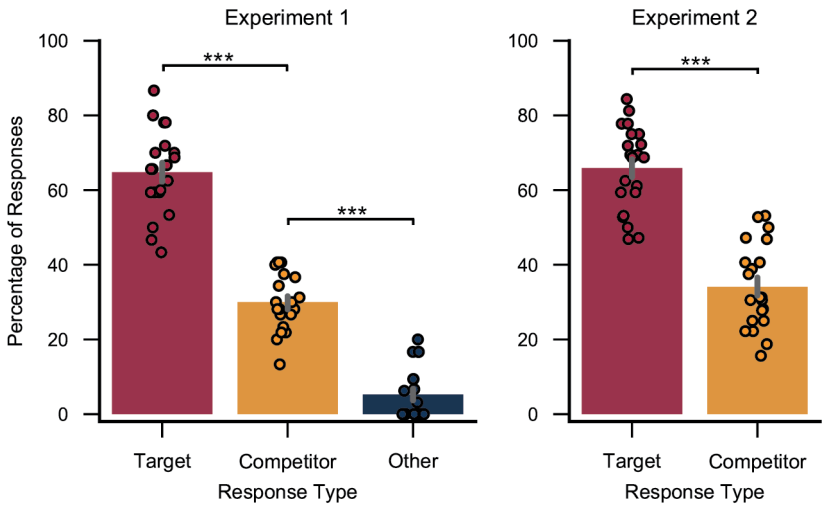

Figure S1. Related to *Figure 2*. Behavioral results from the post-scan picture test in each fMRI Experiment. After finishing all 14 learning rounds and the localizer scan subjects exited the scanner and completed the picture test. On each trial, subjects were shown a static picture drawn from one of the route stimuli (see STAR Methods). Directly below each picture was a set of destination names. Subjects were instructed to select the destination name corresponding to the route picture. In Experiment 1, each trial had 4 destination options corresponding to: the target destination; the overlapping route destination ('competitor'); and the two non-overlapping route destinations ('other'). In Experiment 2 all of the routes studied by each subject ended in one of two possible destinations. Therefore, on each trial, the two destination options corresponded to either the target destination or the overlapping route destination ('competitor'). Analyses were restricted to pictures drawn from Segment 1 of each route (i.e., the segments that contained overlap) in order to test discrimination of the overlapping routes. In both experiments subjects successfully learned to discriminate between the overlapping routes as evidenced by a higher percentage of target responses than competitor responses (Experiment 1: $t_{19} = 8.59$, $p = 0.00000006$; Experiment 2: $t_{20} = 6.44$, $p= 0.000003$). In Experiment 1, subjects were more likely to select the competitor destination than one of the 'other' destinations ($t_{19} = 11.52$, $p < 0.00000001$), indicating that route overlap contributed to memory interference. Error bars reflect +/- SEM. *** $p < 0.001$

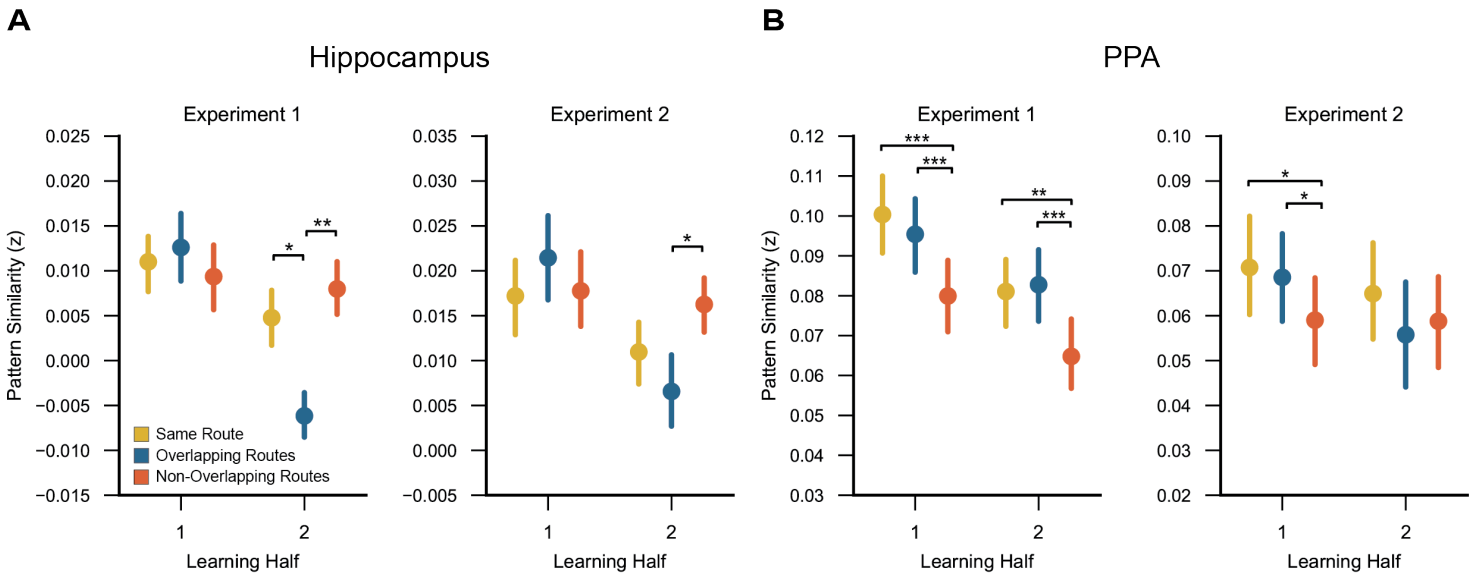

Figure S2. Related to *Figure 3*. Learning-related changes in spatiotemporal pattern similarity (Segment 1 only) for each fMRI Experiment. (A) In each fMRI Experiment, there was a significant learning-related decrease in the similarity of hippocampal representations of overlapping routes relative to non-overlapping routes (Experiment 1: $F_{1,19} = 5.99$, $p = 0.024$; Experiment 2: $F_{1,20} = 8.02$, $p = 0.010$). Furthermore, the reversal effect (overlapping route similarity < non-overlapping route similarity) was significant in the 2nd half of learning for each Experiment (Experiment 1: $t_{19} = 3.03$, $p = 0.007$; Experiment 2: $t_{20} = 2.28$, $p = 0.034$). (B) Within PPA, the interaction between learning half (1st vs. 2nd) and overlap (overlapping vs. non-overlapping routes) was not significant in Experiment 1 ($F_{1,19} = 0.45$, $p = 0.51$). In both the 1st and 2nd halves of learning, overlapping route similarity was significantly greater than non-overlapping route similarity (1^{st} half: $t_{19} = 4.56$, $p = 0.0002$; 2^{nd} half: $t_{19} = 4.76$, $p = 0.0001$). In Experiment 2, however, the interaction between learning half (1st vs. 2nd) and overlap (overlapping vs. non-overlapping routes) was significant ($F_{1,20} = 5.19$, $p = 0.034$), reflecting a relative decrease in overlapping route similarity across learning. Whereas overlapping route similarity was greater than non-overlapping route similarity in the 1^{st} half of learning ($t_{20} = 2.27$, $p = 0.034$), there was no difference between overlapping and non-overlapping route similarity in the 2nd half of learning ($t_{20} = 0.55$, $p = 0.59$). Error bars reflect +/- SEM. * $p < 0.05$, ** $p < 0.01$, *** $p < 0.001$.

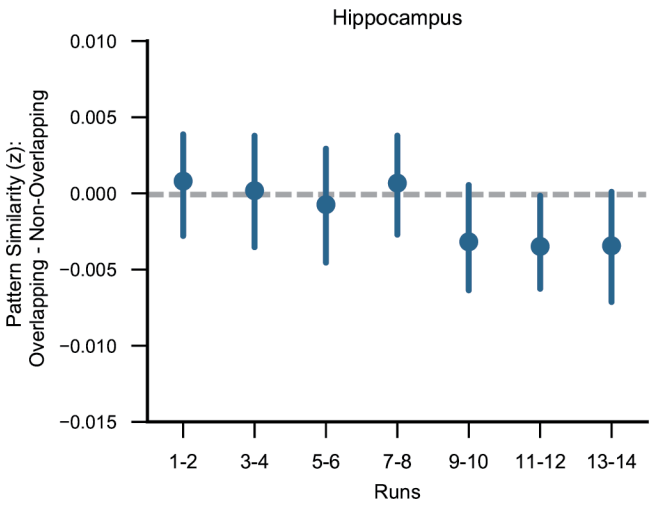

Figure S3. Related to *Figure 3*. Hippocampal spatiotemporal pattern similarity (Segment 1 only) computed every two runs. Qualitatively, there was no evidence for a reversal effect (overlapping route similarity < non-overlapping route similarity) until run 9. However, because each run contained only two repetitions of each route, this analysis was under-powered relative to the main analyses split by learning half.

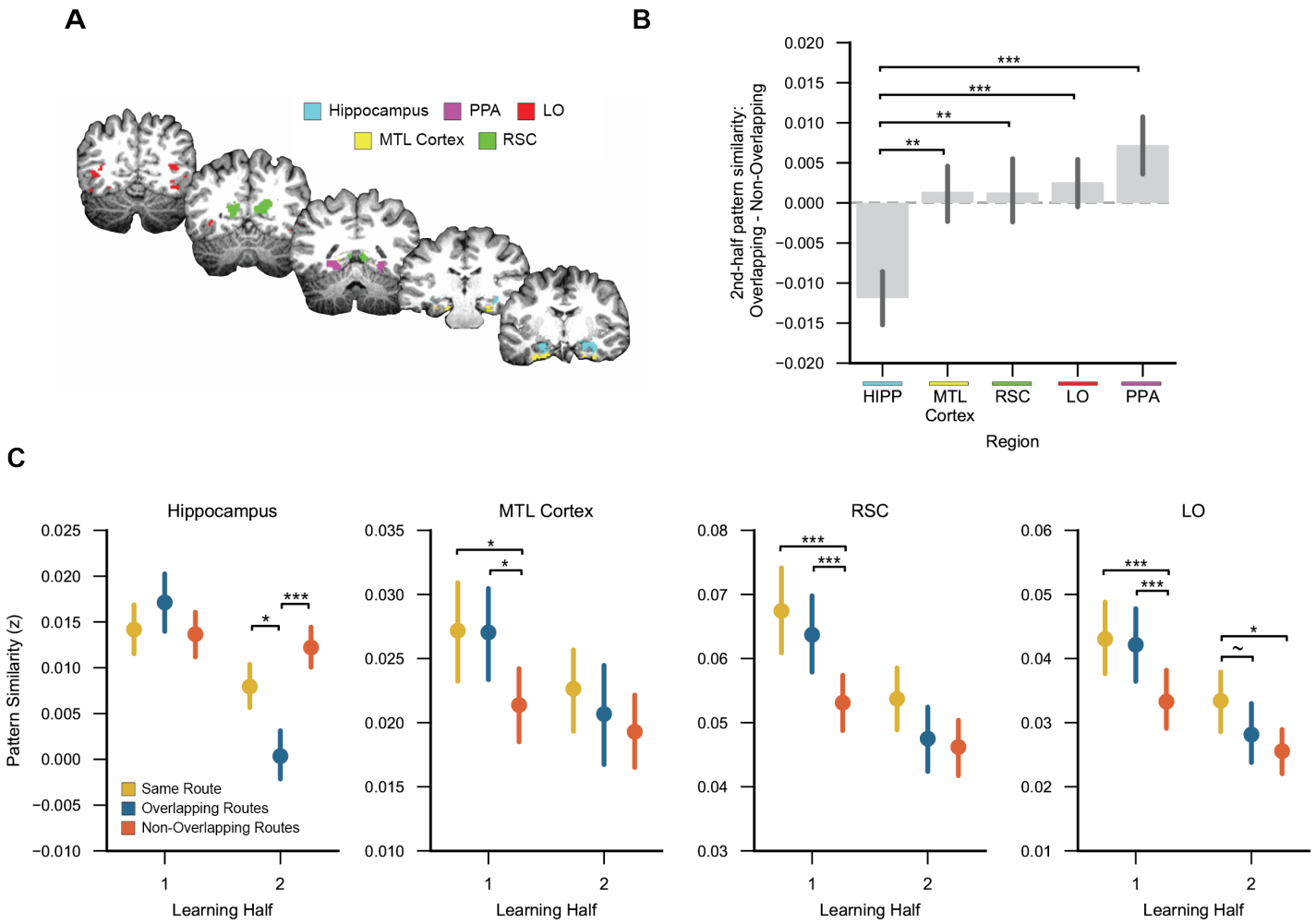

Figure S4. Related to *Figure 3*. Comparison of learning-related changes in spatiotemporal pattern similarity (Segment 1 only) for hippocampus vs. cortical regions. In addition to PPA (our primary control region), we also measured learning-related changes in spatiotemporal pattern similarity in cortical regions involved in spatial navigation [retrosplenial cortex (RSC)], object processing [lateral occipital cortex (LO)], and medial temporal lobe cortex more generally (MTL cortex). (A) Cortical regions of interest from a sample subject are displayed on the subject's T1 anatomical scan. (B) The hippocampus is the only region in which overlapping route similarity dropped below non-overlapping route similarity (reversal effect) in the 2nd half of learning. This was confirmed by significant interactions between overlap (2nd half similarity for overlapping vs non-overlapping routes) and regions of interest [hippocampus vs. MTL: $F_{1,39}$ = 10.84, p = 0.002 ; hippocampus vs. RSC: $F_{1,39}$ = 10.23, p = 0.003; hippocampus vs. LO: $F_{1,39}$ = 14.70, p = 0.0004; hippocampus vs. PPA: $F_{1,39}$ = 22.18, p = 0.00003]. Thus, the representational end-states of learning were qualitatively different in the hippocampus compared to cortex. (C) Segment 1 pattern similarity for each condition and learning half for the hippocampus (identical to **Figure 3C,** but shown for comparison), MTL cortex, RSC, and LO. Among the cortical regions, RSC was the only region that showed a significant decrease in overlapping route similarity across learning, relative to non-overlapping route similarity ($F_{1,39}$ = 4.45, p = 0.041). This effect was marginal in LO ($F_{1,39}$ = 3.26, p = 0.079) and not significant in MTL cortex ($F_{1,39}$ = 1.36, p = 0.25). In all three regions, overlapping route similarly was significantly greater than non-overlapping route similarity in the 1st half of learning [MTL cortex: $F_{1,39}$= 6.27, p = 0.016; RSC: $F_{1,39}$ = 16.10, p = 0.0003; LO: $F_{1,39}$ = 12.87, p = 0.0009] and there was no difference between overlapping and non-overlapping route similarity in the 2nd half of learning [MTL cortex: $F_{1,39}$ = 0.16, p = 0.069; RSC: $F_{1,39}$ = 0.11, p = 0.74; LO: $F_{1,39}$ = 0.76, p = 0.39]. Error bars reflect +/- SEM. ~ p <0.1 * p < 0.05, ** p < 0.01, *** p < 0.001

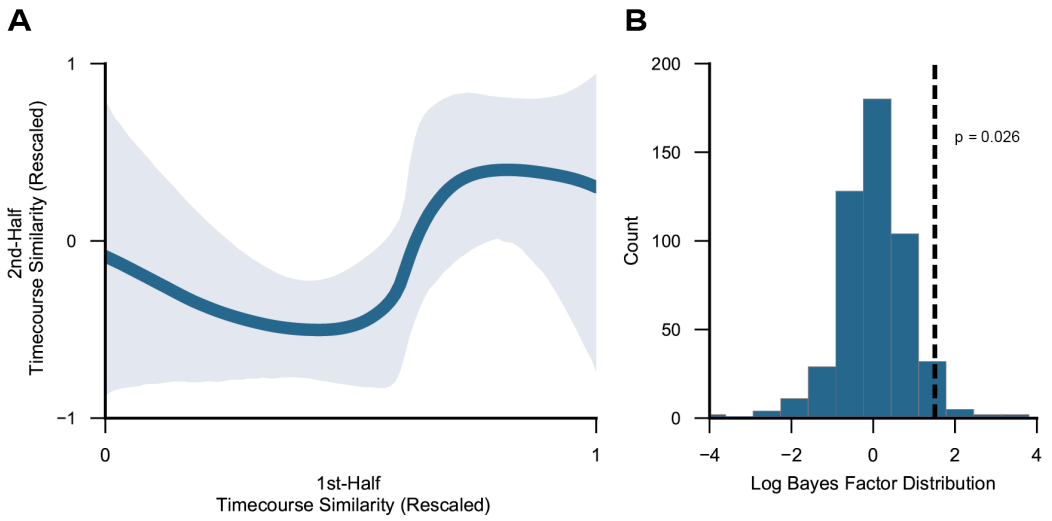

Figure S5. Related to *Figure 6*. Bayesian curve-fitting analysis. To more formally assess the non-monotonic relationship between 1st half and 2nd half timecourse similarity, we used a Bayesian curve-fitting algorithm–the Probablistic Curve Induction and Testing Tooblbox (P-CIT) [S1] that was specifically developed to test for non-monotonic plasticity. Relative to quadratic trend analyses, the P-CIT algorithm allows for a more detailed specification of a predicted curve shape, by explicitly including a set of curve parameters. In our case, the parameters describe the relationship between first-half timecourse similarity (x-axis) and second-half timecourse similarity (y-axis). We parameterized the predicted curve shape using previously described parameters that reflect the prediction of non-monotonic plasticity [S2]. Specifically, the predicted curve was defined as one in which the function, when moving from left to right, drops below the initial start value and then rises above the start value. The first step of the P-CIT algorithm is to estimate a curve shape given the data. To accomplish this, the algorithm estimates a probability distribution over possible curves, conditional on the observed data, by randomly sampling curve shapes and then assigning each sampled curve an importance weight indicating how well the curve's shape fit the observed data. It then estimates a curve by averaging the sampled curves together, weighted by their importance values. The next goal of the algorithm is to evaluate the level of evidence in favor of the predicted curve shape. It does so by labeling each sample curve as theory consistent (in our case, if it drops below the starting value and then rises above the starting value) or inconsistent, and then computes a log Bayes factor value that represents the log ratio of evidence in favor of or against the predicted shape [S3]. Positive log Bayes factor values indicate greater evidence in favor of the theory. For this analysis, we re-binned all of the 1st-half timecourse similarity values into 60 bins (5 voxels per bin) in order to allow for greater variability in the observed curve shape. This analysis used data aggregated across all subjects. (A) The estimated curve was consistent with the predicted curve shape (log Bayes factor = 1.51) and explained a significant amount of variance in the actual (X_2 = 11.13, p = 0.0008). Shaded area reflects the 90% credible interval. (B) We next ran a permutation test to estimate the null distribution of log Bayes Factor values. Out of 500 permutations, only 2.6% yielded log Bayes factor values that matched or exceeded the value obtained from the un-permuted data, indicating that it was unlikely to obtain this level of support for the predicted curve shape by chance. Finally, to assess the population-level reliability of the non-monotonic curve we ran a bootstrap resampling test in which we iteratively resampled data from subjects with replacement and then computed the log Bayes factor value for each iteration. Four-hundred and eighty-eight of the 500 bootstrap iterations (97.6%) yielded positive log Bayes factor values. Thus, the curve-fitting analyses provided additional evidence for a non-monotonic relationship between voxel overlap at the beginning vs. end of learning: that is, hippocampal voxels that were 'moderately shared' across overlapping routes at the beginning of learning were the 'least shared' by the end of learning.

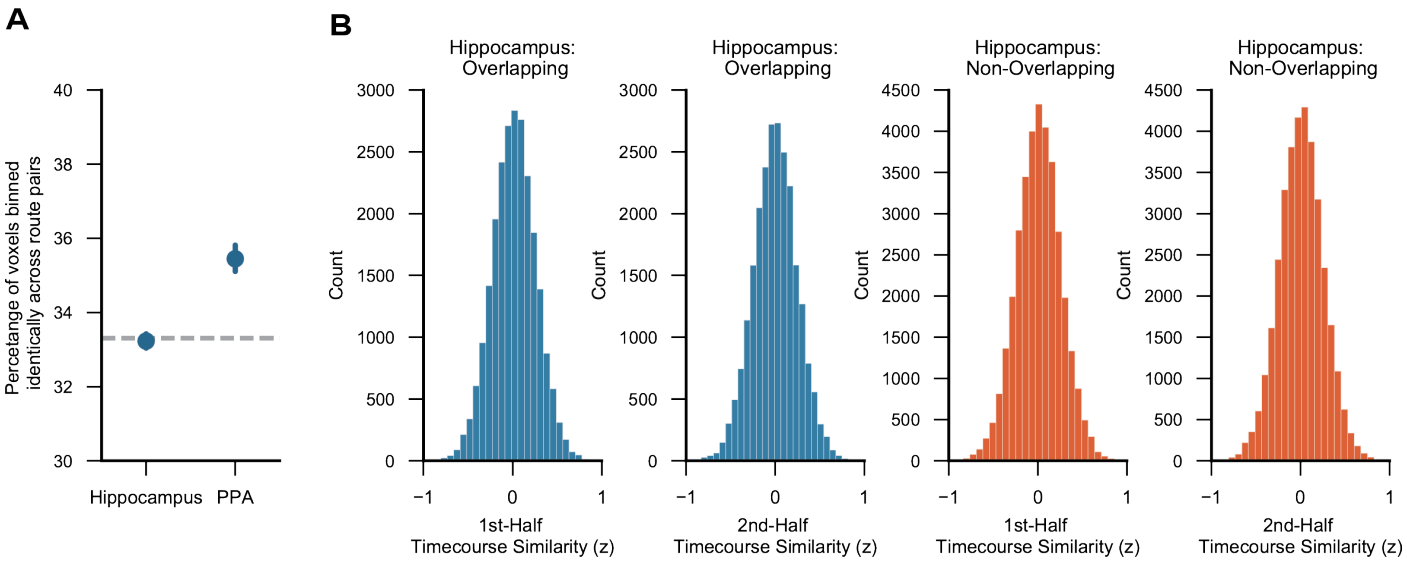

Figure S6. Related to Figure 6. Consistency of voxel timecourse similarity binning across overlapping route pairs. (A) Voxels within each ROI (hippocampus, PPA) were binned into three groups (weak, moderate, or strong) based on their 1st half timecourse similarity across overlapping pairs. Importantly, this binning was independently repeated for each pair of overlapping routes. Thus, a given voxel might be in the "weak" bin (low timecourse similarity) for one pair of overlapping routes, but in the "strong" bin (high timecourse similarity) for a different pair of overlapping routes. To measure the consistency of voxel bins across pairs (i.e. to test whether a voxel's bin for one pair of overlapping routes predicted its bin for another pair of overlapping routes) we computed the percentage of voxels within each ROI, for each subject, that were placed in the same bin across overlapping route comparisons. [Note: for each subject, there were exactly two pairs of overlapping routes]. Chance 'performance' for this measure would be 33.3%–anything above chance would indicate some degree of consistency in voxel binning across route pairs. Hippocampal voxels were not consistently binned across overlapping routes, with the measure of consistency falling right at chance. The consistency values were slightly higher in PPA. However, even in PPA, the values were below 36%, which means the overall consistency of binning was quite low. This indicates that the binning of voxels according to timecourse similarity was idiosyncratic for each overlapping route pair. Error bars reflect +/- SEM. (B) Histograms showing distributions of timecourse similarity values, pooling across all subjects, route pairs, and voxels. Note: count values are higher for the non-overlapping routes because there are more non-overlapping route pairs than overlapping route pairs.

Supplemental References

S1. Detre, G. J., Natarajan, A., Gershman, S. J., and Norman, K. A. (2013). Moderate levels of activation lead to forgetting in the think/no-think paradigm. Neuropsychologia *51*, 2371–2388.

S2. Kim, G., Lewis-Peacock, J. A., Norman, K. A., and Turk-Browne, N. B. (2014). Pruning of memories by context-based prediction error. Proc. Natl. Acad. Sci. USA *111*, 8997–9002.

S3. Lewis-Peacock, J. A., and Norman, K. A. (2014). Competition between items in working memory leads to forgetting. Nat. Commun. *5*, 5768.