brain high-resolution T1-weighted magnetization prepared rapid acquisition gradient-echo anatomical volume (1

× 1 × 1 mm voxels) and a high-resolution (coronal direction) T2-weighted scan (0.43 × 0.43 × 1.8 mm voxels) to facilitate segmentation of hippocampal subfields.

Anatomical data preprocessing

Preprocessing was performed *using fMRIPrep* 23.2.0 (RRID:SCR_016216), which is based on *Nipype* 1.8.6 (RRID:SCR_002502). The T1-weighted (T1w) image was corrected for intensity nonuniformity (INU) with N4BiasFieldCorrection54 (ANTs 2.5.0, RRID: SCR_004757), and used as the T1w reference throughout the workflow. The T1w reference was skull-stripped with the antsBrainExtraction.sh workflow (ANTs) in Nipype, using OASIS30ANTs as the target template. Brain tissue segmentation of cerebrospinal fluid (CSF), white-matter (WM), and gray-matter (GM) was performed on the brain-extracted T1w using fast (FSL; RRID:SCR_002823). T2-weighted image was used to improve pial surface refinement. Brain surfaces were reconstructed using recon-all (FreeSurfer 7.3.2, RRID:SCR_001847). Volume-based spatial normalization to one standard space (MNI152NLin2009cAsym) was performed through nonlinear registration with antsRegistration (ANTs 2.5.0), using brain-extracted versions of both T1w reference and the T1w template. *ICBM 152 Nonlinear Asymmetrical template version 2009c* was selected for spatial normalization (RRID:SCR_008796; TemplateFlow ID: MNI152NLin2009cAsym).

Functional data preprocessing

fMRIPrep

For each of the 8 BOLD scans per participant, the following preprocessing was performed. First, a reference volume was generated, for use in head motion correction. Head-motion parameters with respect to the BOLD reference (transformation matrices,

and six corresponding rotation and translation parameters) are estimated. Fieldmap was collected and estimated based on two (or more) echo-planar imaging (EPI) references with topup (FSL 6.0.5.1:57b01774). The estimated *fieldmap* was then aligned with rigid-registration to the target EPI (echo-planar imaging) reference run. The field coefficients were mapped on to the reference EPI using the transform. The BOLD reference was then co-registered to the T1w reference using bbregister (FreeSurfer) which implements boundary-based registration. Co-registration was configured with six degrees of freedom. Several confounding time-series were calculated based on the preprocessed BOLD: framewise displacement (FD), DVARS and three region-wise global signals. Additionally, a set of physiological regressors were extracted to allow for component-based noise correction (CompCor, Behzadi et al. 2007). Principal components are estimated after high-pass filtering the preprocessed BOLD time-series (using a discrete cosine filter with 128s cut-off) for the two CompCor variants: temporal (tCompCor) and anatomical (aCompCor). Frames that exceeded a threshold of 0.5 mm FD or 1.5 standardized DVARS were annotated as motion outliers. The BOLD time-series were resampled onto fsaverage6. Gridded (volumetric) resamplings were performed using nitransforms, configured with cubic B-spline interpolation. Non-gridded (surface) resamplings were performed using mri_vol2surf (FreeSurfer).

fMRI first-level general linear model (GLM) analyses

Eight brain masks were generated by fMRIPrep for each of the eight functional scans. The intersection of all eight masks was used to perform brain extraction. The *processed BOLD* was then scaled at a mean equal 100, with an upper bound of 200 and a lower bound of 0. For the eight scans, all first-level GLMs were performed in

participants' native space with FSL using a Double-Gamma HRF with temporal derivatives, implemented with Python3.10. GLMs were calculated with AFNI's 3dREMLfit using the Least Squares-Separate method: a separate GLM was calculated for each of the 48 images (24 beaches and 24 gazebos) for each scan. For each GLM, there was one regressor of interest (representing a single scene or object image across its two repetitions per scan). All other trials (including lure images), FD, xyz translation, xyz rotation, aCompCor 00-05, and csf were represented with nuisance regressors. This model resulted in 48 beta-maps per scan (one map per image) which were converted to t-stats maps that represented the pattern of activity elicited by each scene for each scan.

Regions of interest

A region of interest (ROI) for EVC was created from the probabilistic maps of Visual Topography63 in the MNI space with a 0.5 threshold. This ROI was transformed into each participant's native space using inverse T1w-to-MNI nonlinear transformation. For each participant. An ROI for the PPA was created by first using an automated metanalyses in Neurosynth with the key term "place". Then, clusters were created using voxels with a z-score >2 based on the Neurosynth associative tests. Since these clusters were generated through an automated meta-analysis and were not anatomically exclusive to PPA, we visually inspected the results and manually selected the two largest clusters that were spatially consistent with PPA. One cluster was in the right hemisphere (voxel size = 247) and one cluster was in the left hemisphere (voxel size = 163). These clusters were combined into a single PPA mask. This mask was then transformed into each participant's native space using the inverse T1w-to-MNI

transformation. For each participant, a final PPA ROI was generated by averaging the t-maps of all scene exposure phase scans. To create hippocampal ROIs, we used the Automatic Segmentation of Hippocampal Subfields (ASHS) 64 toolbox with the upenn2020 (10.1016/j.media.2022.102683) atlas to generate subfield ROIs in each participant's hippocampal body, including CA3/dentate gyrus (which included CA2, CA3, and dentate gyrus) and CA1. The most anterior and posterior slices of the hippocampal body were manually determined for each participant based on the T2-weighted anatomical structure (see Supplementary fig.7 from the first chapter for a sample demarcation). Each participant's subfield segmentations were also manually inspected to ensure the accuracy of the segmentation protocol. Then, each subfield ROI was transformed into each participant's native space using the T2-to-T1w transformation, calculated with FLIRT (fsl) with six degrees of freedom, implemented with Nipype. All ROIs were again visually inspected following the transformation to native space to ensure the ROIs were anatomically correct.

fMRI pattern similarity analyses

Similarity Scores

Pattern similarity was calculated as the Pearson correlation between t-maps within each ROI. All pattern similarity analyses were performed by correlating the t-maps for stimuli across scans (i.e., correlations were never performed within the same scan). For our primary analyses related to pattern similarity between scene images, of critical interest was the similarity *within* category scenes (within-category similarity) relative to the similarity between *across* category scenes (across-category similarity). Specifically, we calculated the similarities between every possible pair of scenes within

a category, then use the mean to subtract from all possible pairs of scenes across a category, resulting in a similarity score. When the similarity score was positive, it indicated that within-category scenes were represented more similarly to each other than across-category scenes.

To define early and late phase of the experiment, we separated the eight rounds of Exposure into first 4 rounds (early) and the last 4 rounds (late). For the early phase, only correlations using trials that were both from the first 4 rounds were included. For the late phase, only correlations using trials that were both from the last 4 rounds were included.

GENERAL DISCUSSION

Daily events often share overlapping elements, making it sometimes hard to distinguish one from another. However, memory confusions are also solvable with learning over time. The neural mechanisms underlying the resolution of memory interference remains a question to be investigated. Previous literature established that the hippocampus plays an important part in forming new memories and resolving memory interference[1–3]. In the current dissertation, I characterized the role of the hippocampus, with an emphasis on its 'repulsion' of representations between similar events, during the process of memory interference resolution.

Abrupt Repulsion when resolving memory interference.

Previous literatures[4–6] have argued that competing memories can lead to memory interference. In particular, repulsion only occurs when two memories share overlapping elements and does not occur for two memories that were distinct in the first place. Computational models[7,8] and empirical studies[9] also suggested that moderate levels of similarity can lead to memory interference. However, the relationship between hippocampal repulsion and behavioral performance was never established directly.

One of the major contributions of the current dissertation is to directly link repulsion with the behavioral expression of resolving memory interference. In both Chapter 1 and Chapter 2, results showed that at the moment when participants indicated they resolved memory interference between two similar events, hippocampus representational patterns of those events were also the most repulsed away from each other. These temporal connections between repulsion and resolving memory interference strongly suggested the crucial role of repulsion in facilitating the

distinguishing of similar memories. Further, these results also suggested that while participants are actively trying to distinguish similar memories, the hippocampus is also actively trying to push similar memories away from each other.

Moreover, results in all three chapters also established the transient nature of the repulsion effect. In the first two experiments, after interference was resolved, the repulsion ceased to be. In the third experiment, results also showed a transient repulsion effect for the high training group. These results indicated that repulsion cannot be understood as a linear accumulation of experiences to form new stable representations. Instead, repulsion is more likely to be an effortful attempt to resolve the interferences between similar memories and subsides after this mission is accomplished.

One followup question is how the hippocampus accomplishes repulsion at the neural level. Previous computational models[7,8] modeled repulsion as a 'stable state' that the hippocampus patterns changed into, support by the Nonmonotonic Plasticity Hypothesis. However, as stated earlier, evidence in the current dissertation suggested that the repulsion effect is a transient state, which cannot be supported by long term synaptic changes. Instead, the hippocampus is more likely to be temporarily suppressing certain neurons from firing, which therefore increased the differences between similar memories. New computational models are still needed to understand the underlying neural mechanisms for the repulsion effect.

Another natural follow-up question is that, if after interference is resolved the hippocampus does not repulse anymore, are there any other brain regions that take the

responsibility? Future studies targeted at cortex representations are needed to answer this interesting question.

Repulsion with distinct internal states.

In rodent remapping literature, one recent hypothesis is that remapping can represent not only spatial locations[10–12], but also their internal models of the external environments[13–15].

Borrowing ideas from rodent remapping literature, part of this dissertation focused on how internal representations rather than external stimuli change hippocampal repulsion. Chapter 1 showed hippocampal repulsion can change with participants' different behavioral state, while the external visual stimuli stayed the same. Chapter 2 directly tested the hypothesis that hippocampus representation can change dramatically with changes in internal state, independent to changes in the external environment. Specifically, in Chapter 2, our results showed that repulsion can happen even if the external stimuli were identical, given that participants believed they were distinct. However, if participants held the same belief, the repulsion effect became absent.

These results are highly consistent with the rodent remapping literature, arguing that the change of hippocampus representations reflects participants' internal states. However, what is novel in the current dissertation, comparing to existing rodent research, is the following 2 points: 1) the second chapter, using the manipulation of cues, directly showed the difference in hippocampal representations with the same or distinct internal representations. In our knowledge, this direct contrast is yet to be done in rodent research; 2) the report of repulsion effect is also yet to be reported in rodent

empirical studies. Even though some computational models[16] have argued the possibility of the repulsion effect, no rodent studies to our knowledge have showed further distance between similar memories than distinct memories.

Overall, the current dissertation sheds new lights on the mechanisms underlying how and why hippocampus representational patterns change as a function of participants' internal states.

Repulsion happened within CA3/DG subfields

Different from the previous fMRI research that analyzed the hippocampus as a whole[4,6,17,18], this dissertation utilized high-resolution anatomical T2 scans and was therefore able to focus on distinct subfields within the hippocampus.

Across the three studies in the current dissertation, we consistently found repulsion in CA3/DG but not CA1. This is consistent with the literature which indicates that CA1 performs more memory integration and completion, whereas CA3/DG is more active in separating overlapping memories[2].

However, even though the coarse coding nature of the DG makes it able to represent overlapping memories as orthogonal to each other, but also makes it harder to explain the repulsion phenomenon. The orthogonalization of similar memories should end with the same distance similar memories and distinct memories, whereas the results from the current dissertation showed that similar memories were represented *further away* from each other than distinct memories. Therefore, the sparse coding nature in the DG is likely not the ideal neural mechanisms underlying hippocampal repulsion.

On the other hand, the attractor dynamics in CA3[19,20] could potentially explain the abrupt occurrence of the repulsion better than the sparse coding in DG. In particular, the non-linear nature of the attractor dynamics[21,20,22,23] corresponds well with the non-linear nature of the repulsion effect.

Conclusions

This dissertation investigated the hippocampal mechanisms underlying the resolution of memory interference, with a focus on the repulsion effect. The current dissertation provided consistent evidence, across 3 studies, regarding the mechanisms underlying hippocampal repulsion effect. The results showed the direct temporal relationship between hippocampal repulsion and behavioral interference resolution, the effect of distinct internal beliefs in triggering repulsion between identical visual inputs, as well as measuring the level of repulsion as a function of experience. However, more questions regarding repulsion and memory interference remain to be answered, such as whether other brain region perform the interference resolution after repulsion subsides, how the intensity level of repulsion relates to memory confusability, and whether/how repulsion changes with aging. Answers for these questions will be important to further our understanding in human cognition.

REFERENCES

Introduction

1. McClelland, J. L., McNaughton, B. L. & O'Reilly, R. C. Why there are complementary learning systems in the hippocampus and neocortex: Insights from the successes and failures of connectionist models of learning and memory. *Psychological Review* **102**, 419 (19951101).
2. Yassa, M. A. & Stark, C. E. L. Pattern separation in the hippocampus. *Trends Neurosci* **34**, 515–525 (2011).
3. Ritvo, V. J. H., Turk-Browne, N. B. & Norman, K. A. Nonmonotonic Plasticity: How Memory Retrieval Drives Learning. *Trends in Cognitive Sciences* **23**, 726–742 (2019).
4. Norman, K. A., Newman, E. L. & Detre, G. A neural network model of retrieval-induced forgetting. *Psychological Review* **114**, 887–953 (2007).
5. O'Reilly, R. C. & Norman, K. A. Hippocampal and neocortical contributions to memory: advances in the complementary learning systems framework. *Trends in Cognitive Sciences* **6**, 505–510 (2002).
6. Colgin, L. L., Moser, E. I. & Moser, M.-B. Understanding memory through hippocampal remapping. *Trends in Neurosciences* **31**, 469–477 (2008).
7. Kyle, C. T., Stokes, J. D., Lieberman, J. S., Hassan, A. S. & Ekstrom, A. D. Successful retrieval of competing spatial environments in humans involves hippocampal pattern separation mechanisms. *eLife* **4**, e10499 (2015).
8. Steemers, B. *et al.* Hippocampal Attractor Dynamics Predict Memory-Based Decision Making. *Current Biology* **26**, 1750–1757 (2016).
9. Julian, J. B. & Doeller, C. F. Remapping and realignment in the human hippocampal formation predict context-dependent spatial behavior. *Nature Neuroscience* 1–10 (2021) doi:10.1038/s41593-021-00835-3.
10. Bostock, E., Muller, R. U. & Kubie, J. L. Experience-dependent modifications of hippocampal place cell firing. *Hippocampus* **1**, 193–205 (1991).
11. Fyhn, M., Hafting, T., Treves, A., Moser, M.-B. & Moser, E. I. Hippocampal remapping and grid realignment in entorhinal cortex. *Nature* **446**, 190–194 (2007).
12. Latuske, P., Kornienko, O., Kohler, L. & Allen, K. Hippocampal Remapping and Its Entorhinal Origin. *Front. Behav. Neurosci.* **11**, (2018).
13. Lever, C., Wills, T., Cacucci, F., Burgess, N. & O'Keefe, J. Long-term plasticity in hippocampal place-cell representation of environmental geometry. *Nature* **416**, 90–94 (2002).
14. Duncan, K. D. & Schlichting, M. L. Hippocampal representations as a function of time, subregion, and brain state. *Neurobiology of Learning and Memory* **153**, 40–56 (2018).
15. Chanales, A. J. H., Oza, A., Favila, S. E. & Kuhl, B. A. Overlap among Spatial Memories Triggers Repulsion of Hippocampal Representations. *Current Biology* **27**, 2307-2317.e5 (2017).
16. Favila, S. E., Chanales, A. J. H. & Kuhl, B. A. Experience-dependent hippocampal pattern differentiation prevents interference during subsequent learning. *Nat Commun* **7**, 11066 (2016).

17. Hulbert, J. C. & Norman, K. A. Neural Differentiation Tracks Improved Recall of Competing Memories Following Interleaved Study and Retrieval Practice. *Cereb. Cortex* **25**, 3994–4008 (2015).
18. Sanders, H., Wilson, M. A. & Gershman, S. J. Hippocampal remapping as hidden state inference. *eLife* **9**, e51140 (2020).
19. Dimsdale-Zucker, H. R., Ritchey, M., Ekstrom, A. D., Yonelinas, A. P. & Ranganath, C. CA1 and CA3 differentially support spontaneous retrieval of episodic contexts within human hippocampal subfields. *Nat Commun* **9**, 294 (2018).
20. Keinath, A. T., Nieto-Posadas, A., Robinson, J. C. & Brandon, M. P. DG–CA3 circuitry mediates hippocampal representations of latent information. *Nat Commun* **11**, 3026 (2020).
21. Lai, C., Tanaka, S., Harris, T. D. & Lee, A. K. Volitional activation of remote place representations with a hippocampal brain-machine interface. *Science* **382**, 566–573 (2023).

Chapter 1

1. Eichenbaum, H. A cortical–hippocampal system for declarative memory. *Nat. Rev. Neurosci.* **1**, 41–50 (2000).
2. Squire, L. & Zola-Morgan, S. The medial temporal lobe memory system. *Science* **253**, 1380–1386 (1991).
3. O'Keefe, J. & Nadel, L. *The hippocampus as a cognitive map*. (Clarendon Press ; Oxford University Press, 1978).
4. Schapiro, A. C., Turk-Browne, N. B., Botvinick, M. M. & Norman, K. A. Complementary learning systems within the hippocampus: a neural network modelling approach to reconciling episodic memory with statistical learning. *Philos. Trans. R. Soc. B Biol. Sci.* **372**, 20160049 (2017).
5. O'Reilly, R. C. & Norman, K. A. Hippocampal and neocortical contributions to memory: advances in the complementary learning systems framework. *Trends Cogn. Sci.* **6**, 505–510 (2002).
6. Bostock, E., Muller, R. U. & Kubie, J. L. Experience-dependent modifications of hippocampal place cell firing. *Hippocampus* **1**, 193–205 (1991).
7. Muller, R. U. & Kubie, J. L. The effects of changes in the environment on the spatial firing of hippocampal complex-spike cells. *J. Neurosci. Off. J. Soc. Neurosci.* **7**, 1951–1968 (1987).
8. Colgin, L. L., Moser, E. I. & Moser, M.-B. Understanding memory through hippocampal remapping. *Trends Neurosci.* **31**, 469–477 (2008).
9. Colgin, L. L. *et al.* Attractor-Map Versus Autoassociation Based Attractor Dynamics in the Hippocampal Network. *J. Neurophysiol.* **104**, 35–50 (2010).
10. Leutgeb, S., Leutgeb, J. K., Moser, E. I. & Moser, M.-B. Fast rate coding in hippocampal CA3 cell ensembles. *Hippocampus* **16**, 765–774 (2006).
11. Wills, T. J. Attractor Dynamics in the Hippocampal Representation of the Local Environment. *Science* **308**, 873–876 (2005).
12. Lee, I., Rao, G. & Knierim, J. J. A Double Dissociation between Hippocampal Subfields: Differential Time Course of CA3 and CA1 Place Cells for Processing Changed Environments. *Neuron* **42**, 803–815 (2004).

13. Lever, C., Wills, T., Cacucci, F., Burgess, N. & O'Keefe, J. Long-term plasticity in hippocampal place-cell representation of environmental geometry. *Nature* **416**, 90–94 (2002).
14. Plitt, M. H. & Giocomo, L. M. Experience-dependent contextual codes in the hippocampus. *Nat. Neurosci.* 1–10 (2021) doi:10.1038/s41593-021-00816-6.
15. Sanders, H., Wilson, M. A. & Gershman, S. J. Hippocampal remapping as hidden state inference. *eLife* **9**, e51140 (2020).
16. Keinath, A. T., Nieto-Posadas, A., Robinson, J. C. & Brandon, M. P. DG–CA3 circuitry mediates hippocampal representations of latent information. *Nat. Commun.* **11**, 3026 (2020).
17. Molitor, R. J., Sherrill, K. R., Morton, N. W., Miller, A. A. & Preston, A. R. Memory reactivation during learning simultaneously promotes dentate gyrus/CA2,3 pattern differentiation and CA1 memory integration. *J. Neurosci.* (2020) doi:10.1523/JNEUROSCI.0394-20.2020.
18. Chanales, A. J. H., Oza, A., Favila, S. E. & Kuhl, B. A. Overlap among Spatial Memories Triggers Repulsion of Hippocampal Representations. *Curr. Biol.* **27**, 2307-2317.e5 (2017).
19. Kim, G., Norman, K. A. & Turk-Browne, N. B. Neural Differentiation of Incorrectly Predicted Memories. *J. Neurosci.* **37**, 2022–2031 (2017).
20. Favila, S. E., Chanales, A. J. H. & Kuhl, B. A. Experience-dependent hippocampal pattern differentiation prevents interference during subsequent learning. *Nat. Commun.* **7**, 11066 (2016).
21. Schlichting, M. L., Mumford, J. A. & Preston, A. R. Learning-related representational changes reveal dissociable integration and separation signatures in the hippocampus and prefrontal cortex. *Nat. Commun.* **6**, 8151 (2015).
22. Schapiro, A. C., Kustner, L. V. & Turk-Browne, N. B. Shaping of Object Representations in the Human Medial Temporal Lobe Based on Temporal Regularities. *Curr. Biol.* **22**, 1622–1627 (2012).
23. Kyle, C. T., Stokes, J. D., Lieberman, J. S., Hassan, A. S. & Ekstrom, A. D. Successful retrieval of competing spatial environments in humans involves hippocampal pattern separation mechanisms. *eLife* **4**, e10499 (2015).
24. Copara, M. S. *et al.* Complementary Roles of Human Hippocampal Subregions during Retrieval of Spatiotemporal Context. *J. Neurosci.* **34**, 6834–6842 (2014).
25. Hulbert, J. C. & Norman, K. A. Neural Differentiation Tracks Improved Recall of Competing Memories Following Interleaved Study and Retrieval Practice. *Cereb. Cortex* **25**, 3994–4008 (2015).
26. Duncan, K. D. & Schlichting, M. L. Hippocampal representations as a function of time, subregion, and brain state. *Neurobiol. Learn. Mem.* **153**, 40–56 (2018).
27. Guzowski, J. F., Knierim, J. J. & Moser, E. I. Ensemble Dynamics of Hippocampal Regions CA3 and CA1. *Neuron* **44**, 581–584 (2004).
28. Yassa, M. A. & Stark, C. E. L. Pattern separation in the hippocampus. *Trends Neurosci.* **34**, 515–525 (2011).
29. McHugh, T. J. *et al.* Dentate Gyrus NMDA Receptors Mediate Rapid Pattern Separation in the Hippocampal Network. *Science* **317**, 94–99 (2007).

30. Leutgeb, S., Leutgeb, J. K., Treves, A., Moser, M.-B. & Moser, E. I. Distinct Ensemble Codes in Hippocampal Areas CA3 and CA1. *Science* **305**, 1295–1298 (2004).
31. Vazdarjanova, A. & Guzowski, J. F. Differences in Hippocampal Neuronal Population Responses to Modifications of an Environmental Context: Evidence for Distinct, Yet Complementary, Functions of CA3 and CA1 Ensembles. *J. Neurosci.* **24**, 6489–6496 (2004).
32. van Dijk, M. T. & Fenton, A. A. On How the Dentate Gyrus Contributes to Memory Discrimination. *Neuron* **98**, 832-845.e5 (2018).
33. Dimsdale-Zucker, H. R., Ritchey, M., Ekstrom, A. D., Yonelinas, A. P. & Ranganath, C. CA1 and CA3 differentially support spontaneous retrieval of episodic contexts within human hippocampal subfields. *Nat. Commun.* **9**, 294 (2018).
34. Bakker, A., Kirwan, C. B., Miller, M. & Stark, C. E. L. Pattern separation in the human hippocampal CA3 and dentate gyrus. *Science* **319**, 1640–1642 (2008).
35. Yassa, M. A. *et al.* Pattern separation deficits associated with increased hippocampal CA3 and dentate gyrus activity in nondemented older adults. *Hippocampus* **21**, 968–979 (2011).
36. Kriegeskorte, N. Representational similarity analysis – connecting the branches of systems neuroscience. *Front. Syst. Neurosci.* (2008) doi:10.3389/neuro.06.004.2008.
37. Mumford, J. A., Davis, T. & Poldrack, R. A. The impact of study design on pattern estimation for single-trial multivariate pattern analysis. *NeuroImage* **103**, 130–138 (2014).
38. Leutgeb, J. K., Leutgeb, S., Moser, M.-B. & Moser, E. I. Pattern Separation in the Dentate Gyrus and CA3 of the Hippocampus. *Science* **315**, 961–966 (2007).
39. Steemers, B. *et al.* Hippocampal Attractor Dynamics Predict Memory-Based Decision Making. *Curr. Biol.* **26**, 1750–1757 (2016).
40. Julian, J. B. & Doeller, C. F. Remapping and realignment in the human hippocampal formation predict context-dependent spatial behavior. *Nat. Neurosci.* 1–10 (2021) doi:10.1038/s41593-021-00835-3.
41. Hindy, N. C., Ng, F. Y. & Turk-Browne, N. B. Linking pattern completion in the hippocampus to predictive coding in visual cortex. *Nat. Neurosci.* **19**, 665–667 (2016).
42. Jiang, J., Wang, S.-F., Guo, W., Fernandez, C. & Wagner, A. D. Prefrontal reinstatement of contextual task demand is predicted by separable hippocampal patterns. *Nat. Commun.* **11**, 2053 (2020).
43. Ritvo, V. J. H., Turk-Browne, N. B. & Norman, K. A. Nonmonotonic Plasticity: How Memory Retrieval Drives Learning. *Trends Cogn. Sci.* **23**, 726–742 (2019).
44. Norman, K. A., Newman, E. L. & Detre, G. A neural network model of retrieval-induced forgetting. *Psychol. Rev.* **114**, 887–953 (2007).
45. Rouhani, N. & Niv, Y. Signed and unsigned reward prediction errors dynamically enhance learning and memory. *eLife* **10**, e61077 (2021).
46. Kim, G., Lewis-Peacock, J. A., Norman, K. A. & Turk-Browne, N. B. Pruning of memories by context-based prediction error. *Proc. Natl. Acad. Sci.* **111**, 8997–9002 (2014).

47. DuBrow, S., Rouhani, N., Niv, Y. & Norman, K. A. Does mental context drift or shift? *Curr. Opin. Behav. Sci.* **17**, 141–146 (2017).
48. Rebola, N., Carta, M. & Mulle, C. Operation and plasticity of hippocampal CA3 circuits: implications for memory encoding. *Nat. Rev. Neurosci.* **18**, 208–220 (2017).
49. Peirce, J. *et al.* PsychoPy2: Experiments in behavior made easy. *Behav. Res. Methods* **51**, 195–203 (2019).
50. Esteban, O. *et al.* fMRIPrep: a robust preprocessing pipeline for functional MRI. *Nat. Methods* **16**, 111–116 (2019).
51. Esteban, Oscar, Ross Blair, Christopher J. Markiewicz, Shoshana L. Berleant, Craig Moodie, Feilong Ma, Ayse Ilkay Isik, et al. 2018. “FMRIPrep.” *Software*. Zenodo. https://doi.org/10.5281/zenodo.852659.
52. Gorgolewski, K. *et al.* Nipype: A Flexible, Lightweight and Extensible Neuroimaging Data Processing Framework in Python. *Front. Neuroinformatics* **5**, (2011).
53. Gorgolewski, Krzysztof J., Oscar Esteban, Christopher J. Markiewicz, Erik Ziegler, David Gage Ellis, Michael Philipp Notter, Dorota Jarecka, et al. 2018. “Nipype.” *Software*. Zenodo. https://doi.org/10.5281/zenodo.596855.
54. Tustison, N. J. *et al.* N4ITK: Improved N3 Bias Correction. *IEEE Trans. Med. Imaging* **29**, 1310–1320 (2010).
55. Avants, B. B., Epstein, C. L., Grossman, M. & Gee, J. C. Symmetric diffeomorphic image registration with cross-correlation: Evaluating automated labeling of elderly and neurodegenerative brain. *Med. Image Anal.* **12**, 26–41 (2008).
56. Zhang, Y., Brady, M. & Smith, S. Segmentation of brain MR images through a hidden Markov random field model and the expectation-maximization algorithm. *IEEE Trans. Med. Imaging* **20**, 45–57 (2001).
57. Fonov, V., Evans, A., McKinstry, R., Almli, C. & Collins, D. Unbiased nonlinear average age-appropriate brain templates from birth to adulthood. *NeuroImage* **47**, S102 (2009).
58. Cox, R. W. & Hyde, J. S. Software tools for analysis and visualization of fMRI data. *NMR Biomed.* **10**, 171–178 (1997).
59. Greve, D. N. & Fischl, B. Accurate and robust brain image alignment using boundary-based registration. *NeuroImage* **48**, 63–72 (2009).
60. Jenkinson, M., Bannister, P., Brady, M. & Smith, S. Improved Optimization for the Robust and Accurate Linear Registration and Motion Correction of Brain Images. *NeuroImage* **17**, 825–841 (2002).
61. Power, J. D. *et al.* Methods to detect, characterize, and remove motion artifact in resting state fMRI. *NeuroImage* **84**, 320–341 (2014).
62. Mumford, J. A., Turner, B. O., Ashby, F. G. & Poldrack, R. A. Deconvolving BOLD activation in event-related designs for multivoxel pattern classification analyses. *NeuroImage* **59**, 2636–2643 (2012).
63. Wang, L., Mruczek, R. E. B., Arcaro, M. J. & Kastner, S. Probabilistic Maps of Visual Topography in Human Cortex. *Cereb. Cortex N. Y. N 1991* **25**, 3911–3931 (2015).
64. Yushkevich, P. A. *et al.* Automated volumetry and regional thickness analysis of hippocampal subfields and medial temporal cortical structures in mild cognitive impairment. *Hum. Brain Mapp.* **36**, 258–287 (2015).

65. Newman, E. L. & Norman, K. A. Moderate Excitation Leads to Weakening of Perceptual Representations. *Cereb. Cortex* **20**, 2760–2770 (2010).
66. Wanjia, G., & Kuhl, B. A. Abrupt hippocampal remapping signals resolution of memory interference. Retrieved from openneuro.org/datasets/ds003707 (2021).
67. Wanjia, G., & Kuhl, B. A. Abrupt hippocampal remapping signals resolution of memory interference. Retrieved from osf.io/vpq2x (2021).

Chapter 2

1. Lai, C., Tanaka, S., Harris, T. D. & Lee, A. K. Volitional activation of remote place representations with a hippocampal brain-machine interface. *Science* **382**, 566–573 (2023).
2. Keinath, A. T., Nieto-Posadas, A., Robinson, J. C. & Brandon, M. P. DG–CA3 circuitry mediates hippocampal representations of latent information. *Nat Commun* **11**, 3026 (2020).
3. Gauthier, J. L. & Tank, D. W. A Dedicated Population for Reward Coding in the Hippocampus. *Neuron* **99**, 179-193.e7 (2018).
4. Sanders, H., Wilson, M. A. & Gershman, S. J. Hippocampal remapping as hidden state inference. *eLife* **9**, e51140 (2020).
5. Julian, J. B. & Doeller, C. F. Remapping and realignment in the human hippocampal formation predict context-dependent spatial behavior. *Nature Neuroscience* 1–10 (2021) doi:10.1038/s41593-021-00835-3.
6. Chanales, A. J. H., Oza, A., Favila, S. E. & Kuhl, B. A. Overlap among Spatial Memories Triggers Repulsion of Hippocampal Representations. *Current Biology* **27**, 2307-2317.e5 (2017).
7. Favila, S. E., Chanales, A. J. H. & Kuhl, B. A. Experience-dependent hippocampal pattern differentiation prevents interference during subsequent learning. *Nat Commun* **7**, 11066 (2016).
8. Wood, E. R., Dudchenko, P. A., Robitsek, R. J. & Eichenbaum, H. Hippocampal Neurons Encode Information about Different Types of Memory Episodes Occurring in the Same Location. *Neuron* **27**, 623–633 (2000).
9. Eichenbaum, H. Still searching for the engram. *Learn Behav* **44**, 209–222 (2016).
10. Yassa, M. A. & Stark, C. E. L. Pattern separation in the hippocampus. *Trends Neurosci* **34**, 515–525 (2011).
11. Duncan, K. D. & Schlichting, M. L. Hippocampal representations as a function of time, subregion, and brain state. *Neurobiology of Learning and Memory* **153**, 40–56 (2018).
12. Wills, T. J. Attractor Dynamics in the Hippocampal Representation of the Local Environment. *Science* **308**, 873–876 (2005).
13. Colgin, L. L. *et al.* Attractor-Map Versus Autoassociation Based Attractor Dynamics in the Hippocampal Network. *Journal of Neurophysiology* **104**, 35–50 (2010).
14. Steemers, B. *et al.* Hippocampal Attractor Dynamics Predict Memory-Based Decision Making. *Current Biology* **26**, 1750–1757 (2016).
15. Leutgeb, S., Leutgeb, J. K., Moser, E. I. & Moser, M.-B. Fast rate coding in hippocampal CA3 cell ensembles. *Hippocampus* **16**, 765–774 (2006).

16. Lee, I., Rao, G. & Knierim, J. J. A Double Dissociation between Hippocampal Subfields: Differential Time Course of CA3 and CA1 Place Cells for Processing Changed Environments. *Neuron* **42**, 803–815 (2004).
17. Kyle, C. T., Stokes, J. D., Lieberman, J. S., Hassan, A. S. & Ekstrom, A. D. Successful retrieval of competing spatial environments in humans involves hippocampal pattern separation mechanisms. *eLife* **4**, e10499 (2015).

Chapter 3

1. Chanales, A. J. H., Oza, A., Favila, S. E. & Kuhl, B. A. Overlap among Spatial Memories Triggers Repulsion of Hippocampal Representations. Current Biology 27, 2307-2317.e5 (2017).
2. Favila, S. E., Chanales, A. J. H. & Kuhl, B. A. Experience-dependent hippocampal pattern differentiation prevents interference during subsequent learning. Nat Commun 7, 11066 (2016).
3. Hulbert, J. C. & Norman, K. A. Neural Differentiation Tracks Improved Recall of Competing Memories Following Interleaved Study and Retrieval Practice. Cereb. Cortex 25, 3994–4008 (2015).
4. Yassa, M. A. & Stark, C. E. L. Pattern separation in the hippocampus. Trends Neurosci 34, 515–525 (2011).
5. Duncan, K. D. & Schlichting, M. L. Hippocampal representations as a function of time, subregion, and brain state. Neurobiology of Learning and Memory 153, 40–56 (2018).
6. Chen, J., Olsen, R. K., Preston, A. R., Glover, G. H. & Wagner, A. D. Associative retrieval processes in the human medial temporal lobe: Hippocampal retrieval success and CA1 mismatch detection. Learn Mem 18, 523–528 (2011).
7. Duncan, K., Ketz, N., Inati, S. & Davachi, L. Evidence for area CA1 as a match/mismatch detector: A high-resolution fMRI study of the human hippocampus. Hippocampus 22, 389–398 (2012).
8. Hasselmo, M. E., Wyble, B. P. & Wallenstein, G. V. Encoding and retrieval of episodic memories: Role of cholinergic and GABAergic modulation in the hippocampus. Hippocampus 6, 693–708 (1996).
9. Lisman, J. E. & Grace, A. A. The Hippocampal-VTA Loop: Controlling the Entry of Information into Long-Term Memory. Neuron 46, 703–713 (2005).
10. Deshmukh, S. S. & Knierim, J. J. Representation of Non-Spatial and Spatial Information in the Lateral Entorhinal Cortex. Front. Behav. Neurosci. 5, (2011).
11. Guzowski, J. F., Knierim, J. J. & Moser, E. I. Ensemble Dynamics of Hippocampal Regions CA3 and CA1. Neuron 44, 581–584 (2004).
12. Leutgeb, S., Leutgeb, J. K., Treves, A., Moser, M.-B. & Moser, E. I. Distinct Ensemble Codes in Hippocampal Areas CA3 and CA1. Science 305, 1295–1298 (2004).
13. Vazdarjanova, A. & Guzowski, J. F. Differences in Hippocampal Neuronal Population Responses to Modifications of an Environmental Context: Evidence for Distinct, Yet Complementary, Functions of CA3 and CA1 Ensembles. J Neurosci 24, 6489–6496 (2004).
14. Lacy, J. W., Yassa, M. A., Stark, S. M., Muftuler, L. T. & Stark, C. E. L. Distinct pattern separation related transfer functions in human CA3/dentate and CA1

revealed using high-resolution fMRI and variable mnemonic similarity. Learn. Mem. 18, 15–18 (2011).
15. Ritvo, V. J. H., Turk-Browne, N. B. & Norman, K. A. Nonmonotonic Plasticity: How Memory Retrieval Drives Learning. Trends in Cognitive Sciences 23, 726–742 (2019).
16. Wammes, J., Norman, K. A. & Turk-Browne, N. Increasing stimulus similarity drives nonmonotonic representational change in hippocampus. eLife 11, e68344 (2022).
17. Wanjia, G., Favila, S. E., Kim, G., Molitor, R. J. & Kuhl, B. A. Abrupt hippocampal remapping signals resolution of memory interference. Nat Commun 12, 4816 (2021).

Conclusions

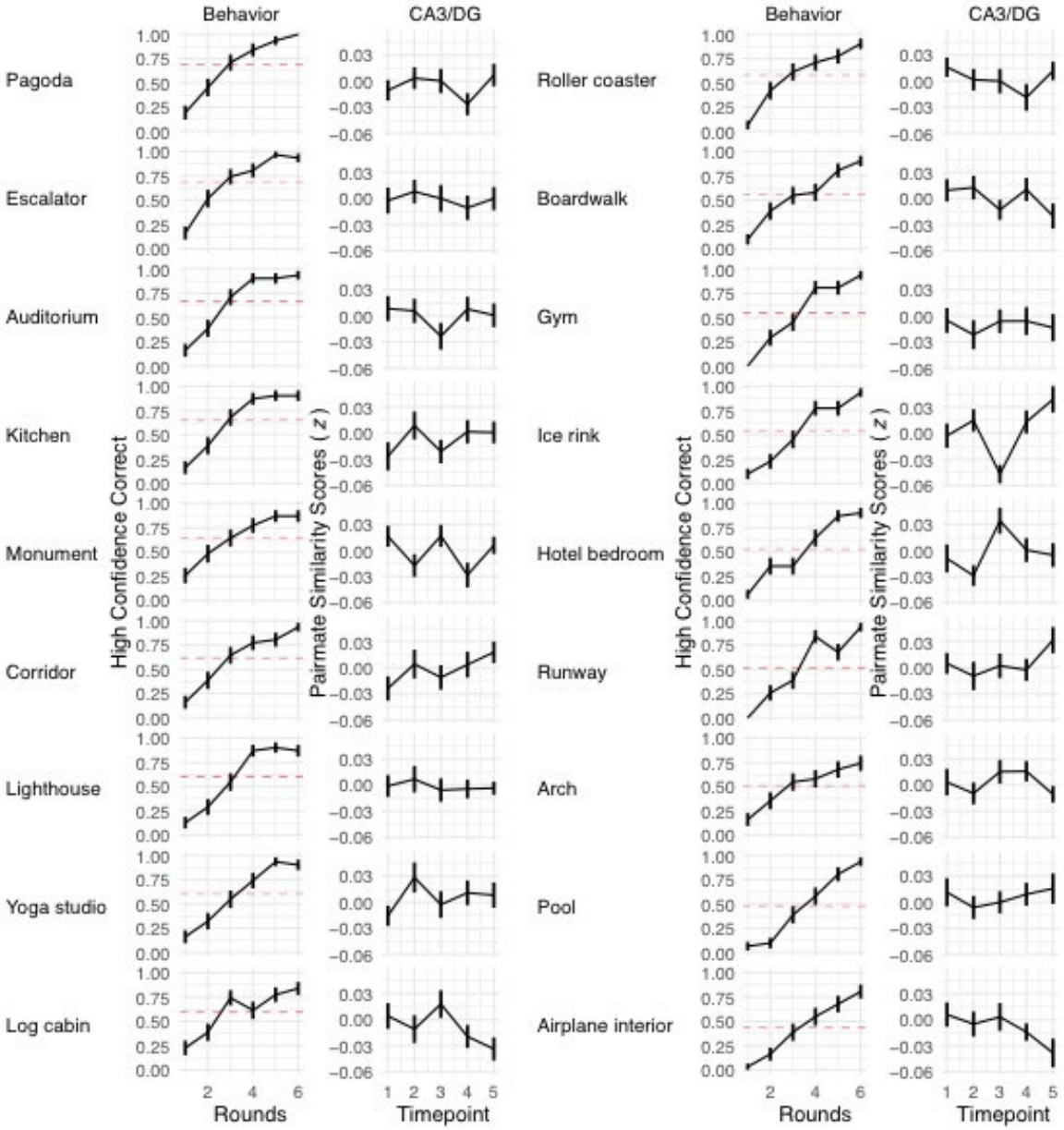
1. Eichenbaum, H., Dudchenko, P., Wood, E., Shapiro, M. & Tanila, H. The Hippocampus, Memory, and Place Cells: Is It Spatial Memory or a Memory Space? Neuron 23, 209–226 (1999).
2. Duncan, K. D. & Schlichting, M. L. Hippocampal representations as a function of time, subregion, and brain state. Neurobiology of Learning and Memory 153, 40–56 (2018).
3. Ekstrom, A. D. & Ranganath, C. Space, time, and episodic memory: The hippocampus is all over the cognitive map. Hippocampus 28, 680–687 (2018).
4. Chanales, A. J. H., Oza, A., Favila, S. E. & Kuhl, B. A. Overlap among Spatial Memories Triggers Repulsion of Hippocampal Representations. Current Biology 27, 2307-2317.e5 (2017).
5. Dimsdale-Zucker, H. R., Ritchey, M., Ekstrom, A. D., Yonelinas, A. P. & Ranganath, C. CA1 and CA3 differentially support spontaneous retrieval of episodic contexts within human hippocampal subfields. Nat Commun 9, 294 (2018).
6. Favila, S. E., Chanales, A. J. H. & Kuhl, B. A. Experience-dependent hippocampal pattern differentiation prevents interference during subsequent learning. Nat Commun 7, 11066 (2016).
7. Norman, K. A., Newman, E. L. & Detre, G. A neural network model of retrieval-induced forgetting. Psychological Review 114, 887–953 (2007).
8. Ritvo, V. J. H., Turk-Browne, N. B. & Norman, K. A. Nonmonotonic Plasticity: How Memory Retrieval Drives Learning. Trends in Cognitive Sciences 23, 726–742 (2019).
9. Wammes, J., Norman, K. A. & Turk-Browne, N. Increasing stimulus similarity drives nonmonotonic representational change in hippocampus. eLife 11, e68344 (2022).
10. Latuske, P., Kornienko, O., Kohler, L. & Allen, K. Hippocampal Remapping and Its Entorhinal Origin. Front. Behav. Neurosci. 11, (2018).
11. Fyhn, M., Hafting, T., Treves, A., Moser, M.-B. & Moser, E. I. Hippocampal remapping and grid realignment in entorhinal cortex. Nature 446, 190–194 (2007).
12. Muller, R. A Quarter of a Century of Place Cells. Neuron 17, 813–822 (1996).
13. Sanders, H., Wilson, M. A. & Gershman, S. J. Hippocampal remapping as hidden state inference. eLife 9, e51140 (2020).
14. Keinath, A. T., Nieto-Posadas, A., Robinson, J. C. & Brandon, M. P. DG–CA3 circuitry mediates hippocampal representations of latent information. Nat Commun 11, 3026 (2020).

15. Gauthier, J. L. & Tank, D. W. A Dedicated Population for Reward Coding in the Hippocampus. Neuron 99, 179-193.e7 (2018).
16. Benna, M. K. & Fusi, S. Place cells may simply be memory cells: Memory compression leads to spatial tuning and history dependence. Proc. Natl. Acad. Sci. U.S.A. 118, e2018422118 (2021).
17. Kim, G., Norman, K. A. & Turk-Browne, N. B. Neural Differentiation of Incorrectly Predicted Memories. J. Neurosci. 37, 2022–2031 (2017).
18. Hulbert, J. C. & Norman, K. A. Neural Differentiation Tracks Improved Recall of Competing Memories Following Interleaved Study and Retrieval Practice. Cereb. Cortex 25, 3994–4008 (2015).
19. Steemers, B. et al. Hippocampal Attractor Dynamics Predict Memory-Based Decision Making. Current Biology 26, 1750–1757 (2016).
20. Wills, T. J. Attractor Dynamics in the Hippocampal Representation of the Local Environment. Science 308, 873–876 (2005).
21. Lacy, J. W., Yassa, M. A., Stark, S. M., Muftuler, L. T. & Stark, C. E. L. Distinct pattern separation related transfer functions in human CA3/dentate and CA1 revealed using high-resolution fMRI and variable mnemonic similarity. Learn. Mem. 18, 15–18 (2011).
22. Leutgeb, S., Leutgeb, J. K., Treves, A., Moser, M.-B. & Moser, E. I. Distinct Ensemble Codes in Hippocampal Areas CA3 and CA1. Science 305, 1295–1298 (2004).
23. Vazdarjanova, A. & Guzowski, J. F. Differences in Hippocampal Neuronal Population Responses to Modifications of an Environmental Context: Evidence for Distinct, Yet Complementary, Functions of CA3 and CA1 Ensembles. J Neurosci 24, 6489–6496 (2004).

APPENDICES

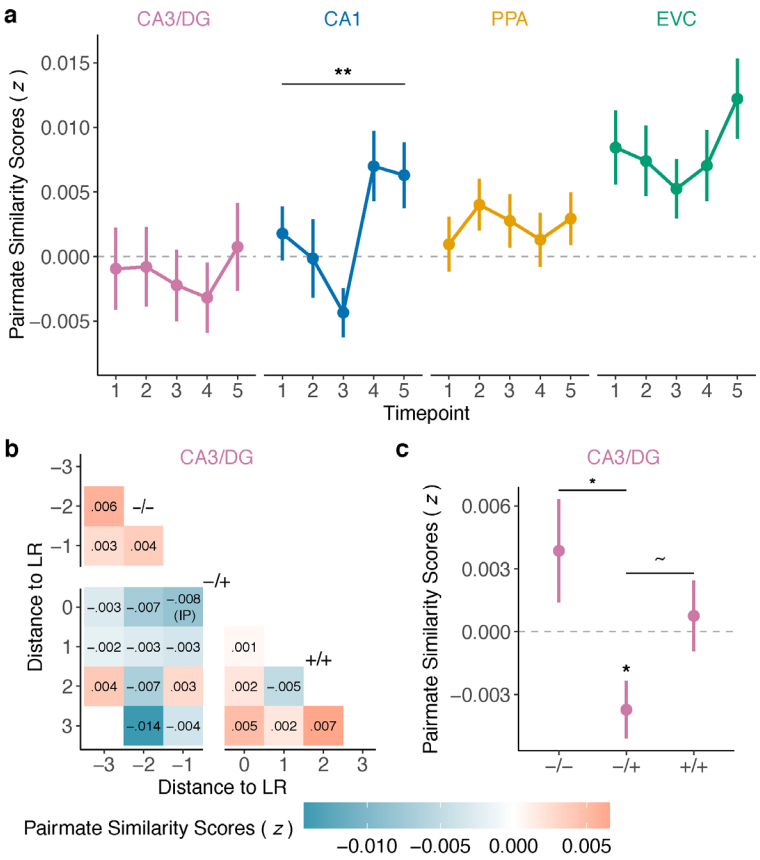
Supplementary Table 1. Number of scene pairmates that transitioned to 'learned' status at each round and for each participant. 'Learned' was defined as high-confidence, correct associative memory for both pairmates. The round at which pairmates transitioned to 'learned' is referred to as the 'learned round' (LR). Note: pairmates that were learned in the first round or never learned were excluded from fMRI analyses.

Participant # \ Round	1	2	3	4	5	6	Never Learned
1	1	7	6	4	0	0	0
2	1	1	4	4	6	2	0
3	1	7	5	5	0	0	0
4	0	3	0	5	4	3	3
5	0	2	3	6	4	2	1
6	3	6	2	6	0	1	0
7	0	6	4	3	3	1	1
8	0	2	5	4	5	1	1
9	0	1	1	2	2	2	10
10	0	0	8	2	5	2	1
11	3	3	4	3	2	2	1
12	0	1	2	5	2	5	3
13	1	1	2	4	7	2	1
14	0	0	3	4	4	5	2
15	1	6	7	2	1	1	0
16	1	2	6	1	2	4	2
17	2	3	3	5	3	2	0
18	5	3	2	3	4	0	1
19	0	0	2	7	6	2	1
20	0	1	6	2	1	4	4
21	0	1	3	3	4	7	0
22	1	3	4	2	3	1	4
23	0	6	5	4	1	2	0
24	3	4	7	1	2	1	0
25	1	10	4	3	0	0	0
26	0	0	2	9	2	1	4
27	3	0	4	2	2	1	6
28	1	8	4	3	0	0	2
29	0	6	2	1	1	2	6
30	2	6	6	1	0	2	1
31	1	1	3	6	3	3	1



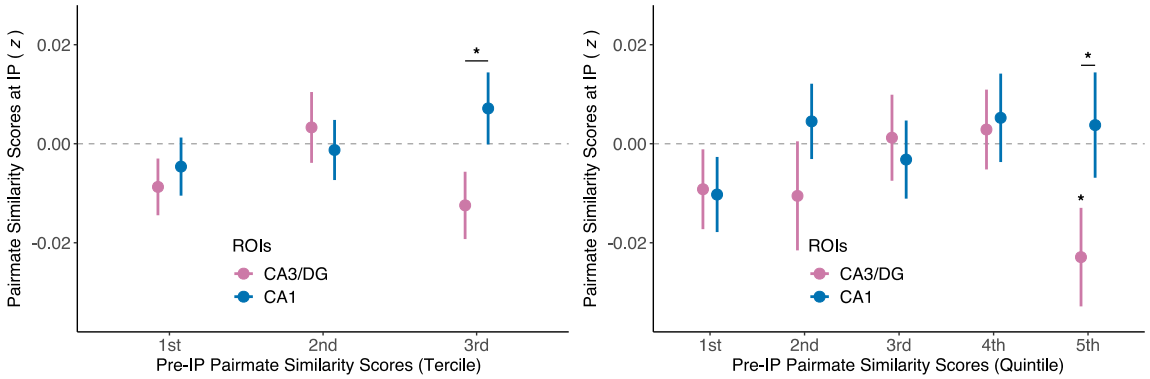
Supplementary Figure 1. Behavioral accuracy and CA3/dentate gyrus pairmate similarity scores for each set of scene pairmates. Scene pairmates are rank-ordered from highest (top, left column) to lowest (bottom, right column) mean accuracy on the associative memory test rounds. Individual plots of 'behavior' show mean associative memory test accuracy by learning round (1-6) for each set of scene pairmates. Mean test accuracy across all rounds is denoted by the red dashed line. A repeated measures ANOVA with factors of stimulus pair and round revealed a significant main effect of stimulus pair ($F_{17,510} = 4.08$, $p < 0.001$, $\eta^2 = 0.08$) as well as a significant interaction between stimulus pair and round ($F_{17,510} = 2.18$, $p = 0.004$, $\eta^2 = 0.02$). Individual plots of 'CA3/DG' show mean pairmate similarity scores in CA3/dentate gyrus by timepoint (1-5) for each set of scene pairmates. Each timepoint reflects similarity in CA3/dentate gyrus across successive learning rounds (1-2, 2-3, etc.). A repeated measures ANOVA with

factors of timepoint and stimulus pair did not reveal a significant main effect of stimulus pair ($F_{17,510} = 0.74$, $p = 0.760$, $\eta^2 = 0.01$) or an interaction between stimulus pair and timepoint in ($F_{17,510} = 1.49$, $p = 0.093$, $\eta^2 = 0.02$). Note: data are presented as mean values ± S.E.M., $n = 31$ independent participants. Source data are provided as a Source Data file.

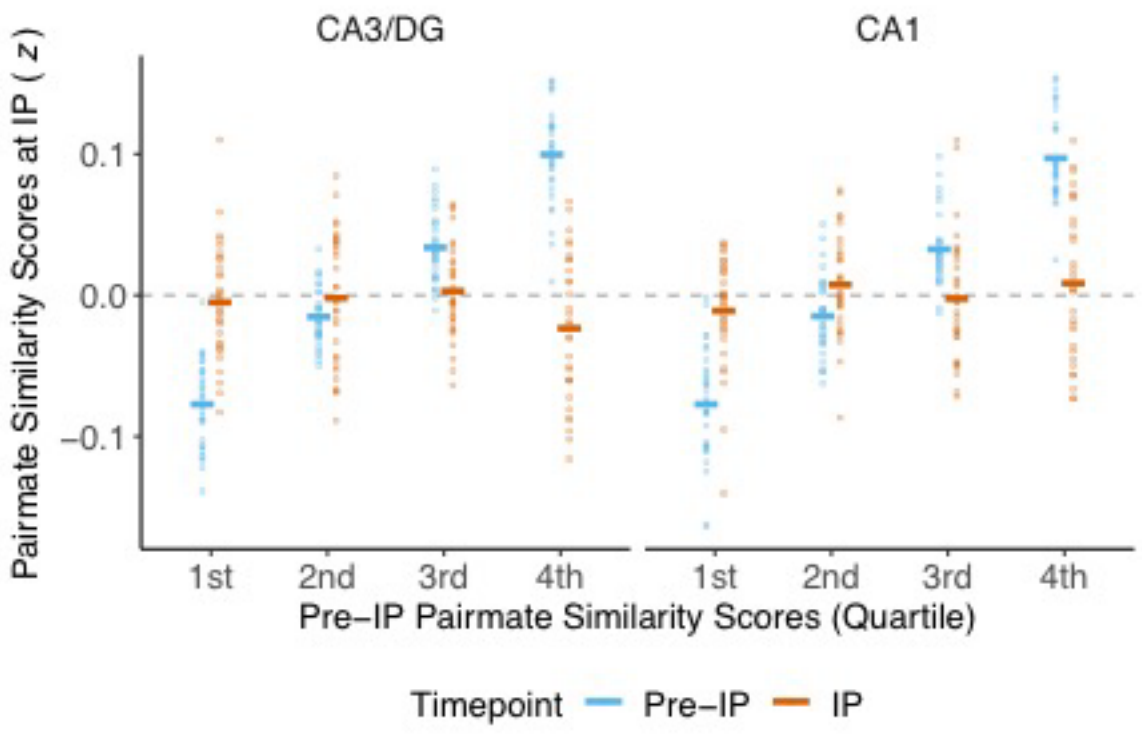


Supplementary Figure 2. Pairmate similarity scores as a function of timepoints and learning. a. Pairmate similarity scores at each timepoint for each region of interest (ROI). Each timepoint reflects correlations between successive scene exposure rounds [i.e., timepoint 1 = *r*(round 1, round 2), timepoint 2 = *r*(round 2, round 3), etc.]. CA1 is the only ROI that showed a significant main effect of timepoint (CA3/DG: $F_{4,120} = 0.24$, $p = 0.913$, $\eta^2 = 0.006$; CA1: $F_{4,120} = 3.89$, $p = 0.005$, $\eta^2 = 0.09$; PPA: $F_{4,120} = 0.34$, $p = 0.848$, $\eta^2 = 0.01$; EVC: $F_{4,120} = 0.82$, $p = 0.517$, $\eta^2 = 0.02$; repeated measures ANOVAs). **b.** Pairmate similarity scores in CA3/dentate gyrus (CA3/DG) calculated by correlating all possible combinations of the scene exposure rounds, expressed in terms of distance relative to the learned round (LR) for each pairmate. Rounds that preceded the LR reflect rounds before learning occurred, whereas the LR and following rounds reflect rounds after learning occurred (i.e., high confidence correct performance on the associative memory test). Thus, the correlations can be grouped into 3 categories: correlations among 'before' rounds (-/-), correlations between 'before' and 'after' rounds (-/+), and correlations among 'after' rounds (+/+). **c.** CA3/dentate gyrus pairmate similarity scores averaged across all of the cells within each of the three categories (-/-, -/+, +/+). Pairmate similarity scores in the -/+ category were significantly below 0 ($t_{30} = -2.70$, $p = 0.011$, $d = 0.48$, CI = [-0.004 ± 0.003], two-tailed one sample *t*-test) and significantly lower than the -/- category ($t_{30} = 2.49$, $p = 0.018$, $d = 0.45$, CI = [0.008 ± 0.006], two-tailed paired samples *t*-test). There was a trend toward lower pairmate similarity scores in -/+ category compared to the +/+ category ($t_{30} = 1.98$, $p = 0.057$, $d = 0.36$, CI = [0.004 ± 0.005], two-tailed paired samples *t*-test). Notes: ** $p < .01$, * $p < .05$,

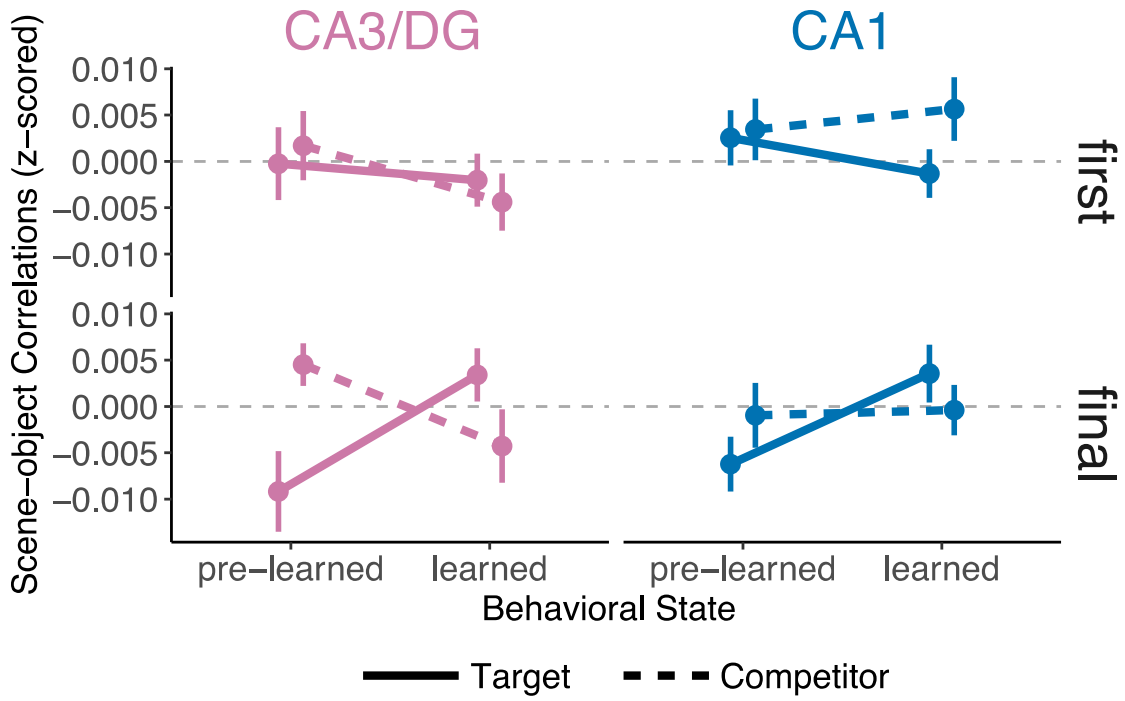
~ $p < .10$. No correction for multiple comparisons was applied given the a priori predictions for CA3/DG. Data are presented as mean values ± S.E.M. and all data reflect $n = 31$ independent participants. Source data are provided as a Source Data file.



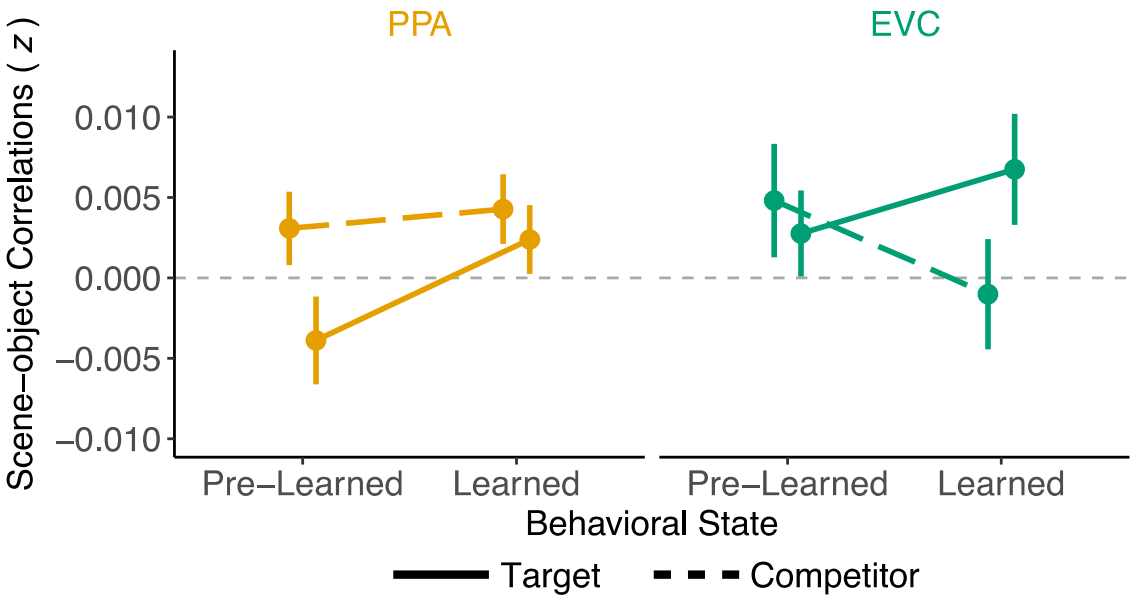
Supplementary Figure 3. Pairmate similarity scores at the inflection point (IP) as a function of relative pairmate similarity scores at the pre-inflection point (pre-IP) (related to Fig. 3c). Left: pre-IP pairmate similarity scores binned into terciles (from lowest to highest). Right: pre-IP pairmate similarity scores binned into quintiles (from lowest to highest). For both analyses, binning was performed within-subject and separately for CA3/dentate gyrus (CA3/DG) and CA1. When binned by terciles or quintiles, pairmate similarity scores in CA3/dentate gyrus were significantly lower than in CA1 at the highest pre-IP bin (terciles: t_{30} = -2.22, p = .034, d = 0.40, CI = [-0.020 ± 0.018]; quintiles: t_{30} = -2.18, p = .037, d = 0.39, CI = [-0.027 ± 0.025]; two-tailed paired samples t-tests). When binned by terciles, pairmate similarity scores in CA3/dentate gyrus were marginally below 0 for the highest pre-IP bin (t-test vs. 0: t_{30} = -1.83, p = .077, d = 0.33, CI = [-0.012 ± 0.014]; two-tailed one sample t-test). When binned by quintiles, pairmate similarity scores in CA3/dentate gyrus were significantly below 0 for the highest pre-IP bin (two-tailed one sample t-test vs. 0: t_{30} = -2.30, p = 028, d = 0.41, CI = [-0.023 ± 0.020]). Note: * p < .05. No correction for multiple comparisons was applied given the a priori predictions for CA3/DG. Data are presented as mean values ± S.E.M. and all data reflect n = 31 independent participants. Source data are provided as a Source Data file.



Supplementary Figure 4. Distributions of pairmate similarity scores at the pre-inflection point (pre-IP) and inflection point (IP), as a function of pre-IP similarity (related to Fig. 3c). Mean pairmate similarity scores at the pre-IP were binned into quartiles, separately for each subject and for CA3/dentate gyrus (CA3/DG) and CA1. Pre-IP data (blue dots) show the distribution (across subjects) of the pairmate similarity scores at each pre-IP bin. IP data (orange dots) show the distribution (across subjects) of pairmate similarity scores at the inflection point as a function of the pre-IP pairmate similarity level (1^{st} quartile = lowest pre-IP similarity, 4^{th} quartile = highest pre-IP similarity). Note: direct comparison of pre-IP versus IP values at each bin is not statistically valid given that the pre-IP data, but not the IP data, were binned by value (quartiles). Notes: all data reflect n = 31 independent participants. Source data are provided as a Source Data file.

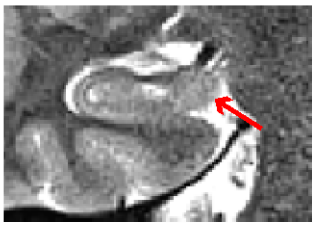


Supplementary Figure 5. Scene-object similarity as a function of object relevance (target, competitor), ROI (CA3/dentate gyrus, CA1), behavioral state (pre-learned, learned), and object run (first, final). The 3-way interaction between object relevance, behavioral state, and object run was not significant for CA3/dentate gyrus (CA3/DG: $F_{1,30}$ = 3.14, p = 0.086, η^2 = 0.01, repeated measures ANOVA) or for CA1 ($F_{1,30}$ = 3.90, p = 0.057, η^2 = 0.01, repeated measures ANOVA). When considering each object run separately, CA3/dentate gyrus exhibited a significant interaction between object relevance and behavioral state for data from the final object run ($F_{1,30}$ = 12.65, p = 0.001, η^2 = 0.07, repeated measures ANOVA), but not from the first object run ($F_{1,30}$ = 0.48, p = 0.495, η^2 = 0.003, repeated measures ANOVA). CA1 did not exhibit a significant interaction between object relevance and behavioral state for data from either object run (first run: $F_{1,30}$ = 1.34, p = 0.255, η^2 = 0.01; final run: $F_{1,30}$ = 3.85, p = 0.059, η^2 = 0.18; repeated measures ANOVAs). Note: no correction for multiple comparisons was applied given the a priori predictions for CA3/DG. Data are presented as mean values ± S.E.M. and all data reflect n = 31 independent participants. Source data are provided as a Source Data file.



Supplementary Figure 6. Scene-object similarity in PPA and EVC as a function of behavioral state. Scene-object similarity as a function of object relevance (target, competitor), ROI (PPA, yellow; EVC, green), and behavioral state (pre-learned round, learned round). There was no significant interaction between behavioral state and object relevance for either ROI (PPA: $F_{1,30} = 1.97$, $p = 0.170$, $\eta^2 = 0.01$; EVC: $F_{1,30} = 3.23$, $p = 0.082$, $\eta^2 = 0.02$, repeated measures ANOVAs). Note: No correction for multiple comparisons was applied. Data are presented as mean ± S.E.M. and all data reflect $n = 31$ independent participants. Source data are provided as a Source Data file.

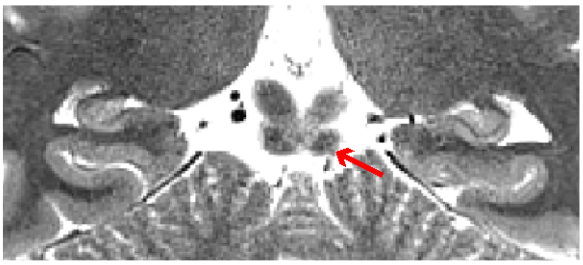
Anterior Boundary



Slice 32: Uncus present, slice not included in the hippocampus body.

Slice 33: The first slice that uncus is absent marks the the first slice of the hippo-campus body.

Posterior Boundary



Slice 40: Last slice with visible colliculi marks the last slice for the hippocampus body.

Slice 41: Colliculi disappeared, slice not included in the hippocampus body.

Supplementary Figure 7. Illustration of anterior and posterior boundaries for the hippocampal body from a sample participant.