# Increasing stimulus similarity drives nonmonotonic representational change in hippocampus

*Jeffrey D. Wammes[1,2], Kenneth A. Norman[3,4] and Nicholas B. Turk-Browne[1]

[1]Department of Psychology, Yale University, New Haven, CT 06520

[2]Department of Psychology, Queen's University, Kingston, ON K7L 3N6

[3]Department of Psychology, Princeton University, Princeton, NJ 08544

[4]Princeton Neuroscience Institute, Princeton University, Princeton, NJ 08544

## Abstract

Studies of hippocampal learning have obtained seemingly contradictory results, with manipulations that increase coactivation of memories sometimes leading to differentiation of these memories, but sometimes not. These results could potentially be reconciled using the nonmonotonic plasticity hypothesis, which posits that representational change (memories moving apart or together) is a U-shaped function of the coactivation of these memories during learning. Testing this hypothesis requires manipulating coactivation over a wide enough range to reveal the full U-shape. To accomplish this, we used a novel neural network image synthesis procedure to create pairs of stimuli that varied parametrically in their similarity in high-level visual regions that provide input to the hippocampus. Sequences of these pairs were shown to human participants during high-resolution fMRI. As predicted, learning changed the representations of paired images in the dentate gyrus as a U-shaped function of image similarity, with neural differentiation occurring only for moderately similar images.


**Keywords:** plasticity; statistical learning; image synthesis; deep neural networks; model-based analysis

# 1 Introduction

Humans constantly learn new facts, encounter new events, and see and hear new things. Successfully managing this incoming information requires accommodating the new with the old, reorganizing memory as we learn from experience. How does learning dynamically shape representations in the hippocampus? Our experiences are encoded in distributed representations (Polyn, Natu, Cohen, & Norman, 2005; Johnson, McDuff, Rugg, & Norman, 2009), spanning populations of neurons that are partially reused across multiple memories, leading to overlap. As we learn, the overlapping neural populations representing different memories in the hippocampus can shift, leading to either *integration*, where memories become more similar to one other, or *differentiation*, where memories become more distinct from one another (for reviews, see Duncan & Schlichting, 2018; Brunec, Robin, Olsen, Moscovitch, & Barense, 2020; Ritvo, Turk-Browne, & Norman, 2019).

Whether memories integrate or differentiate depends on whether the synapses that are common across the different memories strengthen or weaken. Traditional Hebbian learning models hold that synaptic connections strengthen when the pre-synaptic neuron repeatedly stimulates the post-synaptic neuron, causing them to fire together (Buonomano & Merzenich, 1998; Caporale & Dan, 2008; Feldman, 2009; Hebb, 1949). In other words, coactivation of neurons leads to strengthened connections between these neurons. This logic can scale up to the level of many synapses among entire populations of neurons, comprising distributed representations. A greater degree of coactivation among representations will strengthen shared connections and lead to integration. Consistent with this view, arbitrary pairs of objects integrate in the hippocampus following repeated temporal or spatial co-occurrence (e.g., Deuker, Bellmund, Schröder, & Doeller, 2016; Schapiro, Kustner, & Turk-Browne, 2012). Moreover, new information that builds a link between two previously disconnected events can lead the representations of the events to integrate (Collin, Milivojevic, & Doeller, 2015; Milivojevic, Vicente-Grabovetsky, & Doeller, 2015; Tompary & Davachi, 2017). In other cases, however, coactivation produces the exact opposite outcome — differentiation. For example, hippocampal representations of two faces with similar associations (Favila, Chanales, & Kuhl, 2016) and of two navigation events with similar routes (Chanales, Oza, Favila, & Kuhl, 2017) differentiate as a result of learning. Further complicating matters, some studies have found that the same experimental conditions can lead to integration in some subregions of the hippocampus and differentiation in other subregions (Dimsdale-Zucker, Ritchey, Ekstrom, Yonelinas, & Ranganath, 2018; Molitor, Sherrill, Morton, Miller, & Preston, 2020; Schlichting, Mumford, & Preston, 2015).

Such findings challenge Hebbian learning as a complete or parsimonious account of hippocampal plasticity. They suggest a more complex relationship between coactivation and representational change than the linear positive relationship predicted by classic Hebbian learning. We recently argued (Ritvo et al., 2019) that

this complex pattern of data could potentially be explained by the nonmonotonic plasticity hypothesis (NMPH; Detre, Natarajan, Gershman, & Norman, 2013; Hulbert & Norman, 2015; Newman & Norman, 2010; Ritvo et al., 2019), which posts a 'U-shaped' pattern of representational change as a function of the degree to which two memories coactivate (Fig. 1). According to the NMPH, low levels of coactivation between two memories will lead to no change in their overlap; high levels of coactivation will strengthen mutual connections and lead to integration; and moderate levels of coactivation (where one memory is strongly active and the unique parts of the other memory are only moderately active) will weaken mutual connections and lead to differentiation, thereby reducing competition between the memories for later retrieval attempts (Hulbert & Norman, 2015; Ritvo et al., 2019; Wimber, Alink, Charest, Kriegeskorte, & Anderson, 2015). As discussed in Ritvo et al. (2019), the NMPH can explain findings showing that a shared associate leads to differentiation (Schlichting et al., 2015; Favila et al., 2016; Chanales et al., 2017) by positing that the shared associate results in a moderate level of coactivation. Likewise, the NMPH can explain findings showing that a shared associate leads to integration (Schlichting et al., 2015; Collin et al., 2015; Milivojevic et al., 2015) by positing that the shared associate results in a high level of coactivation.

Importantly, although the NMPH is compatible with findings of both differentiation and integration, the explanations provided above are *post hoc* and do not provide a principled test of the NMPH's core claim that the function relating coactivation to representational change is U-shaped. If there were a way of knowing where on the x-axis of this function an experimental condition was located (note that Fig. 1 has no units), we could make *a priori* predictions about the learning that should take place, but practically speaking this is impossible: A wide range of neural findings on *metaplasticity* (summarized by Bear, 2003) suggest that the transition point on the U-shaped curve between synaptic weakening (leading to differentiation) and synaptic strengthening (leading to integration) can be shifted based on experience. In light of this constraint, Ritvo et al. (2019) argue that the key to robustly testing the NMPH account of representational change is to obtain samples from the full x-axis of the U-shaped curve and to look for a graded transition where differentiation starts to emerge at higher levels of memory coactivation and then disappears for even higher levels of memory coactivation.

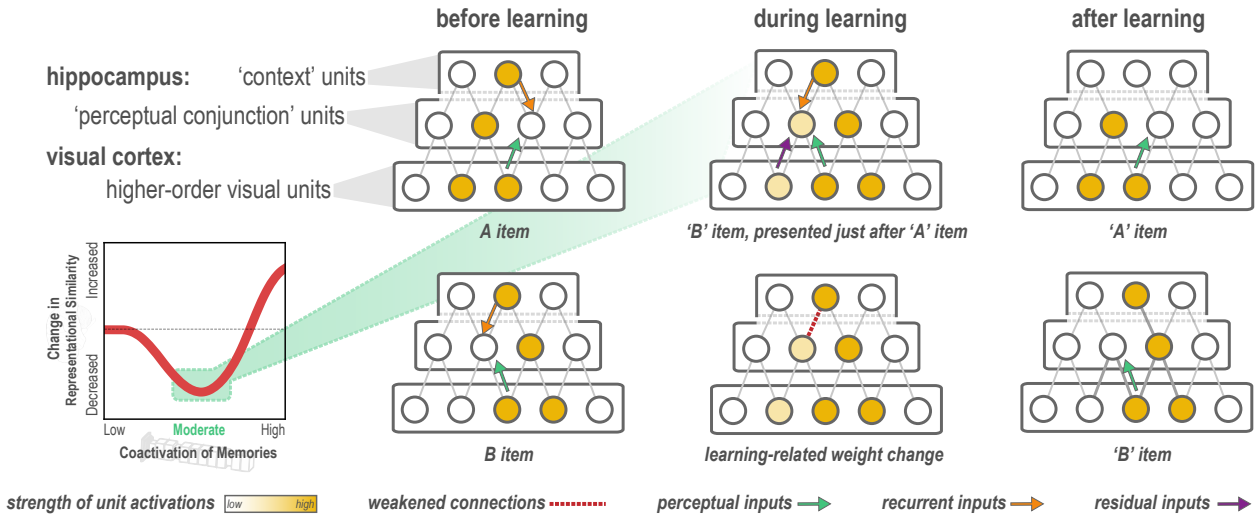


Figure 1: **Explanation of why moderate levels of visual similarity lead to differentiation.** Inset (bottom left) depicts the hypothesized nonmonotonic relationship between coactivation of memories and representational change from pre- to post-learning in the hippocampus. Low coactivation leads to no representational change, moderate coactivation leads to differentiation, and high coactivation leads to integration. Network diagrams show activity patterns in high-level visual cortex and the hippocampus evoked by two stimuli (A and B) with a moderate level of visual similarity that are presented as a 'pair' in a statistical learning procedure (such that B is reliably presented after A). Note that the hippocampus is hierarchically organized into a layer of *perceptual conjunction units* that respond to conjunctions of visual features and a layer of *context units* that respond to other features of the experimental context (McKenzie et al., 2014). Before statistical learning (left-hand column), the hippocampal representations of A and B share a context unit (because the items appeared in the same experimental context) but do not share any perceptual conjunction units. The middle column (top) diagram shows network activity during statistical learning, when the B item is presented immediately following an A item; the key consequence of this sequencing is that there is residual activation of A's representation in visual cortex when B is presented. The colored arrows are meant to indicate different sources of input converging on the unique part of each item's hippocampal representation (in the perceptual conjunction layer) when the other item is presented: green = perceptual input from cortex due to shared features (this is proportional to the overlap in the visual cortex representations of these items); orange = recurrent input within the hippocampus; purple = input from residual activation of the unique features of the previously-presented item. The purple input is what is different between the pre-statistical-learning phase (where A is not reliably presented before B) and the statistical learning phase (where A is reliably presented before B). In this example, the orange and green sources of input are not (on their own) sufficient to activate the other item's hippocampal representation during the pre-statistical-learning phase, but the combination of all three sources of input is enough to moderately activate A's hippocampal representation when B is presented during the statistical learning phase. The middle column (bottom) diagram shows the learning that will occur as a result of this moderate activation, according to the NMPH: The connection between the (moderately activated) item-A hippocampal unit and the (strongly activated) hippocampal context unit is weakened (note that this is not the only learning predicted by the NMPH in this scenario, but it is the most relevant learning and hence is highlighted in the diagram). As a result of this weakening, when item A is presented after statistical learning (right-hand column, top), it does not activate the hippocampal context unit, but item B still does (right-hand column, bottom), resulting in an overall decrease in the overlap of the hippocampal representations of A and B from pre-to-post learning.

No existing study has demonstrated the full U-shaped pattern for representational change; that is what we set out to do here, using a statistical learning paradigm. Fig. 1 illustrates the NMPH's predictions regarding how pairing two items (A and B) in a visual statistical learning paradigm (such that B reliably follows A) can affect the similarity of the hippocampal representations of A and B. The figure depicts a situation where items A and B have moderate visual similarity, and statistical learning leads to differentiation of their hippocampal

representations (because item A's hippocampal representation is moderately activated during the presentation of item B). Crucially, the figure illustrates that there are three factors that influence how strongly the hippocampal representation of item A co-activates with the hippocampal representation of item B during statistical learning: (1) overlap in the high-level visual cortex representations of items A and B; (2) recurrent input from overlapping features within the hippocampus; and (3) residual activation of item A's representation in visual cortex (because item A was presented immediately before item B). Thus, if we want to parametrically vary the coactivation of the hippocampal representations (to span the full axis of Fig. 1 and test for a full 'U' shape), we need to vary at least one of these three factors. In our study, we chose to focus on the first factor (overlap in visual cortex). Specifically, by controlling the visual similarity of paired items (c.f. Molitor et al., 2020), we sought to manipulate overlap in visual cortex and (through this) parametrically vary the coactivation of memories in the hippocampus.

To accomplish this goal, we developed a novel approach for synthesizing image pairs using deep neural network (DNN) models of vision. These models provide a link from pictures to rich quantitative descriptions of visual features, which in turn approximate some key principles of how the visual system is organized (e.g., Güçlü & van Gerven, 2015; Cichy, Khosla, Pantazis, Torralba, & Oliva, 2016; Luo, Li, Urtasun, & Zemel, 2016; Zeiler & Fergus, 2014; Cichy et al., 2016; de Beeck, Torfs, & Wagemans, 2008; Khaligh-Razavi & Kriegeskorte, 2014; Kriegeskorte, 2009, 2015; Kubilius, Bracci, & de Beeck, 2016). Most critically, later DNN layers correspond most closely to higher-order, object-selective visual areas (Eickenberg, Gramfort, Varoquaux, & Thirion, 2017; Güçlü & van Gerven, 2015; Jozwik, Kriegeskorte, Cichy, & Mur, 2019; Khaligh-Razavi & Kriegeskorte, 2014), and when neural networks are optimized to match human performance, their higher layers predict neural responses in higher-order visual cortex (Cadieu et al., 2014; Yamins et al., 2014). We reasoned that synthesizing pairs of stimuli that parametrically varied in their feature overlap in the upper layers of a DNN (Szegedy et al., 2015) would also parametrically vary their *neural* overlap in the high-level visual regions that provide input to the hippocampus.

Image pairs spanning the range of possible representational overlap values were synthesized according to the procedure shown in Fig. 2A and B and embedded in a statistical learning paradigm (Schapiro et al., 2012). During fMRI, participants were given a pre-learning templating run (where the images were presented in a random order, allowing us to record the neural activity evoked by each image separately), followed by six statistical learning runs (where the images where presented in a structured order, such that the first image in a pair was always followed by the second image), followed by a post-learning templating run (Fig. 2C). We hypothesized that manipulating the visual similarity of the paired images would allow us to span the x-axis of Fig. 1 and reveal a full U-shaped curve going from no change to differentiation to integration. An important caveat in this regard

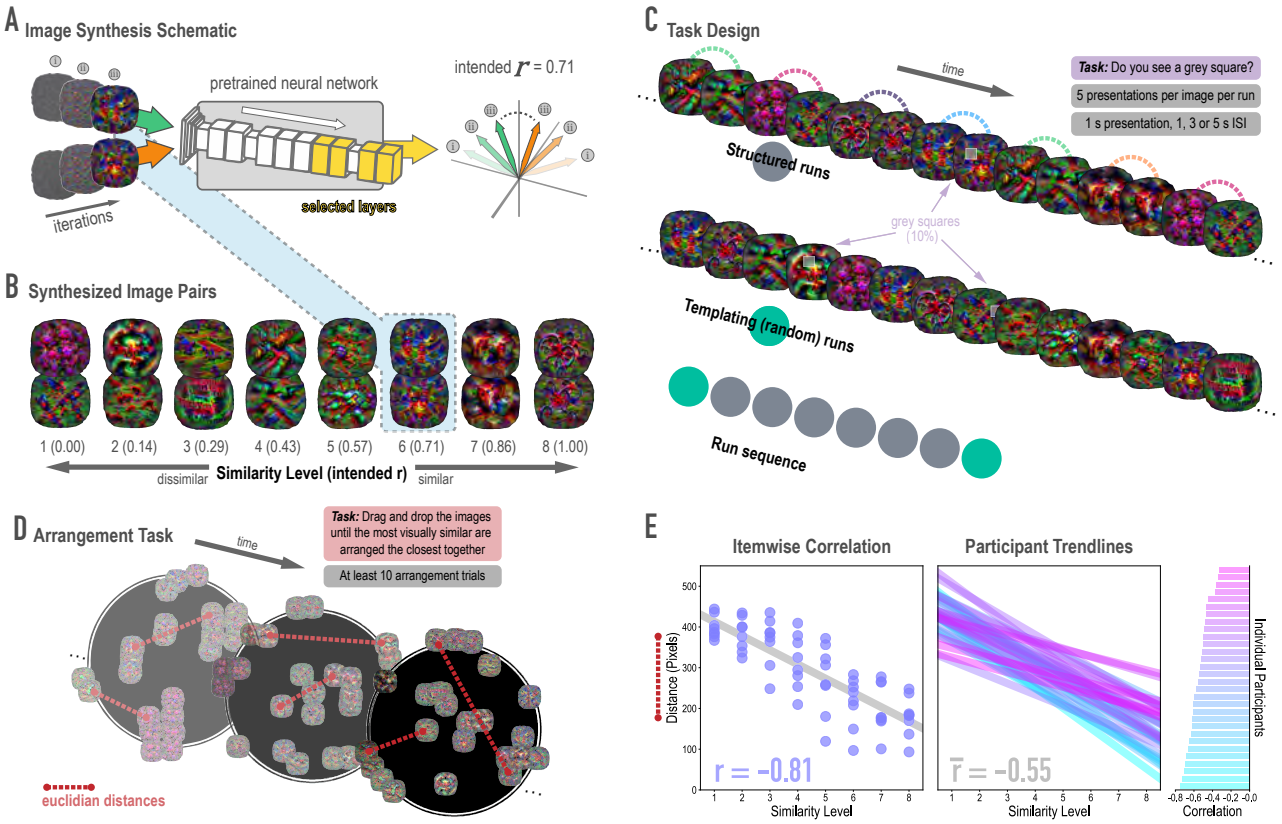


Figure 2: **Schematic of image synthesis algorithm, fMRI task design, and behavioral validation.** **(A)** Our image synthesis algorithm starts with two visual noise arrays that are updated through many iterations (only three are depicted here: i, ii, and iii), until the feature activations from selected neural network layers (shown in yellow) achieve an intended Pearson correlation (*r*) value. **(B)** The result of our image synthesis algorithm was eight image pairs, that ranged in similarity from completely unrelated (similarity level 1, intended *r* among higher-order features = 0) to almost identical (similarity level 8, intended *r* = 1.00). **(C)** An fMRI experiment was conducted with these images to measure neural similarity and representation change. Participants performed a monitoring task in which they viewed a sequence of images, one at a time, and identified infrequent (10% of trials) grey squares in the image. Unbeknownst to participants, the sequence of images in structured runs contained the pairs (i.e., the first pairmate was *always* followed by the second pairmate); the images in templating runs were pseudo-randomly ordered with no pairs, making it possible to record the neural activity evoked by each image separately. **(D)** A behavioral experiment was conducted to verify that these similarity levels were psychologically meaningful. Participants performed an arrangement task in which they dragged and dropped images in a workspace until the most visually similar images were closest together. From the final arrangements, pairwise Euclidean distances were calculated as a measure of perceived similarity. **(E)** Correlation between model similarity level and distance between images (in pixels) in the arrangement task. On the left, each point represents a pair of images, with distances averaged across participants. In the center, each trendline represents the relationship between similarity level and an individual participants' distances. The rightmost plot shows the magnitude of the correlation for each participant.

is that some hippocampal subregions might be better suited to show the U-shape than others: Prior work has shown that there is extensive variation in overall activity (sparsity) levels across hippocampal subregions, with CA2/3 and DG showing much sparser codes than CA1 (Duncan & Schlichting, 2018; Barnes, McNaughton, Mizumori, Leonard, & Lin, 1990). To the extent that overall levels of representational coactivation are higher in CA1 than CA2/3 and DG (i.e., falling on the rightmost side of Fig. 1), it might be harder to observe the 'dip' in the U-shaped curve corresponding to differentiation in CA1. Consistent with this idea, CA1 is relatively biased toward integration and CA2/3/DG are relatively biased toward differentiation (Kim, Norman, & Turk-Browne, 2017; Dimsdale-Zucker et al., 2018; Molitor et al., 2020). Based on these findings, we expected that the U-shaped pattern would be more prominent in CA2/3 and DG, and perhaps strongest in DG, given that it has higher sparsity (Barnes et al., 1990) and greater pattern separation (Berron et al., 2016) than CA3. As discussed below, this prediction was borne out in the data: Using synthesized image pairs varying in similarity, we demonstrate the full U-shape (transitioning into and out of differentiation, as a function of similarity) in DG, thereby providing direct evidence that hippocampal plasticity is nonmonotonic.

# 2 Results

## 2.1 Stimulus synthesis

### 2.1.1 Model validation

Before looking at the effects of statistical learning on hippocampal representations, we wanted to verify that our model-based synthesis approach was effective in creating graded levels of feature similarity in the targeted layers of the network (corresponding to high-level visual cortex): Specifically, our goal was to synthesize images that varied parametrically in their similarity in higher layers while not differing systematically in lower and middle layers of the network. To assess whether we were successful in meeting this goal, we fed the final image pairs (Fig. 2B) back through the neural network that generated them (GoogLeNet/Inception; Szegedy et al., 2015), and computed the actual feature correlations at the targeted layers. We found that the intended and actual similarity levels of the images (in terms of model features) showed a close correspondence (Supplementary Fig. 2): In the highest four layers (4D-5B), the intended and actual feature correlations were strongly associated ($r(62)$ = .970, .983, .977, .985, respectively). In the lower and middle layers, feature correlations did not vary across pairs, as intended.

### 2.1.2 Behavioral validation

Because deep neural networks can be influenced by visual features to which humans are insensitive (Nguyen, Yosinski, & Clune, 2015), we also sought to validate that the differences in similarity levels across image pairs were perceptually meaningful to human observers. We employed a behavioral task in which participants ($n$ = 30) arranged sets of images (via dragging and dropping in a 2-D workspace; Fig. 2D), with the instruction to place images that are visually similar close together and images that are visually dissimilar far apart (Kriegeskorte & Mur, 2012). Participants completed at least 10 arrangement trials and the distances for each synthesized image pair were averaged across these trials. When further averaged across participants, perceptual distance was strongly negatively associated with the intended model similarity ($r(62)$ = -.813, $p < .0001$; Fig. 2E). In other words, image pairs at the highest similarity levels were placed closer to one another. In fact, every individual participant's correlation was negative (mean $r$ = -.552, 95% CI = [-0.593 -0.512]).

### 2.1.3 Neural validation

Because we were synthesizing image pairs based on features from the highest model layers, we hypothesized that model similarity would be associated with representational similarity in high-level visual cortical regions such as lateral occipital (LO) and inferior temporal (IT) cortices. We also explored ventral temporal regions parahippocampal cortex (PHC) and fusiform gyrus (FG), and early visual regions V1 and V2. Based on separate viewing of the 16 synthesized images during the initial templating run (prior to statistical learning), we calculated an image-specific pattern of BOLD activity across voxels in each anatomical ROI. We then correlated these patterns across image pairs as a measure of neural similarity (Fig. 3A). Model similarity level was positively associated with neural similarity in LO (mean $r$ = .163, 95% CI = [.056 .270], randomization $p$ = .024) and PHC (mean $r$ = .127, 95% CI = [.016 .239], $p$ = .031). No other region showed a significant positive relationship to model similarity (V2: mean $r$ = -.046, 95% CI = [-.163 .069], $p$ = .720; IT: $r$ = .090, 95% CI = [-.049 .226], $p$ = .095; FG: $r$ = .058, 95% CI = [-.059 .179], $p$ = .195); V1 showed a negative relationship (mean $r$ = -.141, 95% CI = [-.261 -.025], $p$ = .020). The correspondence between the similarity of image pairs in the model and in LO and PHC is consistent with our use of the highest layers of a neural network model for visual object recognition in image synthesis. The fact that this correspondence was observed in LO and PHC but not in earlier visual areas further validates that similarity was based on high-level features.

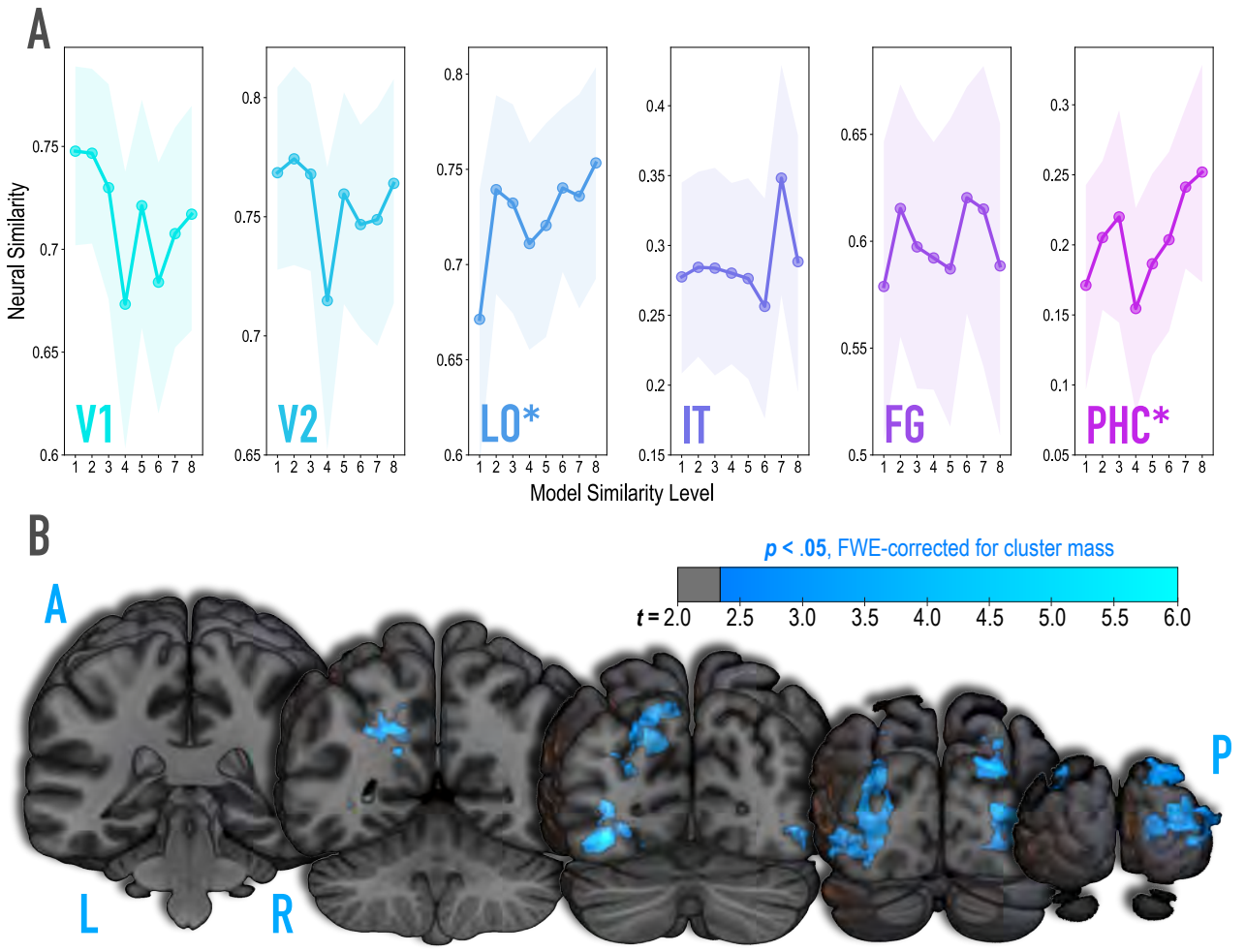


Figure 3: **Analysis of where in the brain representational similarity tracked model similarity, prior to statistical learning.** **(A)** Correlation of voxel activity patterns evoked by pairs of stimuli (before statistical learning) in different brain regions of interest, as a function of model similarity level (i.e., how similar the internal representations of stimuli were in the targeted layers of the model). Neural similarity was reliably positively associated with model similarity level only in LO and PHC. Shaded areas depict bootstrap resampled 95% confidence intervals at each model similarity level. **(B)** Searchlight analysis. Brain images depict coronal slices viewed from a posterior vantage point. Clusters in blue survived correction for family-wise error (FWE) at $p < .05$ using the null distribution of maximum cluster mass. L = left hemisphere, R = right hemisphere, A = anterior, P = posterior.

It is unlikely that any given model layer(s) will map perfectly and exclusively to a single anatomical region. Accordingly, although we targeted higher-order visual cortex (e.g., LO, IT), the layers we manipulated may have influenced representations in other regions, or alternatively, a subset of the voxels within a given anatomical ROI. To explore this possibility, we performed a searchlight analysis (Fig. 3B) testing where in the brain neural similarity was positively associated with model similarity. This revealed two large clusters of voxels ($p < .05$ corrected): left ventral and dorsal LO extending into posterior FG (3722 voxels; peak $t$-value = 5.60; MNI coordinates of peak = -37.5, -72.0, -10.5; coordinates of center = -26.7, -71.7, 17.4) and right

ventral and dorsal LO extending into occipital pole (3107 voxels; peak $t$-value = 4.82; coordinates of peak = 33.0, -88.5, 10.5; coordinates of center = 30.9, -86.5, 10.6).

## 2.2 Representational change

### 2.2.1 Hippocampus

We hypothesized that learning-related representational change in the hippocampus would follow a nonmonotonic curve. That is, we predicted a cubic function wherein low levels of model similarity would yield no neural change, moderate levels of model similarity would dip toward neural differentiation, and high levels of model similarity would climb back toward neural integration (Fig. 1 inset). We predicted that this nonmonotonic pattern would be observed in the CA2/3 and DG subfields, given the predisposition of these subfields (especially DG) to sparse representations and pattern separation.

To test this hypothesis, we extracted spatial patterns of voxel activity associated with each image from separate runs that occurred before and after statistical learning (pre- and post-learning templating runs, respectively). In the templating runs, images were presented individually in a completely random order to evaluate how their representations were changed by learning. The response to each image was estimated in every voxel using a GLM. The voxels from each individual's hippocampal subfield ROIs were extracted to form a pattern of activity for each image and subfield. We then calculated the pattern similarity between images in a pair using Pearson correlation, both before and after learning, and subtracted before-learning pattern similarity from after-learning pattern similarity to index the direction and amount of representational change. A separate representational change score was computed for each of the eight model similarity levels.

To test for the U-shaped curve predicted by the NMPH, we fit a theory-constrained cubic model to the series of representational change scores across model similarity levels (Fig. 4A). Specifically, the leading coefficient was forced to be positive to ensure a dip, followed by a positive inflection — the characteristic shape of the NMPH (Fig. 1 inset). The predictions of this theory-constrained cubic model were reliably associated with representational change in DG ($r$ = .134, 95% CI = [.007 .267], randomization $p$ = .022). The fit was not reliable in CA2/3 ($r$ = .082, 95% CI = [-.027 .191], $p$ = .13), CA1 ($r$ = .116, 95% CI = [-.001 .231], $p$ = .10), or the hippocampus as a whole ($r$ = .084, 95% CI = [-.018 .186], $p$ = .15).

We followed up on the observed effect in DG and determined that there was reliable differentiation at model similarity levels 5 ($\Delta r$ = -.093, 95% CI = [-.177 -.007], $p < .0001$) and 6 ($\Delta r$ = -.090, 95% CI = [-.179 -.004], $p$ = .016). This trough in the center of the U-shaped curve was also reliably lower than the peaks

preceding it at level 4 ($\Delta r$ = .129, 95% CI = [.019 .243], $p$ = .015) and following it at level 8 ($\Delta r$ = .150, 95% CI = [.025 .271], $p$ = .005). The curve showed a trend toward positive representational change, suggestive of integration, for model similarity level 8 ($\Delta r$ = .057, 95% CI = [-.034 .147], $p$ = .078).

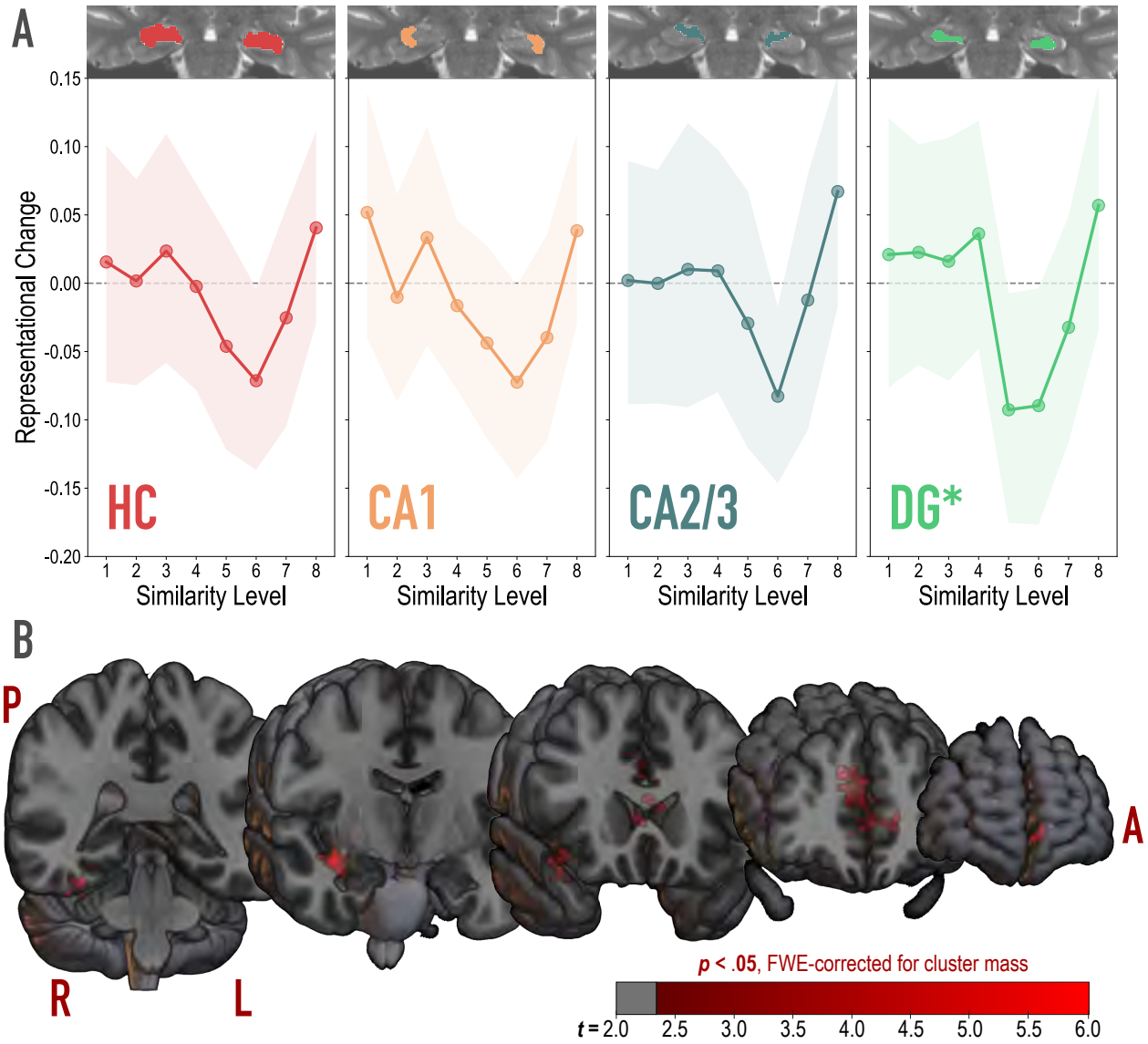


Figure 4: **Analysis of representational change predicted by the nonmonotonic plasticity hypothesis. (A)** Difference in correlation of voxel activity patterns between paired images after minus before learning at each model similarity level, in the whole hippocampus (HC) and in hippocampal subfields CA1, CA2/3 and DG. Inset image shows an individual subject mask for the ROI in question, overlaid on their T2-weighted anatomical image. The nonmonotonic plasticity hypothesis reliably predicted representational change in DG. Shaded area depicts bootstrap resampled 95% CIs. **(B)** Searchlight analysis. Brain images depict coronal slices viewed from an anterior vantage point. Clusters in red survived correction for family-wise error (FWE) at $p < .05$ using the null distribution of maximum cluster mass. L = left hemisphere, R = right hemisphere, A = anterior, P = posterior.

### 2.2.2 Whole-brain searchlight

To determine whether nonmonotonic learning effects were specific to the hippocampus, we ran an exploratory searchlight analysis in which we repeated the above cubic model-fitting analysis over the whole brain (Fig. 4B). This analysis revealed two reliable clusters ($p < .05$ corrected): right hippocampus extending into PHC and FG (832 voxels; peak $t$-value = 4.97; MNI coordinates of peak = 37.5, -7.5, -15.0; coordinates of center = 32.8, -18.7, -17.1) and anterior cingulate, extending into medial prefrontal cortex (1604 voxels; peak $t$-value = 5.43; coordinates of peak = -7.5, 28.5, 21.0; coordinates of center = -5.0, 28.5, 10.2).

# 3 Discussion

We set out to determine how learning shapes representations in the hippocampus and found that the degree of overlap in visual features determined the nature of representational change in DG. The pattern of results was U-shaped: with low or high overlap, object representations did not reliably change with respect to one another, whereas with moderate overlap, they pushed apart from one another following learning. This is consistent with the predictions of the NMPH (Ritvo et al., 2019) and related theories (e.g. Bienenstock, Cooper, & Munro, 1982). Although previous studies have reported evidence consistent with the NMPH (e.g., manipulations that boost coactivation of hippocampal representations lead to differentiation; Schlichting et al., 2015; Chanales et al., 2017; Favila et al., 2016; Kim et al., 2017), these studies generally compared only two or three conditions and their results can also be explained by competing hypotheses (e.g., a monotonic increase in differentiation with increasing shared activity). Crucially, the present study is the first to span coactivation continuously in order to reveal the full U-shape predicted by the NMPH, whereby differentiation emerges as coactivation grows from low to moderate and dissipates as coactivation continues from moderate to high.

To measure the impact of the degree of coactivation across a broad range of possible values, we developed a novel method of synthesizing experimental image pairs using DNN models. The intent of this approach was to precisely control the overlap among visual features at one or more layers of the model. In this case, we targeted higher layers of the model to indirectly control representational similarity in higher-order visual regions that provide input to the hippocampus. We found that the imposed visual feature relationships between images influenced human similarity judgments and were associated with parametric changes in neural similarity in higher-order visual cortex (i.e., LO, PHC). This is the first demonstration of the efficacy of stimulus synthesis in manipulating high-level representational similarity in targeted brain regions in humans. These results resonate with recent advances in stimulus synthesis designed to target individual neurons in primates (Bashivan, Kar,

& DiCarlo, 2019; Ponce et al., 2019). Although fMRI does not allow for targeting of individual neurons, our findings show that it is feasible to use this method to target distribute representations in different visual cortical regions.

Our approach of manipulating the overlap of visual inputs to the hippocampus, rather than manipulating hippocampal codes directly, was a practical one, based on the fact that we have much better computational models of visual coding than hippocampal coding. Numerous studies have shown that, while hippocampal neurons are indeed 'downstream' from visual cortex, they additionally encode complex information from multiple sensory modalities (Lavenex & Amaral, 2000), as well as information about reward (Wimmer & Shohamy, 2012), social relevance (Olson, Plotzker, & Ezzyat, 2007), context (Turk-Browne, Simon, & Sederberg, 2012), and time (Hsieh, Gruber, Jenkins, & Ranganath, 2014; Schapiro et al., 2012), to name a few. So, although we controlled visual inputs to the hippocampus, there were many additional non-visual inputs that were free to vary and could play a role in determining the overall relational structure of the representational space. Our work here demonstrates that controlling the visual features alone was sufficient to elicit non-monotonic learning effects, raising the possibility that controlling additional dimensions might yield greater differentiation (or integration). Future work could explore combining models of vision with hippocampal models and attempt to directly target hippocampal representations with image synthesis.

Importantly, our study allowed us to examine representational change in specific hippocampal subfields. We found that the differentiation 'dip' (creating the U shape) was most reliable in DG. This fits with prior studies that found differentiation in a combined CA2/3/DG ROI (Dimsdale-Zucker et al., 2018; Kim et al., 2017; Molitor et al., 2020), though note that the U-shaped pattern was trending but not significant in CA2/3 in our study. The clearer effects in DG may suggest that sparse coding (and the resulting low activation levels) is necessary to traverse the full spectrum of coactivation from low to moderate to high that can reveal nonmonotonic changes in representational similarity; regions with less sparsity (and higher baseline activation levels) may restrict coactivation to the moderate to high range, resulting in a bias toward integration and monotonic increases in representational similarity. Indeed, we had expected that CA1 might show integration effects due to its higher overall levels of activity (Barnes et al., 1990), consistent with prior studies emphasizing a role for CA1 in memory integration (Dimsdale-Zucker et al., 2018; Molitor et al., 2020; Schlichting, Zeithamova, & Preston, 2014; Duncan & Schlichting, 2018; Brunec et al., 2020). One speculative possibility is that the hippocampus is affected by feature overlap in earlier stages of visual cortex in addition to later stages (e.g., Huffman & Stark, 2017). Our paired stimuli were constructed to have high overlap at the top of the visual hierarchy but low overlap earlier on in the hierarchy; it is possible that allowing stimuli to have higher overlap throughout the visual hierarchy would lead to even greater coactivation in the hippocampus, resulting in integration.

Although our results are broadly consistent with prior findings that increasing the coactivation of memories can lead to differentiation (Schlichting et al., 2015; Chanales et al., 2017; Favila et al., 2016; Kim et al., 2017), they are notably inconsistent with results from Schapiro et al. (2012), who reported memory integration for arbitrarily paired images as a result of temporal co-occurrence; pairs in our study with comparable levels of visual similarity (roughly model similarity level 3) showed no evidence of integration. This difference between studies may relate to the fact that the visual sequences in Schapiro et al. (2012) contained a mix of strong and weak transition probabilities, whereas we used strong transition probabilities exclusively; moreover, our study had a higher baseline of visual feature overlap among pairs. Contextual and task-related factors (Brunec et al., 2020), as well as the history of recent activation (Bear, 2003), can bias the hippocampus toward integration or differentiation, similar to the remapping based on task context that occurs in rodent hippocampus (Anderson & Jeffery, 2003; Colgin, Moser, & Moser, 2008; McKenzie et al., 2014). Speculatively, the overall higher degree of competition in our task — from stronger transition probabilities and higher baseline similarity — may have biased the hippocampus toward differentiation (Ritvo et al., 2019).

### 3.1 Conclusion

Overall, these results highlight the complexity of learning rules in the hippocampus, showing that moderate levels of visual feature similarity lead to differentiation following a statistical learning paradigm, but higher and lower levels of visual similarity do not. From a theoretical perspective, these results provide the strongest evidence to date for the NMPH account of hippocampal plasticity. From a methodological perspective, our results provide a proof-of-concept demonstration of how image synthesis, applied to neural network models of specific brain regions, can be used to test how representations in these regions shape learning. As neural network models continue to improve, we expect that this kind of model-based image synthesis will become an increasingly useful tool for studying neuroplasticity.

## 4 Methods

### 4.1 Participants

For the fMRI study, we recruited 42 healthy young adults participants (18-35 years old, 25 females) with self-reported normal (or corrected to normal) visual acuity and good color vision. All participants provided informed consent to a protocol approved by the Yale IRB and were compensated for their time ($20 per hour). Five participants did not complete the task because of technical errors and/or time constraints, though their data

could still be used for the visual templating analyses, as this only required the initial pre-learning templating run. One additional participant's data quality precluded segmentation of hippocampal subfields. As such, our final sample for the learning task was 36 participants, with a total of 41 participants available for the visual templating analyses.

For the behavioral validation study, we recruited 30 naive participants through Amazon Mechanical Turk (mTurk). All participants provided informed consent to a protocol approved by the Yale IRB, and were compensated for their time ($6 per hour).

## 4.2 Stimulus synthesis

Image pairs were generated via a gradient descent optimization using features extracted from *GoogLeNet* (a version of *Inception*; Szegedy et al., 2015), a deep neural network (DNN) architecture. This particular instantiation had been pretrained on ImageNet (Deng et al., 2009), which contains over one million images of common objects. Accordingly, the learned features reflect information about the real-world features of naturalistic objects. Our approach drew heavily from *Deepdream* (Mordvintsev, Olah, & Tyka, 2015; *DeepDreaming with TensorFlow*, n.d.), an approach used to visualize the learned features of pretrained neural networks. Deepdream's optimization uses gradient ascent to iteratively update input pixels such that activity in a given unit, layer, or collection of layers is maximized. Different from Deepdream, the core of our approach was controlling the correlation between the features of two images at a given layer *j*, as a means of controlling visual overlap at a targeted level of complexity. We prioritized image optimization over features in a subset of network layers: the early convolutional layers and the output layers of later inception modules (i.e. 12 total layers).

Because we were interested in targeting higher-order visual representations (e.g., in LO), our intention was to produce pairs of images whose higher-layer (top four layers) features were correlated with one another at a specified value, ranging from 0 (not at all similar), to 1 (almost exactly the same). As such, we produced pairs of images that fell along an axis between two 'endpoints', where the endpoints were pairs of images designed to have a correlation of 0. Each subsequent pair of increasing similarity can be thought of as sampling two new points by stepping inward along the axis from each side. Because it was our aim to have some specificity in the level of representation we were targeting, we sought to fix the feature correlations between all of the pairs in the lower and middle layers (i.e., the bottom 8 layers) at 0.25. Altogether, our stimulus synthesis procedure was composed of three phases, described below: (1) endpoint channel selection, (2) image initialization, and (3) correlation tuning (Supplementary Fig. 1).

The purpose of the endpoint channel selection phase was to select higher-layer feature channels that, when optimized, were maximally different from one another. To do this, we generated an optimized image (*DeepDreaming with TensorFlow*, n.d.) that maximally expressed each of the 128 feature channels in layer mixed 4E (yielding 128 optimized images). We fed these images back through the network, and extracted the pattern of activation in the top four selected layers for each image. We then computed Pearson correlations among the extracted features and selected the 16 optimized channel images whose activation patterns were least inter-correlated with one another. These 16 channels were formed into 8 pairs, which became the endpoint channels for a spectrum of image pairs (Supplementary Fig. 1A).

During image initialization, the endpoint channels served as the starting point for generating sets of image pairs with linearly increasing visual similarity. For every pair of endpoint channels, eight image pairs were synthesized, varying in intended higher-layer feature correlation from 0 to 1. These are also referred to as model similarity levels 1 through 8. Every AB image pair began with two randomly generated visual noise arrays, and an 'endpoint' was assigned to each – for example, channel 17 to image A and channel 85 to image B. If optimizing image A, channel 17 was always maximally optimized, while the weighting for the optimization of channel 85 depended on the intended correlation. For example, if the intended correlation was 0.14, channel 85 was weighted at 0.14. If optimizing image B for a correlation of 0.14, channel 85 was maximally optimized and channel 17's optimization was weighted at 0.14. These two weighted channel optimizations were added together, and served as the cost function for gradient descent in this phase (Supplementary Fig. 1B). On each iteration, the gradient of the cost function was computed with respect to the input pixels. In this way, the pixels were updated at each iteration, working toward this weighted image initialization objective. Eight different AB image pairs were optimized for each of the eight pairs of endpoints, for a total of 64 image pairs (128 images in total). After 200 iterations, the resulting images were fed forward into the correlation tuning phase.

The correlation tuning phase more directly and precisely targeted correlation values. On each iteration, the pattern of activations to image A and image B were extracted from every layer of interest, and a correlation was computed between image A's pattern and image B's pattern (Fig. 2A, Supplementary Fig. 1C). The cost function for this phase was the squared difference between the current iteration's correlation, and the intended correlation. Our aim was to equate the image pairs in terms of their similarity at lower layers, so the intended correlation for the first eight layers was always 0.25. For the highest four layers, the intended correlation varied from 0 to 1. Similar to the endpoint initialization phase, the gradient of the new cost function was computed with respect to the input pixels, and so the images iteratively stepped toward exhibiting the exact feature correlation properties specified. After 200 iterations, we were left with eight pairs of images, whose

correlations (i.e. similarity in higher-order visual features) varied linearly from 0 to 1 (Fig. 2B, Supplementary Fig. 1C, right side).

Feature channel optimization tends to favor the expression of high frequency edges. To circumvent this, as in Deepdream (Mordvintsev et al., 2015), we utilized Laplacian pyramid regularization to smooth the image and allow low-frequency features and richer color to be expressed. Also, we wanted to ensure that the final feature correlations were not driven by eccentric pixels in the image, and that the feature correlation was reasonably well represented in various parts of the images and at various scales. To accomplish this, we performed the entire 400 iteration optimization procedure three times, magnifying the image by 40% for each volley. We also computed the gradient over a subset of the input pixels, which were selected using a moving window slightly smaller than the image, in the center of the image. As a result of this procedure, pixels toward the outside of the image were not updated as often. We then cropped the images such that pixels that had been iterated over very few times were removed. Although this was intended to avoid feature correlations driven by the periphery, it had the added benefit of giving the images an irregular ragged edge, rather than a sharp square frame which can make the images appear more homogenous. However, this cropping procedure did remove some pixels that were contributing in part to the assigned correlation value. For this reason, the final inter-image correlations are closely related to but do not *exactly* match the assigned values.

## 4.3 Model validation

We produced a large set of image pairs using the stimulus synthesis approach detailed above. For each of the eight sets of image endpoints, we generated a pair of images at each of eight model similarity levels, yielding a total of 128 ($8 \times 8 \times 2$) images. Each image was then fed back through the network, resulting in a pattern of activity across units in relevant layers that was correlated with the pattern from its pairmate. We also fed these images through a second commonly used DNN, VGG19 (Simonyan & Zisserman, 2014), and applied the same procedure. This was done to verify that the feature overlaps we set out to establish were not specific to one model architecture. In the selected layers of both networks, we computed second-order correlations between the intended and actual image-pair correlations. In the network used to synthesize the image pairs, we averaged these second-order correlations in each of the top four target layers, to provide an estimate of the extent to which the synthesis procedure was effective. Despite varying in similarity in the target layers, the differences in similarity in the lower and middle layers (intended to be uniformly $r = 0.25$) should be minimal. To confirm this, we calculated differences between the actual correlations across image pairs, as well as the standard deviation of these differences (Supplementary Fig. 2). Second-order correlations and standard deviations were also computed for each layer in VGG19, the alternate architecture (Supplementary Fig. 2).

## 4.4 Behavioral validation

Online participants consented to participate and then were provided with task instructions (see next section). The most critical instruction was as follows: "It will be your job to drag and drop those images into the arena, and arrange them so that the more visually similar items are placed closer together, and the more dissimilar items are placed farther apart." After confirming that they understood, participants proceeded to the task. At the start of each trial, they clicked a button labeled “Start trial”. They were then shown a black screen with a white outlined circle that defined the arena. The circle was surrounded by either 24 (20% of trials) or 26 images (80%), pseudo-randomly selected from the broader set of 128 images (8 pairs from each of 8 sets of endpoints; 64 pairs). Sets were selected such that there were no duplicates, the two images from a given target pair were always presented together, and every image was presented at least twice. Trials were self-paced, but could not last more than 5 minutes each. Warnings were provided in orange text and then red text when there was 60 and 30 seconds remaining, respectively. There was no minimum time, except that a trial could not be completed unless every image had been placed. This timing structure ensured that we would get at least 10 trials per subject in no longer than 50 minutes. On the right side of the arena, there was a button labeled “Click here when finished”. If this button was clicked prior to placing all of the images, a warning appeared, “You have not placed all of the images”. If all images had been placed, additional buttons appeared, giving the participant the option to confirm completion (“Are you sure? Click here to confirm.”) or return to sorting (“...or here to go back”). We used the coordinates of the final placement location of each image to compute pairwise Euclidean distances between images (Fig. 2D). We averaged across trials within participant to get one distance metric for each image pair for each participant. We then computed the Pearson correlation between the model similarity level (1 through 8) defined in the stimulus synthesis procedure and the distance between the images based on each participant’s placements. We also averaged Euclidian distances across participants for each of the 64 image pairs, and then computed a Pearson correlation between model similarity level and these group-averaged distances (Fig. 2E).

## 4.5 Arrangement task instructions

**Instructions (1/5):**

Thank you for signing up!

In this experiment, you will be using the mouse to click and drag images around the screen. At the beginning, you will click the ‘Start trial’ button in the top left corner of your browser window, and a set of images will appear, surrounding the border of a white circular arena. It will be your job to drag and drop those

images into the arena, and arrange them so that the more visually similar items are placed closer together, and the more dissimilar items are placed farther apart.

**Instructions (2/5):**

You can move each image as many times as you'd like to make sure that the arrangement corresponds to the visual similarity. The images you will be viewing will be abstract in nature, and will not be animals, but the following example should clarify the instructions:

If you had moved images of a wolf, a coyote, and a husky into the arena, you might think that they look quite similar to one another, and therefore place them close together. However, if the next image you pulled in was a second image of a husky, you may need to adjust your previous placements so that the two huskies are closer together than a husky and a wolf.

**Instructions (3/5):**

You may find that some clusters of similar items immediately pop out to you. Once you have a set of several of these somewhat similar items grouped together, you might notice more fine-grained differences between them. Please make sure that you take the time to tinker and fine-tune in these situations. The differences within these clusters is just as important.

If two images are exactly the same, they should be placed on top of one another, and if they are ALMOST exactly the same, feel free to overlap one image with the other. By that same logic, if images are totally dissimilar, place them quite far apart, as far as on opposite sides of the arena.

Depending on your screen resolution and zoom settings, the entire circle may not initially be in your field of view. This is okay. Feel free to scroll around and navigate the entire space as you arrange the images. However, it is important that the individual images are large enough that you can make out their details.

**Instructions (4/5):**

There will be multiple trials to complete, each with its own set of images. You will have a maximum of 5 minutes to complete each trial. When there is one minute remaining, an orange message will appear in the center of the circle to inform you of this. You will receive another message, this time in red, when there are 30 seconds remaining. When time runs out, your arrangement will be saved, and the next trial will be prepared. If you are satisfied with your arrangement before 3 minutes has elapsed, you can submit it using the 'Click here when finished' button. This will bring up two buttons; one to confirm, and one to go back to sorting. You will not be able to submit your arrangement unless you have placed every image.

When the trial ends, whether it was because you submitted your response, or because time ran out, the screen will be reset, revealing a new set of images and empty arena.

The task will take just under an hour to complete, no matter how quickly or how slowly you complete each trial, so there is no benefit to rushing.

**Instructions (5/5):**

We would like to thank you for taking the time to participate in our research. The information that you provide during this task is very valuable to us, and will be extremely helpful in developing our research. With this in mind, we ask that you really pay attention to the details of the images, and complete each trial and arrangement carefully and conscientiously.

***[printed in red text]*** We would also like to remind you that because there is a set time limit, not a set number of trials, there is no benefit to rushing through the trials. In general, we have found that it tends to take 3.5 minutes at minimum to complete each trial accurately.

## 4.6 fMRI design

Each functional task run lasted 304.5 seconds and consisted of viewing a series of synthesized abstract images, one at a time. Images were presented for 1 s each. The first image onset occurred after 6 s, and each subsequent onset was presented after an ISI of 1, 3, or 5s (40:40:20 ratio). There were eight pairs, meaning that there were 16 unique images. The order of image presentations was pseudo-randomly assigned in one of two ways. In the first and last (pre- and post-learning) templating runs, the 16 images were presented in a random order for the first 16 trials. For the next 16 trials, the 16 images were presented in a different random order, with the constraint that the same image not be presented twice in a row. This same procedure was repeated until there were 80 total trials (five presentations of each of image). In the six intervening statistical learning runs, image pairs were always presented intact and in the same A to B order. In these runs, the eight pairs were presented in a random order for the first 16 trials. For the next 16 trials, the eight pairs were presented in a different order, with the constraint that the same pair could not be presented twice in a row, and so on. Critically, the images appeared continuously without segmentation cues between pairs, such that participants had to learn transition probabilities in the sequence (i.e., for pair AB, the higher probability of A transitioning to B than B transitioning to any number of other images). One out of every 10 trials was randomly assigned to have a small, partially transparent grey patch overlaid on the image. The participants performed a cover task of pressing a button on a handheld button box when they saw the grey square. This task was designed to encourage participants to maintain attention on the images, but was completely orthogonal to the pair structure.

## 4.7 Data acquisition

Data were acquired using a 3T Siemens Prisma scanner with a 64-channel head coil at the Yale Magnetic Resonance Research Center. We collected eight functional runs with a with a multiband echo-planar imaging (EPI) sequence (TR=1500 ms; TE=32.6 ms; voxel size=1.5mm isotropic; FA=71°; multiband factor=6), yielding 90 axial slices. Each run contained 203 volumes. For field map correction, two spin-echo field map volumes (TR= 8000; TE=66) were acquired in opposite phase encoding directions. These otherwise matched the parameters of our functional acquisitions. We also collected a T1-weighted magnetization prepared rapid gradient echo (MPRAGE) image (TR=2300 ms; TE=2.27 ms; voxel size=1 mm isotropic; FA=8°; 192 sagittal slices; GRAPPA acceleration factor=3), and a T2-weighted turbo spin-echo (TSE) image (TR=11390 ms; TE=90 ms; voxel size=$0.44 \times 0.44 \times 1.5$ mm; FA=150°; 54 coronal slices; perpendicular to the long axis of the hippocampus; distance factor=20%).

## 4.8 fMRI preprocessing

For each functional run, preprocessing was performed using the FEAT tool in FSL (Woolrich, Ripley, Brady, & Smith, 2001). Data were brain-extracted, corrected for slice timing, high-pass filtered (100 s cutoff), aligned to the middle functional volume of the run using MCFLIRT (Jenkinson, Bannister, Brady, & Smith, 2002), and spatially smoothed (3 mm). FSL's topup tool (Smith et al., 2004), in conjunction with the two field maps, was used to estimate susceptibility-induced distortions. The output was converted to radians and used to perform fieldmap correction in FEAT. The functional runs were also aligned to both the participants' T1-weighted anatomical image using boundary-based registration, and to MNI standard space with 12 degrees of freedom, using FLIRT (Jenkinson & Smith, 2001). Analyses within a single run were conducted in native space. For comparisons across participants, analyses were conducted in standard space.

## 4.9 Defining regions of interest

For each participant, their T1- and T2-weighted anatomical images were submitted to the automatic segmentation of hippocampal subfields (ASHS) software package (Yushkevich et al., 2015), to derive participant-specific medial temporal lobe regions of interest. We used an atlas containing 51 manual segmentations of hippocampal subfields (Aly & Turk-Browne, 2015, 2016). The resulting automated segmentations were used to create masks for the CA1, CA2/3, and dentate gyrus (DG) subfields. For visual ROIs, freesurfer (`http://surfer.nmr.mgh.harvard.edu/`) was used to create masks for V1, V2, lateral occipital (LO) cortex, fusiform gyrus (FG), parahippocampal cortex (PHC), and inferior temporal (IT) cortex, for each participant.

## 4.10 General linear model

For the pre- and post-learning templating runs, a regressor was developed for each of the 16 unique synthesized images. This was done by placing a delta function at each image onset and convolving this time course with the double-gamma hemodynamic response function. We then used these 16 regressors to fit a GLM to the time course of BOLD activity using FSL's FILM tool, correcting for local autocorrelation (Woolrich et al., 2001). This yielded parameter estimates for each of the 16 images, which were used for subsequent analyses.

## 4.11 Stimulus synthesis validation analyses

The pre-learning templating run was analyzed to derive an estimate of the baseline representational similarity among the eight target pairs before any learning had taken place. For ROI analyses, the 16 parameter estimates output by the GLM (one per stimulus) were extracted for each voxel in a given ROI and vectorized to obtain the multivoxel pattern of activity for each stimulus. We then computed the Pearson correlation between the two vectors corresponding to the pairmates in each of the eight target image pairs. This yielded eight representational similarity values, one for each image pair. As established through model-based synthesis, each of the eight image pairs also had a corresponding model similarity level. We computed and Fisher transformed the second-order correlation of neural and model similarity across levels. In other words, this analysis tested whether the pattern of similarity built in to the image pairs through the DNN model corresponded to the representational similarity in a given brain region. We constructed 95% confidence intervals (CIs) for estimates of model-brain correspondence for each ROI by bootstrap resampling of participants 50,000 times. As an additional control, we compared the true group average correlation value to a noise distribution, wherein A and B images were paired randomly 50,000 times, obliterating any systematic similarity relationships among them. We did not, however, constrain the random pairing to exclude the true image pairings. This means that the control is especially conservative because some of the resulting shuffled pairs contained the true image pairings. This analysis was used to ensure that the true effect would not be likely to occur due to random noise. In addition to ROI analyses, we also conducted exploratory searchlight analyses over the whole brain. This involved repeating the same representational similarity analyses but over patterns defined from 125-voxel searchlights (radius = 2) centered on every brain voxel. The resulting whole-brain statistical maps were then Fisher transformed, concatenated, and tested for reliability at the group level using FSL's randomise (Winkler, Ridgway, Webster, Smith, & Nichols, 2014). We used a cluster-forming threshold of $t = 2.33$ ($p < .01$, one-tailed) in cluster mass, and corrected for multiple comparisons using the null distribution of maximum cluster mass. Clusters that survived at ($p < .05$) were retained.

### 4.12 Representational change analyses

To test the NMPH, we measured whether learning-related changes in pairmate representational similarity (i.e., changes from pre- to post-statistical-learning) followed a U-shaped, cubic function of model similarity. This involved measuring the degree to which the image pairs at each level of similarity showed integration (i.e., increased representational similarity) or differentiation (i.e., decreased representational similarity). For ROI analyses, the 16 parameter estimates output by the GLM were extracted for each voxel in a given ROI and vectorized. This procedure was performed for both pre- and post-learning templating runs. In each run, Pearson correlations were computed between the vectors for each of the target image pairs, yielding eight representational similarity values. To measure learning-related changes in representational similarity, pre-learning values were subtracted from post-learning values. A positive value on this metric signifies integration, whereas a negative value signifies differentiation. Our predictions were not about any one model similarity level, but rather about a specific nonmonotonic relationship between model similarity and representational change. This hypothesis can be quantified using a cubic model, with the leading coefficient constrained to be positive. To test the efficacy of this model in each MTL ROI, we computed a cross-validated estimate of how well this model predicted the true data. Specifically, we fit the constrained cubic model to the data from all but one held-out participant. We then used the model to predict the held-out participant's data, and computed a correlation between the predicted and actual data. This procedure was repeated such that every participant was held out once, and the resulting correlations were averaged into an estimate of the fit for that ROI. We then constructed 95% confidence intervals (CIs) for estimates of the model fit for each ROI by bootstrap resampling participants 50,000 times. To produce a noise distribution, we randomly re-paired A and B images 50,000 times and repeated the above analysis with our entire sample for each repairing. We compared the true group average model fit statistic to this noise distribution for each ROI. Lastly, we conducted an exploratory searchlight analysis, repeating the analysis above in 125-voxel searchlights (radius = 2) centered on every brain voxel. As in the model similarity searchlight, all subjects' outputs were then Fisher transformed, concatenated, submitted to randomise, thresholded, and corrected for maximum cluster mass.

### 4.13 Quantification and statistical analysis

No statistical methods were used to predetermine sample size, but we aimed to collect at least 36 participants for the primary learning analysis. For all analyses, we used bootstrap resampling methods to analyze our results non-parametrically. Details of these analyses, as well as exact results of statistical tests, 95% confidence

intervals, and p values with respect to noise distributions, are reported alongside each analysis in our Results. Statistical significance was set at $p < 0.05$ unless otherwise specified.

# 5 Resource Availability

Further information and requests for resources or code should be directed to and will be fulfilled by the Lead Contact, Jeffrey Wammes (jeffrey.wammes@queensu.ca). Synthesize image stimuli, fMRI data and analysis code will be released upon publication.

# 6 Supplemental Information

Supplementary information can be found here.

# 7 Acknowledgments

We thank L. Rait for help with recruitment, scheduling, and data collection. This work was supported by NSERC PDF and SSHRC Banting PDF (J.D.W.), NIH R01 MH069456 (K.A.N. and N.B.T.-B.), and the Canadian Institute for Advanced Research (N.B.T.-B.).

# 8 Author Contributions

All authors designed research and planned analyses. J.D.W. developed stimulus synthesis method, collected and analyzed data, and wrote the initial draft of the manuscript. All authors contributed to editing of the manuscript.

# 9 Competing Interests

Authors declare no competing interests.

# References


Aly, M., & Turk-Browne, N. B. (2015). Attention stabilizes representations in the human hippocampus. *Cerebral Cortex*, *26*(2), 783–796.

Aly, M., & Turk-Browne, N. B. (2016). Attention promotes episodic encoding by stabilizing hippocampal representations. *Proceedings of the National Academy of Sciences*, *113*(4), E420–E429.

Anderson, M. I., & Jeffery, K. J. (2003). Heterogeneous modulation of place cell firing by changes in context. *Journal of Neuroscience*, *23*(26), 8827–8835.

Barnes, C. A., McNaughton, B. L., Mizumori, S. J., Leonard, B. W., & Lin, L.-H. (1990). Chapter comparison of spatial and temporal characteristics of neuronal activity in sequential stages of hippocampal processing. In *Progress in brain research* (Vol. 83, pp. 287–300). Elsevier.

Bashivan, P., Kar, K., & DiCarlo, J. J. (2019). Neural population control via deep image synthesis. *Science*, *364*(6439), eaav9436.

Bear, M. F. (2003). Bidirectional synaptic plasticity: from theory to reality. *Philosophical Transactions of the Royal Society of London. Series B: Biological Sciences*, *358*(1432), 649–655.

Berron, D., Schütze, H., Maass, A., Cardenas-Blanco, A., Kuijf, H. J., Kumaran, D., & Düzel, E. (2016). Strong evidence for pattern separation in human dentate gyrus. *Journal of Neuroscience*, *36*(29), 7569–7579.

Bienenstock, E. L., Cooper, L. N., & Munro, P. W. (1982). Theory for the development of neuron selectivity: orientation specificity and binocular interaction in visual cortex. *Journal of Neuroscience*, *2*(1), 32–48.

Brunec, I. K., Robin, J., Olsen, R. K., Moscovitch, M., & Barense, M. D. (2020). Integration and differentiation of hippocampal memory traces. *Neuroscience & Biobehavioral Reviews*.

Buonomano, D. V., & Merzenich, M. M. (1998). Cortical plasticity: from synapses to maps. *Annual Review of Neuroscience*, *21*(1), 149–186.

Cadieu, C. F., Hong, H., Yamins, D. L., Pinto, N., Ardila, D., Solomon, E. A., ... DiCarlo, J. J. (2014). Deep neural networks rival the representation of primate it cortex for core visual object recognition. *PLoS Computational Biology*, *10*(12), e1003963.

Caporale, N., & Dan, Y. (2008). Spike timing–dependent plasticity: a hebbian learning rule. *Annual Review Neuroscience*, *31*, 25–46.

Chanales, A. J., Oza, A., Favila, S. E., & Kuhl, B. A. (2017). Overlap among spatial memories triggers repulsion of hippocampal representations. *Current Biology*, *27*(15), 2307–2317.

Cichy, R. M., Khosla, A., Pantazis, D., Torralba, A., & Oliva, A. (2016). Comparison of deep neural networks to spatio-temporal cortical dynamics of human visual object recognition reveals hierarchical correspondence. *Scientific Reports*, *6*, 27755.

Colgin, L. L., Moser, E. I., & Moser, M.-B. (2008). Understanding memory through hippocampal remapping. *Trends in Neurosciences*, *31*(9), 469–477.

Collin, S. H., Milivojevic, B., & Doeller, C. F. (2015). Memory hierarchies map onto the hippocampal long axis in humans. *Nature Neuroscience*, *18*(11), 1562.

de Beeck, H. P. O., Torfs, K., & Wagemans, J. (2008). Perceived shape similarity among unfamiliar objects and the organization of the human object vision pathway. *Journal of Neuroscience*, *28*(40), 10111–10123.

*Deepdreaming with tensorflow.* (n.d.). `https://github.com/tensorflow/tensorflow/blob/master/tensorflow/examples/tutorials/deepdream/deepdream.ipynb`.

Deng, J., Dong, W., Socher, R., Li, L.-J., Li, K., & Fei-Fei, L. (2009). Imagenet: A large-scale hierarchical image database. In *2009 ieee conference on computer vision and pattern recognition* (pp. 248–255).

Detre, G. J., Natarajan, A., Gershman, S. J., & Norman, K. A. (2013). Moderate levels of activation lead to forgetting in the think/no-think paradigm. *Neuropsychologia*, *51*(12), 2371–2388.

Deuker, L., Bellmund, J. L., Schröder, T. N., & Doeller, C. F. (2016). An event map of memory space in the hippocampus. *Elife*, *5*, e16534.

Dimsdale-Zucker, H. R., Ritchey, M., Ekstrom, A. D., Yonelinas, A. P., & Ranganath, C. (2018). Ca1 and ca3 differentially support spontaneous retrieval of episodic contexts within human hippocampal subfields. *Nature Communications*, *9*(1), 1–8.

Duncan, K. D., & Schlichting, M. L. (2018). Hippocampal representations as a function of time, subregion, and brain state. *Neurobiology of Learning and Memory*, *153*, 40–56.

Eickenberg, M., Gramfort, A., Varoquaux, G., & Thirion, B. (2017). Seeing it all: Convolutional network layers map the function of the human visual system. *NeuroImage*, *152*, 184–194.

Favila, S. E., Chanales, A. J., & Kuhl, B. A. (2016). Experience-dependent hippocampal pattern differentiation prevents interference during subsequent learning. *Nature Communications*, *7*(1), 1–10.

Feldman, D. E. (2009). Synaptic mechanisms for plasticity in neocortex. *Annual Review of Neuroscience*, *32*, 33–55.

Güçlü, U., & van Gerven, M. A. (2015). Deep neural networks reveal a gradient in the complexity of neural representations across the ventral stream. *Journal of Neuroscience*, *35*(27), 10005–10014.

Hebb, D. O. (1949). *The organization of behavior: a neuropsychological theory*. J. Wiley; Chapman & Hall.

Hsieh, L.-T., Gruber, M. J., Jenkins, L. J., & Ranganath, C. (2014). Hippocampal activity patterns carry information about objects in temporal context. *Neuron*, *81*(5), 1165–1178.

Huffman, D. J., & Stark, C. E. (2017). The influence of low-level stimulus features on the representation of contexts, items, and their mnemonic associations. *NeuroImage*, *155*, 513–529.

Hulbert, J., & Norman, K. A. (2015). Neural differentiation tracks improved recall of competing memories following interleaved study and retrieval practice. *Cerebral Cortex*, *25*(10), 3994–4008.

Jenkinson, M., Bannister, P., Brady, M., & Smith, S. (2002). Improved optimization for the robust and accurate linear registration and motion correction of brain images. *NeuroImage*, *17*(2), 825–841.

Jenkinson, M., & Smith, S. (2001). A global optimisation method for robust affine registration of brain images. *Medical Image Analysis*, *5*(2), 143–156.

Johnson, J. D., McDuff, S. G., Rugg, M. D., & Norman, K. A. (2009). Recollection, familiarity, and cortical reinstatement: a multivoxel pattern analysis. *Neuron*, *63*(5), 697–708.

Jozwik, K., Kriegeskorte, N., Cichy, R. M., & Mur, M. (2019). Deep convolutional neural networks, features, and categories perform similarly at explaining primate high-level visual representations.

Khaligh-Razavi, S.-M., & Kriegeskorte, N. (2014). Deep supervised, but not unsupervised, models may explain it cortical representation. *PLoS Computational Biology*, *10*(11).

Kim, G., Norman, K. A., & Turk-Browne, N. B. (2017). Neural differentiation of incorrectly predicted memories. *Journal of Neuroscience*, *37*(8), 2022–2031.

Kriegeskorte, N. (2009). Relating population-code representations between man, monkey, and computational models. *Frontiers in Neuroscience*, *3*, 35.

Kriegeskorte, N. (2015). Deep neural networks: a new framework for modeling biological vision and brain information processing. *Annual Review of Vision Science*, *1*, 417–446.

Kriegeskorte, N., & Mur, M. (2012). Inverse mds: Inferring dissimilarity structure from multiple item arrangements. *Frontiers in Psychology*, *3*, 245.

Kubilius, J., Bracci, S., & de Beeck, H. P. O. (2016). Deep neural networks as a computational model for human shape sensitivity. *PLoS Computational Biology*, *12*(4), e1004896.

Lavenex, P., & Amaral, D. G. (2000). Hippocampal-neocortical interaction: A hierarchy of associativity. *Hippocampus*, *10*(4), 420–430.

Luo, W., Li, Y., Urtasun, R., & Zemel, R. (2016). Understanding the effective receptive field in deep convolutional neural networks. In *Advances in neural information processing systems* (pp. 4898–4906).

McKenzie, S., Frank, A. J., Kinsky, N. R., Porter, B., Rivière, P. D., & Eichenbaum, H. (2014). Hippocampal representation of related and opposing memories develop within distinct, hierarchically organized neural schemas. *Neuron*, *83*(1), 202–215.

Milivojevic, B., Vicente-Grabovetsky, A., & Doeller, C. F. (2015). Insight reconfigures hippocampal-prefrontal memories. *Current Biology*, *25*(7), 821–830.

Molitor, R. J., Sherrill, K. R., Morton, N. W., Miller, A. A., & Preston, A. R. (2020). Memory reactivation during learning simultaneously promotes dentate gyrus/ca2, 3 pattern differentiation and ca1 memory integration. *Journal of Neuroscience*.

Mordvintsev, A., Olah, C., & Tyka, M. (2015). Deepdream-a code example for visualizing neural networks. *Google Research*, *2*(5).

Newman, E. L., & Norman, K. A. (2010).

*Cerebral Cortex*, *20*(11), 2760–2770.

Nguyen, A., Yosinski, J., & Clune, J. (2015). Deep neural networks are easily fooled: High confidence predictions for unrecognizable images. In *Proceedings of the ieee conference on computer vision and pattern recognition* (pp. 427–436).

Olson, I. R., Plotzker, A., & Ezzyat, Y. (2007). The enigmatic temporal pole: a review of findings on social and emotional processing. *Brain*, *130*(7), 1718–1731.

Polyn, S. M., Natu, V. S., Cohen, J. D., & Norman, K. A. (2005). Category-specific cortical activity precedes retrieval during memory search. *Science*, *310*(5756), 1963–1966.

Ponce, C. R., Xiao, W., Schade, P. F., Hartmann, T. S., Kreiman, G., & Livingstone, M. S. (2019). Evolving images for visual neurons using a deep generative network reveals coding principles and neuronal preferences. *Cell*, *177*(4), 999–1009.

Ritvo, V. J., Turk-Browne, N. B., & Norman, K. A. (2019). Nonmonotonic plasticity: How memory retrieval drives learning. *Trends in Cognitive Sciences*.

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

Schlichting, M. L., Mumford, J. A., & Preston, A. R. (2015). Learning-related representational changes reveal dissociable integration and separation signatures in the hippocampus and prefrontal cortex. *Nature Communications*, *6*(1), 1–10.

Schlichting, M. L., Zeithamova, D., & Preston, A. R. (2014). Ca1 subfield contributions to memory integration and inference. *Hippocampus*, *24*(10), 1248–1260.

Simonyan, K., & Zisserman, A. (2014). Very deep convolutional networks for large-scale image recognition. *arXiv*.

Smith, S. M., Jenkinson, M., Woolrich, M. W., Beckmann, C. F., Behrens, T. E., Johansen-Berg, H., ... others (2004). Advances in functional and structural mr image analysis and implementation as fsl. *NeuroImage*, *23*, S208–S219.

Szegedy, C., Liu, W., Jia, Y., Sermanet, P., Reed, S., Anguelov, D., ... Rabinovich, A. (2015). Going deeper with convolutions. In *Proceedings of the ieee conference on computer vision and pattern recognition* (pp. 1–9).

Tompary, A., & Davachi, L. (2017). Consolidation promotes the emergence of representational overlap in the hippocampus and medial prefrontal cortex. *Neuron*, *96*(1), 228–241.

Turk-Browne, N. B., Simon, M. G., & Sederberg, P. B. (2012). Scene representations in parahippocampal cortex depend on temporal context. *Journal of Neuroscience*, *32*(21), 7202–7207.

Wimber, M., Alink, A., Charest, I., Kriegeskorte, N., & Anderson, M. C. (2015). Retrieval induces adaptive forgetting of competing memories via cortical pattern suppression. *Nature neuroscience*, *18*(4), 582–589.

Wimmer, G. E., & Shohamy, D. (2012). Preference by association: how memory mechanisms in the hippocampus bias decisions. *Science*, *338*(6104), 270–273.

Winkler, A. M., Ridgway, G. R., Webster, M. A., Smith, S. M., & Nichols, T. E. (2014). Permutation inference for the general linear model. *NeuroImage*, *92*, 381–397.

Woolrich, M. W., Ripley, B. D., Brady, M., & Smith, S. M. (2001). Temporal autocorrelation in univariate linear modeling of fmri data. *NeuroImage*, *14*(6), 1370–1386.

Yamins, D. L., Hong, H., Cadieu, C. F., Solomon, E. A., Seibert, D., & DiCarlo, J. J. (2014). Performance-optimized hierarchical models predict neural responses in higher visual cortex. *Proceedings of the National Academy of Sciences*, *111*(23), 8619–8624.

Yushkevich, P. A., Pluta, J. B., Wang, H., Xie, L., Ding, S.-L., Gertje, E. C., ... Wolk, D. A. (2015). Automated volumetry and regional thickness analysis of hippocampal subfields and medial temporal cortical structures in mild cognitive impairment. *Human Brain Mapping*, *36*(1), 258–287.

Zeiler, M. D., & Fergus, R. (2014). Visualizing and understanding convolutional networks. In *European conference on computer vision* (pp. 818–833).

# **Supplementary Information:** Increasing stimulus similarity drives nonmonotonic representational change in hippocampus

*Jeffrey D. Wammes[1,2], Kenneth A. Norman[3,4] and Nicholas B. Turk-Browne[1]

[1]Department of Psychology, Yale University, New Haven, CT 06520

[2]Department of Psychology, Queen's University, Kingston, ON K7L 3N6

[3]Department of Psychology, Princeton University, Princeton, NJ 08544

[4]Princeton Neuroscience Institute, Princeton University, Princeton, NJ 08544

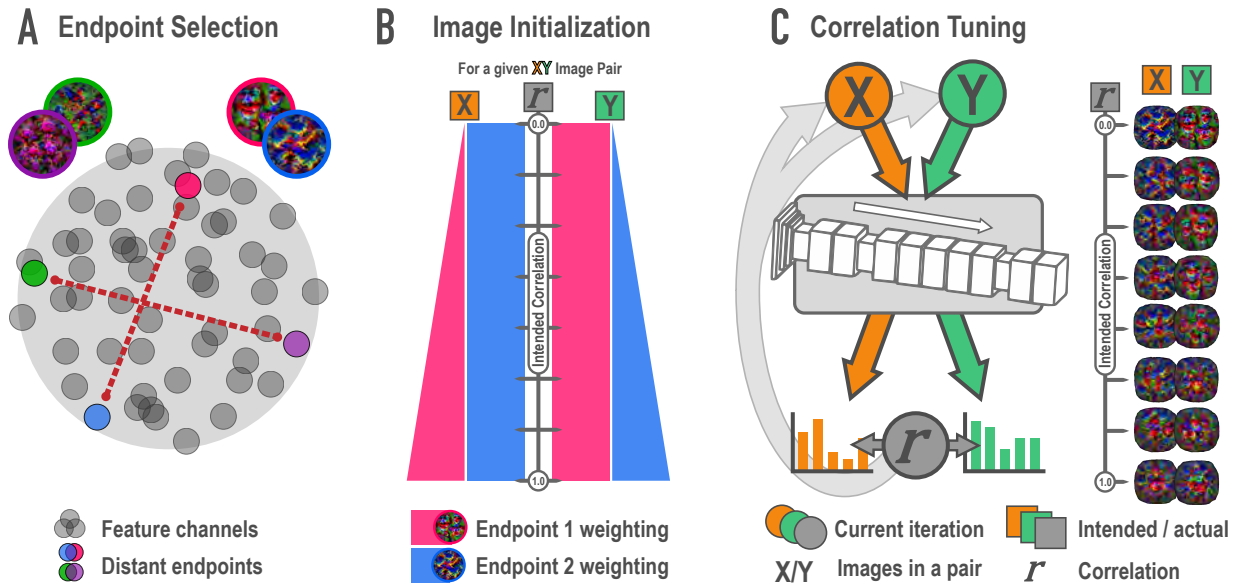


Figure S1: **Schematic of synthesis algorithm, related to Figure 1A. (A)** In the endpoint selection phase, images were generated that maximally express each of the feature channels in a later layer of a deep neural network. These feature activations were then correlated across images and pairs of endpoint channels with the lowest possible correlation were selected. Gray circles represent feature channels and closer proximity denotes higher correlations. Colored circles highlight examples of distant, low-correlation channel pairs. **(B)** In the image initialization phase, paired channels served as endpoints for a weighted optimization procedure. For each endpoint pair, eight new pairs of images were created with the intention that the correlation among pairmates would span a correlation of 0 (top) to 1 (bottom) across the eight pairs. The image pairs began as simple visual noise arrays. Each image pairmate had its own endpoint, which was always maximally optimized. The other image pairmate's endpoint was weighted according to the intended correlation. The pixels in the visual noise arrays were iteratively updated to meet this weighted goal, creating intermediate images which proceeded to the next phase. **(C)** In the correlation tuning phase, the feature activations for the two intermediate images of a pair were extracted from the 12 layers of interest. The correlation between these activations was computed at each iteration. The absolute difference between the intended correlation and the current iteration's correlation was used to iteratively update the images to minimize this difference. For a single set of endpoints, image pairs were made at various intended correlations (eight shown here).

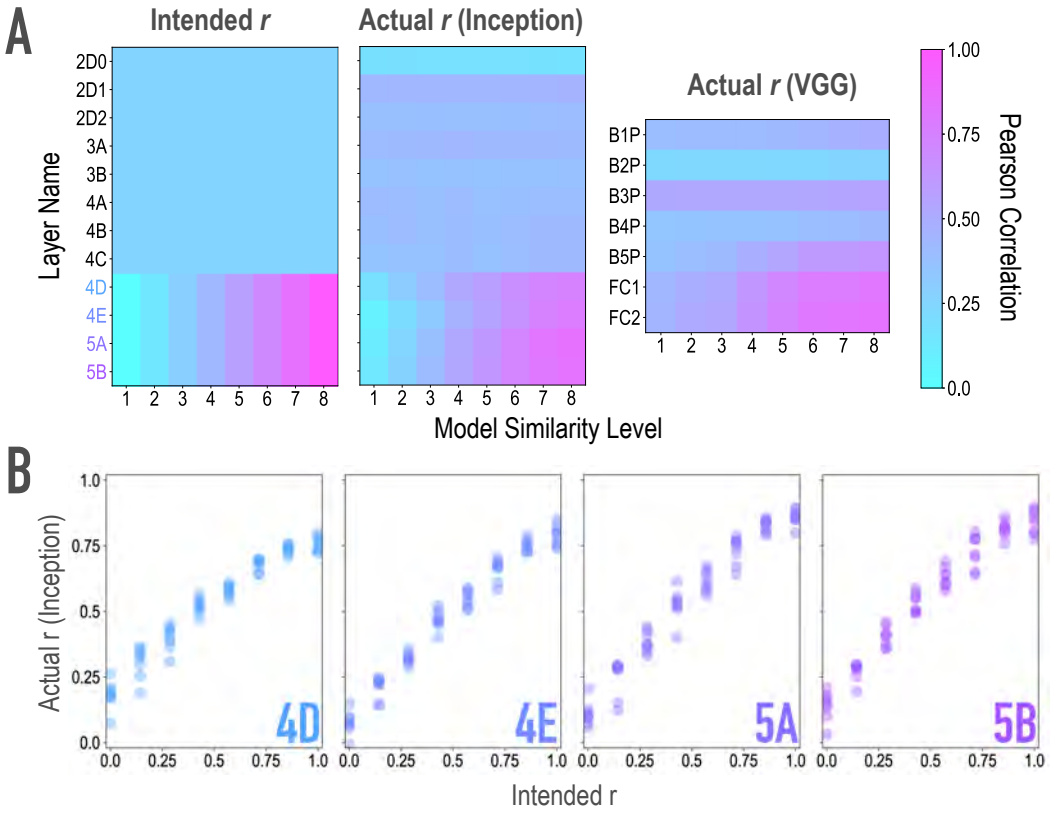


Figure S2: **Model validation, related to Figure 1A, B.** To ensure that our model-based synthesis approach was effective in constraining the shared features among image pairs, we fed the final image pairs (Fig. 1B) back through the neural network that generated them (GoogleNet/Inception; Szegedy et al., 2015; middle), as well as through an additional architecture (VGG19; Simonyan & Zisserman, 2014; right) to ensure that the similarity space produced was not dependent on the characteristics of one specific model. We extracted features from these networks for each of the generated images, and computed representational similarity matrices (RSMs) at the targeted layers to ensure that they reflected the intended similarity space. **(A)** Intended (left) and actual (center) correlations (r) of features across paired images as a function of similarity level and layer of the Inception model. We aimed for and achieved uniform correlations across pairs in lower and middle layers (2D0-4C), and linearly increasing correlations across pairs in higher layers (4D-5B, see panel B). Also shown are the actual feature correlations across layers of the alternate VGG19 model (right). In this alternate architecture (VGG19), the three highest model layers (B5P-FC2) mirrored the intended similarity for the highest model layers in the generating network ($r(62)$ = .861, .849, .856 for the three highest model layers). The four low/middle layers (B1P-B4P) did as well ($r(62)$ = .592, .542, .396, .455 for the four low/middle layers), but to a lesser extent (Steiger test $ps < .001$), and with lower variability across pairs ($SD$ = .050, .017, .025, .052 for the four low/middle layers; for comparison, for the three highest layers: .116, .164, .169). **(B)** Comparison of feature correlations in each of the targeted higher layers of Inception (4D-5B). In the four highest layers (4D-5B), across all eight pairs of each of the eight endpoint axes (64 pairs total), the intended and actual feature correlations were strongly associated ($r(62)$ = .970, .983, .977, .985, respectively, for the four highest layers). Considering each endpoint axis separately, the minimum feature correlations across the eight pairs of that axis remained high ($r(6)$ = .945, .965, .966, .967, respectively, for the four highest layers). In the lower and middle layers (2D0-4C), feature correlations did not vary across pairs. This cannot be quantified by relating intended and actual feature correlations, given the lack of variance in intended feature correlations across pairs. Instead, the average differences between any two pairs in these eight lower/middle layers were small ($M$ = .012, .011, .012, .041, .025, .045, .043, .053 for the eight lower/middle layers; for comparison, for the four highest layers $M$ = 0.14, 0.19, 0.22, 0.23). The standard deviations of the feature correlations across pairs in these layers were also low ($SD$ = .010, .010, .011, .035, .022, .039, .039, .048 for the eight lower/middle layers; for comparison, for the highest four layers $SD$ = .199, .244, .259, .243).

# References

Simonyan, K., & Zisserman, A. (2014). Very deep convolutional networks for large-scale image recognition. *arXiv*.

Szegedy, C., Liu, W., Jia, Y., Sermanet, P., Reed, S., Anguelov, D., ... Rabinovich, A. (2015). Going deeper with convolutions. In *Proceedings of the ieee conference on computer vision and pattern recognition* (pp. 1–9).