Semantic representations in episodic memory enhance recall and compositional consolidation

Albert Albesa-González[1] and Claudia Clopath[1]

[1]Department of Bioengineering, Imperial College London, London, UK

October 17, 2025

Abstract

Episodic and semantic memory are classically thought to play distinct roles: episodic memory encodes unique experiences, while semantic memory generalizes across them. Current conceptualizations of episodic and semantic memory interactions emphasize a one-way consolidation from episodic traces in the medial temporal lobe (MTL) to semantic knowledge in neocortex (CTX). However, this tradition has left largely unexplored how semantic memory may affect episodic encoding of new memories. Here we introduce a cognitive model in which the code used in episodic memories shifts from purely sensorial to including explicit semantic representations of stored events. Simultaneously, we propose a computational circuit model of how such a cognitive strategy could be implemented in the brain using biologically-plausible learning rules. We show that increased sparsity during replay enables neocortex to extract compositional structure from overlapping episodes, which creates a dictionary of inter-connected concepts in semantic memory. Furthermore, we show that spontaneous activity in neocortical areas can imprint the abstracted representations into the medial temporal lobe, giving rise to concept-like cells. This bidirectional interaction improves episodic recall and replay fidelity, and facilitates the consolidation of higher-order representations based on previous semantic knowledge. The model accounts for behavioural advantages of schema-congruent learning, the emergence of concept neurons, and enhanced memory performance for semantically familiar stimuli. Together, our results provide a mechanistic account of how episodes and concepts reinforce each other, extending standard consolidation theories and suggesting a cooperative framework where semantic knowledge scaffolds episodic encoding, which in turn favours compositional abstraction.

Introduction

The distinction between episodic and semantic memory has been central to cognitive neuroscience since its first formalization (Tulving, 1972). In its original view, episodic memory captures individual experiences, while semantic memory accumulates generalized knowledge that emerges from repeated exposure to similar episodes (Squire, 2004). One influential computational explanation for this division is provided by the Complementary Learning Systems (CLS) framework. This theory posits that a fast-learning system (the hippocampus and the rest of the medial temporal lobe, MTL) specializes in rapidly encoding episodic memories. In contrast, a slow-learning neocortical (CTX) system works in concert with the MTL to consolidate and generalize memories (McClelland et al., 1995; O'Reilly et al., 2014).

Complementary Learning Systems and other consolidation models have historically emphasized a unidirectional transfer of knowledge from episodic traces (MTL) to semantic schemas (CTX) (Moscovitch & Gilboa, 2021). In this sense, models of the opposite (cortex-to-hippocampus teaching) have remained widely unexplored, as recently pointed out in Liu et al. (2024). However, humans often remember well-structured material more reliably than random or unfamiliar content, suggesting that top-down, schema-like influences can guide the encoding and retrieval of new events (Bartlett, 1932; Alba & Hasher, 1983; Lin et al., 2021). On the neurobiological side, research has identified so-called *concept cells* in the human MTL. These are neurons that fire selectively for particular semantic categories or even specific concepts (such as the famously named *Jennifer Aniston neuron*) (Quiroga et al., 2005; Quiroga, 2012; Courellis et al., 2024). Furthermore, regions in the medial temporal lobe such as the hippocampus and the entorhinal cortex have been shown to

contain neurons coding for latent elements in the environment. These especially include latents related to space (O'Keefe & Dostrovsky, 1971; Fyhn et al., 2004; Moser et al., 2008), but also other forms of complex task structure (Aronov et al., 2017; C. Sun et al., 2020; Zhao et al., 2022). Finally, the existence of long-range prefrontal-to-hippocampal inhibitory connections suggests that abstract information represented in neocortical areas can be directly conveyed to subcortical regions (Malik et al., 2022), with these regions exerting top-down control of subcortical areas during sleep (Swanson et al., 2020; Shin & Jadhav, 2024), typically associated with consolidation. All in all, these results suggest that abstract semantic information can be strongly represented in the Medial Temporal Lobe regions, despite its canonical role in episodic memory.

While views challenging a strict dichotomy between semantic and episodic memory are almost as old as the semantic-episodic distinction itself (McKoon et al., 1986; Greenberg & Verfaellie, 2010; Gentry & Buckner, 2024), we still lack a mechanistic understanding of how or why these two memory systems interact. First, most consolidation models assume that replay from hippocampus to cortex is a faithful reinstatement of cortical activity, often overlooking the biological reality that such replays are inherently noisy and the resulting cortical representations are actively constructed rather than simply copied (Swanson et al., 2020). Furthermore, highly semantized pre-existing representations are usually assumed (Saxe et al., 2019; Chrysanthidis et al., 2022) sidestepping how such mappings might arise through local, biologically plausible learning rules in the first place. Another key challenge is reconciling the presence of semantic representations in hippocampus and rest of MTL with attractor models of episodic memory. Semantic representations naturally induce overlapping patterns, which have been known for a very long time to reduce memory capacity. For this reason, consolidation models many times assume hash-like codes or other forms of perfect pattern separation, even when the episodes are conceptually highly overlapping (W. Sun et al., 2023; Chandra et al., 2025).

Here, we introduce a model of episodic and semantic memory interaction with two key processes: (i) episodic-to-semantic consolidation and (ii) semantic-to-episodic consolidation. Episodic-to-semantic consolidation is associated with the standard gradual transfer of information of episodic regularities into semantic memory. Conversely, episodic-to-semantic consolidation transfers these regularities into the episodic code, actively shaping the representations of future episodes based on previously formed abstractions. This cognitive model is presented together with a biologically-motivated circuit model of how such a cognitive algorithm could be implemented. The circuit, based on canonical models of subcortical-cortical consolidation, learns using only local signals via Hebbian and homeostatic plasticity. We use compositional input with a predefined statistical structure, which the neocortical component of the model is able to extract and decompose. These compositional representations are then transferred into the episodic memory system (Medial Temporal Lobe), increasing the semantic content of episodic representations. Our model can explain the formation of concept-like cells in subcortical areas (Quiroga et al., 2005; Quiroga, 2012), the advantage of blocked learning in humans (Flesch et al., 2018), and the better recall of memories with semantic content (Lin et al., 2021). Altogether, our model suggests that semantic memory can enhance episodic memory by transferring into it semantic representations. Furthermore, the presence of these semantic representations in episodic encoding facilitate the consolidation of concepts of higher complexity in semantic memory, supporting learning of higher-order abstractions. These results extend previous models of unidirectional transfer of information from hippocampus to cortex, pointing to a cooperative mechanism where both memory systems enhance each other.

Results

A Dual Cognitive-Circuit Model of Semantic and Episodic Memory

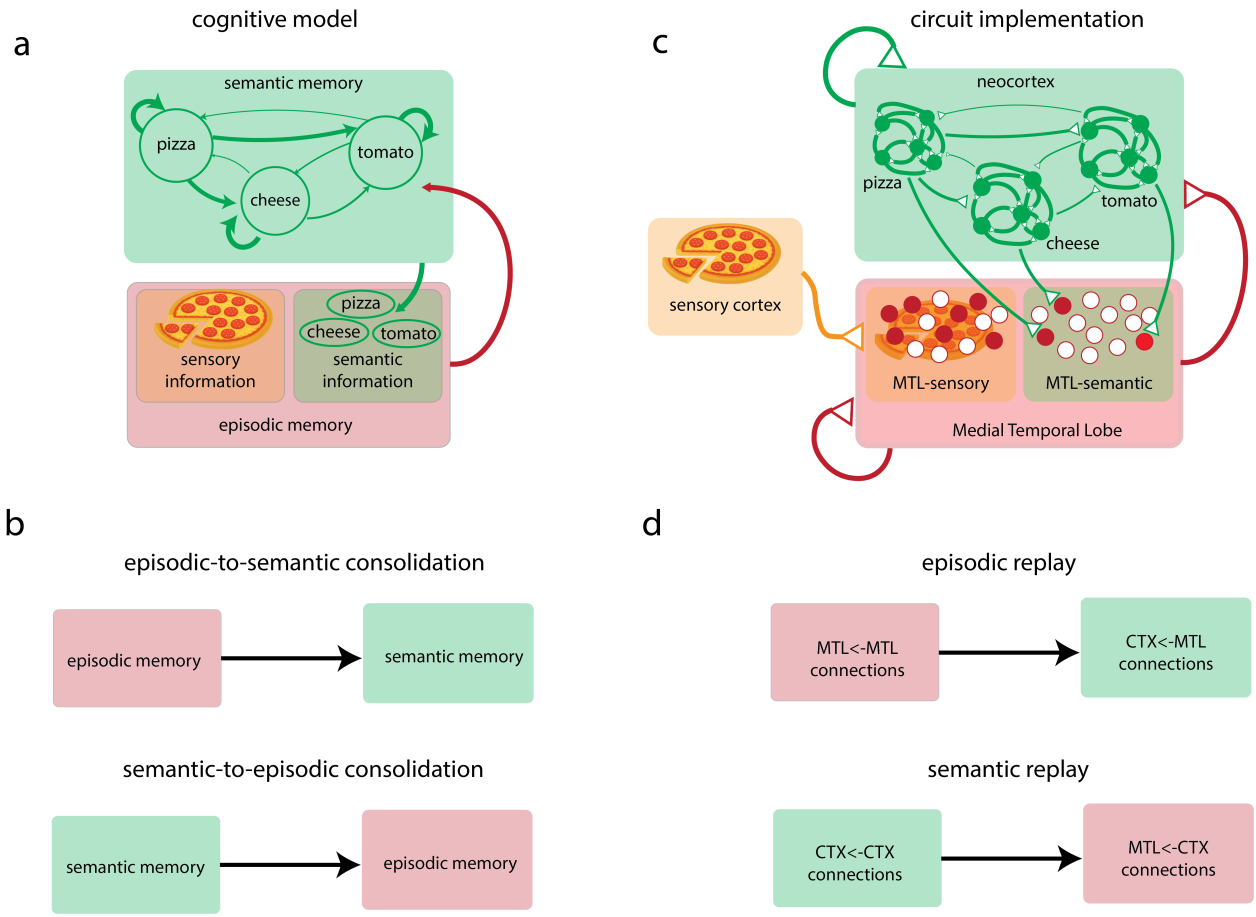

Figure 1: **We propose a cognitive model in which semantic memory shapes representations in episodic memory, and an associated circuit implementation.** **a**: Cognitive model of episodic (red) and semantic (green) memory. Episodic memories integrate both sensory information (orange) and semantic content (green), while semantic memory extracts regularities across episodes. **b**: Schematic of bidirectional consolidation. Episodic-to-semantic consolidation transfers regularities from episodic memory to semantic memory, while semantic-to-episodic consolidation builds semantic representations in episodic memory. **c**: Circuit implementation of the model. Neocortex (CTX) stores semantic memories and Medial Temporal Lobe (MTL) stores episodic memories. MTL is split into neurons primarily receiving input from sensory cortex (MTL-sensory, coding for low-level sensory features) and neurons primarily receiving input from neocortex (MTL-semantic, coding for high-level abstract features). **d**: Mapping of cognitive consolidation processes onto replay mechanisms. Episodic replay corresponds to CTX←MTL interactions (MTL-to-CTX connections trained during replay), while semantic replay corresponds to MTL-semantic←CTX interactions (semantic assemblies in CTX driving concept representations in MTL-semantic).

We introduce a cognitive model of semantic and episodic memory, along with an associated neural circuit that could implement it (Fig. 1). We begin with a simple observation: episodic memories often contain both sensory and semantic components. Consider the experience of being served your favorite food: pizza. Across different restaurants and recipes, the visual appearance, smell, and taste may vary widely, leading to diverse sensory inputs and memories. Yet, all of these episodes have something in common: the semantic notion of *pizza* and its main ingredients (crust, cheese, and tomato).

Our model captures this idea by positing that episodic memories contain two forms of representations: sensory and semantic. In the previous example, an episode involving a pizza (Fig. 1a) integrates the sen-

sory experience (orange), as well as its conceptual representation (green). Crucially, even though the model assumes that episodic memory contains semantic representations, the associations formed remain episodic: they link specific events in time.

In contrast, semantic memory is in charge of extracting regularities across episodes, gradually constructing a dictionary of abstract components. For these two systems to work in concert, we assume two memory transformation processes. First, the semantic system engages in *episodic-to-semantic* consolidation, extracting structure across experiences. Second, semantic memory shapes the representations used by episodic memory, undergoing *semantic-to-episodic* consolidation (Fig. 1b). Through this interaction, early episodic traces, initially dominated by sensory detail, contribute to building semantic knowledge. In turn, semantic knowledge is later used to guide and enrich the formation of new episodic memories. This bidirectional exchange gradually enables episodic representations to combine both sensory and semantic features -as our opening *pizza* example illustrates.

For the circuit implementation, we build upon classic mappings of semantic memory onto neocortical areas (CTX) and episodic memory onto hippocampal and parahippocampal regions (Medial Temporal Lobe, MTL). Following the cognitive model, we subdivide the MTL (episodic memory) into two functional subregions (Fig. 1c), one receiving input from sensory cortex (MTL-sensory), and one from higher-order neocortical areas (MTL-semantic).

An important question in the neuroscience of consolidation is how, starting from sensory input, semantic representations can be extracted in CTX (which in the cognitive model corresponds to episodic-to-semantic consolidation). In this sense, the first contribution of our model is proposing that, during *episodic replay*, connections from MTL to CTX are trained to extract repeated elements across episodes, building compositional representations in CTX. Once these representations have formed, CTX forms highly recurrently connected assemblies. These connections learn both what are the different semantic components and what is their causal relationship, completing the process of semantic extraction.

The second question we explore here is how semantic representations can later be transferred to MTL-semantic (semantic-to-episodic consolidation), forming concept-like representations. To answer this question, in addition to MTL spontaneous activity driving CTX (episodic replay), we include periods in which CTX spontaneous activity drives MTL (*semantic replay*, Fig. 1d), inspired by recent work on top-down control during sleep (Swanson et al., 2020; Shin & Jadhav, 2024). By doing this, neurons in MTL-semantic learn to reproduce the compositional structure of CTX.

In summary, we propose a model of how episodes and semantics can interact within the brain, supporting both the extraction of compositional knowledge from experiences and the imprinting of conceptual structure onto episodic traces. Next, we will delve deeper into the underlying circuit dynamics that enable these processes, and explore the biological implications and predictions arising from the model.

Episodic replay integrates sensory information into semantic memory

We first examine how episodic-to-semantic consolidation (Fig. 2a, left) might be mediated by episodic replay (Fig. 2a, right). Returning to our pizza example we ask: how can the experience of many episodes containing cheese lead to the formation of a *cheese* representation in CTX? Our circuit model proposes that, by increasing sparsity during replay, MTL can recover common sub-patterns across stored episodes. (Fig. 2a, right). This promotes the formation of decorrelated neural representations that can be easily learned by neocortex. After abstract cortical mappings have been established in CTX, semantic structure can be further extracted in its recurrent connections.

To test this idea, we generate sensory input following an Episode Generation Protocol as in Albesa-González and Clopath (2025) (also see Methods). Each episode consists of the simultaneous presentation of a stimulus A (one of A_1 to A_5) and a stimulus B (B_1 to B_5) (Fig. 2b). Each A_i and B_j is considered a distinct concept, so an episode is defined by a pair (A_i, B_j) -for example, in our illustrative example, A_1 could be crust and B_2 tomato, so (A_1, B_2) would correspond to an episode with a very basic pizza containing only these two ingredients. To enrich the semantic structure across concepts, episodes in which A and B share the same sub-index -such as (A_3, B_3)- occur 50% of the time. The remaining combinations are equally distributed in

the remaining 50%. This makes stimulus A_2, for example, more semantically related to B_2 than to other B_j's. According to the cognitive model, semantic memory should therefore build a *dictionary* of concepts in CTX (the full set of A_i and B_j) while also capturing the statistical regularities between them (e.g. A_1 is more likely to co-occur with B_1).

During episode presentation (*wake*, Algorithm 7), MTL←MTL connections undergo Hebbian plasticity (Algorithm 3) . During *sleep* (Algorithm 8) random initial activity in MTL evolves via sparse pattern completion (Algorithm 2). Higher sparsity helps uncover sub-patterns of neural activity that correspond to overlaps across different episodes (Fig. 2c). By reducing the number of active neurons during pattern completion dynamics driving replay, subsets of repeatedly co-active neurons are retrieved (for example those coding for A_1, see top-left corner in Fig. 2c). This activity in MTL is projected to CTX via CTX←MTL connections, which are initially random ($t = 0$ in Fig. 2d). Rapid one-shot plasticity induces an initial receptive field ($t = 200$ in Fig. 2d) that connects all-to-all between the recovered pattern in MTL and the resulting projection onto CTX. Subsequently, affected postsynaptic neurons follow slow Hebbian learning combined with homeostatic plasticity, which refines the original receptive field to match the statistical structure of the input ($t > 200$, Fig. 2d). After training, CTX←MTL connections converge to cover the entire set of episodic concepts A_1 to B_5 (Fig. 2e), making neurons in CTX replicate a de-noised and compositional representation of MTL-sensory.

Pairs of neurons that have formed a receptive field in CTX also follow Hebbian and outgoing homeostatic plasticity (see Methods) in CTX←CTX connections during *wake*. Importantly, under this form of plasticity corticocortical connections converge to the conditional firing probabilities (Albesa-González & Clopath, 2025). Because cortical neurons become tuned to individual concepts in episodes (Figs. 2e and 2f), capturing conditional firing probabilities is equivalent to capturing the semantic structure (see Albesa-González and Clopath (2025) and Methods) across concepts (Fig. 2g).

Together, these results illustrate how in our circuit model the semantic system (CTX) extracts regularities across overlapping episodic patterns, thereby implementing episodic-to-semantic consolidation. At the synaptic level, cortical feed-forward receptive fields tune neurons to abstract sensory representations, while recurrent connections organize into a block-like structure. This structure contains the semantic information, with intra-block connections representing concepts, and inter-block connections reflecting how much semantically related concepts are. In terms of our pizza example, blocks represent concepts like cheese or tomato, and connections between blocks represent the semantic connection between the ingredients (e.g. how likely is to have cheese in a dish if tomato is also present).

Semantic replay integrates semantic information into episodic memory

We now investigate the role of semantic replay in semantic-to-episodic consolidation. During semantic-to-episodic consolidation (Fig. 3a, left), semantic memory should transfer to episodic memory the abstracted representations. In our *pizza* example, after the different ingredients have been consolidated in CTX, these are mapped to MTL-semantic. For example, for someone who has never seen *tomato* sauce, the first episode would not contain a *tomato* explicit representation. However, after semantic-to-episodic consolidation, MTL-semantic would contain neurons explicitly coding for the presence of tomato. From the circuit perspective, MTL-semantic is expected to develop receptive fields that are highly selective to the different assemblies representing concepts in CTX (Fig. 3a, right).

To support this process, we now equip the simulated model with semantic replay (Algorithm 8). Semantic replay initializes CTX at random and allows it to recover its own attractors. As in episodic replay, sparsity is increased to isolate patterns corresponding to individual concepts (Fig. 3b). These patterns are then projected to MTL-semantic, which initially contains random connections. The rationale mirrors that of episodic replay: increasing sparsity produces non-overlapping patterns that can be efficiently learned via competitive learning. As a result, receptive fields in MTL-semantic come to represent individual concepts identified by CTX (Fig. 3c). Consequently, while MTL-sensory presents a mixed selectivity to different concepts, MTL-semantic neurons show an absolute selectivity to these concepts (Fig. 3d). This absolute selectivity is reminiscent of concept cells found in the human MTL (Quiroga et al., 2005; Rey et al., 2025).

We next hypothesize that this type of encoding may support more efficient downstream learning. To test this, we train a classifier to decode the concepts present in an episode from either MTL-sensory or MTL-

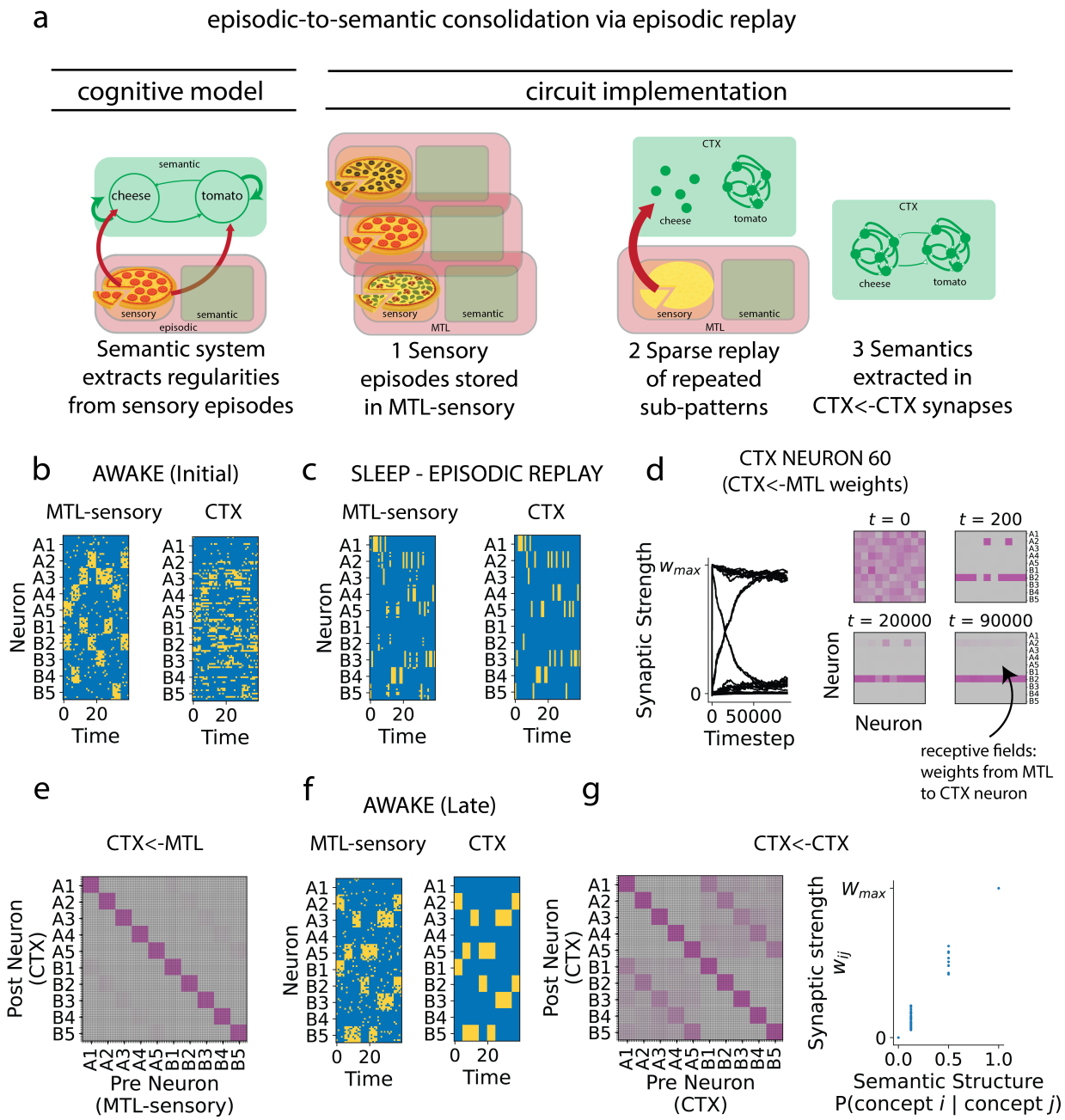

Figure 2: **Sparse episodic replay creates compositional representations in semantic memory. a**: Episodic-to-semantic consolidation via episodic replay. Cognitive model (left) and circuit implementation (middle/right). Sensory episodes (no semantics extracted yet) are stored in MTL-sensory, replayed sparsely during sleep to uncover overlapping sub-patterns, and then projected to CTX, where compositional semantic structure is extracted in CTX←MTL (replay) and CTX←CTX synapses (awake). **b**: Example neuronal activity during wake. Input episodes are sampled following a protocol as described in Albesa-González and Clopath (2025). To determine MTL-sensory activity, two different concepts (one $A_i \in A$ and one $B_j \in B$) are simultaneously sampled. 50% of episodes are of form $(A_i, B_{j=i})$, and 50% $(A_i, B_{j\neq i})$. This makes A_i more semantically related to B_i than the rest of B's. 10 neurons in MTL-sensory code for each concept, and each episode lasts 5 timesteps. At each timestep, $N_{\text{swaps}} = 4$ random neurons that code for the present concepts are turned off, and $N_{\text{swaps}} = 4$ random neurons that code for non-present concepts are turned on (this introduces noise with a consistent signal-to-noise ratio). CTX activity is driven by random initial CTX←MTL-sensory connections. **c**: Example neuronal activity during episodic replay. Sparse replay in MTL-sensory reveals concept-specific sub-patterns. These create structure in CTX←MTL-sensory connections via one-shot followed by slow learning (Fig. 2d).

Figure 2: *(continued)* **d**: Example CTX neuron receptive field (CTX←MTL synapses) over training. Images show weights associated to each presynaptic MTL-sensory neuron rearranged in a rectangular pattern according to the concept they code for (each row represents weights from neurons coding for a particular concept in MTL-sensory). Initially random connectivity ($t = 0$) is one-shot driven to represent a replayed presynaptic pattern ($t = 200$). Then, slow Hebbian and homeostatic plasticity drive statistical learning, uncovering receptive fields that represent semantic components ($t = 200 - 90000$). Right: weight dynamics for all synapses of the same neuron. **e**: Final CTX←MTL weight matrix. Neurons in CTX become tuned to individual concepts (A_1–A_5, B_1–B_5), forming a dictionary of abstract components. **f**: Example activity during late wake. CTX neurons now respond selectively to semantic components, showing compositional coding aligned with MTL-sensory input. **g**: Final CTX←CTX weight matrix (left) and relationship between synaptic strength and semantic structure (right). Recurrent connectivity captures conditional firing probabilities across concepts, organizing into block-like structure encoding semantic associations.

semantic representations (Fig. 3f). As a baseline, we also train a model using a one-hot encoding of the concepts present in each episode (*labels* in Figs. 3f-3i). We then compare the classification accuracy after a single epoch, using either sensory or semantic representations, under the assumption that semantic coding may facilitate rapid learning. Initially, the same episode statistics are maintained, and we refer to this as the *In-Distribution* condition. The classifier trained on MTL-semantic representations has a similar performance to training using labels (Fig. 3g), which is expected given both representations are highly aligned. In contrast, classification using MTL-sensory representations is less accurate, suggesting an advantage of using highly semantic codes in learning.

We then address the generalization capacity of MTL representations by introducing novel episode statistics. Specifically, we now exclude episodes of the form $(A_i, B_{j=i})$ during learning, and then sample MTL representations exclusively from this type of episode. Thus, the MTL activity used to train the classifiers is obtained with episodes that are *Compositional Out-Of-Distribution*. Again, classifiers trained on MTL-semantic activity match the performance of those trained on labels, while those using MTL-sensory activity perform worse (Fig. 3h). This highlights the compositional generalization enabled by semantic representations.

Finally, we examine a more extreme form of generalization by presenting a *Semantic Out-Of-Distribution* condition. Here, episodes are (fixed) random permutations of those shown during training. For this reason, episodes are no longer compositions of A and B, and no longer fall under Compositional Out-Of-Distribution. Under these conditions, MTL-semantic representations effectively become a lossy random projection, and no longer provide a reliable basis for decoding (Fig. 3i). Instead, MTL-sensory becomes essential for tracking novel input. This shows the two forms of representation serve complementary roles: MTL-semantic supports abstraction and compositional generalization, while MTL-sensory enables flexible adaptation to (semantically) unfamiliar sensory experiences.

Overall, these results demonstrate that our circuit model, via semantic replay, implements semantic-to-episodic consolidation, as proposed in the cognitive model. This provides a mechanistic account for how semantic representations can be transferred from cortical to subcortical areas, supporting the presence of concept-like cells in MTL (Quiroga et al., 2005; Quiroga, 2012). Moreover, MTL-semantic representations facilitate learning both In-Distribution and Compositional Out-Of-Distribution, while MTL-sensory remains essential for adapting to novel stimuli lacking semantic structure (Semantic Out-Of-Distribution).

The model explains the formation of concept cells starting from highly mixed representations, which is favored by blocked learning

We now explore whether the representations formed in the model are consistent with experimental observations in the literature. To this end, we modify the structure of MTL-sensory, which is now a random projection of the patterns used in previous sections (Fig. 4a and Methods). In the previous MTL-sensory representations, overlaps were completely non-systematic and always due to noise. By randomly projecting the original input (SEN in Fig. 4a), neurons can be activated by different concepts in a persistent manner, with MTL-sensory neurons becoming partially tuned to multiple concepts. While the previous arrangement

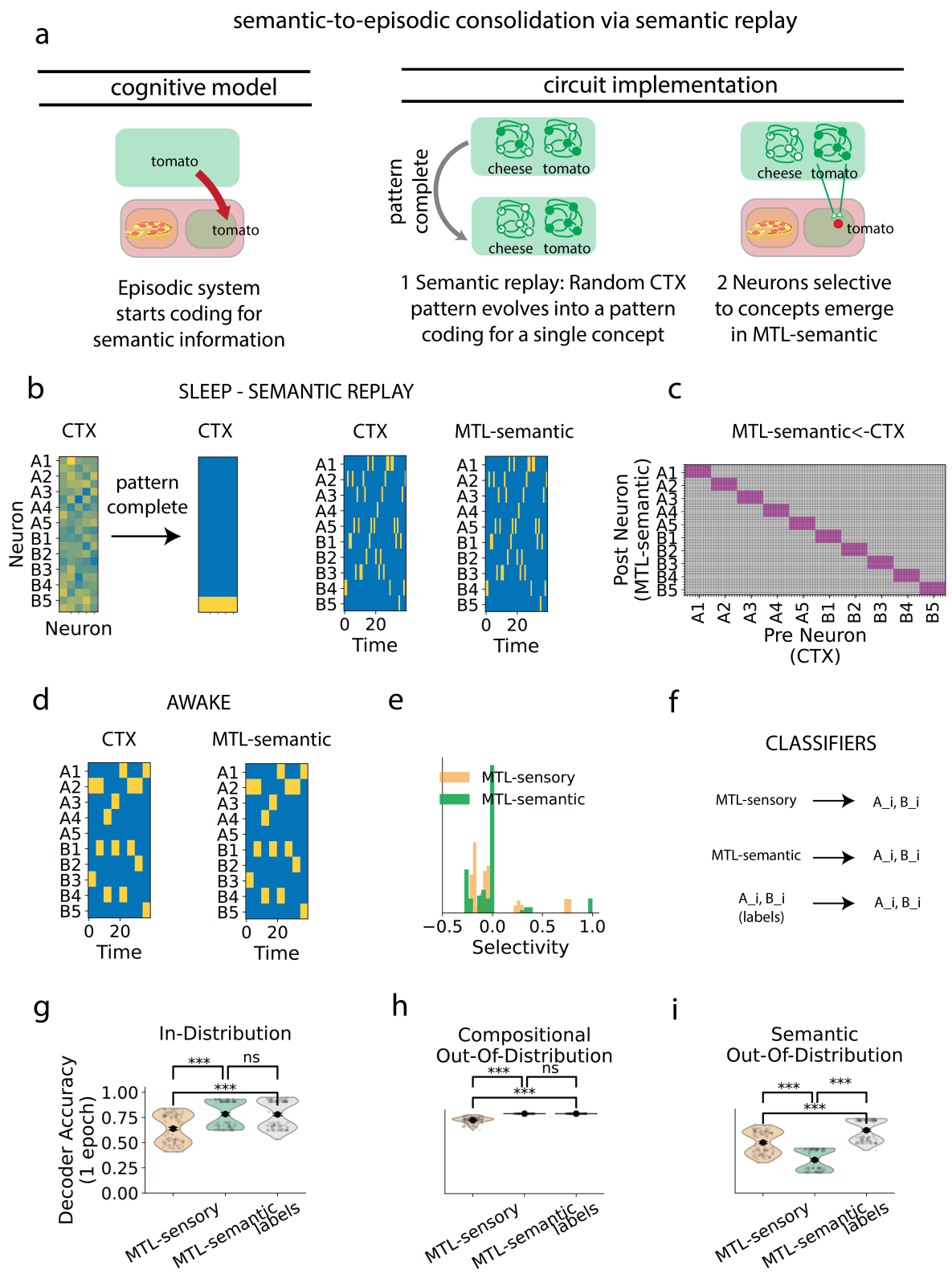

Figure 3: **Semantic replay forms abstract representations in episodic memory.** **a**: Semantic-to-episodic consolidation via semantic replay. Cognitive model (left) and circuit implementation (right). During semantic replay, a random CTX pattern converges to a concept-specific attractor, which is then projected to MTL-semantic, creating semantic-specific receptive fields in MTL-semantic and enabling episodic traces to incorporate semantic information. **b**: Example neuronal activity during semantic replay. A random initial CTX state evolves into a complete pattern coding for a single concept, which drives activity in MTL-semantic. A combination of one-shot and slow learning again make neurons that are randomly slightly more sensitive to cortical patterns selective to cortical semantic representations. **c**: Final MTL-semantic←CTX weight matrix, showing selective mapping from cortical semantic codes to MTL-semantic neurons.

Figure 3: *(continued)* **d**: Example activity during wake after semantic replay. MTL-semantic neurons now code selectively for concepts, complementing MTL-sensory representations. **e**: Selectivity index of neurons in MTL-sensory and MTL-semantic populations, showing that semantic replay induces sharper concept selectivity in MTL-semantic. **f**: Schematic of classifier training. Classifiers decode concept identity (A_i, B_j) from either MTL-sensory or MTL-semantic activity. **g–i**: Decoder performance across conditions. **g**: In-distribution decoding shows both MTL-sensory and MTL-semantic reliably code for concepts. **h**: Compositional out-of-distribution generalization. MTL-semantic neurons enable successful decoding of unseen concept combinations, while MTL-sensory neurons fail. **i**: Semantic out-of-distribution generalization. MTL-semantic neurons outperform MTL-sensory neurons in decoding novel semantic relationships. Statistical comparisons: *** $p < 0.001$, ns = not significant.

facilitated analysis and visualization, the distribution of neuronal selectivity in the medial temporal lobe is considerably broader than in Fig. 3e. By projecting the original input, we obtain a more realistic range of selectivity (Fig. 4c, left, green).

Our first question is whether concept-like representations still emerge in MTL-semantic with this type of sensory input. We find a clear increase in MTL-semantic selectivity when compared to MTL-sensory (Fig. 4c, left). However, some concepts collapse into a single representational pattern (see Fig. S1a, right), which hinders the formation of the associated concept neurons. This is a known problem in correlation-based learning, which is still substantially bypassed here (due to the decorrelation from sparse episodic replay), but not completely.

Following this result, we test the hypothesis that -as in humans (Flesch et al., 2018)- *blocked learning* promotes the acquisition of compositional structures (Figs. 4b and 4f, Methods). Blocked learning involves presenting successive *blocks* of episodes in which one concept is held fixed while the others vary. In contrast, *interleaved learning* maintains the full feature distribution consistently across training. In our simulations, the blocked condition begins with a brief phase in which the network is exposed to sequential batches, each corresponding to a fixed concept (from A_1 to B_5, Fig. 4b). We observe that in the model this training structure enhances the selectivity of learned representations (Fig. 4d, 4e and Figs. S1a and S1b) and improves accuracy (Fig. 4f) when using MTL-semantic activity directly as a classifier (see Methods).

To better understand these effects, we study the replayed representations in MTL-sensory at the beginning of learning (when semantic representations have not formed yet). Then, we define *concept prototype* (Methods) as a pattern representing the average neuronal activation that corresponds to a fixed concept. By comparing replayed MTL-sensory activity and concept prototypes (Fig. 4g) we can see how blocked learning promotes replaying activity that is closer to concept prototypes (Fig. 4h). This makes the initial one-shot receptive fields in CTX←MTL more selective and less susceptible to representational collapse.

Our findings demonstrate that our model not only supports the emergence of concept-like representations in MTL, but also highlights the importance of input structure during learning. In particular, blocked learning facilitates the formation of more selective prototype-aligned representations during replay, enhancing the formation of abstract representations in CTX.

Semantic representations in episodic memory explain semantic facilitation of recall and predict more faithful experience replay

Then, we study the impact of semantic representations on episodic recall. Empirical studies have shown that reducing the semantic component of a stimulus, for example, by scrambling the phase of an image (Lin et al., 2021) - significantly impairs its memorability (Fig. 5a). Conceptually, this is akin to the greater difficulty of recalling a random sequence of letters compared to a familiar word of the same length.

To demonstrate this behaviour in our model, we choose to mimic the original *Intact* vs *Phase-Scrambled* conditions in Lin et al. (2021). We compared episodic recall performance between a pre-trained network (containing semantic representations, which we call *Intact*), and one in which the MTL-semantic←CTX connections have been permuted (condition *Scrambled*, to otherwise maintain any other statistic the same,

Methods). In both networks, we presented 5 distinct episodes, each lasting 5 timesteps. We then initialized MTL-sensory with one of its states during episode presentation, and allowed all of MTL (sensory and semantic) to follow attractor dynamics, while maintaining sparsity levels to enable recall of full episodes. Recall performance was quantified as the cosine similarity between the recovered pattern and the original (Fig. 5b and Methods).

We observe episodes with a clear semantic component are recalled significantly more accurately than those without it, even though both networks have identical sparsity levels (Fig. 5c). This behaviour persists as the number of stored episodes increases (*Intact* vs. *Scrambled* in Fig. 5d). It should be noted how, on the one hand, recall within the semantic component is higher than in the sensory component. However, the sensory component is also recalled better in the presence of intact semantic representations than in the scrambled case (compare orange curves for *Intact* and *Scrambled* in Fig. 5d). This suggests that semantic representations can act as an error-correcting mechanism in MTL-sensory, given that connections span all-to-all across MTL (both sensory and semantic). Furthermore, remarkably, the presence of structured correlations in MTL deviates the system from classic memory capacity regimes (Dubreuil et al., 2014; Kang & Toyoizumi, 2023; Chandra et al., 2025). In this sense, here there is no memory cliff (i.e., no sudden drop to near-zero recall after a specific number of episodes). In both *Scrambled* and *Intact*, recall decreases and then starts increasing with the number of stored episodes, consistent with a transition from memorization to generalization. This U-shaped curve, however, is more pronounced for MTL-semantic than MTL-sensory (compare gap between green and orange dashed lines), pointing to semantic representations facilitating generalization. It should be noted, however, this also results in incorrect recall when the number of episodes is very high and episodes are of the form $(A_i, B_{j\neq i})$ (Fig. S2), resulting in over-generalization.

Given that semantic representations error-correct pattern completion during episodic recall, we next hypothesize MTL-semantic could also improve the quality of replayed activity. Until now, during episodic replay, sparsity was increased to decorrelate replay dynamics, yielding sub-patterns shared across episodes rather than full episodes. However, the presence of semantic structure may allow the recovery of episodic patterns, where the number of active neurons matches that observed during *wake*. We tested this by following a similar approach to the *recall* analysis, but starting from random noise instead of a previously seen episode. Because no specific episode was to be recalled, replay performance was measured taking the pattern with the maximum cosine similarity (interpreting the associated pattern as the recalled one). Simulations confirmed the hypothesis, showing a higher overlap between replayed activity and the stored episodes in the *Intact* condition (Fig. 5e). Furthermore, including MTL-semantic representations during episodic replay increased their own accuracy (Fig. 5f), supporting the idea of an abstraction loop: initial semantic representations in MTL improve replay quality, which in turn refines cortical semantic representations, finally reflected back in MTL-semantic representations).

Starting from a validation of previous experimental data (Figs. 5a and 5c), we obtain two key predictions of our model: one behavioural and one physiological. The behavioural prediction (derived from Fig. 5d) is that when episodes contain semantically familiar content, recall initially decreases with increasing memory load but eventually recovers, plateauing at a higher level. The physiological prediction is that replay of semantically structured episodes is more strongly correlated with awake activity than replay of semantically incomprehensible material (Fig. 5e).

Semantic representations during replay favour consolidation of higher-order compositional concepts

Until now, episodic replay has been constrained to increased levels of sparsity, which can decorrelate episodic sub-patterns and uncover compositional structure. However, by increasing the number of active neurons one could in principle recover patterns that correspond to higher-order associations of concepts. For example, when presented with multiple episodes containing pizza, the semantic components *crust*, *cheese*, and *tomato* can be replayed together by allowing 3 times more active neurons (Fig. 6a). Our hypothesis is that, following the results in the previous section, given that less sparse replay is stabilized by semantic representations, CTX can form new concepts that correspond to combinations of previously learned concepts.

We test the network's ability to use variable sparsity rates during replay to consolidate in CTX categories that are compositions of other categories. To this aim, we define the *Semantic Load* as the ratio between

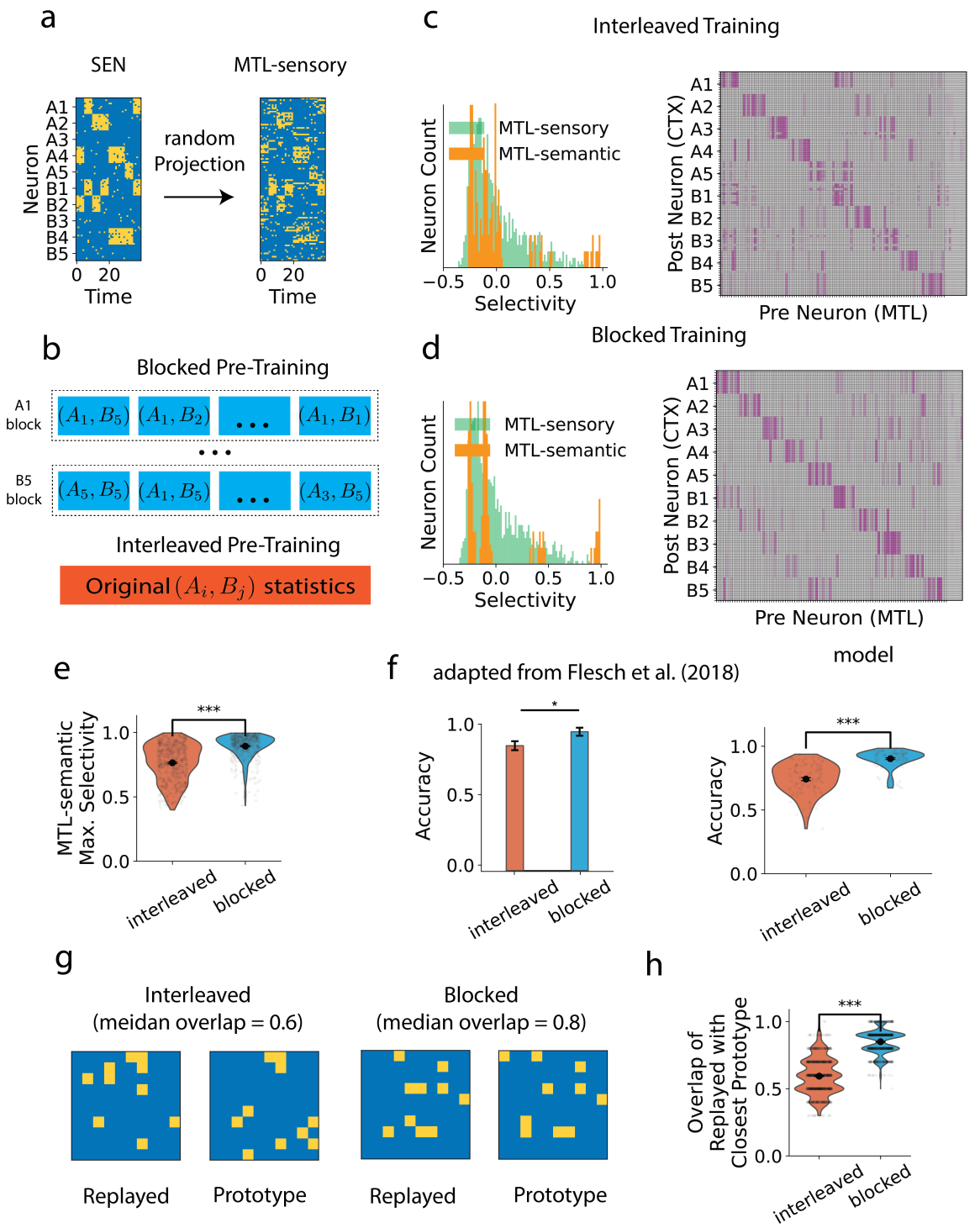

Figure 4: **Semantic abstraction is favoured by blocked training a**: Example sensory-to-MTL transformation. Sensory input (SEN) is projected through random connections into MTL-sensory, producing noisy but structured population codes. **b**: Training protocols. Interleaved training presents episodes with original (A_i, B_j) statistics preserved. Blocked training presents episodes grouped by concept pairs $(A_1, B,\ A_2, B,$ ...), disrupting original statistics. **c**: Interleaved training results. Left: selectivity distributions for neurons in MTL-sensory (green) and MTL-semantic (orange). Right: final CTX←MTL-sensory weight matrix. **d**: Same as **c** for blocked training. **e**: Quantification of MTL-semantic maximum selectivity across training conditions. Interleaved training produces higher selectivity than blocked training (*** $p < 0.001$). **f**: Behavioral and model comparison. Left: human data adapted from Flesch et al. (2018), showing increased accuracy under blocked training. Right: model reproduces this effect, with higher accuracy in blocked vs. interleaved training (*** $p < 0.001$).

Figure 4: *(continued)* **g**: Example replayed patterns in MTL-sensory for interleaved (left) and blocked (right). At the right of each replayed pattern, the conceptual prototype activity that is closer to the replayed representation. Under blocked training, replayed patterns align closely with semantic prototypes; under interleaved training, replay is noisier and less aligned. **h**: Overlap of replay with closest prototype. Blocked training yields higher overlap than interleaved training (*** $p < 0.001$).

the number of active neurons during replay at a given time and the maximum possible level of sparsity in that region. Intuitively, the semantic load is a measure of how many concepts are present in a representation with a specific sparsity. Then, at each replay event there is a random probability of the semantic load being 1 (one concept replayed) or 2 (two simultaneous concepts replayed) (Fig. 6b). Thus, an initial MTL pattern follows in any case the attractor dynamics driven by recurrent connections, but depending on the maximum number of allowed active neurons falls in an attractor coding for one single or two concepts (top and bottom panels in (Fig. 6b). MTL activity is further projected to CTX, where a new subregion has been added to drive the discovery of patterns across pairs of concepts.

As previously, replay with a semantic load of 2, also induces a competition between patterns corresponding to full episodes -specific combinations (A_i, B_j). Eventually, selective cortical receptive fields emerge for specific episodes (Fig. 6d). This modifies cortical representations, which now contains during *wake* neurons coding for each A_i, for each B_j, and for each (A_i, B_j) simultaneously (Fig. 6c). This completes the process of semantic extraction in CTX, which now stores in its recurrent synapses the full structure of the different compositional units (Fig. 6e). Because these synapses represent conditional probabilities across concepts, and now there are different levels of abstraction (for example concept A_1B_1 implies concept A_1, but given concept A_1 there is a probability of 1/5 of concept A_1B_1), the resulting matrix is highly asymmetric. As has been recently pointed in (Albesa-González & Clopath, 2025), the asymmetric semantic relationships that exist across hierarchies of concepts cannot be captured by classic symmetric Hebbian-based rules (Hopfield, 1982; Tsodyks & Feigel'man, 1988), but can be obtained by including homeostatic plasticity.

Finally, we investigate whether semantic representations during replay are causally involved in higher-order learning. This is suggested by our previous results (Fig. 5e and Fig. 5f), which showed replay and consolidation are enhanced by the presence of semantic representations. Thus, we again randomly permute MTL-semantic←CTX connections during higher-order concept learning. Then, we measure the selectivity of formed CTX representations to particular episodes (combinations of two concepts). We observe a significant decrease in episode selectivity in the region of CTX forming higher-order representations (Fig. 6f), showcasing the relevance of semantic information in consolidating compositions of previously learned concepts.

In summary, these results demonstrate that varying sparsity during replay enables the network to flexibly traverse the abstraction hierarchy, from single-concept representations to higher-order compositions of multiple concepts (akin to consolidating more episodic-like patterns). Crucially, the causal perturbation analysis highlights that semantic representations actively scaffold the emergence of higher-order structure. This matches the intuition that learning and retaining new combinations of previously learned concepts is much easier than when the concepts are novel. For example, long-term remembering the ingredients of a pizza that contains never-seen-before regional vegetables is much more complicated than remembering one containing chocolate and honey.

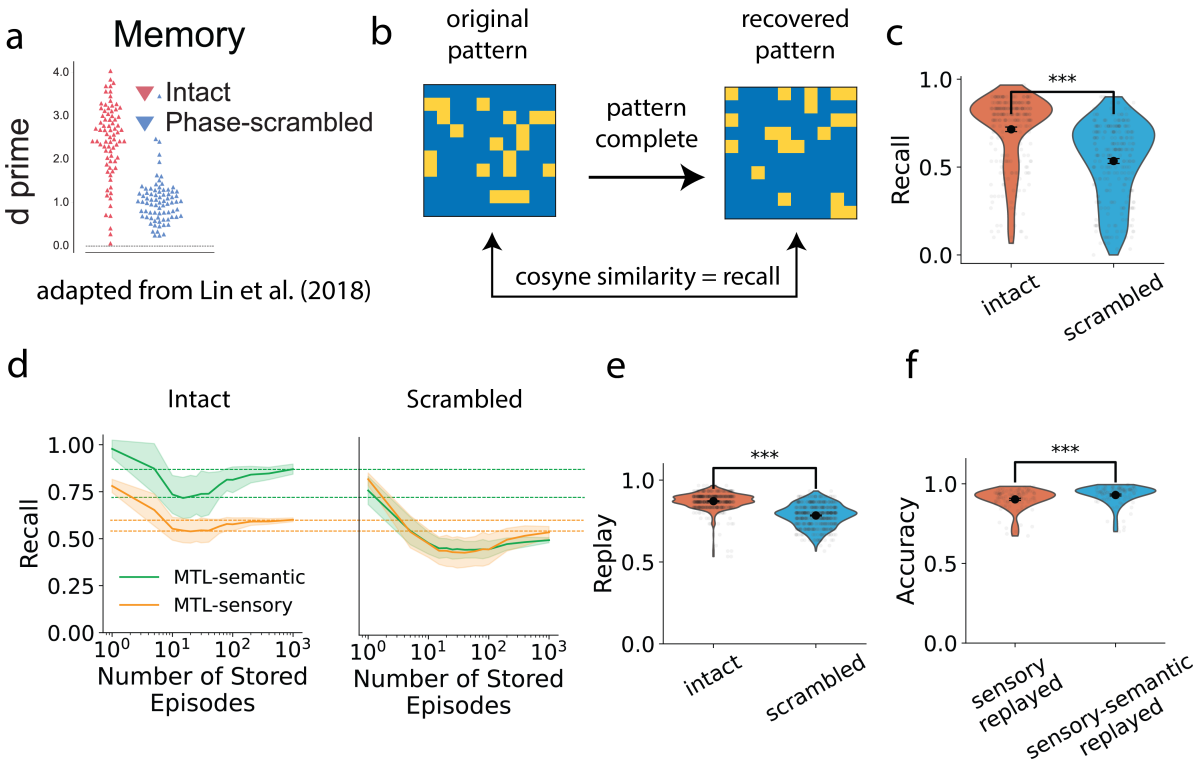

Figure 5: **Semantic representations structure pattern completion dynamics. a**: Behavioral benchmark. Human data adapted from Lin et al. (2018) show reduced memory performance (d-prime) for phase-scrambled compared to intact stimuli. **b**: Schematic of recall measure in the model. Cosine similarity between the original input pattern and the recovered pattern after pattern completion using MTL recurrent connections defines recall accuracy. **c**: Model recall performance. Recall is significantly higher for intact than scrambled patterns (*** $p < 0.001$). **d**: Recall as a function of memory load. Due to the correlation structure across patterns, the weight matrix reaches a stable fixed point, resulting in a converging recall performance. In the case of *Intact*, convergence is at a higher value than the performance trough, as a result of semantic representations capturing episodic structure and error-correcting pattern completion dynamics. **e**: Replay quality. Cosine similarity of replayed patterns to the closest input episode is higher for intact than scrambled stimuli (*** $p < 0.001$). **f**: Accuracy of MTL-semantic representations after extending episodic replay to full (sensory-semantic) replay, compared to initial MTL-semantic accuracy (only sensory representations used during episodic replay). (*** $p < 0.001$). Figs. 5c-e are obtained for the network with best accuracy in Fig. 4f, see Fig. S3 for results on the network with median accuracy (key results still hold).

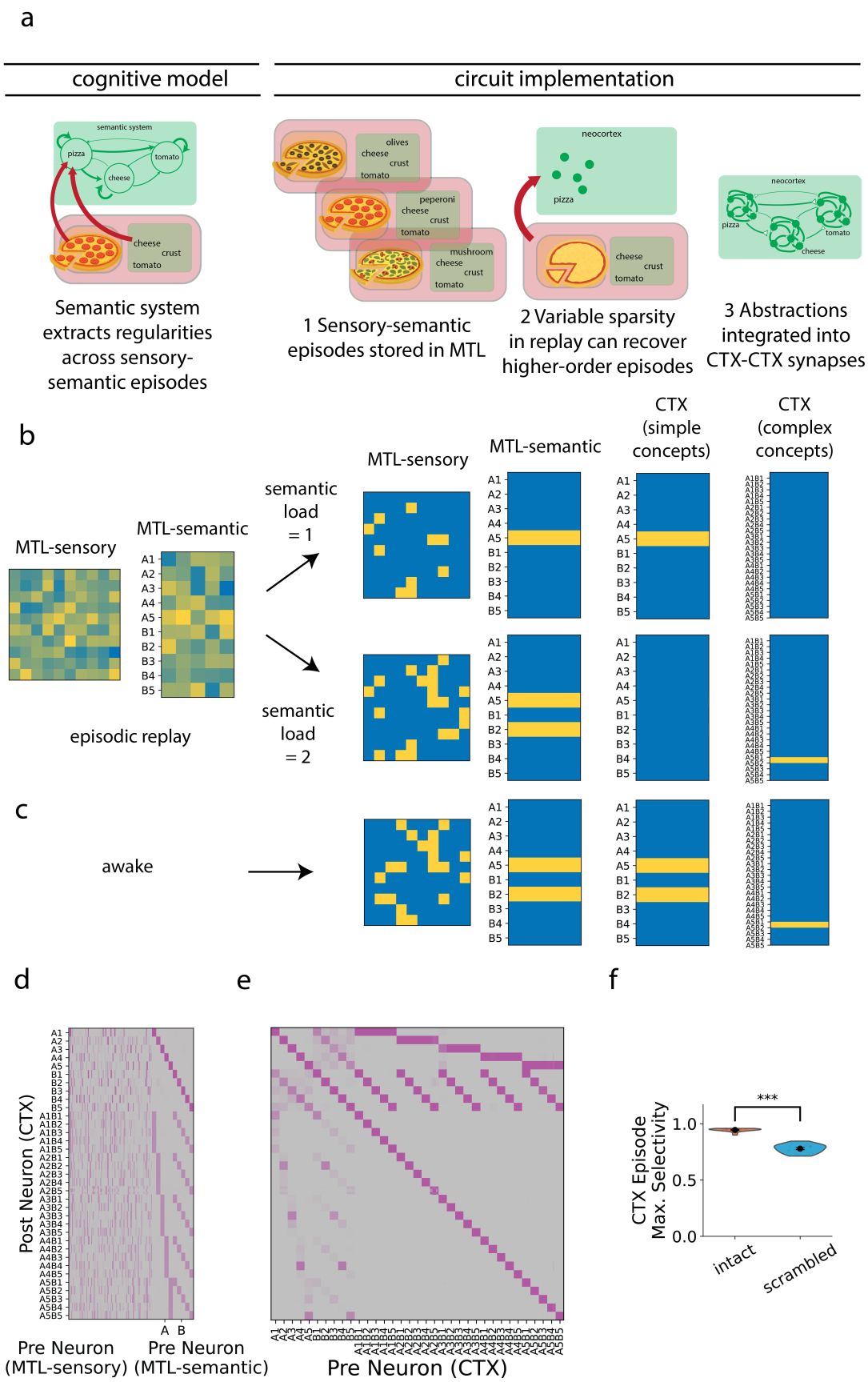

Figure 6: **Learning higher-order concepts based on simpler representations is facilitated by the presence semantic representations during replay.**

Figure 6: *(continued)* **A**: Cognitive model (left) and circuit implementation (right) of higher-order concept learning. Multiple sensory–semantic episodes are stored in MTL. Variable sparsity during replay (semantic load) allows the recovery of higher-order associations, which are then consolidated into CTX←MTL and CTX←CTX connections.**B**: Example activity during replay with semantic load 1 (top) or 2 (bottom). When one concept is replayed, CTX develops selective coding for single concepts. When two concepts are replayed simultaneously, CTX develops receptive fields for their combinations, supporting higher-order abstractions. **C**: Example activity during wake after training. CTX contains neurons selective to individual concepts A_i and B_j, as well as to combinations (A_i, B_j). **D**: Final CTX←MTL weight matrix. Cortical neurons integrate inputs from both MTL-sensory and MTL-semantic at different levels of abstraction. **E**: Final CTX←CTX weight matrix, showing asymmetric recurrent connectivity consistent with conditional probabilities across concepts. Block-like and diagonal substructures capture both simple and compositional units. **F**: Causal role of semantic replay. Random permutation of MTL-semantic←CTX connections during higher-order learning significantly reduces episode selectivity in CTX (*** $p < 0.001$), demonstrating that semantic representations actively scaffold the emergence of compositional structure. Figs. 5b-e are obtained for the network with best accuracy in Fig. 5f, see Fig. S4 for results on the network with median accuracy (key results still hold).

Discussion

In this study, we introduce a cognitive strategy by which episodic memory is improved through leveraging previously extracted semantic representations. Our model posits that, following an episodic-to-semantic consolidation process, semantic-to-episodic memory transfer reshapes the code of episodic memory itself. We further propose a biologically plausible neural circuit that implements such a cognitive model, successfully capturing the associated learning algorithm. Our circuit model aligns with and explains experimental data, and highlights the behavioural advantages of such a bidirectional learning mechanism.

First, we demonstrated how episodic replay from hippocampal and para-hippocampal (medial temporal lobe, MTL) regions effectively supports semantic consolidation in cortical (CTX) networks. While this initial step of consolidation corresponds to the most widely studied phase of learning across subcortical-cortical interactions (Alvarez & Squire, 1994; McClelland et al., 1995; O'Reilly et al., 2014; W. Sun et al., 2023; Guerreiro & Clopath, 2024; Spens & Burgess, 2024) (to name just a few), our model exhibits key differences from previous models. To start, most models assume a faithful replay of awake subcortical and cortical activity, with abstraction happening directly in CTC←CTX (neocortex to neocortex) connections. This is problematic for two reasons: first, the assumption that subcortical replay can reinstate perfectly previous cortical activity is not given (Swanson et al., 2020). Secondly, this relies on meaningful (to some degree) representations already existing in neocortex, when it is precisely via semantic consolidation that these representations should be created in the first place. For example, in Chrysanthidis et al. (2022), semantics are also extracted by forming block-like structure of recurrently connected neurons, but the high specificity of these neurons is already given. Furthermore, many studies assume a copy-paste mechanism between cortex and hippocampus, either directly via one-to-one connections (Alvarez & Squire, 1994; Tomé et al., 2022; Guerreiro & Clopath, 2024; Spens & Burgess, 2024) or perfect pattern-separating random connections (W. Sun et al., 2023), with less emphasis placed on how synapses mapping one to the other can be formed with biologically plausible learning mechanisms (although previous effort has been made in the context of backpropagation-like error-driven learning rules (Singh et al., 2022)). In contrast, here we have proposed that increasing sparsity levels during replay can lead to CTX←MTL synapses becoming selective to frequent episodic sub-patterns. This effectively decorrelates the training input provided to these connections, which could otherwise lead to representational collapse using standard Hebbian mechanisms.

Next, we proposed a mechanism whereby the semantic structure imprinted in CTX←CTX connections can be transferred into the coding scheme of MTL. Both the notion that semantics are extracted in CTX (McClelland et al., 1995) and that the MTL encodes a latent representation of the environment (Quiroga et al., 2005; Aronov et al., 2017; Behrens et al., 2018) have been present in the literature for a long time, but a mechanistic account of how these representations emerge in MTL remains unknown. We put forward the hypothesis that, after extracted in CTX, spontaneous cortical activity could transfer information encoded in cortical synapses to MTL←CTX synapses, inducing a semantic code in MTL. We called this process semantic-to-episodic consolidation, and conceptualized it as a mirror of episodic-to-semantic consolidation. As such, instead of transferring episodic information contained in recurrent MTL connections to CTX representations, it transfers semantic information present in recurrent CTX connections to MTL representations. The idea that cortex-to-hippocampus teaching is an underexplored topic in complementary learning systems literature was recently highlighted in Liu et al. (2024), and implicitly explored in Spens and Burgess (2024). In the latter study, the authors explored the possibility of using a Hopfield network to store pairs of sensory-latent representations for episodic recall, in a similar manner that our MTL-sensory and MTL-semantic neurons form joint attractors in MTL recurrent connections. Again, however, the process by which these joint representations were formed was not fully specified, with the model instead employing a copy-paste strategy to generate them. A model that also studied the formation of joint sensory-semantic representations as attractors in hippocampus is the Tolman Eichenbaum Machine (Whittington et al., 2020). In this study, the information in the sensory input was already highly semantized (i.e. using a one-hot encoding of sensory features), and the abstract graph position of the stimuli also given a priori. The mechanism for obtaining these representations in the first place was also not explored. In addition to establishing a mechanism for concept cell formation in MTL, our results show that this form of representations is more suitable for rapid learning than mixed representations, both in In-Distribution and Compositional Out-Of-Distribution environmental statistics. This allows MTL to compositionally encode never-seen episodes as long as these correspond to rearrangements of previously seen semantic representations. This is in agreement to previous results in continual learning in artificial neural networks (Chaudhry et al., 2020). We also showed how main-

taining sensory low-level representations of the environment is crucial for encoding episodes that are Semantic Out-Of-Distribution (that is, when new episodes cannot be encoded using previous semantic representations).

Our model has shown how even in non-trivial, highly overlapping scenarios like a random projection of the original input one can obtain concept cells in MTL. Moreover, the model offers a mechanism via which the formation of these representations is favored by blocked (as opposed to interleaved) learning, as demonstrated in humans (Flesch et al., 2018). In our model, blocked learning provides the MTL with many different examples of episodes containing the same concept, which are averaged out in its recurrent connections, facilitating the replay of its corresponding prototypical representation (Posner & Keele, 1968). The model also explains why recall of episodes with a higher semantic component is better than that of episodes that, from the perspective of the learned semantic structure, appear random (Lin et al., 2021). In classic models of recall based on attractor models, overlaps between episodes is seen as detrimental, reducing the network capacity (Hopfield, 1982; Treves & Rolls, 1994). However, semantic representations have a structured overlap, where the overlap is not stochastic but instead corresponds to the two episodes containing semantic overlaps. Thus, rather than reducing recall, this overlap acts as an error-correcting code (Albesa-González & Clopath, 2025). Moreover, semantic representations not only show better recall, but also actively improve the recall of the sensory component. Similarly, semantic representations induce replaying activity patterns that resemble more activity during wakefulness. These results can help explain why learning schema-congruent information is much faster than that of completely new episodes (Tse et al., 2007). While many models have focused on explaining how faster neocortical learning could happen at the cortex level (McClelland, 2013; Guerreiro & Clopath, 2024), our results suggest that this could be in part because the replayed activity in the first place is less noisy. Understanding the mechanisms of positive transfer learning is one of the cornerstones of modern artificial intelligence, where inducing models to use previously learned information to learn new tasks is still a challenge (Iman et al., 2023).

Following this potential link between replay involving semantic representations and continual learning, we showed that closing the loop (performing episodic-to-semantic consolidation of episodes that contain both sensory and semantic representations) allows the formation of higher-order compositional representations in CTX (Kurth-Nelson et al., 2023), in turn inducing a complete semantic description of the causal links across concepts in different levels of abstraction (Barlow, 1972). Not only higher-order representations (for example pizza is a conjunction of the different ingredients it contains) were developed, but CTX←CTX connections captured the semantic relationships across these different hierarchical levels (such as the fact that pizza is much more related to cheese than cheese is to pizza). Crucially, these forms of asymmetric relationships cannot be captured by classic models of Hebbian learning, as recently shown in Albesa-González and Clopath (2025). Finally, our framework might be related to the phenomenon of infantile amnesia (Alberini & Travaglia, 2017). In the model, the semantic load quantifies how many distinct concepts are bound within a neural representation, ranging from low (highly conceptual) to high (highly episodic) values. Furthermore, we showed that consolidating representations with a high semantic load (those combining many concepts) requires pre-existing semantic structure to scaffold them. It could be that early in development, when semantic representations are still forming, this scaffolding is absent, making the consolidation of rich, episodic memories difficult. Consequently, infants may initially consolidate only simple, concept-like representations (e.g., family members or locations), while conjunctive memories combining these elements fail to stabilize. As semantic representations mature, episodes involving multiple known concepts can finally be consolidated. It would be interesting to study whether this coincides with the end of infantile amnesia.

The model faces several limitations at different levels. For example, the model omits hash-like, episode-specific representations in episodic memory (as have been observed experimentally (Kolibius et al., 2023; Chettih et al., 2024) and included in previous models (W. Sun et al., 2023; Chandra et al., 2025)). An open question in neuroscience is how episodic representations solve the trade-off between perfect pattern separation and using overlapping patterns, more aligned with the hippocampus as a cognitive map that includes behaviourally-relevant latent representations. Further studies using a combination of shared and non-shared neuronal subspaces across similar episodes could explore a combination of both strategies might facilitate learning and optimal behaviour (efforts in this direction have been made before (Schapiro et al., 2017)). Additionally, while we have slightly moved away from completely orthogonal input representations usually used in biologically-plausible mechanistic models of consolidation, the input used here was still relatively simple, with more complex representations still requiring gradient-based methods. How the brain might use local plasticity rules to learn from more realistic input remains an open problem.

The results presented here could finally be of significance for modern artificial neural networks. In the recent years, many authors have highlighted episodic memory as a missing piece in many Artificial Intelligence (AI) paradigms (Fountas et al., 2025; Dong et al., 2025; Pink et al., 2025). Our model suggests that episodic memory should not simply be added as a fast external buffer, but rather be continuously reshaped by semantic knowledge. This bidirectional interaction yields two advantages directly relevant for AI: (i) more robust recall of past experiences through semantic "error correction," and (ii) the ability to rapidly learn new combinations of previously known concepts, supporting compositional generalization. Such designs may allow modern neural networks to leverage prior structure in order to accelerate continual learning and adapt flexibly to new environments.

Altogether, our study shows that bidirectional consolidation can be captured by a biologically plausible circuit model relying only on local plasticity rules. We demonstrated how sparsity during replay enables cortical extraction of compositional structure, how semantic feedback supports the emergence of concept cells in MTL, and how the presence of semantic codes enhances recall, replay quality, and higher-order concept learning. These results extend standard consolidation models by providing a mechanistic explanation for the formation of semantic representations in MTL and their role in scaffolding both new episodic and semantic memories.

Methods

Episode Generation Protocol

The Episode Generation Protocol (EGP) is a generative process that samples neuronal activity in the sensory layer. Our proposed EGP follows a similar approach to that described in Albesa-González and Clopath (2025), but is outlined here again for completion:

Episodes, episode attributes and episode concepts

All *episodes* share a structure, such that every episode contains one *episode concept* per *episode attribute*. This means that, given episode attributes A and B, with each episode attribute representing a set of episode concepts $a \in A, b \in B$, then the set of all possible episodes is defined as:

$$E = \{\{a, b\} \mid a \in A, b \in B\} \tag{1}$$

In other words, an episode $e \in E$ is a collection of episode concepts a, b such that element a belongs to episode attribute A and element b belongs to episode attribute B. In this study we have fixed a to have 5 possible values $a \in \{A_1, A_2, A_3, A_4, A_5\}$ and likewise for $b \in \{B_1, B_2, B_3, B_4, B_5\}$.

Episode Probability Distribution

As a generative process, our EGP samples episodes as a previous step to sampling sensory input. We do this by fixing a probability distribution over episodes:

$$1 = \sum_{A_i, B_j} P(e = \{A_i, B_j\}) \tag{2}$$

such that $P(e = \{A_i, B_j\}) \equiv P(i, j)$ imposes the likelihood of an episode with latents (A_i, B_j) to be sampled when generating episodes.

Episode-Input Mapping

Ultimately, the EGP samples sensory layer activity. For this reason, one also has to define how a sampled episode is mapped into sensory input X^{SEN}. Every concept c in $A \cup B$ has a set of associated neurons SEN_c in X^{SEN} such that, given an episode e

$$X_i^{\text{SEN}} = \begin{cases} 1 & \text{if } \ i \in \text{SEN}_c \ \text{ and } \ c \in e \\ 0 & \text{else} \end{cases} \tag{3}$$

In order to define SEN_c, we specify the number of attributes $N_{\text{attributes}}$, in this case 2 (A and B). Then, we define the number of concepts per attribute ($|A|$ and $|B|$, here 5 and 5, from this follows the fact that $A_{i \in [1,5]}$ and $B_{j \in [1,5]}$). Then, we fix the number of neurons per concept N^A_{concept} and N^B_{concept}, and define SEN_c as:

$$\text{SEN}_c = \{i : \text{start}(c) \leq i < \text{stop}(c)\} \tag{4}$$

with $\text{start}(c) = (i-1) \cdot N^A_{\text{concept}}$ if $c = A_i \in A$ and $\text{start}(c) = |A| \cdot N^A_{\text{concept}} + (j-1) \cdot N^B_{\text{concept}}$ if $c = B_j \in B$, and similarly $\text{stop}(c) = i \cdot N^A_{\text{concept}}$ if $c = A_i \in A$ and $\text{stop}(c) = |A| \cdot N^A_{\text{concept}} + j \cdot N^B_{\text{concept}}$ if $c = B_j \in B$ (i.e. we generate a list of one-hot encodings for each concept c).

To account for variability between different presentations of the same episode (not all episodes, even though they contain the same concepts, will be exactly the same), we further add some stochasticity by randomly picking N_{swap} inactive neurons and N_{swap} active neurons and inverting their activity (a total of $2N_{\text{swap}}$ neurons randomly change their activity). This ensures that activity sparsity in the sensory layer is maintained (the number of swaps from 0 to 1 is the same as the number of swaps from 1 to 0).

Input temporal structure

Generated input X_{input} has dimensions $(N_{\text{days}}, T_{\text{day}})$. Every day, a total of number $T_{\text{day}}/T_{\text{episode}}$ episodes (of T_{episode} length) are sampled, and within each timestep an independent noise realization is sampled.

Semantic Structure of an EGP

Following Albesa-González and Clopath (2025), we refer to *semantic field theory* (Bussmann et al., 2006) in order to define what is the *semantic structure* of an EGP. According to this theory, the meaning of a word is not isolated but dependent on its relation to the rest of the words. While our task is not one of language, we can use this same paradigm to define the *meaning* of episode concepts. In this sense, the meaning (semantics) of our episode concepts depends on how they are related to the rest (also note this is how semantics are established in a language dictionary, where each concept is defined via its relationship with the rest of concepts).

Following this proposal, we use the conditional probabilities of being present in an episode between episode concepts as a proxy for the semantic structure of an EGP. In other words, extracting the semantics of an EGP is equivalent to extracting how likely is one episode concept y to be present in an episode e if an episode concept x is also present. We define the Semantic Structure $\mathcal{SS}$ of an EGP with N^{concepts} different concepts c_i, as as a $N^{\text{concepts}} \times N^{\text{concepts}}$ matrix of the form:

$$\mathcal{SS}_{ij} = P(c_i \in e | c_j \in e) \tag{5}$$

For example, the intuition that pizza has a stronger semantic connection to cheese than cheese to pizza is reflected in their conditional probabilities (it is very common to have cheese in a pizza, but much less to have pizza in any dish containing cheese).

Circuit Model

Architecture and nomenclature

Regions and neural activity The circuit model contains 3 main regions: sensory cortex (SEN), neocortex (CTX), and medial temporal lobe (MTL). In turn, MTL is split into two sub-regions: MTL-sensory and MTL-semantic. We use X to denote the neural activity, which is written as X^{region} (for example X^{CTX}) when the region is to be made explicit. The same follows for pre-activation (synaptic) input, which is denoted by $\hat{X}$ (or $\hat{X}^{\text{region}}$). The activity of neuron i in a region is X_i^{region}. The number of neurons in a region is denoted by N_{region} (e.g. N_{CTX}).

Connectivity Connectivity matrices connecting a region pre to a region post take the form $W^{\text{post}\leftarrow\text{pre}}$, and a synapse connecting neuron j in region pre with neuron i in region post is $W_{ij}^{\text{post}\leftarrow\text{pre}}$. For example, $W_{13}^{\text{CTX}\leftarrow\text{MTL}}$ represents the connection of neuron 3 in MTL with neuron 1 in CTX. Connectivity is classified as either feed-forward (connecting two different regions), which includes $W^{\text{MTL-sensory}\leftarrow\text{SEN}}$, $W^{\text{CTX}\leftarrow\text{MTL}}$, and $W^{\text{MTL-semantic}\leftarrow\text{CTX}}$, or recurrent (connecting a region with itself), which includes $W^{\text{CTX}\leftarrow\text{CTX}}$, and $W^{\text{MTL}\leftarrow\text{MTL}}$. All connections are plastic except $W^{\text{MTL-sensory}\leftarrow\text{SEN}}$, which is fixed. Given a connectivity matrix $W^{\text{post}\leftarrow\text{pre}}$, we call receptive field of a postsynaptic neuron i $\text{RF}_i^{\text{post}\leftarrow\text{pre}}$ the vector representing the connections from region pre to postsynaptic neuron i

$$\text{RF}_i^{\text{post}\leftarrow\text{pre}} = \{{W^{\text{post}\leftarrow\text{pre}}}_{kj}\} \ : \ k = i \tag{6}$$

Synaptic Initialization

Sensory cortex to MTL-sensory In Figs. 2 and 3, $W^{\text{MTL-sensory}\leftarrow\text{SEN}}$ is initialized as the identity, mapping to MTL-sensory an exact copy of X^{SEN}. In the rest of Figures, the matrix is generated randomly with the condition that all receptive fields are of equal size:

$$||RF_i||_0 = |\{j \in \{1, ..., N_{\text{SEN}}\} : W_{ij}^{\text{MTL-sensory}\leftarrow\text{SEN}} \neq 0\}| = W_{\text{size}}^{\text{MTL-sensory}\leftarrow\text{SEN}} \quad \forall i \tag{7}$$

and sum:

$$\sum_j \{RF_i^{\text{MTL-sensory}\leftarrow\text{SEN}}\}_j = \sum_j W_{ij}^{\text{MTL-sensory}\leftarrow\text{SEN}} = W_{\text{max-in}}^{\text{MTL-sensory}\leftarrow\text{SEN}} \quad \forall i \tag{8}$$

where $||RF_i||_0$ denotes the l_0 norm (which counts the number of non-zero entries), $|A|$ denotes the number of elements in set A, $W_{\text{size}}^{\text{MTL-sensory}\leftarrow\text{SEN}}$ is a network parameter determining the number of non-zero connections each receptive field has, and $W_{\text{max-in}}^{\text{MTL-sensory}\leftarrow\text{SEN}}$ another network parameter determining the maximum sum of incoming connections a postsynaptic neuron has. In order to achieve both conditions, we generate $W^{\text{MTL-sensory}\leftarrow\text{SEN}}$ by randomly sampling each receptive field i independently, following:

$$\{RF_i\}_j = \begin{cases} \frac{W_{\text{max-in}}^{\text{MTL-sensory}\leftarrow\text{SEN}}}{W_{\text{size}}^{\text{MTL-sensory}\leftarrow\text{SEN}}}, & \text{if } j \in S, \\ 0, & \text{if } j \notin S, \end{cases} \tag{9}$$

with S a random subset of the presynaptic entries:

$$S \sim \text{Unif}\{T \subseteq \{1, \ldots, N_{\text{SEN}}\} : |T| = W_{\text{size}}^{\text{MTL-sensory}\leftarrow\text{SEN}}\} \Longrightarrow |S| = W_{\text{size}}^{\text{MTL-sensory}\leftarrow\text{SEN}}, \tag{10}$$

Feed-Forward connections Feed-forward connections are sampled from a Gaussian distribution with mean 0 and standard deviation $W_{\sigma}^{\text{post}\leftarrow\text{pre}}$ as a network parameter.

$$W_{ij}^{\text{post}\leftarrow\text{pre}} \sim \mathcal{N}(0, W_{\sigma}^{\text{post}\leftarrow\text{pre}}) \tag{11}$$

Recurrent connections Recurrent connections are initialized at 0. MTL connections are reset every *day* (see Algorithm 6)

Mature and immature neurons Neurons in the model can be either *mature* or *immature*. The intuition is: do these neurons contain meaningful representations via input received from another region (mature) or are mostly non-selective and random (immature)? More details on the dynamics and plasticity of mature and immature neurons can be seen below, but to summarize: (i) immature neurons have higher excitability during feed-forward input processing (not during pattern completion) and (ii) immature neurons have a higher learning rate in their incoming feed-forward connections and its recurrent connections frozen. The intuition is that by having higher excitability, during receptive field formation in feed-forward synapses, immature neurons win the competition for postsynaptic activity only if the presynaptic pattern is too far from the receptive fields that have already formed. By doing this, if a presynaptic pattern is similar to previously formed representations, it slightly changes the already formed representations. If not, a set of immature neurons win the competition, develop receptive field due to higher plasticity, and then become mature. This allows one-shot formation of new postsynaptic representations of presynaptic input when an input pattern is too different from those presented to the network before, and then refining these representations with a slower learning rate. Recurrent connectivity in CTX, which reflects statistical regularities across the representations formed, is only developed between mature neurons that actually code for an environmental variable, to avoid recurrent connections representing spurious correlations. Maturity in a region is denoted by $\text{IM}^{\text{region}}$, where $\text{IM}_i^{\text{region}} = 1$ if neuron i in the region is immature, and 0 otherwise.

Outgoing and incoming homeostasis across types of connections

Below, we formalize two regimes of homeostatic plasticity: outgoing homeostasis and incoming homeostasis. Outgoing homeostasis guarantees that the sum of the outgoing connections from neuron j in a connectivity matrix W does not exceed a certain value $W_{\text{max-out}}$ ($\sum_i W_{ij} \leq W_{\text{max-out}} \forall j$). Similarly, incoming homeostasis guarantees a bound in incoming connections ($\sum_j W_{ij} \leq W_{\text{max-in}} \forall i$). Depending on the type of connectivity we choose:

- outgoing homeostasis for CTX←CTX: this makes connectivity matrix reflect semantic structure as defined in (5), as connection from neuron coding for concept i from neuron connecting for concept j becomes proportional to the conditional probability of neuron i firing given neuron j fires (Albesa-González & Clopath, 2025)

- incoming homeostasis in feed-forward connectivity: in this homeostatic regime, connection from neuron j to neuron i becomes proportional to $p(j\ active|i\ active)$ (Albesa-González & Clopath, 2025). The *trick* is that if an initial one-shot receptive field makes neuron i selective to concept c, receptive fields approximate the average presynaptic activity pattern (prototype) given the concept c is present (neuron i is active), thus reinforcing higher selectivity in noisier regimes.

- Homeostatic reset in MTL←MTL connections. In order to account for symmetric plasticity kernels in hippocampus (Mishra et al., 2016; Li et al., 2023), we simply have Hebbian learning in MTL←MTL. Homeostasis is obtained by resetting MTL recurrent connections to 0 after each sleep cycle.

Network Operations

Neuronal activation

Synaptic input is mapped to neural activity via a *K-winners-take-all* mechanism, in which the K neurons with the highest pre-activation input $\hat{X}$ set their activity to 1, and the rest to 0. This can be written as

$$X = \text{top-}K\big(\hat{X}\big) \tag{12}$$

In order to determine K in a particular region, first we define

$$K^{\text{region}}_{\text{concept}} \leq K^{\text{region}}_{\text{episode}} \tag{13}$$

such that each concept is represented by $K^{\text{region}}_{\text{concept}}$ neurons, and a typical episode during wakefulness contains $K^{\text{region}}_{\text{episode}}$ active neurons. In this context, given K active neurons in a region, we define the semantic load S_L to be the number of concepts represented by K neurons, such that

$$K = S_L \cdot K^{\text{region}}_{\text{concept}} \tag{14}$$

In other words, by fixing the semantic load, one fixes how many neurons in total are active, making K higher if more concepts are being represented. The maximum number of concepts that can be encoded in an episode is thus:

$$N^{\text{region}}_{\text{concept}} = \frac{K^{\text{region}}_{\text{episode}}}{K^{\text{region}}_{\text{concept}}} \tag{15}$$

Therefore, we can define Activate (Algorithm 1) as an algorithm that maps presynaptic input $\hat{X}$, given a semantic load S_L, into neural activation X in the following way:

Algorithm 1 Neuronal Activation

1: **function** Activate($\hat{X}$, region, S_L)
2: Compute number of active neurons:

$$K \leftarrow S_L \times K^{\text{region}}_{\text{concept}}$$

3: Select top-K entries:

$$X \leftarrow \text{top-}K(\hat{X})$$

4: **return** X
5: **end function**

Feed-forward input processing

Input processing is computed for a single pre and post region at a time, with pre-activation input $\hat{X}^{\text{post}}$ obtained via

$$\hat{X}^{\text{post}} = W^{\text{post}\leftarrow\text{pre}} \cdot X^{\text{pre}} + b^{\text{post}}\text{IM}^{\text{post}} \tag{16}$$

b^{post} is the extra excitability that immature neurons have in region post.

Pattern completion (recurrent input processing)

Given a recurrent network, pattern completion is typically defined as an iterative update on neural activity X^{region} that is initiated in an initial state X^{region}_0:

$$X^{\text{region}}_{k+1} = \text{top-}K(W^{\text{region}\leftarrow\text{region}} \cdot X^{\text{region}}_k) \tag{17}$$

with $W^{\text{region}\leftarrow\text{region}}$ the region recurrent connectivity matrix, K dependent on the specified semantic load S_L, and k a timestep operating on a smaller timescale, than the overall timescale t. This means that within a timestep t, there are pattern_complete_iterations iterations on temporal index k. This is used to define PatternComplete as Algorithm 2:

Algorithm 2 Pattern Completion

```
function PatternComplete(X_0, S_L, pattern_complete_iterations)
    X ← X_0
    for iteration = 1 to pattern_complete_iterations do
        X̂ ← W^{region←region} · X
        X ← Activate(X̂, region, S_L)
    end for
    return X
end function
```

Hebbian Learning

The connection between a presynaptic neuron j and a postsynaptic neuron i subject to Hebbian learning is updated as:

$$\Delta W_{ij} = \lambda_{ij}^{\text{post}\leftarrow\text{pre}} X_i^{\text{post}} X_j^{\text{pre}} \tag{18}$$

for a presynaptic region pre and a postsynaptic region post, where $\lambda_{ij}^{\text{post}\leftarrow\text{pre}}$ is the learning rate from neuron j in region pre to neuron i in region post.

Maturity-dependent learning rate (feed-forward connections) In order to combine one-shot and statistical learning in feed-forward synapses (see Albesa-González and Clopath (2025)), the learning rate of feed-forward (post $\neq$ pre) connections depends on the *maturity* of the postsynaptic neuron, which is given by IM. Neurons start in an *immature* state $\text{IM}_i = 1$, allowing them to form relatively selective initial receptive fields. Once this initial receptive field is formed, the learning rate is reduced to capture the statistical regularities of the presynaptic patterns driving the postsynaptic neuron (the neuron is *mature*, $\text{IM}_i = 0$). To do this, we define a quick learning rate

$$\lambda_{\text{quick}}^{\text{post}\leftarrow\text{pre}} = \frac{W_{\text{max-in}}^{\text{post}\leftarrow\text{pre}}}{K^{\text{pre}}} \tag{19}$$

where again $W_{\text{max-in}}^{\text{post}\leftarrow\text{pre}}$ is a network parameter determining the maximum sum of incoming connections the matrix $W^{\text{post}\leftarrow\text{pre}}$ has for every postsynaptic neuron ($\sum_j W_{ij}^{\text{post}\leftarrow\text{pre}} \leq W_{\text{max-in}}^{\text{post}\leftarrow\text{pre}}$). Then, at each Hebbian update, we have

$$\lambda_{ij}^{\text{post}\leftarrow\text{pre}} = (1 - \text{IM}_i^{\text{post}})\lambda_{\text{slow}}^{\text{post}\leftarrow\text{pre}} + \text{IM}_i^{\text{post}}\lambda_{\text{quick}}^{\text{post}\leftarrow\text{pre}} \tag{20}$$

where $\lambda_{\text{slow}}^{\text{post}\leftarrow\text{pre}}$ is a network parameter indicating the slow learning rate (used throughout all simulations except for the first receptive field formed at each postsynaptic neuron). Note how for an initial (before maturity) driving presynaptic pattern X^{pre} this guaranties that the initial receptive field formed is X^{pre} scaled such that the sum of incoming connections is exactly equal to $W_{\text{max-in}}^{\text{post}\leftarrow\text{pre}}$, as the K-winners-share-all mechanism imposes exactly K_{pre} active neurons.

Maturity-dependent learning rate (CTX recurrent connections) Recurrent connections are designed to capture conditional firing probabilities between neurons. When individual neurons become mature and tuned to individual concepts, this allows capturing the environment semantic structure. To avoid capturing the statistical structure of immature neurons, and based on the fact that learning in CTX is favoured by the previous formation of schemas (Tse et al., 2007), we follow an opposite recipe. Hebbian updates on recurrent connections are only effective if both pre- and postsynaptic neurons are mature:

$$\lambda_{ij}^{\text{region}\leftarrow\text{region}} = (1 - \text{IM}_i^{\text{post}})(1 - \text{IM}_j^{\text{post}})\lambda_0^{\text{region}\leftarrow\text{region}} \tag{21}$$

In the case of recurrent connections, there are no slow and quick learning rates, but the parameter $\lambda_0^{\text{post}\leftarrow\text{pre}}$ is used to distinguish it from the maturity-dependent learning rate matrix

Hebbian learning algorithm Altogether, this defines Hebbian (Algorithm 3) as the following update in $W^{\text{post}\leftarrow\text{pre}}$:

Algorithm 3 Hebbian Learning

function HEBBIAN(post, pre)
  Obtain learning rate matrix:
  if post $\neq$ pre **then**

$$\lambda_{ij}^{\text{post}\leftarrow\text{pre}} \leftarrow (1 - \text{IM}_i^{\text{post}})\, \lambda_{\text{slow}}^{\text{post}\leftarrow\text{pre}} + \text{IM}_i^{\text{post}}\, \lambda_{\text{quick}}^{\text{post}\leftarrow\text{pre}}$$

  else if post = pre = CTX **then**

$$\lambda_{ij}^{\text{post}\leftarrow\text{pre}} \leftarrow (1 - \text{IM}_i^{\text{post}})(1 - \text{IM}_j^{\text{pre}})\, \lambda_0^{\text{post}\leftarrow\text{pre}}$$

  else

$$\lambda_{ij}^{\text{post}\leftarrow\text{pre}} \leftarrow \lambda_0^{\text{post}\leftarrow\text{pre}}$$

  end if
  Compute Hebbian update with activity outer product:

$$\Delta W \leftarrow \lambda^{\text{post}\leftarrow\text{pre}} \odot \left(X^{\text{post}} \otimes X^{\text{pre}}\right)$$

  Update weights:

$$W^{\text{pre}\leftarrow\text{post}} \leftarrow W^{\text{pre}\leftarrow\text{post}} + \Delta W$$

end function

Homeostatic Plasticity

We use multiplicative normalization that guarantees, for every weight matrix $W^{\text{post}\leftarrow\text{pre}}$, the total sum of incoming or outgoing connections is capped to a certain value. Following Albesa-González and Clopath (2025), we use *incoming homeostasis* as:

$$W_{ij}^{k+1} = \begin{cases} W_{ij}^k \cdot \frac{W_{\text{max-in}}}{\sum_j W_{ij}^k} & \text{if } \sum_j W_{ij}^k = W_{\text{max}}^{\text{in}} + \epsilon_i, \quad \epsilon_i > 0 \\ W_{ij}^k & \text{else} \end{cases} \tag{22}$$

where W_{ij}^k denotes the value of synapse ij before applying homeostasis and W_{ij}^{k+1} the value after applying homeostasis (for readability, we simply write $W \equiv W^{\text{post}\leftarrow\text{pre}}$). Similarly, we have *outgoing homeostasis* to be defined by:

$$W_{ij}^{k+1} = \begin{cases} \frac{W_{\text{max-out}}}{\sum_i W_{ij}^k} & \text{if } \sum_j W_{ij}^k = W_{\text{max}}^{\text{out}} + \epsilon_j, \quad \epsilon_j > 0 \\ 0 & \text{else} \end{cases} \tag{23}$$

This allows, for a presynaptic region pre and a postsynaptic region post, defining HOMEOSTASIS as Algorithm 4.

Learning fixed points

The combination of *K-winners-take-all* (top-K) activation function with Hebbian and homeostatic plasticity as defined here has been recently shown (Albesa-González & Clopath, 2025) to make synaptic connections converge to conditional firing probabilities. For outgoing homeostasis, the fixed points are:

$$W_{ij}^{\text{pre}\leftarrow\text{post, out}}(\infty) \equiv \frac{W_{\text{max-out}}^{\text{pre}\leftarrow\text{post}}}{K^{\text{post}}} p(X_i^{\text{post}} = 1 | X_j^{\text{pre}} = 1) \tag{24}$$

where K^{post} corresponds to the number of active neurons during plasticity. Given that the conditional probability can be at most 1, we define in this scenario W_{max} as $\frac{W_{\text{max-out}}^{\text{pre}\leftarrow\text{post}}}{K^{\text{post}}}$. Similarly, for incoming homeostasis, the fixed points are

$$W_{ij}^{\text{pre}\leftarrow\text{post, out}}(\infty) \equiv \frac{W_{\text{max-in}}^{\text{pre}\leftarrow\text{post}}}{K^{\text{pre}}} p(X_j^{\text{pre}} = 1 | X_i^{\text{post}} = 1) \tag{25}$$

with $W_{\text{max}} = \frac{W_{\text{max-in}}^{\text{pre}\leftarrow\text{post}}}{K^{\text{pre}}}$. These results have an important associated intuition, which is that if neurons code for individual concept (have a concept selectivity close to 1), synaptic connections reflect the associated

Semantic Structure (Albesa-González & Clopath, 2025) under outgoing homeostasis (as happens in CTX). In MTL-semantic←CTX and CTX←MTL, which follow incoming homeostasis, when initial one-shot receptive fields induce a selectivity to a concept close to 1, feed-forward synapses slowly approximate to the associated prototype, where each synapse is becomes proportional to the contribution of each presynaptic neuron to the concept encoded in the postsynaptic neuron.

Algorithm 4 Homeostasis

function HOMEOSTASIS(post, pre)
  $W \leftarrow W^{\text{post}\leftarrow\text{pre}}$
  for each post-synaptic unit i **do**
    if $\sum_j W[i,j] > W^{\text{post}\leftarrow\text{pre}}_{\text{max-post}}$ **then**
      Normalize:

$$W[i,:] \leftarrow W[i,:] \times \frac{W}{\sum_j W[i,j]}$$

    end if
  end for
  for each pre-synaptic unit j **do**
    if $\sum_i W[i,j] > \text{max_pre_connectivity}$ **then**
      Normalize:

$$W[:,j] \leftarrow W[:,j] \times \frac{\text{max_pre_connectivity}}{\sum_i W[i,j]}$$

    end if
  end for
end function

Replay

We define Replay as a learning mechanism involving a pre and a post region, such that $W^{\text{post}\leftarrow\text{pre}}$ extract statistical structure from $W^{\text{pre}\leftarrow\text{pre}}$ connections. A replay operation takes as argument a specific semantic load S_L, which determines the conceptual complexity of the extracted regularities. To do this, we first pattern complete an initial presynaptic pattern X_0^{pre}, defined as a Gaussian random vector of the same size as the pre region:

$$X_0^{\text{pre}} \sim \mathcal{N}(0, I_{N_{\text{pre}}}) \tag{26}$$

where N_{pre} is the size of region pre and $\mathcal{N}(0, I_{N_{\text{pre}}})$ is a multivariate Gaussian distribution with 0 mean and identity covariance. This initial random pattern is pattern completed to obtain X^{pre}, and then projected to the post region, to give X^{post}, using in both operations the same fixed semantic load. Then, Hebbian and homeostatic plasticity operations as defined in Algorithms 3 and 4 are used in the (post, pre) region pair. Replay is computed in simulations following Algorithm 5. In the text, we call replay from MTL to CTX *Episodic Replay*, because it finds regularities across recurrent synapses that contain episodic information, and replay from CTX to MTL-semantic *Semantic Replay*, because in this case recurrent synapses contain semantic information.

Algorithm 5 Replay

function REPLAY(post, pre, S_L, `pattern_complete_iterations`)
  Sample Gaussian random vector in pre:

$$X_0^{\text{pre}} \leftarrow \text{randn}(N_{\text{pre}})$$

  $X^{\text{pre}} \leftarrow \text{PATTERNCOMPLETE}\big(X_0^{\text{pre}}, S_L, \texttt{pattern_complete_iterations}\big)$
  $X^{\text{post}} \leftarrow W^{\text{post}\leftarrow\text{pre}} \cdot X^{\text{pre}} + b^{\text{post}}\text{IM}^{\text{post}}$
  HEBBIAN(post, pre)
  HOMEOSTASIS(post, pre)
  return X
end function

Learning Algorithm

For each day of experience, assuming an input X_{input} of shape $(N_{\text{days}}, T_{\text{day}})$, we define the following learning algorithm TRAIN (Algorithm 6), where Wake is defined as in Algorithm 7 and the Sleep as in Algorithm 8. All operations used have been previously defined, with these algorithms showing the order in which they are performed. DURATION_PHASE_A represents the number of days in which only episodic replay takes place (Fig. 2), and DURATION_PHASE_B the number of days in which only MTL-sensory synapses are used for episodic replay (but MTL-semantic representations have been formed already, Figs. 3, 4, 5, except Fig. 5f). From day DURATION_PHASE_B on, MTL replays both sensory and semantic representations. Therefore, during DURATION_PHASE_A days, episodic memories containing only sensory information are stored in MTL during Wake (Algorithm 7). During Sleep (Algorithm 8), cortical abstractions of sensory information are formed. After DURATION_PHASE_A days, cortical abstractions are transferred to MTL-semantic via semantic replay during Sleep. Finally, after DURATION_PHASE_B days, episodic memories contain both sensory and semantic information, and so does episodic replay. This allows higher-order abstraction, resulting in the formation of new semantic representations that are based on previous abstractions.

Algorithm 6 Learning Algorithm Per Day

```
function TRAIN(X_input)
    for day = 1 to N_days do
        Extract daily input:
            X_input^day ← X_input[day, :]
        WAKE(X_input^day)
        SLEEP()
    end for
end function
```

Algorithm 7 Wake

function WAKE(X^{day})
  Reset MTL recurrent connections:
$$W^{\text{MTL}\leftarrow\text{MTL}} \leftarrow 0$$
  for $t = 1$ to T_{day} **do**
    $X^{\text{SEN}} \leftarrow \text{ACTIVATION}\big(X^{\text{day}}[t], \text{SEN}\big)$
    $\hat{X}^{\text{MTL-sensory}} \leftarrow W^{\text{MTL-sensory}\leftarrow\text{SEN}} \cdot X^{\text{SEN}}$
    $X^{\text{MTL-sensory}} \leftarrow \text{ACTIVATION}\big(\hat{X}^{\text{MTL-sensory}}, \text{MTL-sensory}\big)$
    $\hat{X}^{\text{MTL-semantic}} \leftarrow \text{randn}(N_{\text{MTL-semantic}})$
    $X^{\text{MTL-semantic}} \leftarrow \text{ACTIVATION}\big(\hat{X}^{\text{MTL_semantic}}, \text{MTL_semantic}\big)$
    $X^{\text{MTL}} \leftarrow [X^{\text{MTL-sensory}},\ X^{\text{MTL-semantic}}]$
    $\hat{X}^{\text{CTX}} \leftarrow W^{\text{CTX}\leftarrow\text{MTL}} \cdot X^{\text{MTL}}$
    $X^{\text{CTX}} \leftarrow \text{ACTIVATION}\big(\hat{X}^{\text{CTX}}, \text{CTX}\big)$
    if day $\geq$ duration_phase_A **then**
      $\hat{X}^{\text{MTL-semantic}} \leftarrow W^{\text{MTL-semantic}\leftarrow\text{CTX}} \cdot X^{\text{CTX}}$
      $X^{\text{MTL-semantic}} \leftarrow \text{ACTIVATION}\big(\hat{X}^{\text{MTL-semantic}}, \text{MTL-semantic}\big)$
      $X^{\text{MTL}}[N_{\text{MTL-sensory}} :] \leftarrow X^{\text{MTL_semantic}}$
      $\text{HEBBIAN}(X^{\text{MTL_semantic}}, X^{\text{MTL_semantic}})$
    end if
    if day $\geq$ duration_phase_B **then**
      $\hat{X}^{\text{CTX}} \leftarrow W^{\text{CTX}\leftarrow\text{MTL}} \cdot X^{\text{MTL}}$
      $X^{\text{CTX}} \leftarrow \text{ACTIVATION}\big(\hat{X}^{\text{CTX}}, \text{CTX}\big)$
    end if
    $\text{HEBBIAN}(\text{MTL}, \text{MTL})$
    $\text{HEBBIAN}(\text{CTX}, \text{CTX})$
    $\text{HOMEOSTASIS}(\text{CTX}, \text{CTX})$
  end for
  Increment day counter
end function

Algorithm 8 Sleep

function SLEEP
  Episodic Replay
  for $t = 1$ to `sleep_duration_A` **do**
    if day $\leq$ duration_phase_B **then**
      $\text{REPLAY}(\text{CTX}, \text{MTL-sensory},\ S_L = 1,\ \text{pattern_complete_iterations} = 10)$
    else
      Sample S_L randomly
      $S_L \leftarrow \text{RANDINT}(1, S_{L-\max})$
      $\text{REPLAY}(\text{CTX}, \text{MTL},\ S_L = S_L,\ \text{pattern_complete_iterations} = 10)$
    end if
  end for
  if day $\geq$ duration_phase_A **then**
    for $t = 1$ to `sleep_duration_B` **do**
      $\text{REPLAY}(\text{MTL-semantic}, \text{CTX},\ S_L = 1,\ \text{pattern_complete_iterations} = 10)$
    end for
  end if
end function

Data Analysis

Neuronal Selectivity

Given a vector X that contains neuronal activity across T timesteps -with dimensions (T, N_{neurons})- and one that contains certain latent variables Z -with dimensions (T, N_{latents}), we do obtain the normalized values:

$$X_{\text{norm}} = \frac{X - \langle X \rangle_{\text{time}}}{\sqrt{\langle X^2 \rangle_{\text{time}} - \langle X \rangle^2_{\text{time}}}} \quad ; \quad Z_{\text{norm}} = \frac{Z - \langle Z \rangle_{\text{time}}}{\sqrt{\langle Z^2 \rangle_{\text{time}} - \langle Z \rangle^2_{\text{time}}}} \tag{27}$$

And then obtain the matrix

$$S = \frac{1}{T} X^T_{\text{norm}} Z_{\text{norm}} \tag{28}$$

which, for neuron i and latent j contains their correlation as:

$$S(i,j) = \frac{1}{T} \sum_t X_{ti} Z_{tj} \tag{29}$$

The Maximum Selectivity of neuron i is the maximum selectivity value of neuron i across all latents j:

$$S_{\max}(i) = \max_j S(i,j) \tag{30}$$

Neuron Ordering

Throughout this study, we re-order the neurons indices for visualization purposes. We do this with respect to a certain latent variable (for example the different *concepts* A_i and B_i or the different episodes $A_i B_j$. This is done by imposing a threshold, and then for each possible discrete value the latent can take we obtain the neurons with a selectivity higher to that value with respect to the given latent. For example, assuming latents A_1 to B_5, the first neurons would be those with a selectivity higher than 0.9 to A_1, then the next neurons would be those selective to A_2, etc. All neurons that are not selective to any latent keep the original order at the end of the new order.

Quick Downstream Learning

In Fig. 3, we test how well can a classifier learn from different representations. These representations are: MTL-sensory, MTL-semantic and *labels*. While the first two are simply the simulated neural activity of the corresponding regions, *labels* are simply lists one-hot encodings of the labels being predicted (i.e. simply the same as the output, repeated to match the size of MTL). This is to obtain a baseline of the most *trivial* task possible, which would be mapping labels to labels. Then, for each panel in 3g, we initialize at random 100 classifiers consisting on a linear layer followed by a softmax activation. For each dataset $(X^{\text{MTL-sensory}}, Y^{\text{latent}})$, $(X^{\text{MTL-semantic}}, Y^{\text{latent}})$, and $(X^{\text{labels}}, Y^{\text{latent}})$ we train the same (initially) classifier to predict Ai and B_j that generated each representation X^{region}. We do this for only one epoch as a proxy for quick downstream learning facilitation. Each epoch consists of 100 days of experience, which amounts to 8000 samples.

While obtaining the neural representations at each region, the model is frozen. This allows us to study 3 different regimes depending on whether the outside world distribution matches the original input distribution or not: (1) In-Distribution (input sampled from the same distribution) (2) Compositional Out-Of-Distribution (episodes had never been seen by the network, but the individual compositional blocks are maintained), and (3) Semantic Out-Of-Distribution (input generated using never-seen blocks, here implemented as a fixed random permutation of original input to maintain rest of statistics).

Random projection of sensory inputs

In Figs. 4 and 5 we forward the initially sampled episode through random connections. This connectivity is obtained by fixing a receptive field size $N^{\text{pre}}_{\text{RF}}$, then for each postsynaptic neuron i randomly picking $N^{\text{pre}}_{\text{RF}}$ synapses $ij \in \text{RF}(i)$, and doing

$$W^{\text{post}\leftarrow\text{pre}}_{ij} = W^{\text{post}}_{\max} / N^{\text{pre}}_{\text{RF}} \tag{31}$$

Thus, every postsynaptic neuron is randomly assigned $N^{\text{pre}}_{\text{RF}}$ presynaptic neurons, and has total incoming weights $W^{\text{post}}_{\max}$. Given our top-k activation, this guarantees a much more distributed firing rate, which depends exclusively on the presynaptic pattern and not on the total sum of weights assigned at random.

Blocked vs interleaved training

In order to model blocked training, we change input statistics during the first 100 days. For each concept A_1 to B_5, we have a 10-day block of 200 timesteps where every episode is sampled with equal probability as long as the corresponding concept is present. During the first block, all episodes with $A_i = A_1$ are equally likely (the rest have probability 0), then second block the same with $A_i = A_2$, until $B_j = B_5$. For interleaved training, we maintain original statistics but also increase the number of timesteps to 200 per day so the total amount of episodes is the same.

Concept prototypes

For a given neural representation X^{region}, and a concept c, one can define the prototype of c in the region as

$$\text{top-}K^{\text{region}}\big(\langle X^{\text{region}}\rangle_{c\in e}\big) \tag{32}$$

that is, the activation of the average neural activity across all episodes in which concept c is present. For example, the prototype in MTL-sensory of A_1 would be the average neural activity across episodes containing A_1. Note how this could also be extended to higher-order concepts (for example the prototype of A_1B_3).

Recall and replay

In Fig. 5 we obtain two measures related to the episodic memories formed in MTL←MTL connections: *recall* and *replay*. In both cases we have a training phase, where multiple episodes are sampled as usual, and all of which are stored in MTL recurrent connections following *wake* (Algorithm 7). Then there is a test phase. For *recall*, we (1) randomly sample the neural activity of MTL-sensory for one of the episodes presented during training, (2) impose that activity and leave MTL-semantic random, and finally obtain $X^{MTL}_{\text{P.Completed}}$ via the *Pattern Complete* operation (Algorithm 2). Recall is defined as the cosine similarity between the encoded and the recovered representation. Where indicated in a legend as MTL-sensory or MTL-semantic, that means the cosine similarity measured only in that subregion (but the previeous procedure is maintained the same). To measure replay, we simply follow Algorithm (5), which means doing pattern completion on a random initial activity pattern. Then, we measure the cosine similarity of MTL-sensory with each of the different prototypes (Eq. (32)), and obtain the highest score across all prototypes. Intuitively, because there was no original pattern presented, and because we are interested in measuring replay as an initial generalization step, we measure how well do replayed activity patterns reflect latent prototypes (as these patterns are the ones that will be used to form semantic representations in CTX).

Simulation parameters

Default input, temporal, and model parameters can be found in Tables 1-3. Figure-specific parameters can be found in Tables 4-8.

Table 1: Summary of Episode Generation Protocol parameters

Variable	**Description**	**Value**
$N_{\text{attributes}}$	Number of episode attributes	2
$\lvert A\rvert, \lvert B\rvert$	Number of concepts per attribute	[5, 5]
N^A_{concept}, N^B_{concept}	Total size (in neurons) per latent variable	[10, 10]
$P(A_i, B_j)$	Episode probability	0.5/5 if $i = j$; 0.5/20 if $i \neq j$
N_{swaps}	Number of swaps (total noise is $2N_{\text{swaps}}$)	2

Table 2: Summary of default input temporal parameters

Variable	**Description**	**Value**
T_{day}	Day length	80
T_{episode}	Episode length	5

Acknowledgments

We would like to thank Rachel Swanson and Colleen J. Gillon, as well as the rest of the members of the Clopath Lab, for discussions and comments on earlier versions of the manuscript. We gratefully acknowledge the financial support of the Wellcome Trust, the Simons Foundation, the Engineering and Physical Sciences Research Council (EPSRC), and the European Research Council (ERC).

Declaration of Interests

The authors declare no competing interests.

Code Availability

All code supporting the findings of this study is available at https://github.com/albesagonzalez/sensory-semantic-episodes

Declaration of generative AI and AI-assisted technologies in the manuscript preparation process

During the preparation of this work the authors used ChatGPT in order to assist with writing at the sentence level, and coding at the snippet level. After using this tool/service, the authors reviewed and edited the content as needed and take full responsibility for the content of the published article.

References

Alba, J. W., & Hasher, L. (1983). Is memory schematic? *Psychological Bulletin*, *93*(2), 203.

Alberini, C. M., & Travaglia, A. (2017). Infantile amnesia: a critical period of learning to learn and remember. *Journal of Neuroscience*, *37*(24), 5783–5795.

Albesa-González, A., & Clopath, C. (2025). Homeostatic binary networks: A simple framework for learning with overlapping patterns. *bioRxiv*.

Alvarez, P., & Squire, L. R. (1994). Memory consolidation and the medial temporal lobe: a simple network model. *Proceedings of the national academy of sciences*, *91*(15), 7041–7045.

Aronov, D., Nevers, R., & Tank, D. W. (2017). Mapping of a non-spatial dimension by the hippocampal–entorhinal circuit. *Nature*, *543*(7647), 719–722.

Barlow, H. B. (1972). Single units and sensation: a neuron doctrine for perceptual psychology? *Perception*, *1*(4), 371–394.

Bartlett, F. C. (1932). *Remembering: A study in experimental and social psychology*. Cambridge university press.

Behrens, T. E., Muller, T. H., Whittington, J. C., Mark, S., Baram, A. B., Stachenfeld, K. L., & Kurth-Nelson, Z. (2018). What is a cognitive map? organizing knowledge for flexible behavior. *Neuron*, *100*(2), 490–509.

Bussmann, H., Kazzazi, K., & Trauth, G. (2006). *Routledge dictionary of language and linguistics*. Routledge.

Chandra, S., Sharma, S., Chaudhuri, R., & Fiete, I. (2025). Episodic and associative memory from spatial scaffolds in the hippocampus. *Nature*, *638*(8051), 739–751.

Chaudhry, A., Khan, N., Dokania, P., & Torr, P. (2020). Continual learning in low-rank orthogonal subspaces. *Advances in Neural Information Processing Systems*, *33*, 9900–9911.

Chettih, S. N., Mackevicius, E. L., Hale, S., & Aronov, D. (2024). Barcoding of episodic memories in the hippocampus of a food-caching bird. *Cell*, *187*(8), 1922–1935.

Chrysanthidis, N., Fiebig, F., Lansner, A., & Herman, P. (2022). Traces of semantization, from episodic to semantic memory in a spiking cortical network model. *Eneuro*, *9*(4).

Courellis, H. S., Minxha, J., Cardenas, A. R., Kimmel, D. L., Reed, C. M., Valiante, T. A., ... Rutishauser, U. (2024). Abstract representations emerge in human hippocampal neurons during inference. *Nature*, *632*(8026), 841–849.

Dong, C. V., Lu, Q., Norman, K. A., & Michelmann, S. (2025). Towards large language models with human-like episodic memory. *Trends in Cognitive Sciences*.

Dubreuil, A. M., Amit, Y., & Brunel, N. (2014). Memory capacity of networks with stochastic binary synapses. *PLoS computational biology*, *10*(8), e1003727.

Flesch, T., Balaguer, J., Dekker, R., Nili, H., & Summerfield, C. (2018). Comparing continual task learning in minds and machines. *Proceedings of the National Academy of Sciences*, *115*(44), E10313–E10322.

Fountas, Z., Benfeghoul, M., Oomerjee, A., Christopoulou, F., Lampouras, G., Ammar, H. B., & Wang, J. (2025). Human-inspired episodic memory for infinite context llms. In *The thirteenth international conference on learning representations*.

Fyhn, M., Molden, S., Witter, M. P., Moser, E. I., & Moser, M.-B. (2004). Spatial representation in the entorhinal cortex. *Science*, *305*(5688), 1258–1264.

Gentry, H., & Buckner, C. (2024). Transitional gradation and the distinction between episodic and semantic memory. *Philosophical Transactions B*, *379*(1913), 20230407.

Greenberg, D. L., & Verfaellie, M. (2010). Interdependence of episodic and semantic memory: Evidence from neuropsychology. *Journal of the International Neuropsychological society*, *16*(5), 748–753.

Guerreiro, I. C., & Clopath, C. C. (2024). Memory's gatekeeper: the role of pfc in the encoding of familiar events. *bioRxiv*, 2024–02.

Hopfield, J. J. (1982). Neural networks and physical systems with emergent collective computational abilities. *Proceedings of the national academy of sciences*, *79*(8), 2554–2558.

Iman, M., Arabnia, H. R., & Rasheed, K. (2023). A review of deep transfer learning and recent advancements. *Technologies*, *11*(2), 40.

Kang, L., & Toyoizumi, T. (2023). Hopfield-like network with complementary encodings of memories. *Physical Review E*, *108*(5), 054410.

Kolibius, L. D., Roux, F., Parish, G., Ter Wal, M., Van Der Plas, M., Chelvarajah, R., ... others (2023). Hippocampal neurons code individual episodic memories in humans. *Nature human behaviour*, *7*(11), 1968–1979.

Kurth-Nelson, Z., Behrens, T., Wayne, G., Miller, K., Luettgau, L., Dolan, R., ... Schwartenbeck, P. (2023). Replay and compositional computation. *Neuron*, *111*(4), 454–469.

Li, Y., Briguglio, J. J., Romani, S., & Magee, J. C. (2023). Mechanisms of memory storage and retrieval in hippocampal area ca3. *biorxiv*, 2023–05.

Lin, Q., Yousif, S. R., Chun, M. M., & Scholl, B. J. (2021). Visual memorability in the absence of semantic content. *Cognition*, *212*, 104714.

Liu, X. L., Ranganath, C., & O'Reilly, R. C. (2024). A complementary learning systems model of how sleep moderates retrieval practice effects. *Psychonomic Bulletin & Review*, 1–14.

Malik, R., Li, Y., Schamiloglu, S., & Sohal, V. S. (2022). Top-down control of hippocampal signal-to-noise by prefrontal long-range inhibition. *Cell*, *185*(9), 1602–1617.

McClelland, J. L. (2013). Incorporating rapid neocortical learning of new schema-consistent information into complementary learning systems theory. *Journal of Experimental Psychology: General*, *142*(4), 1190.

McClelland, J. L., McNaughton, B. L., & O'Reilly, R. C. (1995). Why there are complementary learning systems in the hippocampus and neocortex: insights from the successes and failures of connectionist models of learning and memory. *Psychological review*, *102*(3), 419.

McKoon, G., Ratcliff, R., & Dell, G. S. (1986). A critical evaluation of the semantic-episodic distinction.

Mishra, R. K., Kim, S., Guzman, S. J., & Jonas, P. (2016). Symmetric spike timing-dependent plasticity at ca3–ca3 synapses optimizes storage and recall in autoassociative networks. *Nature communications*, *7*(1), 1–11.

Moscovitch, M., & Gilboa, A. (2021). Systems consolidation, transformation and reorganization: Multiple trace theory, trace transformation theory and their competitors.

Moser, E. I., Kropff, E., & Moser, M.-B. (2008). Place cells, grid cells, and the brain's spatial representation system. *Annu. Rev. Neurosci.*, *31*, 69–89.

O'Keefe, J., & Dostrovsky, J. (1971). The hippocampus as a spatial map: preliminary evidence from unit activity in the freely-moving rat. *Brain research*.

O'Reilly, R. C., Bhattacharyya, R., Howard, M. D., & Ketz, N. (2014). Complementary learning systems. *Cognitive science*, *38*(6), 1229–1248.

Pink, M., Wu, Q., Vo, V. A., Turek, J., Mu, J., Huth, A., & Toneva, M. (2025). Position: Episodic memory is the missing piece for long-term llm agents. *arXiv preprint arXiv:2502.06975*.

Posner, M. I., & Keele, S. W. (1968). On the genesis of abstract ideas. *Journal of experimental psychology*, *77*(3p1), 353.

Quiroga, R. Q. (2012). Concept cells: the building blocks of declarative memory functions. *Nature Reviews Neuroscience*, *13*(8), 587–597.

Quiroga, R. Q., Reddy, L., Kreiman, G., Koch, C., & Fried, I. (2005). Invariant visual representation by single neurons in the human brain. *Nature*, *435*(7045), 1102–1107.

Rey, H. G., Panagiotaropoulos, T. I., Gutierrez, L., Chaure, F. J., Nasimbera, A., Cordisco, S., ... others (2025). Lack of context modulation in human single neuron responses in the medial temporal lobe. *Cell Reports*, *44*(1).

Saxe, A. M., McClelland, J. L., & Ganguli, S. (2019). A mathematical theory of semantic development in deep neural networks. *Proceedings of the National Academy of Sciences*, *116*(23), 11537–11546.

Schapiro, A. C., Turk-Browne, N. B., Botvinick, M. M., & Norman, K. A. (2017). Complementary learning systems within the hippocampus: a neural network modelling approach to reconciling episodic memory with statistical learning. *Philosophical Transactions of the Royal Society B: Biological Sciences*, *372*(1711), 20160049.

Shin, J. D., & Jadhav, S. P. (2024). Prefrontal cortical ripples mediate top-down suppression of hippocampal reactivation during sleep memory consolidation. *Current biology*, *34*(13), 2801–2811.

Singh, D., Norman, K. A., & Schapiro, A. C. (2022). A model of autonomous interactions between hippocampus and neocortex driving sleep-dependent memory consolidation. *Proceedings of the National Academy of Sciences*, *119*(44), e2123432119.

Spens, E., & Burgess, N. (2024). A generative model of memory construction and consolidation. *Nature Human Behaviour*, 1–18.

Squire, L. R. (2004). Memory systems of the brain: a brief history and current perspective. *Neurobiology of learning and memory*, *82*(3), 171–177.

Sun, C., Yang, W., Martin, J., & Tonegawa, S. (2020). Hippocampal neurons represent events as transferable units of experience. *Nature neuroscience*, *23*(5), 651–663.

Sun, W., Advani, M., Spruston, N., Saxe, A., & Fitzgerald, J. E. (2023). Organizing memories for generalization in complementary learning systems. *Nature neuroscience*, *26*(8), 1438–1448.

Swanson, R. A., Levenstein, D., McClain, K., Tingley, D., & Buzsáki, G. (2020). Variable specificity of memory trace reactivation during hippocampal sharp wave ripples. *Current Opinion in Behavioral Sciences*, *32*, 126–135.

Tomé, D. F., Sadeh, S., & Clopath, C. (2022). Coordinated hippocampal-thalamic-cortical communication crucial for engram dynamics underneath systems consolidation. *Nature communications*, *13*(1), 840.

Treves, A., & Rolls, E. T. (1994). Computational analysis of the role of the hippocampus in memory. *Hippocampus*, *4*(3), 374–391.

Tse, D., Langston, R. F., Kakeyama, M., Bethus, I., Spooner, P. A., Wood, E. R., ... Morris, R. G. (2007). Schemas and memory consolidation. *Science*, *316*(5821), 76–82.

Tsodyks, M. V., & Feigel'man, M. V. (1988). The enhanced storage capacity in neural networks with low activity level. *Europhysics Letters*, *6*(2), 101.

Tulving, E. (1972). Episodic and semantic memory. *Organization of memory/Academic Press*.

Whittington, J. C., Muller, T. H., Mark, S., Chen, G., Barry, C., Burgess, N., & Behrens, T. E. (2020). The tolman-eichenbaum machine: unifying space and relational memory through generalization in the hippocampal formation. *Cell*, *183*(5), 1249–1263.

Zhao, X., Hsu, C.-L., & Spruston, N. (2022). Rapid synaptic plasticity contributes to a learned conjunctive code of position and choice-related information in the hippocampus. *Neuron*, *110*(1), 96–108.

Table 3: Summary of network parameters used in the simulations (default unless specified otherwise).

Variable	Description	Value
General		
`duration_phase_A`	Phase A duration (episodic → semantic only)	1000
`duration_phase_B`	Phase B duration (episodic → semantic + semantic → episodic)	1500
`sleep_duration_episodic`	Sleep duration for episodic replay	10
`sleep_duration_semantic`	Sleep duration for semantic replay	10
$S_L^{\max}$	Min semantic load during replay	1
$S_L^{\min}$	Max semantic load during input	2
Pattern Completion		
`mtl_pattern_complete_iterations`	Iterations in MTL pattern completion	10
`ctx_pattern_complete_iterations`	Iterations in CTX pattern completion	10
Region Parameters		
N_{sen}	Number of SEN neurons	100
$K_{\text{concept}}^{\text{SEN}}$	Neurons per concept in SEN	10
$K_{\text{episode}}^{\text{SEN}}$	Neurons per episode in SEN	20
N^{CTX}	Number of CTX neurons (list denotes sub-regions)	[100, 300]
$K_{\text{concept}}^{\text{CTX}}$	Neurons per concept in CTX (list denotes sub-regions)	[10, 10]
$K_{\text{episode}}^{\text{CTX}}$	Neurons per episode in CTX (list denotes sub-regions)	[20, 10]
b^{CTX}	CTX immature neurons excitability (list denotes sub-regions)	[0.3, 0.6]
$N_{\text{MTL-sensory}}$	Number of MTL-sensory neurons	100
$K_{\text{concept}}^{\text{MTL-sensory}}$	Neurons per concept in MTL-sensory	10
$K_{\text{episode}}^{\text{MTL-sensory}}$	Neurons per episode in MTL-sensory	20
$N_{\text{MTL-semantic}}$	Number of MTL-semantic neurons	100
$K_{\text{concept}}^{\text{MTL-semantic}}$	Neurons per concept in MTL-semantic	5
$K_{\text{episode}}^{\text{MTL-semantic}}$	Neurons per episode in MTL-semantic	10
$b^{\text{MTL.semantic}}$	MTL-semantic immature neurons excitability	0.3
Synaptic Parameters		
`MTL-sensory_SEN_projection`	MTL-sensory←SEN random projection	True
$W_{\text{size}}^{\text{MTL-sensory}\leftarrow\text{SEN}}$	MTL-sensory←SEN receptive field size	30
$W_{\text{max-post}}^{\text{MTL-sensory}\leftarrow\text{SEN}}$	Max incoming MTL-sensory←SEN weight	1
$W_{\sigma}^{\text{CTX}\leftarrow\text{MTL}}$	CTX←MTL weight init std	$5 \cdot 10^{-4}$
$\lambda_{\text{slow}}^{\text{CTX}\leftarrow\text{MTL}}$	CTX←MTL learning rate	$5 \cdot 10^{-4}$
$W_{\text{max-pre}}^{\text{CTX}\leftarrow\text{MTL}}$	Max outgoing CTX←MTL weight	∞
$W_{\text{max-post}}^{\text{CTX}\leftarrow\text{MTL}}$	Max incoming CTX←MTL weight	1
$\lambda_0^{\text{MTL}\leftarrow\text{MTL}}$	MTL recurrent learning rate	$5 \cdot 10^{-3}$
$W_{\text{max-pre}}^{\text{MTL}\leftarrow\text{MTL}}$	Max outgoing MTL←MTL weight	∞
$W_{\text{max-post}}^{\text{MTL}\leftarrow\text{MTL}}$	Max incoming MTL←MTL weight	∞
$\lambda_0^{\text{CTX}\leftarrow\text{CTX}}$	CTX recurrent learning rate	$5 \cdot 10^{-4}$
$W_{\text{max-pre}}^{\text{CTX}\leftarrow\text{CTX}}$	Max outgoing CTX←CTX weight	1
$W_{\text{max-post}}^{\text{CTX}\leftarrow\text{CTX}}$	Max incoming CTX←CTX weight	∞
$W_{\sigma}^{\text{MTL-semantic}\leftarrow\text{CTX}}$	MTL-semantic←CTX weight init std	$1 \cdot 10^{-3}$
$\lambda_{\text{slow}}^{\text{MTL-semantic}\leftarrow\text{CTX}}$	MTL-semantic←CTX learning rate	$5 \cdot 10^{-3}$
$W_{\text{max-pre}}^{\text{MTL-semantic}\leftarrow\text{CTX}}$	Max outgoing MTL-semantic←CTX weight	∞
$W_{\text{max-post}}^{\text{MTL-semantic}\leftarrow\text{CTX}}$	Max incoming MTL-semantic←CTX weight	1

Table 4: Figure 2 parameters (different than or not specified in default)

Variable	**Description**	**Value**
`MTL-sensory_SEN_projection`	MTL-sensory←SEN random projection	False
N_{days}	Number of days	1000
N_{swaps}	Number of swaps (total noise is $2N_{\text{swaps}}$)	4

Table 5: Figure 3 parameters (different than or not specified in default)

Variable	**Description**	**Value**
`MTL-sensory _SEN_projection`	MTL-sensory←SEN random projection	False
N_{swaps}	Number of swaps (total noise is $2N_{\text{swaps}}$)	4
Figs. 3b-3e		
N_{days}	Number of days	1500
Fig. 3g-3i		
N_{Days}	Number of days (to get classifier train sample, after model trained)	100
Fig. 3h		
$P(A_i, B_j)$	Episode probability (train)	0 if $i = j$; 1/20 if $i \neq j$
$P(A_i, B_j)$	Episode probability (test)	1/5 if $i = j$; 0 else

Table 6: Figure 4 parameters (different than or not specified in default)

Variable	**Description**	**Value**
Interleaved (pre-training)		
N_{days}	Number of days	100
T_{day}	Day length	200
T_{episode}	Episode length	1
Blocked (pre-training)		
N_{days}	Number of days	10 (per concept)
T_{day}	Day length	200
T_{episode}	Episode length	1
$P(A_i, B_j)$	Episode probability	0.2 if block concept present (0 else)
Training		
N_{days}	Number of days	1400

Table 7: Figure 5 parameters (different than or not specified in default)

Variable	**Description**	**Value**
Fig. 5c and 5e (Start with best accuracy network from Fig. 4)		
N_{days}	Number of days	1
T_{day}	Day length	25
T_{episode}	Episode length	5
Fig. 5d (Start with best accuracy network from Fig. 4)		
N_{days}	Number of days	1
T_{day}	Day length	1 to 5000 (spaced)
T_{episode}	Episode length	5
Fig. 5f		
Sensory replayed 1. Pre-train with blocked (see Table 6) 2. 1400 days 3. 100 days frozen to test accuracy		
Sensory-semantic replayed		
$S_L^{\max}$	Max. semantic load during replay	1
1. Pre-train with blocked (see Table 6) 2. 2000 days 3. 100 days frozen to test accuracy		

Table 8: Figure 6 parameters (different than or not specified in default)

Variable	**Description**	**Value**
Start with best accuracy network from Fig. 5f		
N_{days}	Number of days	500

Supplementary Material

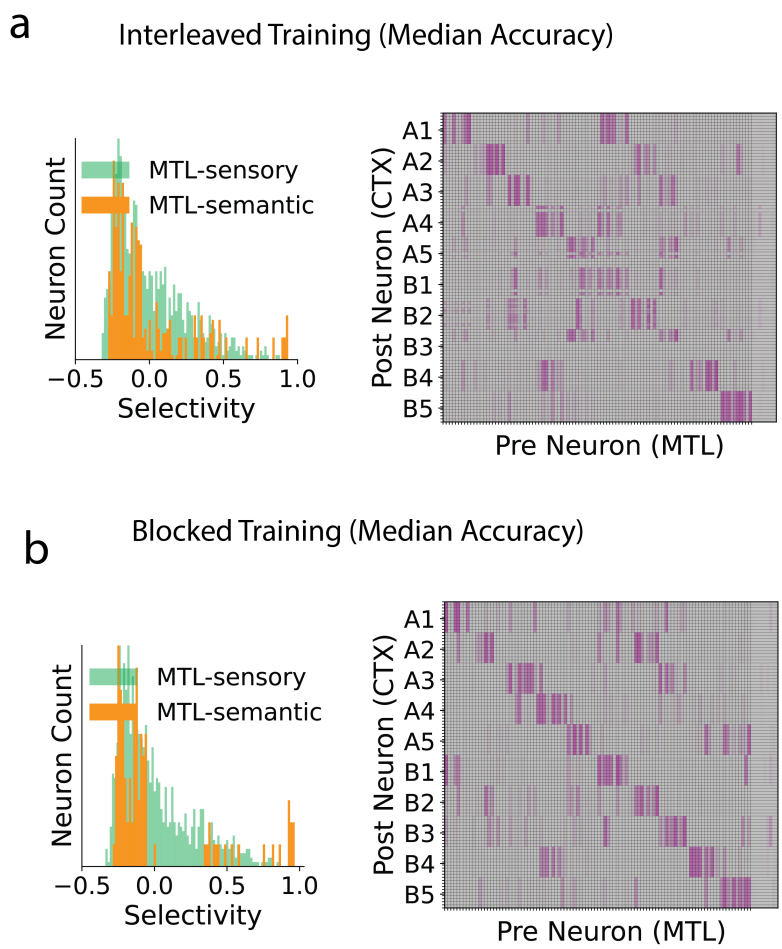

Figure S1: Same as Fig. 4 but for a network with median accuracy in Fig. 4f

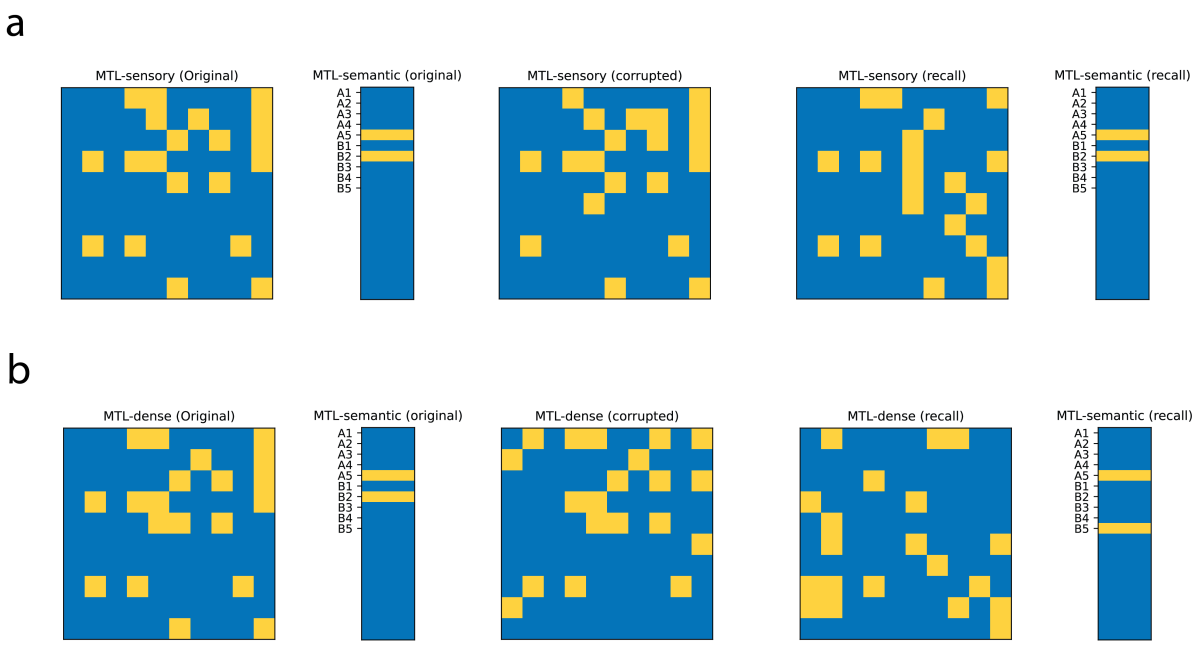

Figure S2: Example recall patterns in Fig. 4d for only one stored episode (**a**) and 1000 stored episodes (**b**). In the limit of infinite stored patterns, a form of over-generalization leads to wrong pattern completion when the presented pattern is of the form $(A_i, B_{j\neq i})$, over-correcting it to the more likely patterns (A_i, B_i) and (A_j, B_j)

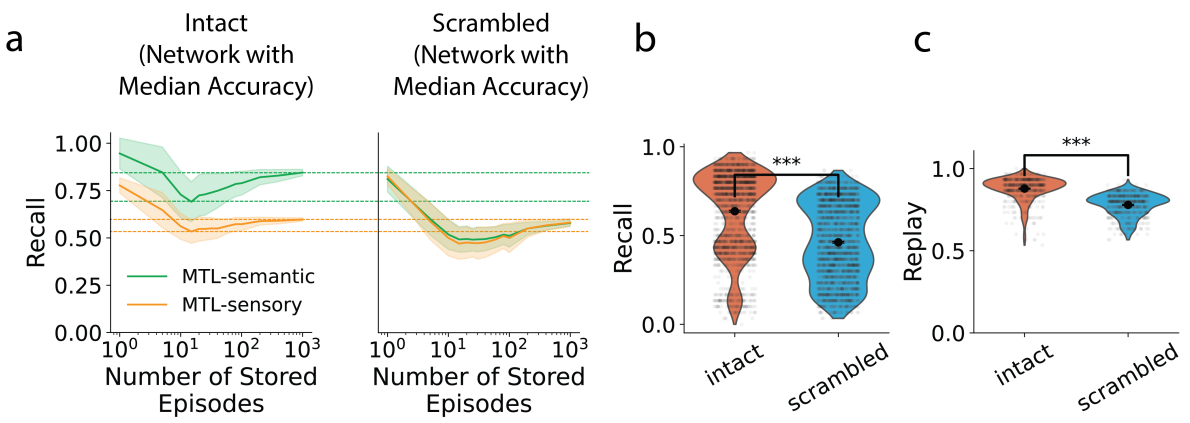

Figure S3: Same as Fig. 5 but for a network with median accuracy in Fig. 4f

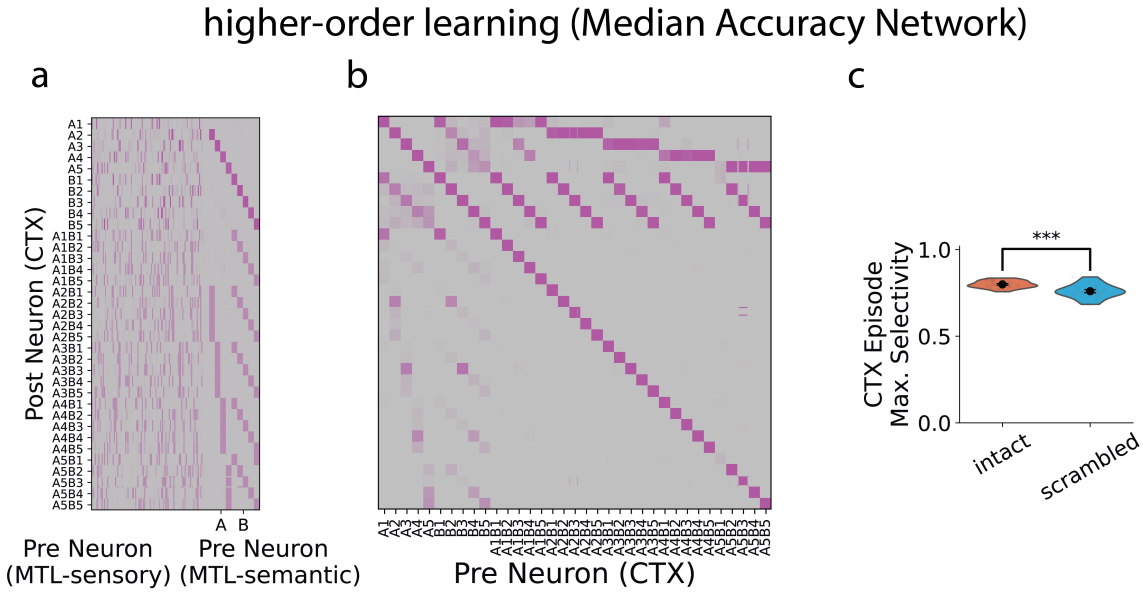

Figure S4: Same as Fig. 6 but starting with a network with median accuracy in Fig. 5f