Contents lists available at ScienceDirect

Neuropsychologia

journal homepage: www.elsevier.com/locate/neuropsychologia

Sparsity and memory constraints interact with training sequence to bias learning of associative maps

Sharon M. Noh [a,1], Dale Zhou [a,b,1], Keiland W. Cooper [b,c], Shuheng Guo [a], Emily T. Dinh [a], Aaron M. Bornstein [a,c,d,*]

[a] Department of Cognitive Sciences, University of California, Irvine, Irvine, CA, USA
[b] Department of Neurobiology and Behavior, University of California, Irvine, Irvine, CA, USA
[c] Center for the Neurobiology of Learning and Memory, University of California, Irvine, Irvine, CA, USA
[d] Center for Theoretical Behavioral Sciences, University of California, Irvine, Irvine, CA, USA

ARTICLE INFO

Keywords:
Cognitive maps
Spatial navigation
Associative inference
Neural representations
Memory capacity
Sparse and distributed coding
Training schedules

ABSTRACT

Cognitive maps support inference and planning by representing associations between experiences encoded in memory. These map-like representations are thought to carry information not only about directly observed links but also about longer paths. The ability to make judgments based on multi-step associations varies with one's experience in an environment and with changes in memory abilities across the lifespan. However, it remains unclear exactly how representations of associative structure are influenced by learning curricula and memory constraints. Prior studies have suggested a tradeoff: memory representations can either be more *integrated* to improve inference, or more *separated* to recall distinct direct associations. Whether overlapping associations are experienced nearby in time (*interleaved*) or spaced apart (*blocked*) can bias memory representations toward integration or separation. However, key recent findings about how blocked versus interleaved experience bias integration or separation have been inconsistent. Here, we introduce a computational framework that helps reconcile these apparent discrepancies. Using neural network simulations of three separate memory-guided inference tasks, we show that variations in memory capacity and the sparsity of neural codes interact with learning sequence to shape network representations. Specifically, blocked learning promotes integration when memory capacity is low, while interleaved learning promotes integration when memory capacity is high. Integration is especially likely to result from representations formed when neural codes are both sparse and distributed. These results offer a principled computational account of how flexible, map-like representations can arise from experience and suggest avenues for individualized memory interventions to improve inference, generalization, and planning.

1. Introduction

Individuals extract both commonalities and distinctions across related experiences. For instance, one may *integrate* similarities across experiences to support inference and generalization (e.g., realizing that the parking spaces nearest to building entrances are usually unavailable). Conversely, one might encode distinct details of an event (e.g., parking under a tall tree) to *separate* this episode from similar ones to achieve a specific goal (e.g., locating your car at the end of the work day). This ability to detect regularities across episodes is thought to be critical for forming higher-order knowledge structures, such as cognitive maps. Just as spatial maps support navigation and path integration, the cognitive map hypothesis proposes that abstracted mental representations can support structural inference about unseen connections and paths (Tolman, 1948; Collett and Graham, 2004; McNaughton et al., 2006). However, it is less clear how such cognitive maps are assembled from distinct experiences. Here, we examine the conditions under which known episodic memory processes of integration and separation can drive the formation of complex knowledge structures such as spatial or cognitive maps.

This article is part of a special issue entitled: Maps in the Brain published in Neuropsychologia.

* Corresponding author. Department of Cognitive Sciences, University of California, Irvine, Irvine, CA, USA
E-mail address: aaron.bornstein@uci.edu (A.M. Bornstein).

[1] Co-first authors.

https://doi.org/10.1016/j.neuropsychologia.2026.109396
Received 28 March 2025; Received in revised form 4 December 2025; Accepted 9 February 2026

The cognitive map framework has been extended to encompass abstract, non-spatial associative networks, referred to as *cognitive graphs* (Chrastil and Warren, 2014; Yoo et al., 2024). According to the cognitive graph hypothesis, cognitive maps can be formally described as graphs with nodes defined as stimulus features and edges defined as transitions or associations between nodes. These nodes and edges can be extracted from experience and refined through episodic memory encoding processes (Yoo et al., 2024). Research has identified distinct neural representations associated with storing episodic details: integrated memories may support generalization and inference, whereas separated memories may reduce interference among individual episodes (Bakker et al., 2008; Kumaran and McClelland, 2012; Schlichting et al., 2014, 2015; Zhou et al., 2023). Specifically, pattern-separated representations may protect against memory interference by storing related memories as distinct, non-overlapping codes (Bakker et al., 2008; Bennett and Stark, 2016). In contrast, integrated representations may facilitate schema and concept formation by emphasizing shared similarities across related events (Mack et al., 2018; Schlichting et al., 2014). Of particular note for the study we present here is that this balance may be mediated by the sparsity of neural codes (Barak et al., 2013).

Prior work in episodic memory also suggests that the sequence of information presentation during learning biases whether neural codes become integrated or separated (Beukers et al., 2024; Schlichting et al., 2015; Zhou et al., 2023). The idea that one can shape the nature of memory representations simply by manipulating study sequences holds promise for developing interventions that can optimize learning in different contexts. For instance (Schlichting et al., 2015), used an associative inference task in which participants encoded overlapping episodes (e.g., A_1B_1 and later B_1C_1, where A_1, B_1, and C_1 represent distinct elements of an episode). They showed that *blocked* learning, during which all AB pairs are presented before BC pairs, promoted integration: A and C items showed increased neural similarity after learning (Fig. 1A). The authors suggested this occurred because blocked learning strengthens AB representations before introducing overlapping (BC) episodes, enabling retrieval and updating of existing memories rather than encoding new ones separately (Morton et al., 2017; Zeithamova et al., 2012). By contrast, *interleaved* learning, in which AB and BC episodes are presented nearby in time and in shuffled order, increased the potential for interference, and was associated with greater neural *differentiation* of A and C after learning, consistent with adaptive separation and interference resolution (Chanales et al., 2021; Schlichting et al., 2015). However, another study (Zhou et al., 2023) using a similar experiment design reported the opposite: blocked learning produced highly specific, *localized* representations, whereas interleaving yielded more *distributed* representations that supported generalization. These conflicting findings complicate efforts to identify the conditions under which integrated versus separated representations emerge.

One possible source of inconsistency lies in how representational changes are examined and measured across studies. Theories suggest that stronger pre-established AB memories increase the likelihood that B items will cue related AB memories (O'Reilly and Rudy, 2001; Kumaran and McClelland, 2012; Schlichting and Preston, 2014; Schlichting et al., 2015; Zeithamova et al., 2012a; Winocur et al., 2010; Leutgeb et al., 2004) and encourage memory updating (integration) via pattern completion. In contrast, other theories suggest that presenting overlapping episodes closer in time promotes integration (Estes, 1955; Howard and Kahana, 2002; Zeithamova and Preston, 2017; Zhou et al., 2023). Thus, blocked learning may support integration by strengthening prior representations, whereas interleaving may do so by placing related episodes in close temporal proximity to emphasize their similarities.

This tension raises a natural question: how should training curricula be optimized to promote integration of related episodes? Blocking strengthens AB memories before BC is introduced, whereas interleaving highlights commonalities across overlapping events, albeit with higher cognitive load. The optimal approach may depend on individual differences in susceptibility to interference. Individuals prone to

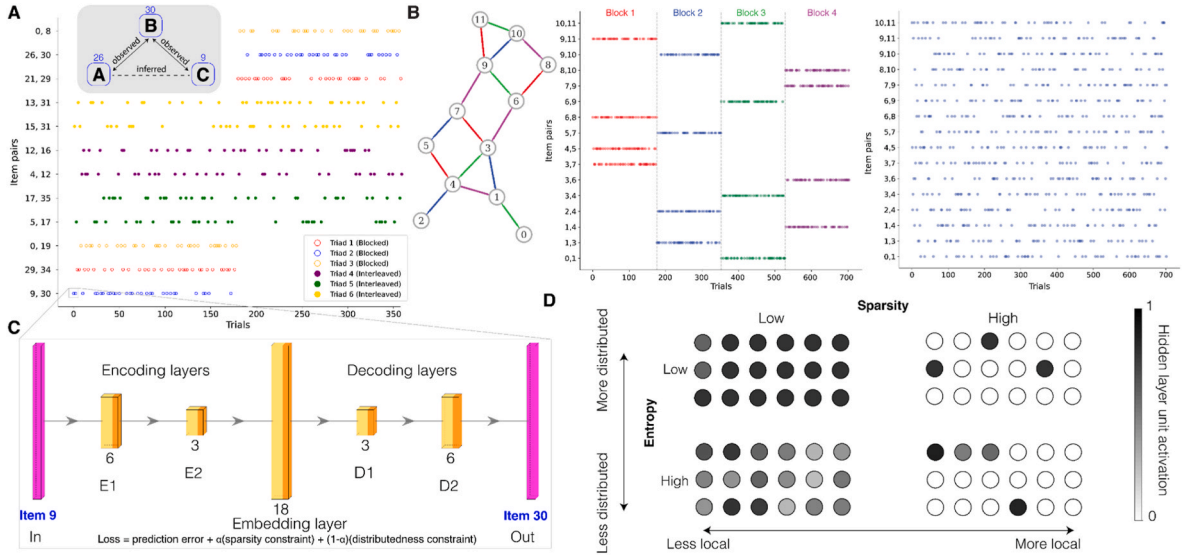


Fig. 1. Tasks and models. (A) Structural inference in simple triad graphs. Models are trained to predict paired sequences then tested on an inference task. *Inset*: Triad with solid arrows showing observed associations and dashed line showing the unobserved association that must be inferred. Items were organized into 6 triads (colors) in a blocked (open circles) and interleaved schedule (filled circles). **(B)** Neural networks were also trained to perform structural inference across an entire graph. A different set of models were trained to learn a sequence of edges drawn from a latent graph structure. Edges are colored according to the blocked schedule, where half the models were trained with a blocked schedule. In this schedule, 4 mini-blocks were created where each block contained 4 edges that did not share any nodes with each other. **(C)** The neural network was trained using a loss function that penalized for errors in predicting the next item, given the current item (Chandak et al., 2024). The loss function also contained a term that encourages sparse representations with low activation strengths, inspired by energy constraints in biology. A scaling parameter, α, controls the degree of sparsity. Different memory capacities were simulated by varying the size of the encoding (E1, E2) and decoding (D1, D2) layers. **(D)** The loss function encourages localist versus distributed codes in the 18 embedding layer units (circles). Sparser activation (lighter colors) characterize more localist versus distributed codes by encouraging fewer units to activate (see Supplementary Materials, section on *Defining sparsity*). We can also quantify the information content by calculating the code's entropy, where higher entropy indicates a more diverse and distributed pattern of activation to encode the same stimulus. (For interpretation of the references to color in this figure legend, the reader is referred to the Web version of this article.)

interference may benefit from blocked training, whereas those with stronger memory capacity may benefit from interleaving. Indeed, prior work using a graph-structured associative inference task showed that individuals with weaker memory abilities performed better on graph-based inference judgments when overlapping edges were learned in a blocked sequence, whereas those with stronger abilities performed better when all edge pairs were interleaved (Noh et al 2026).

Beyond memory capacity, representational coding strategies may also shape outcomes. Schlichting et al. (2015) and Zhou et al. (2023) observed evidence of integration/separation in different neural subregions and pathways, suggesting that coding biases within those regions may have contributed to their divergent findings. Schlichting et al. (2015) found integration in anterior hippocampus and posterior mPFC, but separation in posterior hippocampus and anterior mPFC. Zhou et al. (2023), by contrast, emphasized differences between the monosynaptic (MSP) and trisynaptic (TSP) hippocampal pathways, showing that blocking versus interleaving produced more localist versus distributed codes.

These differences highlight an important point: the way neural populations represent information can vary across individuals and brain regions. Some representations (*sparse* neural codes) emphasize efficiency by recruiting only a small subset of neurons to encode a stimulus, producing distinct, minimally overlapping codes. Other representations (*distributed* neural codes) emphasize generalization by recruiting many neurons, leading to overlapping codes that highlight shared features across experiences. At an individual level, factors such as age and memory ability can bias which coding strategy is favored: older adults, for instance, may be less likely to maintain sparse codes and more likely to rely on distributed, overlapping codes relative to younger adults (Wilson et al., 2006; Yassa and Stark, 2011). Within individuals, different hippocampal subregions also exhibit distinct coding biases. The dentate gyrus, with its dense population of granule cells and strong inhibitory circuitry, is well-suited for sparse coding and pattern separation. In contrast, CA3, with its recurrent collaterals, is more prone to distributed coding that supports pattern completion and generalization (Kumaran and McClelland, 2012; Leutgeb et al., 2007; Neunuebel and Knierim, 2014; Treves and Rolls, 1994). Thus, both individual- and regional-level biases in encoding strategy may influence whether overlapping experiences are prone to integration or separation during learning. Specifically, the *sparsity* of neural activity may further define how separated versus integrated information is represented (Benna and Fusi, 2021; Cayco-Gajic et al., 2017; Cayco-Gajic and Silver, 2019; Chavlis et al., 2017). Sparse coding arises when relatively few, locally clustered neurons are recruited to encode a stimulus. This strategy reduces overlap across memories and helps minimize interference by decorrelating inputs, often through inhibitory feedback mechanisms (Tetzlaff et al., 2012; Wiechert et al., 2010). Distributed coding, by contrast, arises when many neurons spread across a network are recruited, producing overlapping representations that emphasize shared features. This strategy supports generalization and increases overall representational capacity (Hinton, 1984; McClelland and Rumelhart, 1988; Rigotti et al., 2013). Importantly, a balance between sparse and distributed codes may support an optimal tradeoff, capturing complex patterns with both efficiency and robustness (Hinton and Ghahramani, 1997).

In light of these considerations, the present study aims to clarify the mechanisms by which sequencing effects shape memory representations and, ultimately, cognitive map formation. We propose that reconciling conflicting findings regarding sequencing effects requires systematically examining how individual differences in memory capacity and coding strategies interact with learning schedules. Both sources of variability (memory capacity and coding strategy differences) may yield different representational outcomes in response to blocked or interleaved learning sequences. Specifically, we hypothesize that individuals with lower memory capacity are more vulnerable to interference, and therefore may benefit from blocked learning. By first strengthening AB associations before introducing overlapping BC associations, blocked training reduces cognitive load and increases the likelihood that BC episodes update existing AB memories rather than compete with them (Schlichting et al., 2015). In contrast, individuals with higher memory capacity may tolerate greater load and benefit more from interleaved training, which presents overlapping associations in close temporal proximity and encourages structural inference across episodes (Zeithamova and Preston, 2017; Zhou et al., 2023).

Critically, we also predict that capacity effects will be further tuned by differences in coding strategies. Sparse coding, which emphasizes decorrelation and separation, may amplify the benefits of blocking by reducing interference across sequentially presented events. Distributed coding, which emphasizes overlap and generalization, may instead amplify the benefits of interleaving by highlighting similarities across temporally adjacent episodes. Together, these considerations point toward an interaction between learning schedule, memory capacity, and representational coding strategy as a key determinant of whether integration or separation emerges during learning.

To test this framework, we use feedforward neural network simulations of published associative inference tasks (Schlichting et al., 2015; Zhou et al., 2023). Our models systematically manipulate memory capacity and sparsity constraints to evaluate how these factors affect the representations formed under blocked vs. interleaved training. This approach allows us to explain the conflicting findings with regard to sequencing effects observed in prior studies, and to identify conditions under which integration versus separation may be favored. We then compare the qualitative patterns that emerge from model simulations under our framework to performance in a graph-structured multi-step associative inference task, which showed variability across the lifespan and with memory capacity (Noh et al 2026; Rmus et al., 2022). Our models show that differences in memory capacity and coding strategies can be sufficient to generate the kinds of divergent patterns observed in prior empirical work with respect to how training conditions shape cognitive representations. In this way, we demonstrate the potential utility of our framework for interpreting sequencing effects and motivate future empirical tests under this account.

2. Method

2.1. Associative inference task (simple triad graph)

2.1.1. Training with blocked and interleaved schedules

Training datasets were generated to mimic experimental data collected by Zhou et al. (2023) and Schlichting et al. (2015). Specifically, the datasets were constructed with two kinds of stimulus "schedules": *hybrid* and *pure*. The hybrid schedule (Zhou et al., 2023) includes both blocked and interleaved curricula within a single learning phase, whereas the pure schedule (Schlichting et al., 2015) includes either a blocked or an interleaved curriculum in separate, counterbalanced learning phases. In the blocked schedule, all direct associations of one type (A,B) are presented before any overlapping associations (B, C) are introduced. The interleaved schedule shows the (A,B) and (B,C) associations in a random order.

For the hybrid schedule, there were 360 training trials. We one-hot encoded 36 items. From these, 18 were randomly sampled for training. The 18 items were grouped into six triads (A, B, C). During training, triads were presented as overlapping pairs (AB or BC). In the blocked condition, all of one pair type (e.g., AB) were presented before the overlapping pairs (e.g., BC). In the interleaved condition, AB and BC pairs were interleaved throughout the learning phase. Each direct pair type was shown 30 times. The order of A, B, and C within pairs was randomized, and trials were randomized following the blocked or interleaved curriculum. Pairs sharing the same A, B, or C item were never shown consecutively.

For the pure schedules, we created two schedules, each with 360 trials (the same as the hybrid schedule). The key difference is that the

pure schedule separates blocked and interleaved conditions into two distinct, counterbalanced learning phases (blocked first vs. interleaved first). For the pure blocked schedule, we used the 180 blocked trials from the hybrid schedule to improve comparability. Similarly, for the pure interleaved schedule, we used the 180 interleaved trials from the hybrid schedule, allowing direct comparison between formats (pure vs. hybrid). Although raw similarity metrics differed somewhat by format, blocked vs. interleaved training did not produce fundamentally different patterns of results (e.g., blocked > interleaved in pure vs. hybrid). For clarity, we therefore collapsed pure blocked with hybrid blocked and pure interleaved with hybrid interleaved into two conditions–blocked and interleaved–for all main analyses (see Supplementary Fig. 1 for disaggregated values).

2.2. Structural inference task generalized to a complex graph

2.2.1. Training with blocked or interleaved schedules

Models were trained to perform a complex associative inference task designed to approximate shortest-path distance judgments. We trained 16 pairs (edges) of 12 stimuli (nodes) in either a blocked or interleaved schedule using one-hot codes for each stimulus. For both schedules, pairs were drawn from the edges of an underlying, undirected and unweighted, graph (Fig. 1B). The 16 pairs were repeated 44 times each, yielding 704 trials. In the interleaved schedule, trial order was randomized during the learning phase. In the blocked schedule, pairs were grouped into four mini-blocks, each containing four unique object pairs. To reduce potential for memory interference during encoding, the 4 pairs presented in each mini-block shared no overlapping nodes (Fig. 1B).

2.3. Neural networks

To simulate individual differences in sequence learning, we used feed-forward neural networks with five hidden layers (two encoding, one embedding, and two decoding). For the simple triad task, we trained 100 models per schedule type (hybrid and pure). For the pure schedule, training was counterbalanced: 50 models were trained with pure blocked then pure interleaved schedules, and 50 with pure interleaved then pure blocked schedules. We trained 200 models (100 each for hybrid and pure schedules) for three memory capacities–low, medium, and high–and across 11 sparsity constraints (described below), yielding 6600 trained models in total. For the complex graph task, we trained 25 models for each of two schedules, 11 sparsity constraints, and five levels of memory capacity, totaling 2750 models.

Differences in coding strategy were operationalized systematically and quantitatively. Distributed activation can be indexed by the *entropy* of stimulus-evoked population activity, with higher entropy reflecting more distributed coding. Sparsity was indexed by the inverse of activation strengths, with smaller values indicating that only a few units were recruited to encode a stimulus in a more localist manner. Thus, coding style was measured along continuous dimensions, allowing us to test how sparsity and distributedness interact with learning schedule and memory capacity to shape representational outcomes. Below, we detail how we encouraged specific coding strategies using different forms of regularization during training.

2.3.1. Unsupervised learning

The model was first pre-trained to reproduce each item using a loss function similar to an autoencoder model. This stabilized task training and provided pre-task representations of each item, which were later compared to post-task representations to evaluate changes after learning.

Pre-training was performed for 100 epochs with a loss function defined by a mean squared error term plus sparsity constraint:

$$L_{reconstruction} = MSE + \alpha * \sum |w_i| + (1\text{-}\alpha) * \sum (w_i)^2, \quad (1)$$

where $L_{reconstruction}$ is the total reconstruction loss, MSE is the mean-squared error between the reconstruction and original input, w_i are individual network weights, α_1 is the L1 regularization strength (Lasso penalty), α_2 is the L2 regularization strength (Ridge penalty), $\sum|w_i|$ is the L1 norm (sum of absolute weight values), and $\sum(w_i)^2$ is the L2 norm (sum of squared weight values). α ranged from 0.0 to 1.0 in increments of 0.1.

Experimentally, pre-training was needed to provide a baseline for representational similarity analysis (RSA). Given our aim to reconcile discrepancies in prior literature, we followed the analysis approach from Schlichtling et al. (2015), showing how representations change after successful learning. Practically, pre-training and regularization (L1 and L2 norms) were necessary to achieve above-chance performance compared with shallower or non-pretrained networks. These steps allowed the task model to begin with a representation that can at least support accurate reconstruction by initializing the weights in a space that respects the distinct items, instead of randomly projecting the stimulus into an initialization with a noisy space.

2.3.2. Supervised learning

Following unsupervised pre-training, networks were trained on the dataset described above with a batch size of 32 for 100 epochs. Each trial consisted of a pair: the network received the first one-hot encoding as input and was tasked with producing the second one-hot encoding as output. In this supervised learning phase, the network was trained with binary cross-entropy loss.

We selected layer sizes, embedding layer size, learning rates, and weight decay by grid search across five random seeds. To improve performance while reducing overfitting, each layer included batch normalization, ReLU activation, and dropout regularization (0.3). The linear output layer used a softmax function to produce probability distributions for predicted outputs. To prevent vanishing or exploding gradients, weights were initialized with Xavier uniform initialization (and biased to a small constant, 0.01, to utilize more neurons during initial stages of training) with the AdamW optimizer using weight decay = 0.001 and learning rate = 0.001. This provides better regularization for the model. Training efficiency was optimized with a ReduceLROnPlateau learning rate scheduler. This scheduler monitors the loss, and when the loss fails to decrease across epochs (patience = 20 epochs), it reduces the learning rate by a factor of 0.5 to a minimum of 0.00005. This adaptive learning rate mechanism was used to help the model converge in later stages of training.

To model individual differences in memory capacity, we trained low-, medium-, and high-capacity networks. Low capacity models had encoding/decoding layer sizes of (6, 3). Importantly, we selected these layers to be smaller than the input layer, enforcing a many-to-one mapping of incoming information to simulate conditions of increased interference pressure. In contrast, high memory capacity models had sizes of (256, 128), supporting one-to-one mapping of inputs, with ample space for pattern separation. Medium memory models had sizes of (32, 16). All models used an embedding layer of size 18, chosen to match the input size and stabilize integration/separation metrics such as cosine similarity, which are sensitive to dimensionality of the vectors.

To better interpret the effects of our manipulations, we included two separate encoding and decoding layers from the embedding layer that either reduce or expand the input. We decided on the depth of two encoding/decoding layers following prior work comparing models with similar architecture to human performance data (Noh et al 2026). Here we further sought to manipulate the size of the layers to allow for both compression-expansion and expansion-compression dynamics, which have been shown to support learning and generalization (Farrell et al., 2022; Ito and Murray, 2023). Finally, to better compare differences between network representations, we used an additional fixed-size embedding layer to facilitate cross-model comparison while isolating memory capacity manipulations. In practice, the size of the medium and

high capacity neural networks were chosen to improve inference performance across tasks, as prior work using similar but smaller models performed only modestly above chance (Noh et al 2026). For additional details on layer-size choices and how they relate to the human literature, see the **Supplementary Materials** (section on *Memory capacity and representational dimensionality*).

2.3.3. Supervised learning loss function to encourage integration or separation

Both Integrated or separated representations may support successful AC inference. Integrated representations are thought to be useful because they distribute information across representational units, allowing for quick and efficient generalization between integrated AC item representations (Zhou et al., 2023). However, these representations can be susceptible to memory failures and false memories, as integrated representations make it difficult to distinguish individual memory episodes from inferred ones. Separated representations are thought to enhance memory precision because they are encoded with more local and sparse properties that distinguish A from C and improve resistance to interference. However, making the AC association may then require a more explicit and costly retrieval process of separate A and C representations (e.g., retrieving separate AB and BC memories to infer the AC relationship).

We use these neural networks to investigate how representations influence inference and how learning schedules shape those representations. Prior work showed that specific models biased toward either separated (localist) or integrated (distributed) representations improved performance under blocked or interleaved learning, respectively (Zhou et al., 2023). However, those models also differed in other architectural respects. To simplify the comparison, we varied separated versus integrated representation types within the same class of feedforward neural networks. We achieved this by using a loss function that encouraged either more separated or more integrated internal representations, which in turn could help or hinder AC inference under blocked or interleaved schedules. L1 and L2 regularization encourage localist and distributed representations, respectively. Therefore, by combining them, we can manipulate the balance of separation and integration using an elastic net regularization loss function:

$$L_{prediction} = BCE + \alpha * \sum|w_i| + (1\text{-}\alpha) * \sum(w_i)^2, \quad (2)$$

where $L_{prediction}$ is the total prediction loss, BCE is the cross-entropy classification loss, w_i are individual network weights, α_1 is the L1 regularization strength (Lasso penalty), α_2 is the L2 regularization strength (Ridge penalty), $\sum|w_i|$ represents the L1 norm (sum of absolute weight values), and $\sum(w_i)^2$ represents the L2 norm (sum of squared weight values). Eleven values of α were tested (0.0 to 1.0 with a step size of 0.1). Intuitively, as the $\alpha * \sum|w_i|$ term becomes larger, the representations become more localist and separated: this L1 regularization drives many weights to zero and introduces sparsity. In contrast, as the $(1\text{-}\alpha) * \sum(w_i)^2$ term becomes larger, the representations become more distributed and integrated: this L2 regularization discourages large weights but does not drive them to zero, smoothing the weight distribution across units. $L_{prediction}$ defines an error which is propagated to adjust how representations are updated to balance a compression trade-off between efficient representations for association and the fidelity of that representation.

2.4. Measuring separation or integration of representations

We used entropy to quantify whether the loss function shifted stimulus encoding toward more distributed/integrated versus more sparse/separated representations. Entropy measures the "spread" of a representation, with higher entropy indicating more distributed coding. A distribution with maximum entropy has a uniform spread of values, whereas one with minimal entropy has only a single value (Fig. 1D). To calculate entropy, each item was input to the network to produce a vector of activations in the embedding layer. This activation vector was then converted into probabilities using a softmax function: $P(x_i) = exp(x_i) / \sum_j exp(x_j)$. These probabilities were used to calculate entropy:

$$H(X) = -\sum_i P(x_i) \log P(x_i), \quad (3)$$

where X is the representation, x_i are possible items in the representation, and $P(x_i)$ is their probability.

Sparsity was used as a complementary measure, capturing the suppression of activation when representing a stimulus. The higher the sparsity, the lower the activity used to represent an item. To quantify sparsity, items were input into the network to produce activations in the embedding layer. The sparsity is defined as the inverse of collective activity strength (for discussion of alternative definitions of sparsity, see the Supplementary Materials section on *Defining sparsity*).

2.5. Analysis

2.5.1. Indirect AC inference in simple triad graph task

Our primary analysis centered around the integration or separation of AC representations after learning. We input the18 learned items and recorded the network's internal representation of A and C. We then calculated cosine similarity between A and C vectors. Intuitively, higher cosine similarity reflects more integrated representations, whereas lower cosine similarity reflects more separated representations.

2.5.2. Judgment of relative distances in complex graph task

In the complex graph task, models judged which of two target nodes was closer to a source node (Rmus et al., 2022; Noh et al 2026). True distances were defined as the shortest paths in the graph, (i.e., the fewer edges between source and target). Models chose which of the two target nodes was closer to the source object based on the indirect relationships of the graph learned during training. Specifically, the model chose the target node with smaller cosine distance to the source node, thus approximating the shorter distance. Trials varied in difficulty based on the degree to which the target node options differed in the topological distance from the reference node, in which a difference of 1 was the most difficult, 2 had intermediate difficulty, and 3 was the easiest. Accuracy was computed within each distance bin.

Each of the 12 nodes served as a source node 17 times, for a total of 204 trials. The two target node options for each trial were randomly selected with three constraints. First, target pairs at relative distance 3 were deterministically included to ensure sufficient sampling of the easiest trials. For example, when node 0 was the source, the 17 trials necessarily included pairs such as 3-10, 3-11, 4-10, and 4-11, as these corresponded to distance differences of 3. Second, neither target node was directly paired with the source node in the underlying graph. Third, the two target nodes were required to differ in distance from the source node. The same set of 204 trials were used to test all models.

2.5.3. Pre-study versus post-study representational change in the simple triad inference task

Following prior work (Schlichting et al., 2015), we examined how AC representations changed after learning. Neural representations of each item were measured both before and after study. For each schedule (blocked and interleaved), similarity between A and C items was calculated within and between ABC triads. Comparisons of AC similarity across schedules (i.e., between blocked and interleaved conditions) were excluded. This procedure produced representational similarity matrices in which rows corresponded to C items and columns corresponded to A items (Fig. 3A). Cosine similarity values defined each matrix element, measuring the similarity of A and C items both before and after learning. Learning-related change in representational similarity (ΔRS) was calculated as: post-study minus pre-study matrix similarities, with positive values indicating the magnitude of integration and negative values indicating the magnitude of separation. Following methods analogous to

those used in prior work (Schlichting et al., 2015), only successfully learned pairs (accurately discerning that A was more similar to C than a foil) were included in the analysis.

Using this ΔRS approach, the resulting matrices were compared to two hypothesized representational structures. The first hypothesized structure is one in which the blocked schedule leads to more similar AC representations (blocked → integration) whereas the interleaved schedule leads to more dissimilar AC representations (interleaved → separation). The second hypothesized representational structure is the opposite, in which the interleaved schedule leads to more integrated AC representations (interleaved → integration) while the blocked schedule leads to more separated AC representations (blocked → separation).

2.5.4. *Pre-study versus post-study representational change in the complex graph task*

We applied a similar approach to the complex graph task. Pre-study representations were defined as the embedding layer activations of each one-hot input after pre-training with an autoencoder. A cosine similarity matrix was then constructed across all nodes to measure the degree of integration. There were two pertinent post-study representations for each trial in the judgment task. Given two options, the participant chose either the correct or incorrect target. We obtained a representation for both the correct target and the incorrect target. We then tested how integrated the correct target was with the source node, as well as how integrated the incorrect target was with the source node. Using the pre-study and post-study values, we calculated the change in integration for each target-source pair (Δr) as post-study similarity minus pre-study similarity. Higher cosine similarity differences indicated that representations became more integrated after the training phase. Finally, we calculated the difference in representation between the cosine similarity of the correct target with the source node and that of the incorrect target with the source node: ($\Delta r_{\text{correct}}$ - Δr_{source}) - ($\Delta r_{\text{incorrect}}$ - Δr_{source}). We would expect this difference to be larger for easier trials than for harder trials. This prediction followed from the assumption that the correct node (i.e., target closer to the source node) should be represented with greater cosine similarity to the source than is the incorrect target node, particularly in the distance difference = 3 trials compared to the distance difference = 1 trials.

3. Results

We report the results of how memory capacity affects the integration of representations for items A and C following training. Throughout, we operationalize integration as higher cosine similarity between the embedding layer's representation of A and C. Viewing the learning curves across all models, higher memory capacity allowed training to converge more quickly to an asymptote of the AC integration curve than lower memory capacity (Fig. 2A). Importantly, the models were not

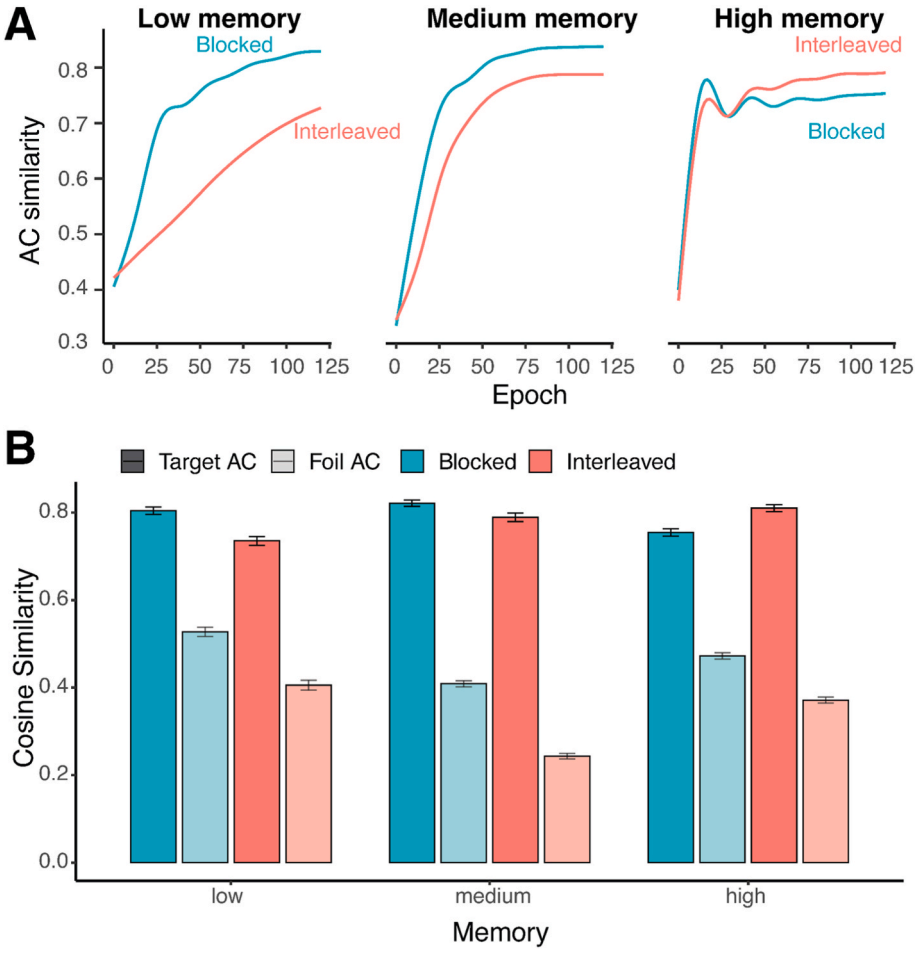
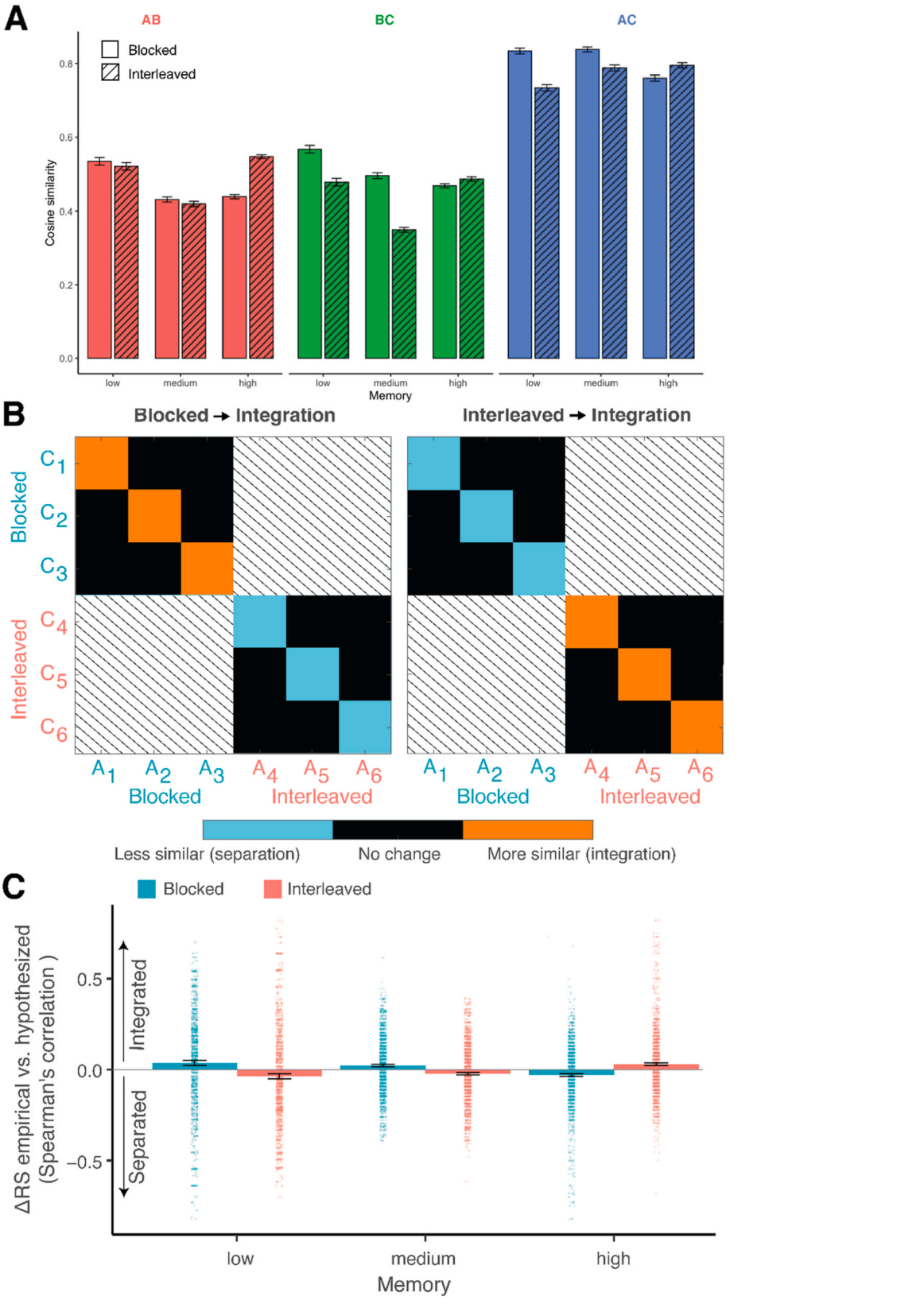


Fig. 2. Encoding the indirect AC association with integrated representations. (A) Integration is operationalized as the cosine similarity between the internal representations of A and C for different memory capacities as the models are exposed to items from different schedules. The cosine similarity between A and C was greater for blocked versus interleaved schedules in the low and medium memory capacity conditions. In contrast, for high memory capacity, the cosine similarity between A and C was greater for interleaved versus blocked schedules. **(B)** AC similarity for every triad and model. AC integration, as measured by cosine similarity between A and C, benefits from blocked schedules when memory capacity is low and interleaved schedules when memory capacity is high. The similarity of AC is consistently greater than foils containing one element from a different ABC triad trained under the same schedule. Error bars depict 95% confidence intervals.

explicitly trained to increase AC integration; rather, integration emerged as a byproduct of predicting the next item in a sequence. A multiple regression analysis indicated that with greater memory capacity, the interleaved schedule produced greater AC integration (adjusted $R^2 = 0.09$, $F(11, 989988) = 9298$, $\beta = 0.31$, $p < 0.001$). However, integration was on average lower than representations produced from blocked learning across memory levels ($\beta = -0.18$, $p < 0.001$). There was also a three-way interaction between epoch, schedule, and memory capacity,



(caption on next page)

Fig. 3. Memory capacity explains differing effects of learning schedule on AC representation integration. (A) Direct AB (green) and BC (blue) and indirect AC (red) pair similarities for models trained using blocked or interleaved schedules. **(B)** To test if integration or separation supports successful inference, we follow prior analysis analyzing the change in similarity after learning versus before learning for only the models with accurate inference. Schematic of the representational similarity analysis of post-study changes in representations compared to pre-study representations. Adapted from Schlichtling et al. (2015). *Left*: Hypothesized changes when the blocked schedule leads to more integrated AC representations. The diagonal entries are AC similarities within distinct ABC triads. The off-diagonal entries are AC similarities across triads but still within either the blocked or interleaved schedule. *Right*: hypothesized representational similarity matrices for when interleaved schedules lead to more integrated AC representations. **(C)** When memory capacity is low, successful performance on the AC inference task arises from the blocked schedule tending to lead to more integrated representations. In contrast, when memory capacity is high, successful performance arises from the interleaved schedule tending to lead to more integrated AC representations. (For interpretation of the references to color in this figure legend, the reader is referred to the Web version of this article.)

such that interleaved learning produced more integrated representations under high memory capacity, whereas blocked learning produced more integrated representations under low memory capacity ($\beta = 0.07$, $p < 0.001$). Together, these findings suggest that exposure to different training sequences affects the degree of integration and separation in learned representations differently as a function of memory capacity.

While effects on separation and integration were evident from the training sequence, we next assess how representations relate to task performance following training. The blocked schedule with low memory capacity is associated with greater AC cosine similarity across all triads and models, whereas an interleaved schedule with high memory capacity is associated with greater AC integration (Fig. 2B). Integration for target pairs (within triads) was greater than the integration for foils (between triads; $t(68240) = 147.18$, $p < 0.001$). The results of a multiple regression analysis suggest that there were main effects of memory capacity (adjusted $R^2 = 0.01$, $F(5, 34140) = 60.35$, $\beta_{memory} = -0.11$, $p < 0.001$) and schedule ($\beta_{schedule} = 0.05$, $p < 0.001$) on AC cosine similarity. Moreover, there was an interaction between schedule and memory capacity on AC cosine similarity ($\beta_{memory*schedule} = 0.27$, $p < 0.001$). The blocked schedules have higher AC integration with low memory capacity, whereas the interleaved schedules have higher AC integration with high memory capacity. These results are consistent with the notion that the type of schedule that best supports AC integration depends on individual differences in memory capacity.

Having shown that memory capacity affects how integrated AC representations become following blocked versus interleaved learning schedules, we next show in more detail how AC representations change after learning for triads where the model performed successful AC inference. For successful trials, we measure representational change as the change in AC cosine similarity within and across triads for each model post-training minus the AC cosine similarity pre-training (Δr). Using the raw cosine similarities of pairs following learning (Fig. 3A), we constructed this cosine similarity matrix for each AC pair within and across triads for the same schedule (Fig. 3B). Positive values suggest learning resulted in integration of AC information, whereas negative values suggest that learning lead to more separation of AC information. Then we compare the empirical matrix to two hypothesized representation matrices, one where blocking leads to integration and interleaving leads to separation (consistent with (Schlichting et al., 2015), and the other where blocking leads to separation and interleaving leads to integration (consistent with (Zhou et al., 2023).

We show that whether blocked or interleaved schedules lead to integrated or separated AC representations depends on memory capacity (Fig. 3C). Specifically, models with low memory capacity produce more integrated representations after blocked training and separated representations after interleaved training, whereas models with high memory capacity showed the opposite pattern: they tend to form integrated representations after interleaved training and separated representations after blocked training. The results from a multiple regression analysis support this observation (adjusted $R^2 = 0.02$, $F(5, 12062)$, $\beta_{schedule*memory} = 0.46$, $p < 0.001$). Notably, this result cannot be fully explained by confounds such as catastrophic interference between representations or poor direct pair learning (**Supplementary Figs. 3 and 4**). Similar to the analysis of AC integration after exposure to different training sequences (Fig. 2), we find an increase in AC integration from pre-study to post-study for the interleaved schedule as memory increases from low to high and *vice versa* for the blocked schedule.

We next show that, beyond the memory capacity of the model, encoding properties of the network also affect how integrated AC representations become after learning. Prior work indicated that interleaved schedules confer their advantages to AC inference by increasing the distributedness of integrated representations (Zhou et al., 2023). Here we ask whether models biased toward using distributed versus sparse representations are more likely to form integrated representations (of related A and C items) after blocked or interleaved training conditions. We operationalize sparsity as the inverse of the activation strengths and the distributedness as the entropy of the activation magnitudes (Fig. 4A). We again analyze the successful inference trials and measure changes in AC integration, but now separate the results by models trained with various sparsity versus distributedness constraints. This constraint was operationalized by the parameter α (Fig. 4B). Here we discretize the models into ones where representations were constrained to be less sparse and more distributed ($\alpha = 0$ to 0.3), a mixture of sparsity and distributedness ($\alpha = 0.4$ to 0.6), and more sparse and less distributed ($\alpha = 0.7$ to 1). We performed a multiple regression analysis to test the effects of sparsity on performance (adjusted $R^2 = 0.03$, $F(11, 6576) = 22.05$, $p < 0.001$). In general, models with greater sparsity tended to have greater representational similarity to either hypothesized integration process ($\beta = 0.07$, $p = 0.0001$). Moreover, the interleaved schedule tended to benefit from greater memory capacity than the blocked schedule ($\beta_{schedule*memory} = 0.51$, $p = 0.0001$). Blocked representations appear to benefit from high sparsity and low distributedness, whereas interleaved representations benefit from high distributedness and low sparsity ($\beta_{schedule*sparsity} = 0.13$, $p < 0.001$; Fig. 4C). This interaction was further moderated by the memory capacity ($\beta_{schedule*sparsity*memory} = 0.13$, $p = 0.002$). These results support the notion that the tendency to form integrated representations after blocked versus interleaved learning not only depends on memory capacity, but also sparsity constraints.

Having demonstrated that our model can capture the effects of memory, sparsity, and distributedness on structural inference in simple triad graph structures, we investigate if these findings generalize to more complex graph tasks that require higher-order structural inference across multiple associations. To do this, we assessed model performance in a structural inference task on a more complex graph consisting of 12 nodes and 16 edges (Fig. 1B; Fig. 5 *insets*). Specifically, the trained models performed a relative distance judgment task. Models were given a source node and two possible target nodes and tasked with determining which target node was closer to the source node based on the cosine similarities between the source and target nodes (for additional task constraints, see **Methods**). We found that the trained models reproduce performance patterns seen in both the triad graph structures above and in human performance on the same complex graph task (Fig. 5) (Noh et al 2026). The results of a multiple regression analysis suggest that performance for both schedules improved with increasing memory capacity (adjusted $R^2 = 0.14$, $F(11, 2988)$, $\beta = 0.22$, $p < 0.001$) and there was a main effect of the schedule ($\beta = 0.25$ $p < 0.001$) but not distance ($\beta = -0.08$, $p = 0.11$) on performance. The interleaved schedule performed better than the blocked schedule as memory capacity increased ($\beta_{schedule*memory} = 0.28$, $p < 0.001$) and this effect was

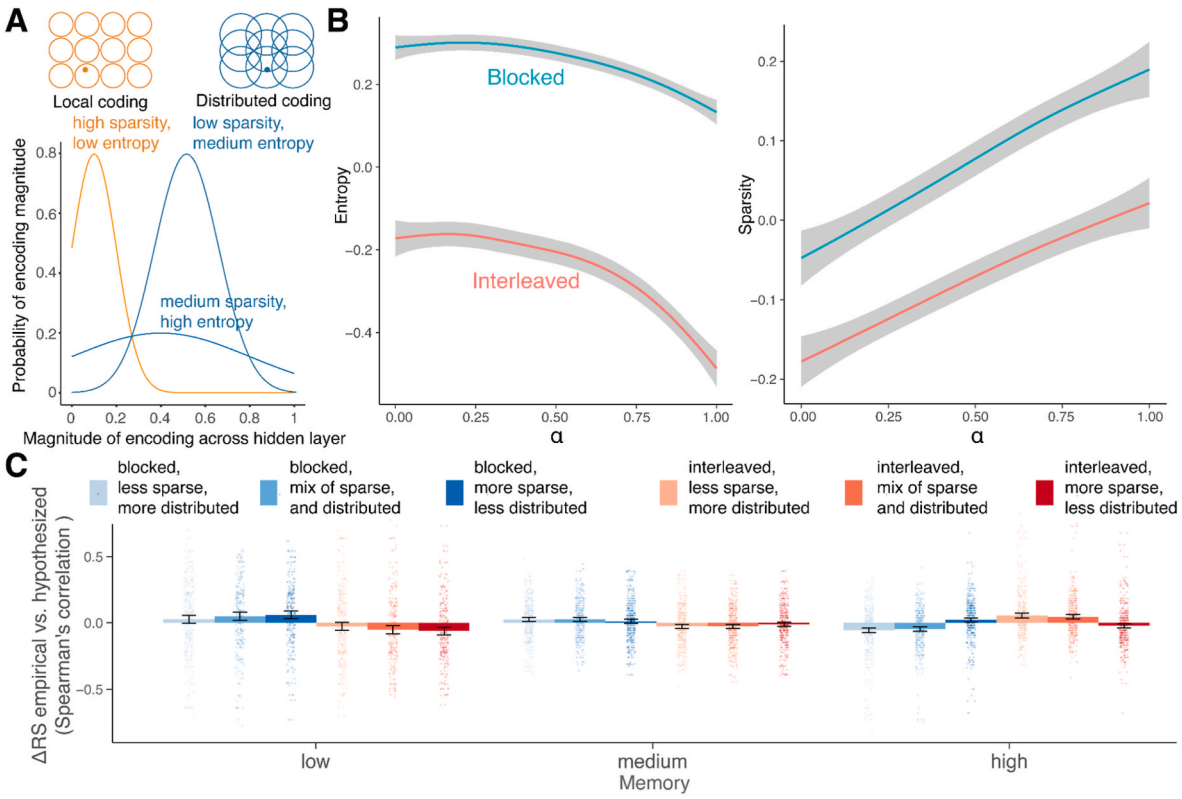


Fig. 4. Effects of sparse and distributed codes on the integration of AC representations. (A) A schematic demonstrating how local and distributed coding can be understood by the receptive field of each unit (circle) that responds to a stimulus (point). We quantify these concepts by measuring how distributedness differs according to the information content (entropy) and magnitude (sparsity) of activation in embedding layer units. The orange distribution is localist because it substantially overlaps with 0, indicating that many units are inactive. The blue curves are more distributed codes because most units are involved in coding for stimuli. The sharper blue curve (low entropy) is less distributed but has lower information content relative to the flatter blue curve (high entropy) that has a higher probability of inactive units and higher information content. **(B)** Different models trained with more constraints on localist memory encoding produced embedding representations with more sparsity (Pearson's $r(39{,}036) = 0.07$, $p < 0.001$) and less entropy ($r(39{,}598) = -0.07$, $p < 0.001$). Models with different sparsity and distributedness constraints mimic mixed representations in hypothesized information processing pathways that use representations with both low and high sparsity or distributedness. **(C)** When representations are more sparse and less distributed, AC inference seems to benefit from the blocked schedule encouraging integrated representations. In contrast, when representations are less sparse and more distributed, successful AC inference tends to benefit from the interleaved schedule encouraging more integrated AC representations. (For interpretation of the references to color in this figure legend, the reader is referred to the Web version of this article.)

further moderated by increases in the relative distance between choice options ($\beta_{schedule*memory*distance} = 0.26$, $p = 0.0002$). Consistent with the models in the simple triad graph structures as well as human performance on the same complex graph structure, the effect of training schedule on performance appears to depend on memory capacity.

Taken together, the above results suggest that training conditions can affect multi-step associative inference differently for neural networks with varying capacity and coding. A series of studies in which humans completed versions of the above tasks has suggested individual differences in performance and neural representations (Noh et al 2026; Schlichting et al., 2015; Zhou et al., 2023). Therefore, we examined whether the degree to which memory encoding was sparse versus distributed during learning can also bias representational change in ways consistent with human behavior in the relative distance judgment task (Noh et al 2026) and in the triad inference task (Schlichting et al., 2015; Zhou et al., 2023). To do this, we trained additional models varying in memory capacity, sparsity, and distributedness across the range described in the triad inference task and applied it to the judgment phase of the complex graph task. For each trial of the judgment of relative distances task, we calculated the representational change (ΔRS) resulting from blocked or interleaved schedules as the post-study minus pre-study change in cosine similarity (Δr) between the correct target with the source node ($\Delta RS_{correct} = \Delta r_{correct}\text{-}\Delta r_{source}$), as well as that for the cosine similarity between the incorrect target with the source node ($\Delta RS_{incorrect} = \Delta r_{incorrect}\text{-}\Delta r_{source}$). We then calculated how different the resulting value was for the correct pair minus the value for the incorrect pair ($\Delta RS_{correct}$ - $\Delta RS_{incorrect}$). Intuitively, the resulting value can be interpreted as the strength of the model's prediction which should be larger for easier versus harder trials.

Using this methodology, we measured the changes in representation for successful performance in the judgment of relative distances task. A multiple regression analysis suggests that across correct trials, there was a main effect of memory capacity (adjusted $R^2 = 0.17$, $F(7, 992) = 29.32$, $\beta = 0.57$, $p < 0.001$) as well as an interaction between schedule and memory capacity ($\beta_{schedule*memory} = -0.54$, $p < 0.001$; Fig. 6A). The interleaved schedule produced more integrated representations when memory capacity and sparsity was high. By contrast, the blocked schedule produced more integrated representations when memory capacity was high and the representations were more distributed. Furthermore, there was an interaction between memory capacity and sparsity, such that the more sparse the encoding, the less memory capacity influenced integration ($\beta_{memory*sparsity} = -1.87$, $p < 0.001$). The schedule also interacted with memory capacity and sparsity to predict integration: increased integration from the interleaved schedule, compared to the blocked schedule, was moderated by memory capacity and sparsity ($\beta_{schedule*memory*sparsity} = 2.64$, $p < 0.001$). These results suggest that how memory capacity is used to perform tasks relates to sparse and distributed properties of memory encoding.

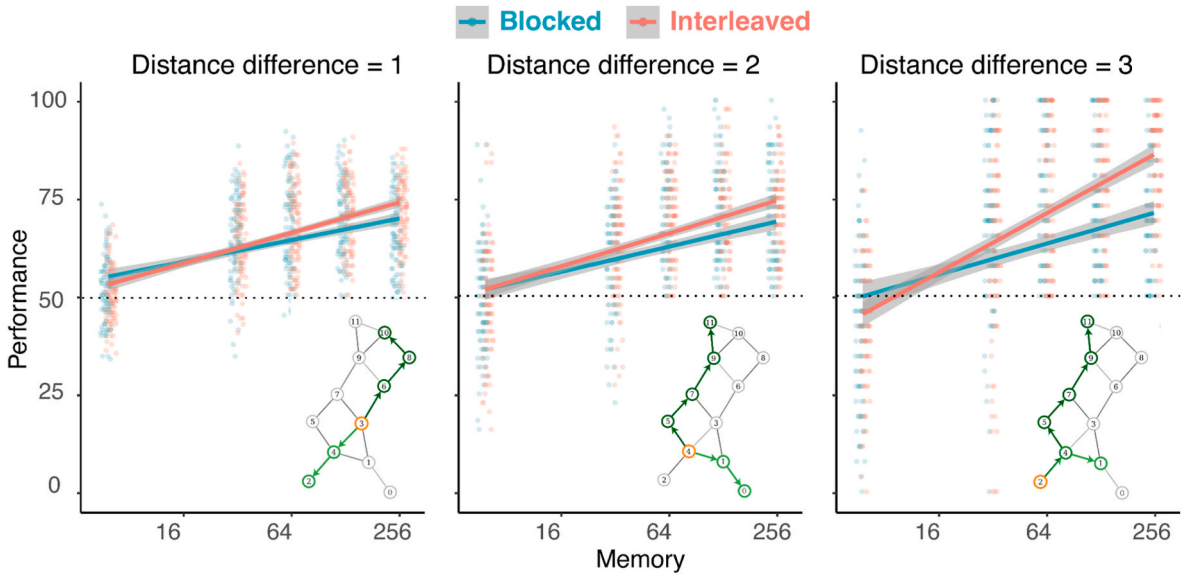


Fig. 5. Effect of schedule on task performance depends on memory. *Insets:* Examples of relative distance 1, 2, and 3 trials. The red node is the source node and the green path is the correct choice because the terminal node is more proximal to the source node than for the blue path. Plots illustrate improving task performance with increased memory capacity (plotted logarithmically) and chance performance marked by the dotted line. There is a main effect of memory capacity, where networks with higher memory tended to perform better. Networks with low capacity benefited from blocking, whereas those with high capacity benefited from interleaving. Easier (longer) distance judgments benefited more from interleaving than blocking. Here we show models with low sparsity constraints ($\alpha = 0$ to 0.3), which prioritize more distributed representations with greater information content. (For interpretation of the references to color in this figure legend, the reader is referred to the Web version of this article.)

Finally, we assessed representational change across different levels of task difficulty. The results of a multiple regression analysis suggest that the integration increased as the relative distances increased between options (e.g., as trials became easier) (adjusted $R^2 = 0.32$, $F(19, 8215) = 203.7$, $\beta = 0.12$, $p < 0.001$; Fig. 6B) and as memory capacity increased ($\beta = 0.43$, $p < 0.001$). In general, the interleaved schedule resulted in more separated representations than the blocked schedule ($\beta = -0.19$, $p < 0.001$). This effect was even stronger at higher memory capacities ($\beta_{schedule*memory} = -0.22$, $p = 0.001$). This apparent conflict with previous results where interleaving generally increased integration can be explained by sparsity limitations. The interleaved schedule led to more integrated representations relative to the blocked schedule when the representations were constrained to have an intermediate mixture of sparse and distributed representations ($\beta_{schedule*sparsity} = 0.44$, $p < 0.001$) and high sparsity ($\beta_{schedule*sparsity} = 0.17$, $p = 0.001$). Across all models, encodings with medium ($\beta = 1.08$, $p < 0.001$) and high ($\beta = 1.54$, $p < 0.001$) sparsity tend to have more separated rather than integrated encoding, consistent with the notion that sparsity mechanisms produce orthogonalized and separated representations. These results suggest that separation does not always result from high sparsity but rather also depends on the schedule and memory capacity. For example, the effect of the interleaved schedule on integrated representations was especially high when representations were both sparse and distributed and memory capacity was higher ($\beta_{schedule*capacity*sparsity} = 0.61$, $p < 0.001$). Together, these results show how sparse and distributed memory encodings permit the formation of separated versus integrated task representations across distinct learning conditions and as a function of individual differences in memory capacity.

4. Discussion

The present study aimed to provide a mechanistic account and framework of how learning sequence, representational capacity, and coding style jointly shape associative maps. When memory capacity is low, we hypothesized that presenting related episodes in a blocked sequence would allow AB associations to stabilize before BC is introduced. Consistent with this hypothesis, our models produce greater AB accuracy during and after blocked training relative to interleaved training (Supplementary Fig. 4), and this stabilization corresponds to increased integration in the low-capacity models. When capacity is high, the system can tolerate greater cognitive load, and interleaving overlapping pairs fosters cross-episode comparisons that allow for integration of similarities across related episodes (Zeithamova and Preston, 2017; Zhou et al., 2023). Coding style further modulates these dynamics: sparse codes reduce overlap and amplify blocking benefits, while distributed codes promote overlap and amplify interleaving benefits (Kumaran and McClelland, 2012). Together, these constraints help determine whether integrated vs. separated representations emerge after learning.

Our framework helps reconcile previously conflicting results in the literature. Schlichting et al. (2015) reported that blocked learning promoted integration, whereas Zhou et al. (2023) reported the opposite (interleaving → integration). While differences in memory capacity were not explicitly measured in the conflicting studies, there is evidence to suggest that these differences may have driven the conflicting findings between the two studies. Schlichting and colleagues used a "pure" schedule in which blocked and interleaved phases were learned separately, thus reducing interference pressures and allowing AB associations to stabilize before BC was introduced. Zhou and colleagues instead used a "hybrid" schedule in which blocked and interleaved pairs were intermixed within a single phase, substantially increasing cognitive load and interference pressures. The hybrid design implemented by Zhou et al. (2023) was associated with higher exclusion rates—up to 65% in some experiments—and substantially lower AC inference accuracy, suggesting that perhaps the results were biased by the exclusion of participants with lower memory capacity. Our model reproduces these patterns, showing that blocked learning generally encourages integration in low-capacity and medium-capacity networks, but interleaving promotes integration in high-capacity networks.

Our results are also consistent with prior work suggesting that implicit temporal structure could play a role during learning in humans, specifically for those with high memory capacity. Because blocking reduces interference by adding temporal distance between AB and BC pairs, the BC learning phase may trigger a new temporal context for

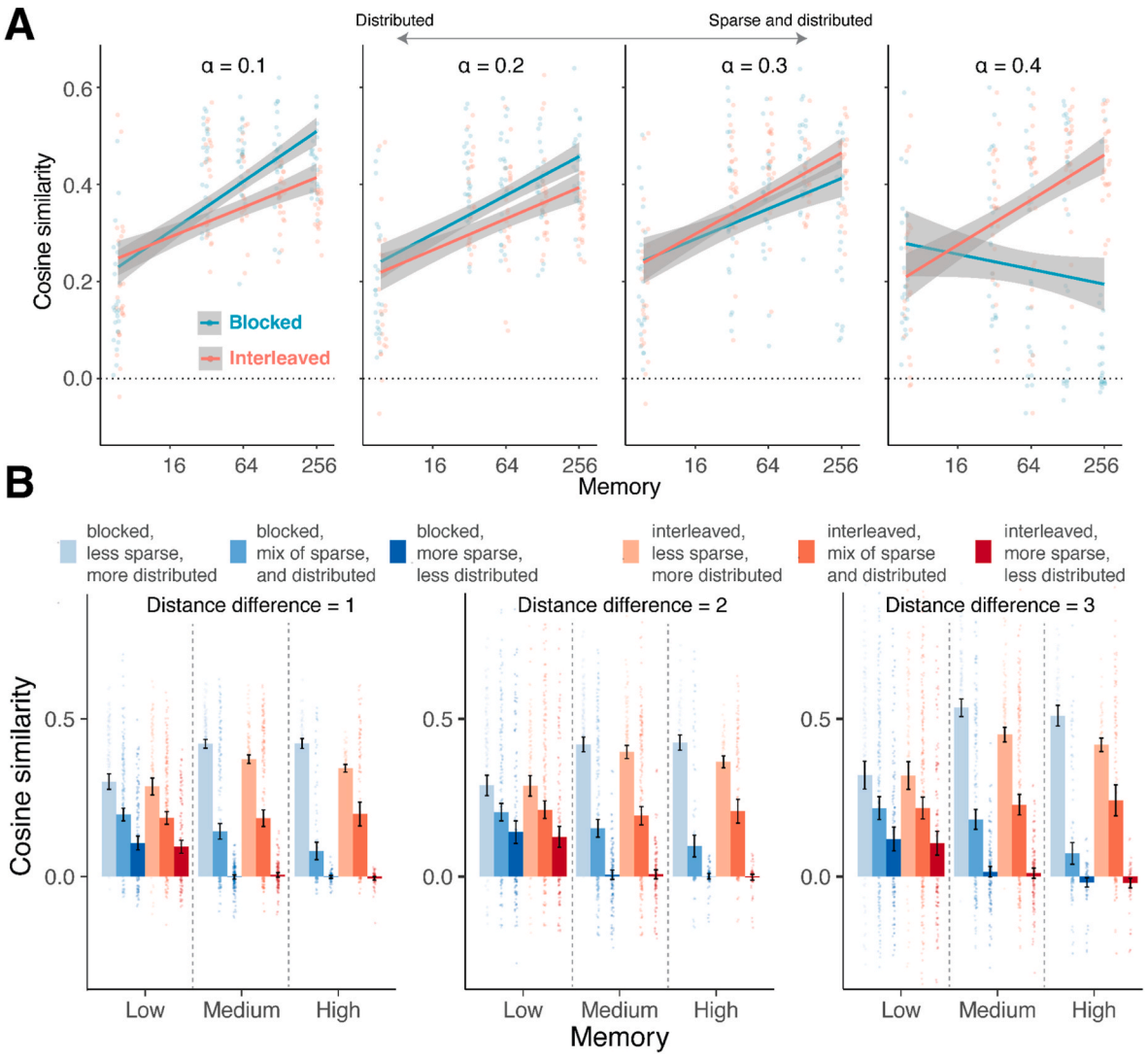


Fig. 6. Effect of sparse and distributed codes on integration of task representations. (A) For correct trials, the integration value is calculated as how much more similar the correct pair became after training compared to the incorrect pair. The relationship between integration and memory differs across different levels of α. As sparsity (α) increases towards a mixture of sparse and distributed representations, integration is more likely to occur in high memory capacity conditions after interleaved training. **(B)** An analysis of the full range of sparsity and entropy levels. When coding strategies have mixed sparsity and distributedness (mixed local and distributed coding), high memory capacity conditions support integration when trained via an interleaved schedule versus a blocked schedule. In contrast, when coding strategies are less sparse and more distributed, higher memory capacity conditions support integration when trained via a blocked schedule versus an interleaved schedule. These effects are consistent across levels of task difficulty (distance differences).

individuals with high memory capacity, which further separates this information from prior learning and make it difficult to draw inferences across episodes encountered across long time delays. On the other hand, interleaving reduces temporal distance between overlapping AB and BC pairs, which can encourage similarity-based integration for individuals with high memory capacity. Prior work also shows that individuals may implicitly encode temporal context and use it to organize representations (Pudhiyidath et al., 2022; Schapiro et al., 2013, 2016), and interestingly, these time-based relationships emerge even without any explicit representations of time in their models (Schapiro et al., 2013, 2016). With that said, because our model's architecture does not include an explicit temporal-context, it cannot be used here to adjudicate between different forms of temporal separation (e.g., elapsed time vs. number of intervening items). Instead, temporal proximity in humans may reflect the ease with which related information can be jointly represented within working memory or a shared neural context. In our networks, while "time" itself is irrelevant, alternating presentation order can serve an analogous computational role, as overlapping associations occupy a shared representational space. Thus, while temporal proximity may support integration in humans by maintaining access to multiple related traces in memory, our models may achieve a comparable outcome via concurrent representational overlap which is facilitated by the interleaved order of presentation. In high-capacity models, interleaving provides an order of presentation which may promote integration by encouraging concurrent representational overlap during training. This order of presentation is an abstraction of the observation that in humans, temporal proximity of related episodes facilitates integration (Zeithamova and Preston, 2017). While related to time, the model abstraction does not encompass the varied effects of temporal dynamics. Beyond the scope of the present models, future models could be developed to better dissociate the influence of time, intervening episodes, and presentation order in shaping memory representations during and after learning. For example, models that explicitly represent temporal context or use sequence models like recurrent neural networks can be used to test if temporal proximity benefits integration and how it may interact with the memory capacity and sparse encoding constraints

of the models (Beukers et al., 2024; Zhou et al., 2023). Extending our framework to include representation of temporal context or proximity as a feature (such as a decaying drift vector) could help to systematically test how the temporal dynamics present within blocked/interleaved schedules interact with capacity and coding to drive relational inference performance. This would be especially relevant given evidence that high-capacity individuals may maintain temporal structure to guide integration, whereas lower-capacity individuals may not (Noh et al 2026).

Beyond reconciling prior discrepancies, the present results generalize to more complex graph learning tasks. The same schedule × capacity × coding interactions that were observed in the simple triad tasks extended to multi-step structural inference performance in a more complex graph-learning design, providing even stronger evidence and application of our framework. Specifically, we find that even in the graph task, blocking promotes integration in low-capacity settings, whereas interleaving promotes integration in high-capacity environments. While the present study does not include behavioral metrics, our simulations are consistent with both behavioral and computational findings using the same graph learning task (Noh et al 2026). It is also important to note that the modeling approach used in prior work had notable differences from the present study, yet still provide converging results in favor of our framework (Noh et al 2026). Conceptually, the interaction between capacity, schedule, and coding style outlined in our framework provides a domain-general mechanism for organizing semantic, spatial, and temporal structure. It can also be loosely applied to predict coding biases at the regional level–such as sparse coding in dentate gyrus versus more distributed coding in CA3 (Treves and Rolls, 1994; Yassa and Stark, 2011), as well as differences at an individual level–such as working memory capacity, to explain divergent representational outcomes across studies and populations.

Limitations of the present work deserve emphasis and we caution against any direct mappings of our findings to specific biological pathways or phenomena. Our feedforward networks are intentionally minimal, isolating capacity and coding constraints and their impacts on resulting representations of relational information. Thus they do not include specific biological phenomena, such as consideration of complementary learning systems, recurrent dynamics, or explicit representations of temporal context. Catastrophic interference, which has been shown to preferentially impact blocked learning (Kumaran and McClelland, 2012), is exaggerated in machine learning compared to human learning. Operationalizations of "integration" and "performance" rely on representational similarity, which do not necessarily translate to better/worse behavioral performance in humans. Our aim was to computationally test whether differences in memory capacity and coding sparsity can interact with training sequences to produce distinct representational changes demonstrated previously in the literature (Zhou et al., 2023; Schlichting et al., 2015). Our results suggest that the neural network models produce the observed effects via a compression-based integration mechanism during which overlapping AB and BC associations become compressed to varying degrees as a function of capacity, sparsity, and schedule. Feed-forward networks with representational bottlenecks or sparsity constraints as implemented here are known to implement a compressive process to efficiently represent inputs, and one well-known example is an autoencoder (Kramer, 1991; Olshausen and Field, 2004). Future work could test the compression and alternative algorithmic processes by directly fitting a suite of such putative process models to human behavioral or neuroimaging data. The relative simplicity of the model, however, also represents a strength: by stripping assumptions to a minimum, our framework isolates the computational role of memory and coding constraints and unifies conflicting empirical findings. These potential limitations are further discussed below.

One potential concern regarding the interpretation of our findings might be that feed-forward networks such as those used here are highly susceptible to catastrophic interference, and especially so for lower capacity models under blocked training conditions (see **Supplementary Materials**, section on *Catastrophic interference as a possible modeling confound*). In this case, the observed pattern of AC integration and subsequent inference performance may simply be an artifact of increased interference or competition, wherein all representations become merged and increase in similarity. To address this possible concern, we further interrogated the input-output mappings of our model using a retrieval-style analysis after training. This analysis probed the likelihood of producing a B item when cued with A (see Supplementary Methods for full analysis details). This retrieval-based analysis confirmed that, although some competition is present after blocked learning (especially in lower capacity models), the direct associations are preserved (target B activations > foil B activations for corresponding A items; Supplementary Figure 5). The models' elevated AC performance in these conditions cannot be explained by catastrophic interference alone and instead points to a compression-based integration process that tolerates, and may even exploit, partial overlap in representations to support AC inference. Thus, catastrophic interference may indeed be a limitation of low-capacity neural networks but does not fully explain nor undermine the key findings. Nevertheless, because humans rarely forget AB pairs entirely, our framework should be interpreted as a proof of principle rather than a direct mapping to human performance. Furthermore, the loss of AB and BC information (on which the models were trained) does not necessarily imply a lack of learning–specifically, if AB and BC become integrated, it is reasonable to think that the specificity of individual AB and BC representations are lost or blurred, as would be expected from compressing these two individual episodes to a single representation (ABC). Thus, a related consideration is that integration may necessarily reduce AB/BC distinctiveness such that representational similarity analyses may misclassify integration as poor direct-pair memory. Our results support this possibility: we show that accurate AC inference can arise even when AB and BC similarity appear weak, consistent with findings that generalized representations persist even as individual memories fade (Brainerd and Reyna, 1990, 2001).

Consistent with existing behavioral studies, our models show that the process of integrating AB and BC representations to support successful AC inference may come at the cost of reduced AB/BC fidelity and discriminability (Carpenter and Schacter, 2017; Carpenter et al., 2021; Zhou et al., 2023; Liu et al., 2024). The feed-forward networks architecture we implement with representational bottlenecks or sparsity constraints are known to implement a compressive process to efficiently represent inputs (Kramer, 1991; Olshausen and Field, 2004). Under this model architecture, compressing representations leads to both updating and losing direct pair fidelity due to the bottleneck and sparsity constraints: ABC representations are compressed during learning to efficiently integrate related AB and BC representations. This compression-based integration is likely to be especially important in low capacity cases, as overlapping AB and BC associations converge onto a more efficient, but limited, latent code (AB + BC → ABC). This updated ABC representation sacrifices pair-specific fidelity (AB, BC) while preserving the relational structure needed for inference. This tradeoff of compressed integration and loss of fidelity—where the integration of ABC to support AC inference is linked to a rise in misattribution of specific item details or greater false memories for AB/BC source information—has been observed in several associative inference tasks within the field of episodic memory (Carpenter and Schacter, 2017; Carpenter et al., 2021; Liu et al., 2024; Zhou et al., 2023). Relatedly, even in the neural network literature, there is increasing recognition that error and lossiness are forms of adaptive distortions which cannot be explained as purely noise (or interference) (Zhao et al., 2021; Lin et al., 2024). These adaptive distortions can be explained by theoretical accounts of generalization via lossy compression (Brainerd and Reyna, 2015; Noh et al., 2024). This highlights the need for experimental designs that jointly measure direct-pair fidelity and inference performance, rather than examining post-training representations and performance. Future work should examine ways to quantify and dissociate these possibilities.

Finally, the learning that occurs in these models is sensitive to initial conditions. This is partly a feature not a bug, because we attempt to interpret the effects of such hyperparameters. For example, we manipulate regularization to operationalize coding schemes which influence the resulting variability in model weights and performance. Sensitivity to the specification of hyperparameters is not unique to our model, and such parameters are often neither explained nor biologically motivated. While prior models are also sensitive to hyperparameter settings (Zhou et al., 2023), the number of parameters is far larger in the neural networks we implement. Hence, the models serve as a useful proof-of-concept of seemingly conflicting computational-level memory and learning phenomena. It may be fruitful to further test these memory and encoding constraints in future work using more sophisticated models, such as implementations of complementary learning systems and biophysical signal propagation in the Leabra framework (O'Reilly et al., 2015).

Taken together, the present study provides a mechanistic explanation for why blocked learning may promote integration under some conditions while interleaving does in others. The results highlight the critical role of individual differences in capacity and coding strategies, as well as design differences across tasks, and how they can shape representational outcomes. Thus, we believe our general framework highlights the importance of considering these potential sources of variability when designing experiments and models to examine learning related representational changes. Potential extensions of our framework might include incorporating some of these ideas into the complementary learning systems framework, or examining the role of temporal context and its potential interactions with individual differences and sequencing effects. Important future directions include linking representational measures to neural and behavioral data in human participants, as well as testing moderators such as working memory capacity and age to predict which schedule optimizes integration, and also exploring interventions that bias coding toward sparse versus distributed representations across individuals. These extensions will help strengthen the link between computational mechanisms and individual variability in cognitive map formation, offering a stronger framework for understanding how episodic learning contributes to inference, generalization, and planning.

CRediT authorship contribution statement

Sharon M. Noh: Writing – review & editing, Writing – original draft, Visualization, Supervision, Project administration, Methodology, Investigation, Conceptualization. **Dale Zhou:** Writing – review & editing, Writing – original draft, Visualization, Supervision, Software, Project administration, Methodology, Investigation, Formal analysis, Data curation, Conceptualization. **Keiland W. Cooper:** Writing – review & editing, Supervision, Software, Methodology, Investigation, Formal analysis, Data curation, Conceptualization. **Shuheng Guo:** Investigation, Formal analysis, Data curation. **Emily T. Dinh:** Investigation, Formal analysis, Data curation. **Aaron M. Bornstein:** Writing – review & editing, Validation, Supervision, Resources, Project administration, Methodology, Funding acquisition, Conceptualization.

Citation Diversity Statement

Recent work in several fields of science has identified a bias in citation practices such that papers from women and other minority scholars are under-cited relative to the number of such papers in the field using bibliometric analyses (Dworkin, 2020; Bertolero, 2020). Using those analysis tools (Zhou, 2020), our references (excluding self-citations) contain 8.99% woman(first)/woman(last), 14.96% man/woman, 10.67% woman/man, and 65.38% man/man citations, as well as 5.96% author of color (first)/author of color(last), 11.58% white author/author of color, 19.85% author of color/white author, and 62.61% white author/white author.

Funding

This work was supported in part by the National Institute on Aging (F32AG072836 to S.M.N., R21AG072673 and R01AG088306 to A.M.B), the National Institute of Neurological Disorders and Stroke (R01NS119468 to A.M.B, PI E.R. Chrastil), the National Science Foundation (Graduate Research Fellowship to K.W.C.), and George E. Hewitt Foundation for Medical Research (to D.Z.).

Declaration of competing interest

The authors declare no competing financial interests.

Appendix A. Supplementary data

Supplementary data to this article can be found online at https://doi.org/10.1016/j.neuropsychologia.2026.109396.

References

Bakker, A., Kirwan, C.B., Miller, M., Stark, C.E.L., 2008. Pattern separation in the human hippocampal CA3 and dentate gyrus. Science 319 (5870), 1640–1642.
Barak, O., Rigotti, M., Fusi, S., 2013. The sparseness of mixed selectivity neurons controls the generalization–discrimination trade-off. J. Neurosci. 33 (9), 3844–3856.
Benna, M.K., Fusi, S., 2021. Place cells may simply be memory cells: memory compression leads to spatial tuning and history dependence. Proceedings of the National Academy of Sciences of the United States of America 118 (51), e2018422118.
Bennett, I.J., Stark, C.E.L., 2016. Mnemonic discrimination relates to perforant path integrity: an ultra-high resolution diffusion tensor imaging study. Neurobiol. Learn. Mem. 129, 107–112.
Bertolero, M.A., Dworkin, J.D., David, S.U., Lloreda, C.L., Srivastava, P., Stiso, J., Zhou, D., Dzirasa, K., Fair, D.A., Kaczkurkin, A.N., Marlin, B.J., Shohamy, D., Uddin, L.Q., Zurn, P., Bassett, D.S., 2020. *Racial and ethnic imbalance in neuroscience reference lists and intersections with gender*. bioRxiv.
Beukers, A.O., Collin, S.H.P., Kempner, R.P., Franklin, N.T., Gershman, S.J., Norman, K. A., 2024. Blocked training facilitates learning of multiple schemas. Communications Psychology 2 (1), 1–17.
Brainerd, C.J., Reyna, V.F., 1990. Gist is the grist: Fuzzy-trace theory and the new intuitionism. Dev. Rev.: DR (Dev. Rev.) 10 (1), 3–47.
Brainerd, C.J., Reyna, V.F., 2001. Fuzzy-trace theory: dual processes in memory, reasoning, and cognitive neuroscience. Adv. Child Dev. Behav. 28, 41–100.
Brainerd, C.J., Reyna, V.F., 2015. Fuzzy-trace theory and lifespan cognitive development. Dev. Rev. 38, 89–121.
Carpenter, A.C., Thakral, P.P., Preston, A.R., Schacter, D.L., 2021. Reinstatement of item-specific contextual details during retrieval supports recombination-related false memories. NeuroImage 236, 118033.
Cayco-Gajic, N.A., Clopath, C., Silver, R.A., 2017. Sparse synaptic connectivity is required for decorrelation and pattern separation in feedforward networks. Nat. Commun. 8 (1), 1116.
Cayco-Gajic, N.A., Silver, R.A., 2019. Re-evaluating circuit mechanisms underlying pattern separation. Neuron 101 (4), 584–602.
Chanales, A.J.H., Tremblay-McGaw, A.G., Drascher, M.L., Kuhl, B.A., 2021. Adaptive repulsion of long-term memory representations is triggered by event similarity. Psychol. Sci. 32 (5), 705–720.
Chandak, S., Shah, P., Borkar, V.S., Dodhia, P., 2024. Reinforcement learning in Non-Markovian environments. Syst. Control Lett. 185 (105751), 105751.
Chavlis, S., Petrantonakis, P.C., Poirazi, P., 2017. Dendrites of dentate gyrus granule cells contribute to pattern separation by controlling sparsity. Hippocampus 27 (1), 89–110.
Chrastil, E.R., Warren, W.H., 2014. From cognitive maps to cognitive graphs. PLoS One 9 (11), e112544.
Collett, T.S., Graham, P., 2004. Animal navigation: path integration, visual landmarks and cognitive maps. Curr. Biol.: CB 14 (12), R475–R477.
Dworkin, J.D., Linn, K.A., Teich, E.G., Zurn, P., Shinohara, R.T., Bassett, D.S., 2020. The extent and drivers of gender imbalance in neuroscience reference lists. Nat. Neurosci. 23 (8), 918–926.
Estes, W.K., 1955. Statistical theory of spontaneous recovery and regression. Psychol. Rev. 62 (3), 145–154.
Farrell, M., Recanatesi, S., Moore, T., Lajoie, G., Shea-Brown, E., 2022. Gradient-based learning drives robust representations in recurrent neural networks by balancing compression and expansion. Nat. Mach. Intell. 4 (6), 564–573.
Hinton, G.E., 1984. Distributed Representations. Carnegie-Mellon University, Computer Science Department.
Hinton, G.E., Ghahramani, Z., 1997. Generative models for discovering sparse distributed representations. Philosophical Transactions of the Royal Society of London. Series B, Biological Sciences 352 (1358), 1177–1190.
Howard, M.W., Kahana, M.J., 2002. A distributed representation of temporal context. J. Math. Psychol. 46 (3), 269–299.

Ito, T., Murray, J.D., 2023. Multitask representations in the human cortex transform along a sensory-to-motor hierarchy. Nat. Neurosci. 26 (2), 306–315.
Kramer, M.A., 1991. Nonlinear principal component analysis using autoassociative neural networks. AIChE J. 37 (2), 233–243.
Kumaran, D., McClelland, J.L., 2012. Generalization through the recurrent interaction of episodic memories: a model of the hippocampal system. Psychol. Rev. 119 (3), 573–616.
Leutgeb, J.K., Leutgeb, S., Moser, M.-B., Moser, E.I., 2007. Pattern separation in the dentate gyrus and CA3 of the hippocampus. Science (New York, N.Y.) 315 (5814), 961–966.
Leutgeb, S., Leutgeb, J.K., Treves, A., Moser, M.B., Moser, E.I., 2004. Distinct ensemble codes in hippocampal areas CA3 and CA1. Science 305 (5688), 1295–1298.
Lin, Q., Li, Z., Lafferty, J., Yildirim, I., 2024. Images with harder-to-reconstruct visual representations leave stronger memory traces. Nat. Hum. Behav. 8 (7), 1309–1320.
Liu, Z., Johansson, M., Johansson, R., Bramão, I., 2024. The effects of episodic context on memory integration. Scientific Reports 14 (1), 30159.
Mack, M.L., Love, B.C., Preston, A.R., 2018. Building concepts one episode at a time: the hippocampus and concept formation. Neurosci. Lett. 680, 31–38.
McClelland, J.L., Rumelhart, D.E., 1988. Explorations in Parallel Distributed Processing. MIT Press.
McNaughton, B.L., Battaglia, F.P., Jensen, O., Moser, E.I., Moser, M.-B., 2006. Path integration and the neural basis of the "cognitive map.". Nat. Rev. Neurosci. 7 (8), 663–678.
Morton, N.W., Sherrill, K.R., Preston, A.R., 2017. Memory integration constructs maps of space, time, and concepts. Curr. Opin. Behav. Sci. 17, 161–168.
Neunuebel, J.P., Knierim, J.J., 2014. CA3 retrieves coherent representations from degraded input: direct evidence for CA3 pattern completion and dentate gyrus pattern separation. Neuron 81 (2), 416–427.
Noh, S.M., Bjork, R.A., Preston, A.R., 2024. General knowledge and detailed memory benefit from different training sequences. J. Appl. Res. Mem. Cognit. 13 (3), 329.
Noh, S.M.*, Cooper, K.W.*, Guo, S., Zhou, D., Stark, C., Bornstein, A., in press. Multi-step inference can be improved across the lifespan with individualized memory interventions. The Journals of Gerontology, Series B: Psychological Sciences and Social Sciences.
Olshausen, B.A., Field, D.J., 2004. Sparse coding of sensory inputs. Curr. Opin. Neurobiol. 14 (4), 481–487.
O'Reilly, R.C., Hazy, T.E., Herd, S.A., 2015. In: Chipman, S.E.F. (Ed.), The Leabra Cognitive Architecture. Oxford University Press.
O'Reilly, R.C., Rudy, J.W., 2001. Conjunctive representations in learning and memory: principles of cortical and hippocampal function. Psychol. Rev. 108 (2), 311.
Pudhiyidath, A., Morton, N.W., Viveros Duran, R., Schapiro, A.C., Momennejad, I., Hinojosa-Rowland, D.M., Molitor, R.J., Preston, A.R., 2022. Representations of temporal community structure in hippocampus and precuneus predict inductive reasoning decisions. J. Cognit. Neurosci. 34 (10), 1736–1760.
Rigotti, M., Barak, O., Warden, M.R., Wang, X.-J., Daw, N.D., Miller, E.K., Fusi, S., 2013. The importance of mixed selectivity in complex cognitive tasks. Nature 497 (7451), 585–590.
Rmus, M., Ritz, H., Hunter, L.E., Bornstein, A.M., Shenhav, A., 2022. Humans can navigate complex graph structures acquired during latent learning. Cognition 225, 105103.
Schapiro, A.C., Rogers, T.T., Cordova, N.I., Turk-Browne, N.B., Botvinick, M.M., 2013. Neural representations of events arise from temporal community structure. Nat. Neurosci. 16 (4), 486–492.
Schapiro, A.C., Turk-Browne, N.B., Norman, K.A., Botvinick, M.M., 2016. Statistical learning of temporal community structure in the hippocampus. Hippocampus 26 (1), 3–8.
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. Nat. Commun. 6, 8151.
Schlichting, M.L., Preston, A.R., 2014. Memory reactivation during rest supports upcoming learning of related content. Proc. Natl. Acad. Sci. 111 (44), 15845–15850.
Schlichting, M.L., Zeithamova, D., Preston, A.R., 2014. CA1 subfield contributions to memory integration and inference. Hippocampus 24 (10), 1248–1260.
Tetzlaff, T., Helias, M., Einevoll, G.T., Diesmann, M., 2012. Decorrelation of neural-network activity by inhibitory feedback. PLoS Comput. Biol. 8 (8), e1002596.
Tolman, E.C., 1948. Cognitive maps in rats and men. Psychological review 55 (4), 189.
Treves, A., Rolls, E.T., 1994. Computational analysis of the role of the hippocampus in memory. Hippocampus 4 (3), 374–391.
Wiechert, M.T., Judkewitz, B., Riecke, H., Friedrich, R.W., 2010. Mechanisms of pattern decorrelation by recurrent neuronal circuits. Nat. Neurosci. 13 (8), 1003–1010.
Wilson, I.A., Gallagher, M., Eichenbaum, H., Tanila, H., 2006. Neurocognitive aging: prior memories hinder new hippocampal encoding. Trends Neurosci. 29 (12), 662–670.
Winocur, G., Moscovitch, M., Bontempi, B., 2010. Memory formation and longterm retention in humans and animals: convergence towards a transformation account of hippocampal-neocortical interactions. Neuropsychologia 48, 2339–2356.
Yassa, M.A., Stark, C.E.L., 2011. Pattern separation in the hippocampus. Trends Neurosci. 34 (10), 515–525.
Yoo, J., Chrastil, E.R., Bornstein, A.M., 2024. Cognitive graphs: representational substrates for planning. Decision 11 (4), 537–556.
Zeithamova, D., Dominick, A.L., Preston, A.R., 2012a. Hippocampal and ventral medial prefrontal activation during retrieval-mediated learning supports novel inference. Neuron 75 (1), 168–179.
Zeithamova, D., Preston, A.R., 2017. Temporal proximity promotes integration of overlapping events. J. Cognit. Neurosci. 29 (8), 1311–1323.
Zeithamova, D., Schlichting, M.L., Preston, A.R., 2012b. The hippocampus and inferential reasoning: building memories to navigate future decisions. Front. Hum. Neurosci. 6, 70.
Zhou, D., Cornblath, E.J., Stiso, J., Teich, E.G., Dworkin, J.D., Blevins, A.S., Bassett, D.S., Feb 2020. Gender diversity statement and code note-book v1.0. Zenodo.
Zhou, Z., Singh, D., Tandoc, M.C., Schapiro, A.C., 2023. Building integrated representations through interleaved learning. J. Exp. Psychol. Gen. 152 (9), 2666–2684.