Hippocampal indexing alters the stability landscape of synaptic weight space allowing life-long learning

Oscar C. González[1,2,*], Ryan Golden[2,3,*], Erik Delanois[2,4],

Bruce L. McNaughton[5,6], Maxim Bazhenov[2,3]

[1]Department of Psychiatry and Behavioral Sciences, Stanford University

[2]School of Medicine, University of California, San Diego

[3]Neurosciences Graduate Program, University of California, San Diego

[4]Department of Computer Science and Engineering, University of California, San Diego

[5]Department of Neurobiology and Behavior, University of California, Irvine

[6]Canadian Centre for Behavioral Neuroscience, University of Lethbridge

[*]Co-first authors

ABSTRACT

Systems-level consolidation holds that the hippocampus rapidly encodes new information during wakefulness, and that coordinated cortico-hippocampal replay during subsequent sleep transfers and stabilizes and refines those traces in cortex. This idea captures key learning principles, but exactly how replay reshapes the synaptic-weight landscape - creating new representations while preserving old ones - remains unclear. To address this, we used a biophysically realistic network model to probe the effects of slow-wave sleep (SWS) on synaptic-weight space. We show that previously learned memories are stable attractors in that space, and that hippocampus-driven interactions between sharp-wave ripples and cortical slow waves push the system into new attractor states that jointly encode old and new memories. As a result, replay allows recently acquired information to be incorporated without degrading prior memories. Our results offer a novel mechanistic - and conveniently "geometric" - framework for understanding how sleep-driven replay sculpts synaptic weights during consolidation.

SIGNIFICANCE STATEMENT

Storing, processing, and retrieving information underpins intelligent behavior. Sleep extracts invariant features from prior experience, promoting the emergence of explicit knowledge and insight. Yet despite abundant empirical findings, our understanding of how sleep reshapes memory representations across brain networks remains limited. Here we present a novel framework that describes how memories are encoded in synaptic-weight space and how sleep dynamics reorganize that landscape. These results advance our understanding of how the brain solves core problems of lifelong learning.

INTRODUCTION

Continual learning is a foundation of human intelligence. To survive in constantly changing environments, human and animal brains must continuously encode and assimilate new memories to guide behavior. Minimizing memory interference is therefore a priority. Not only can the human brain learn without interference, but prior related knowledge often improves new learning - new experiences can build on and refine what we already know (Brod et al., 2016). Our limited understanding of how the brain accomplishes this is highlighted by ongoing attempts to achieve scalable continual learning in artificial neural networks (ANNs), which suffer severe retroactive interference known as catastrophic forgetting (McCloskey and Cohen, 1989; McClelland et al., 1995; French, 1999; Hayes et al., 2021; Kudithipudi et al., 2022).

Sleep has been hypothesized to play an important role in memory consolidation (Walker and Stickgold, 2004; Ji and Wilson, 2007; Lewis and Durrant, 2011). During sleep, neurons are spontaneously active without external input and generate complex patterns of synchronized activity across brain regions (Steriade et al., 1993a). Two critical components which are believed to underlie memory consolidation during sleep are spontaneous replay of memory traces and local unsupervised synaptic plasticity (Wilson and McNaughton, 1994; Stickgold and Walker, 2007; Wei et al., 2016a). Using biophysical models of a thalamocortical network implementing sleep-wake transition (Krishnan et al., 2016), we previously showed that replay of recently learned memories together with relevant old memories can improve learning and enable the network to form orthogonal memory representations which allow overlapping neuronal populations to store competing memories (Wei et al., 2016b; Wei et al., 2018; Gonzalez et al., 2020; Wei et al., 2020; Golden et al., 2025). Related work also revealed that implementing sleep-

like replay in ANNs can mitigate catastrophic forgetting and improve generalization (Tadros et al., 2020b; Tadros et al., 2020a; Tadros and Bazhenov, 2022; Tadros et al., 2022; Delanois et al., 2023; Bazhenov et al., 2024).

While progress has been made, we still lack understanding of the fundamental principles governing sleep-induced dynamics in the synaptic-weight space. There is little doubt that both awake learning and sleep consolidation change the synaptic-weight landscape; however, the extreme dimensionality of that space hinders progress in understanding the principles behind learning- and consolidation-induced synaptic dynamics. In this new study, we used biophysical neural network models to reveal how slow-wave sleep (SWS) alters the synaptic-weight landscape. We found that trained memories converge towards stable attractors under SWS dynamics. We further found that coordinated hippocampal sharp-wave ripples and cortical slow oscillations during offline consolidation keeps the system within a subspace of synaptic weights that supports old memories while guiding it toward intersections of subspaces that support multiple memories, thereby avoiding forgetting. We therefore propose the concept of a “memory manifold” - a subspace of synaptic-weight space that supports strong cued recall - and show that a dual hippocampo-cortical memory system, as proposed by systems-consolidation theory (Wilson and McNaughton, 1994; Rasch and Born, 2013), can efficiently select trajectories along these manifolds in synaptic-weight space to enable continual learning.

RESULTS

The paper is organized as follows. First, we revisit some of our previously published results on the role of sleep in preventing catastrophic forgetting after sequential training (Gonzalez et al., 2020), framed from the perspective of model dynamics in synaptic-weight space – the high-dimensional space composed of all synaptic weight configurations which the network can support. Next, we expand those results to the case of hippocampal indexing and show how indexing can set up learning trajectories so that the network progressively acquires new memories without erasing the old ones. Finally, we discuss a generalization of this idea, proposing that continual learning requires the system to remain near intersections of the memory manifolds - subspaces of synaptic weights which can support cued recall of a particular memory.

Network model design

The network model utilized throughout the study was built upon thalamocortical models developed in our earlier work (Krishnan et al., 2016; Gonzalez et al., 2020). In short (see Methods for details), the basic circuit (Figure 1A) consists of a single cortical layer with excitatory pyramidal cells (PYs) and inhibitory interneurons (INs), and a single thalamic layer with excitatory thalamocortical cells (TCs) and inhibitory reticular interneurons (REs). All neurons were modeled according to the Hodgkin-Huxley formalism, and synaptic connections between cells were set deterministically within a local radius and held at a constant weight value, except for PY-PY synapses. These intracortical synapses were set probabilistically within a local radius with the initial weight values Gaussian distributed (Figure 1B). During awake and sleep states the strength of these synapses was allowed to vary according to the local spike-timing-dependent plasticity (STDP) rules. The model implemented effects of neuromodulators

(Krishnan et al., 2016) to enable transitions between awake state, characterized by asynchronous low-frequency firing, and slow-wave sleep (SWS), characterized by slow oscillations (<1 Hz, (Steriade et al., 1993b)). In this model, cortical network is responsible for generating the sleep slow oscillations (Timofeev et al., 2000; Bazhenov et al., 2002) however, thalamus contributes to synchronization properties of sleep slow-waves (Lemieux et al., 2014).

Differential replay of sequence memories during sleep improves recall performance

An example of an experimental simulation paradigm consisting of testing and training in awake state and a period of SWS is shown in Figure 2A. During the awake state, the network was trained with one of two sequence memories (S1 or S2), each represented by five sequentially ordered cell groups of ten neurons per group (S1=EDCBA or S2=ABCDA, respectively; see Figure 2B, left and middle). These sequences were chosen to elicit maximal interference in simulations where both memories are embedded to the same network. Training proceeded by simulating DC current injections to activate each cell group, with a small delay between groups allowing for STDP to strengthen connections. The awake state was followed by SWS, when the network exhibited spontaneous slow oscillations with silent Down states and active Up states (Figure 2B, right). Additionally, during the awake state, recall was tested to measure performance at baseline, after training the sequence memory, and after sleep (Figure 2C). Recall was assessed by activating the first cell group and testing for pattern completion of the trained sequence (See Figure 2C insets). For single memory training (S1 or S2), recall performance was found to reliably increase for S1 (Figure 2C, top) or S2 (bottom) following both training and subsequent sleep. The coloring for the performance bars indicated which memories the network is capable of recalling: gray – neither; red – S1 only; blue – S2 only; purple – both.

To evaluate differential neural reactivation, we applied PCA to reduce the dimensionality of firing-rate data during Up states in each training condition (see Supp. Figure 1A; see Methods). The Up-state trajectories following training of S1 (red) or S2 (blue) are separable in the first two principal components, indicating robust, reliable differential reactivation. We used this separation to train a linear SVM (one trial per sequence as training data; see Methods) to predict the probability that each Up state reflected an S1 versus S2 replay. Figure 2D shows these replay probabilities for each memory across SWS Up states, averaged over trials. When S1 was trained during wake, S1 replay dominated while S2 replay was rare (Figure 2D, top), and the converse occurred when S2 was trained (Figure 2D, bottom). Replay probabilities varied across individual Up states, but the majority of Up states replayed the memory trained before sleep, and this bias increased over the course of sleep (see Supp. Figure 2A for single-trial replay probabilities).

Sleep can rescue interference effects induced by sequentially training memories

Next, we simulated two memories trained sequentially (S1 → S2) followed by a period of sleep. Under this paradigm, the model exhibited retroactive interference on S1 recall (Fig. 2E, left): S1 performance decreased after S2 training. We also observed prospective interference on S2 recall (Fig. 2E, right), with S2 baseline recall reduced by prior S1 training. Importantly, sleep after S1→S2 training rescued S1 from the damage caused by S2 and further improved S2 performance (first reported in (Gonzalez et al., 2020)). Figure 2F shows that the average replay probability for each sequence during sleep oscillated around 0.5, indicating both memories were replayed approximately evenly. This is well reflected by the average firing rate trajectory of the population during Up states (see Supp. Figure 1A; purple trajectory), which fell directly between

the single memory trajectories. The slight cross-over of the replay probabilities about halfway through sleep suggests the network may have slight biases towards one memory over the other as SWS progresses. Importantly, though, for all trials individual Up states were almost always robustly classified as S1 or S2 (Supp. Fig. 2B), suggesting that different Up states were dominated by distinct memory replays in this case.

Interference can be characterized by the network escaping "memory attractor" in synaptic weight space

Next, we analyzed training- and sleep-induced model dynamics in synaptic-weight space. From the full weight space we constructed a hyperplane defined by three points - the initial weight state (O), the weight state U_1 after training sequence S1, and the weight state U_2 after training S2 - and evaluated recall performance for each sequence (S1, S2) at every grid point on that hyperplane. We then projected the system trajectories (synaptic-weight evolution during training and sleep) and this hyperplane into PC space and assessed S1/S2 performance along those trajectories (Fig. 3). Recall-performance analysis produced contour maps on the hyperplane - the "synaptic-performance landscape" - showing S1 performance (red) and S2 performance (blue) independently (Fig. 3) and joint S1&S2 performance (purple; Supp Fig. 3), which is the average performance on the two memorizes with the absolute difference between the scores subtracted off to penalize asymmetry in performance. See Methods for PCA and landscape details.

Figure 3A shows that initial training of S1 (red arrow) or S2 (blue arrow) pushed the network from its initial location (O) to a new state (U_1 or U_2) in synaptic-weight space, leading to greater recall performance for the corresponding memory. The coloring of the trajectories

reflects which memory can be robustly recalled: gray - neither; red - S1 only; blue - S2 only; purple - both. Subsequent sleep pushes the network further along the same direction as during training, arriving at new synaptic-weight states (S_1 or S_2) with even greater single memory recall performance (Fig. 3B); these states behave as stable attractors under SWS dynamics, with additional sleep leading to small random fluctuations around each. By contrast, if competing-memory training follows U_1 or U_2 instead of sleep (Fig. 3C), the system is driven in the opposite direction in weight space into a "region of ignorance" (gray) where neither sequence can be recalled. This low-performance state (U_{12}) is nevertheless distinct from the initial state (O), because synaptic weights still preserve traces of both memories. Consequently, subsequent sleep (Fig. 3D) can move the network into a region with high recall for both memories - the joint-memory attractor (S_{12}). Importantly, in all cases sleep reveals stable attractor dynamics: the terminal states reached after sufficient sleep (S_1 , S_2 , S_{12}) are stable, such that further sleep produces negligible movement around these attractors.

Sleep moves the network towards stable attractors

Next, we tested basins of attraction of these states (S_1 , S_2 , S_{12}) in respect to the sleep induced dynamics. Figure 3E shows examples of under/overtraining of S2 (blue) memory following initial S1 (red) training. The first undertraining example (Tr_1) halted S2 training before the network left the basin of attraction of S_1 memory attractor, and subsequent sleep pushed the network back towards the S_1 attractor. Supp. Figure 3A shows the average replay probability for this case, with each memory initially replaying at roughly equal probability before S1 replay comes to dominate by the end of sleep. The second undertraining example (Tr_2) halted S2 training just after the network left the basin of attraction of S_1 attractor. In this case, subsequent

sleep pushed the network towards the joint S_{12} memory attractor. Supp. Figure 3B shows that replay initially becomes transiently biased towards S1 before slowly returning towards replay of both memories. Similarly, the first overtraining example (Tr_3) halts the network just after it passes the intersection in the region of ignorance, and subsequent sleep pushed it towards the joint S_{12} memory attractor (Supp. Figure 3C). Finally, in the last two overtraining examples (Tr_4 and Tr_5), the network was trained on S2 until it reached the basin of attraction of S_2 memory attractor, where subsequent sleep pushed the network towards S_2 . Here the average replay probabilities are first balanced before S2 comes to dominate by the end of sleep (Supp. Figure 3D).

Can sleep recover memory states damaged by random synaptic deletion? Biological brains preserve memories over a lifetime despite constant synapse turnover, and sleep may contribute to that resilience. To test this, we started from a model trained on S1 and consolidated in sleep, then reset a fraction of the model's synapses back to their initial (naive) values. Memory loss was gradual, and the memory was relatively resilient to synaptic damage, e.g., when 40% of synapses were reset, performance fell from ~90% to ~65% (Supplementary Fig. 4A). Subsequent sleep recovered a large portion of the lost accuracy: performance rose to ~80% after a 40% synapse reset. Plotting sleep trajectories in synaptic-weight space showed the system moved back toward the attractor representing the damaged memory but settled in a nearby, slightly different state (Supplementary Fig. 4B,C), consistent with the idea that multiple weight configurations can support a given memory trace (Javed and White, 2019; Golden et al., 2022).

Taken together, these results suggest that the synaptic-weight space of the model contains at least three attractors supporting S1, S2, and the joint S1/S2 state. Furthermore, multiple attractor states represented by different synaptic configurations may support the same memory. In the simulations presented here, transitions between attractors required passing through a region of low task performance. Whether the attractor basins are fully disjoint or connected in parts of synaptic-weight space remains unclear, because it is impractical to systematically probe the model's very high-dimensional weight space.

Hippocampal indexing during sleep allows for new memory consolidation without interference

Although sleep could rescue retroactive and proactive interference by moving the network toward a joint memory attractor, it relied on the network passing through a region of ignorance (low task performance on both memories) to do so. The authors are unaware of any study where animals displayed the following behavioral dynamics: underwent catastrophic forgetting on the initial task without achieving significant performance improvements on the competing task but then displayed robust performance on both tasks following sleep. There are two possible solutions. First is to follow each short episode of training by sleep to rescue old memories (Golden et al., 2022). Indeed, when we interleaved short periods of S2 training with short periods of sleep, new sequence S2 was learned without damage to the old memory S1 (Supp. Figure 5). However, this approach was sensitive to the ratio and durations of sleep and S2 training phases – longer training episodes compared to sleep led to S1 forgetting, shorter training failed to learn S2. Thus, in practice this strategy would require multiple training–sleep cycles (i.e., multiple day/night cycles) and relatively low learning rates, consistent with the gradual acquisition of procedural, hippocampus-independent memories (Hong et al., 2019).

Can a new memory be trained fast but without inducing significant interference to old memories? We recently showed (Golden et al., 2025) that dual hippocampo-cortical memory system (McClelland et al., 1995; Born, 2010) provides such solution, so the hippocampal indexing during sleep allows the network to encode the competing memory with minimal interference to the initial memory. To test the synaptic weights dynamics during indexing, we considered hippocampal indexing model from (Golden et al., 2025). The example simulation paradigm (Figure 4A) shows that, following a baseline test, S1 was trained (Figure 4B, left) before conducting another post-training test. Following this, the network was transitioned into SWS and hippocampal indexing of S2 memory was introduced near Down-to-Up transition (Isomura et al., 2006; Sanda et al., 2021) by detecting the onset of each Up State (i.e. the Down-to-Up transition) and applying input to sequentially activate cell groups in the S2 order (Figure 4B, right). A 5 ms delay was applied between stimuli. Hippocampal indexing was chosen to only last for the first half of total sleep duration since rodent studies indicate that hippocampal replay for recent tasks is more robust in early than late sleep (Ji and Wilson, 2007), and declarative memory consolidation is thought to be more strongly associated with early sleep (Plihal and Born, 1999; Mednick et al., 2011; Rasch and Born, 2013). Finally, network performance for both tasks was tested after sleep.

Figure 4C shows that indexing produced a significant increase in S2 recall (right) that persisted during subsequent sleep without indexing, while recall performance for S1 (left) continued to rise across all stages of the simulation. The average replay probabilities (Figure 4D) indicate that the network became progressively more likely to replay S2 over the course of indexing. Stopping

indexing when the two memories had roughly equal replay probability allowed both to continue being replayed during subsequent sleep. Individual Up states from single-network simulations were typically classified robustly as either S1 (red) or S2 (blue) replays (Supp. Figure 2C), consistent with the sequential-training/SWS case. If, however, indexing was allowed to continue throughout the entire sleep period, catastrophic retroactive interference occurred: the old memory S1 degraded (Figure 4E, left) while the new competing memory S2 became extremely robust (Figure 4E, right). Figure 4F shows that this happened because S2 came to dominate replay well before the end of sleep.

Supp. Figure 1B shows that average firing-rate trajectories during Up states with indexing also fell between the single-memory trajectories described earlier. When we classified Up states from indexing and SWS into three groups by SVM-derived reactivation probability (>95% S1 = red; >95% S2 = blue; otherwise = purple), robust S1 and S2 replays were linearly separable in 2D PC space, whereas less-confident states followed intermediate trajectories consistent with mixtures of both memories.

Hippocampal indexing moves the network to its joint memory attractor without damaging old memories

Examining system dynamics in synaptic weight space in the case of indexing was found to require a third principal component in addition to the previous two needed to visualize the single memory and sequential memory training paradigms. This fact prohibited generation of the synaptic performance landscape, as we cannot visualize three dimensional contours. On account of this, Figure 5 displays two different perspective (top and bottom plots) of a three-dimensional

principal component representation of the paths of the networks through synaptic weight space, with the trajectory coloring encoding which memory can be robustly recalled: gray - neither; red - S1 only; blue - S2 only; purple - both. As previously, naive network (O) was initially trained by S1 (or S2) memory, that moved it to U_1 (or U_2) state in synaptic weight space. Next, indexing of alternative memory S2 (or S1) moved the system negatively (resp. positively) in the third principal component dimension towards a newly discovered stable attractor $S_{2/1}$ (or $S_{1/2}$) (Fig. 5A). In the case of over indexing (Fig. 5B), the network was pushed past the new stable regions and continued onto the opposite single memory attractor S_1' (or S_2') . Note that these new single memory attractors (S_1' and S_2') were different from original ones (S_1 and S_2), suggesting coexisting multiple memory attractors in synaptic weights space, as was predicted in another study (Golden et al., 2022). Importantly, with either indexing or over-indexing, the network never passes through a "region of ignorance" and can always robustly recall at least one of the two memories.

Hippocampal indexing results in sparser stable solutions

Both sequential training of two memories (S1 → S2) in wake followed by sleep (Fig. 3) and training one memory in wake followed by indexing of another during sleep (Fig. 5) produced synaptic-weight states that supported both memories (S_{12}, $S_{2/1}$, $S_{1/2}$; see "purple" attractors in Figure 5). What differences exist between these states beyond their specific weight configurations? Figures 6A-B plot the PY→PY synaptic weights for all pairs of bi-directionally connected neurons for the sequential-training case (Fig. 6A) and the indexing case (Fig. 6B). We focus on bi-directional connections because pairs without bi-directionality are not subject to competition under our STDP rules. In Figs. 6A–B, the colored corner regions mark pairs in

which one synapse had been prioritized for S1 (red) or S2 (blue), effectively disconnecting the opposing synapse.

In both paradigms the initial weight values were roughly Gaussian (Figs. 6A–B, left), and the bulk of the density shifted toward the red corner after S1 training (Figs. 6A–B, middle-left). After this point the two paradigms diverged. Both sequential training and indexing produced more symmetric density distributions overall, indicating a roughly equal allocation of synaptic resources to the two memories. However, S2 training during wake (Fig. 6A, middle-right) pushed the density much farther into the top-right corner than S2 indexing during sleep (Fig. 6B, middle-right), which kept the density localized around the diagonal connecting the red and blue corners ($Y = -X$). This pattern implies that sequential training yields many bi-directional pairs that are strong in both directions - an ambiguity that prevents those pairs from contributing to differential memory encoding. By contrast, hippocampal indexing preferentially segregates (orthogonalizes) synaptic resources between the two memories. In both paradigms, subsequent sleep without indexing amplified any preexisting S1/S2 bias in each bi-directional pair, pushing pairs further into either the red or blue corner (Figs. 6A–B, right).

Figure 6C shows the synaptic-weight distributions obtained after the sequential-training (orange) and indexing (green) paradigms for the corresponding simulation phases above. The distribution reached by indexing during sleep is noticeably sparser than the distribution produced by sequential training (Fig. 6C, middle-right), consistent with greater segregation of synaptic resources under indexing.

To quantify the density of learned synaptic representations, we used a measure originally developed to compute the sparseness of neural population codes (Treves and Rolls, 1994) (see **Methods** for more details). Low values indicate that information is carried by a few synapses (low-density synaptic representation), whereas values near 1 indicate that information is distributed broadly across synapses (high-density representation). Figure 6D plots this density at the same four time points shown in Fig. 6A–C. In both paradigms, initial S1 training reduced representational density. However, sequential S2 training (orange bars) increased density, while S2 indexing during SWS (green bars) reduced it — a pattern consistent with stronger interference following sequential training than with indexing. Finally, subsequent sleep after either paradigm produced less dense representations overall, but indexing located regions of weight space that were significantly sparser (and thus more resource-efficient) than those reached by sequential training.

DISCUSSION

The human brain contains an estimated 10^{14} synapses and is in constant interaction with the environment, leading to continual perturbations of its synaptic weight matrix. As demonstrated in artificial neural networks (ANNs), continuous acquisition of new information can induce interference and catastrophic forgetting (Mccloskey and Cohen, 1989; McClelland et al., 1995; French, 1999; Hayes et al., 2021), and can also drive synaptic saturation and loss of plasticity (Dohare et al., 2024). How, then, does the vast pool of memories accumulated over a lifetime persist despite ongoing perturbations, and without requiring explicit relearning?

The critical role of sleep in learning and memory is supported by a vast, interdisciplinary literature spanning psychology and neuroscience (Paller and Voss, 2004; Walker and Stickgold, 2004; Oudiette et al., 2013; Rasch and Born, 2013; Stickgold, 2013). The central idea is that sleep contributes to memory consolidation through repeated reactivation, or replay, of specific traces acquired during learning (Hennevin et al., 1995; Paller and Voss, 2004; Mednick et al., 2011; Oudiette et al., 2013; Rasch and Born, 2013; Lewis et al., 2018; Wei et al., 2018). In other words, replay during sleep strengthens weak, potentially labile representations encoded in wake, promoting long-term stabilization. However, how sleep dynamics reshape synaptic representations of old and new memories in high-dimensional synaptic weight space - so that new knowledge is embedded while interference is minimized - remains unknown.

We propose an intriguing idea: memories are stable attractors in synaptic-weight space with respect to autonomous cortical slow-wave sleep dynamics. Thus, while daytime experiences nudge the network away from its memory-supporting configuration, sleep-induced replay tends to return the system to the attractor state representing previously learned memories. A sufficiently strong perturbation - e.g., learning a new task similar to previous experiences - may push the system toward the basin of attraction of a new memory, inducing forgetting, or toward an attractor that supports both memories (a joint state), enabling continual incremental learning. Indeed, we recently successfully implemented a similar approach in feedforward ANNs, where sleep-like phase prevented catastrophic forgetting and improved generalization (Tadros et al., 2020a; Tadros et al., 2022; Delanois et al., 2023).

Figure 7 summarizes these concepts with illustrative schematics. Two states (S1 and S2) represent stable attractors of individual memories (M1 and M2); the third state, S12, corresponds to an attractor supporting both memories. When initial training brings the system into the vicinity of a task-specific attractor (e.g., S1), subsequent sleep replay induces convergence toward that attractor. Limited perturbations of the S1 attractor are compensated by sleep dynamics that return the system to S1. Learning a new task during wake shifts the system away from the old attractor (S1) toward the new attractor (S2) (Figure 7A-C). With subsequent sleep, it is possible to transition to the joint-task attractor (S12) (Figure 7A), but this requires a finely tuned training duration so the system lands in the basin of S12; otherwise it may fail to consolidate the new memory (proactive interference; return to S1, Figure 7B) or induce catastrophic forgetting (retroactive interference; transition to S2, Figure 7C). Is it possible to avoid this sensitivity and enables reliable continual learning that benefits from sleep dynamics?

Our current study and previous (Golden et al., 2022) work revealed that more than one synaptic weight configuration may support a giving memory. Furthermore, analysis of the loss landscape in artificial neural networks (ANNs) has revealed that independently trained stochastic-gradient-descent solutions can be connected with simplicial complexes that form multi-dimensional manifolds of low loss; any point on that manifold will deliver high-accuracy (low-loss) solutions for a given task (Benton et al., 2021). Expanding this idea to brain dynamics, we suggest that memory attractors may be connected by continuum of states all supporting a given memory trace, thus forming a memory manifold. While not all the points on these manifolds are attractors in respect to sleep dynamics, all of them support given memory. Thus, optimal continual learning strategy needs to keep the model near the manifold representing the old

memories while the new memory is learned, effectively converging to intersections of old and new memory manifolds.

One solution is to interleave new training with periods of sleep to rescue any damage to old tasks (Figure 7D). For procedural, hippocampus-independent memories, new learning is slow; therefore interference with old memories is likely minimal after each training session. In such cases, each subsequent sleep can easily mitigate this minor damage, maintaining high performance on old memories. As we report here and in our previous work (Golden et al., 2022), this solution can work but is sensitive to the relative durations of training and sleep episodes, again requiring fine-tuning of the learning-period duration.

We report here that a memory system involving a fast-learning hippocampus and a slow-learning cortex, as in declarative memory, provides an alternative solution (Figure 7E). In such a case, slow waves co-replay new hippocampus-driven memory and relevant older cortical memories. This dynamic allows the system to remain near the M1 task manifold while converging toward its intersection with the M2 manifold; once it reaches the basin of attraction of the joint state S12, sleep without indexing can move it further into S12 (Figure 7F). While over-indexing can still damage the old cortical memory, the duration of the indexing process is independent of learning duration and is controlled by intrinsic hippocampal mechanisms that limit hippocampo-cortical dialogue to the first 30–60 min of post-learning sleep (Wilson and McNaughton, 1994; Ji and Wilson, 2007; Eschenko et al., 2008).

It is interesting to note that our model suggests the complementary learning-systems framework can be seen as a maximally compressed version of the interleaved-training strategy used for procedural memory consolidation, with the indexed and post-indexed phases corresponding to training and sleep, respectively. From that perspective, both implement interleaved training—a long-proposed general solution to catastrophic forgetting. Indeed, our results align with a large literature showing that interleaved training mitigates catastrophic forgetting in ANNs (McClelland et al., 1995; Hasselmo, 2017; Flesch et al., 2018) and SNNs (Evans and Stringer, 2012; Higgins et al., 2017), which inspired replay-like algorithms that store a subset of prior veridical inputs and mix them with recent inputs during updates (reviewed in (Hayes et al., 2021)). We propose that the brain avoids this data-intensive storing/retraining approach either by separating slow, wake learning from subsequent spontaneous reactivation during sleep (procedural memories), or by embedding spontaneous reactivation of older traces within the training process itself via a hippocampo–cortical dual-memory system (declarative memory), achieving off-line–like cortical consolidation without explicit storage of old data.

In sum, our study suggests that interactions between the hippocampus and neocortex during slow-wave sleep provide optimal dynamics to move the system between stable memory attractors along the memory manifolds. We found complex attractor dynamics during SWS, with multiple distinct attractors representing individual memories and joint states supporting the coexistence of a few competing memory traces. These attractors are stable with respect to autonomous sleep dynamics. Importantly, as hippocampal indexing pushes the system between basins of attraction toward attractor states supporting coexisting memories, it never loses the old memory.

446

447

448 **METHODS**

449 *Thalamocortical network model*

450

451 *Network architecture*. Throughout this study, we make use of a modified version of a
452 thalamocortical network which has been previously described in detail (Wei et al., 2016b;
453 Gonzalez et al., 2020). In brief, the network consisted of a cortical module containing 500
454 excitatory pyramidal neurons (PYs) and 100 inhibitory interneurons (INs), and a thalamic
455 module containing 100 excitatory thalamocortical neurons (TCs) and 100 inhibitory reticular
456 interneurons (REs). Connectivity in the network was determined by cell type and a local radius
457 (see Fig. 1), and excitatory synapses were mediated by AMPA and/or NMDA currents, while
458 inhibitory synapses were mediated by $GABA_A$ and/or $GABA_B$ currents.

459

460 In the cortex, PYs synapsed onto PYs and INs with a radii of $R_{AMPA(PY-PY)} = 20$, $R_{NMDA(PY-PY)} =$
461 5, $R_{AMPA(PY-IN)} = 1$, and $R_{NMDA(PY-IN)} = 1$. All connections were deterministic within these radii,
462 expect for AMPA synapses between PYs, which had a 60% probability of connection.
463 Additionally, INs synapsed onto PYs with a radius of $R_{GABA-A(IN-PY)} = 5$. In the thalamus, TCs
464 synapsed onto REs with a radius of $R_{AMPA(TC-RE)} = 8$ and REs synapsed onto REs and TCs with
465 radii of $R_{GABA-A(RE-RE)} = 5$, $R_{GABA-A(RE-TC)} = 8$, and $R_{GABA-B(RE-TC)} = 8$. Between the cortex and
466 thalamus, TCs synapsed onto PYs and INs with radii of $R_{AMPA(TC-PY)} = 15$, $R_{AMPA(TC-IN)} = 3$,
467 while PYs synapsed onto TCs and REs with radii of $R_{AMPA(PY-TC)} = 10$, and $R_{AMPA(PY-RE)} = 8$.

468

Wake – Sleep transitions. To model the state transitions between awake and N3 sleep, we modulated the intrinsic and synaptic currents of our neuron models to account for differing concentrations of neuromodulators that partially govern these arousal state transitions. As these mechanisms have been described in detail in (Krishnan et al., 2016), here we will simply outline the approach. The model included the effects of changing acetylcholine (ACh), histamine (HA), and GABA concentrations as follows: ACh – by modulating the potassium leak current in all cell types, as well as excitatory AMPA synapses within the cortex; HA – by modulating the hyperpolarization-activated cation current in TC cells; and GABA – by modulating inhibitory GABAergic synapses within the cortex and thalamus. To transition the network from awake to sleep, we modeled the effects of reduced ACh and HA but increased GABA concentrations to reflect experimental observations (Vanini et al., 2012).

Intrinsic currents. All cell types were modeled using the Hodgkin-Huxley formalism, and cortical PYs and INs contained dendritic and axo-somatic compartments that have been previously described (Wei et al., 2018). The dynamics of the membrane potential were modeled according to:

$$C_m \frac{dV_D}{dt} = -I_D^{Na} - I_D^{NaP} - I_D^{Km} - I_D^{KCa} - ACh_{gkl} I_D^{KL} - I_D^{HVA} - I_D^{L} - g(V_D - V_S) - I^{syn},$$

$$g(V_D - V_S) = -I_S^{Na} - I_S^{NaP} - I_S^{K},$$

where C_m is the membrane capacitance, $V_{D,S}$ are the dendritic and axo-somatic membrane voltages respectively, I^{Na} is the fast sodium (Na^+) current, I^{NaP} is the persistent Na^+ current, I^{Km} is the slow voltage-dependent non-inactivating potassium (K^+) current, I^{KCa} is the slow calcium (Ca^{2+})-dependent K^+ current, ACh_{gkl} represents the change in K^+ leak current I^{KL} which

is dependent on the level of ACh during the different arousal states, I^{HVA} is the high-threshold Ca^{2+} current, I^{L} is the chloride (Cl^{-}) leak current, g is the conductance between the dendritic and axo-somatic compartments, and I^{syn} is the total synaptic current input to the neuron. IN neurons contained all intrinsic currents present in PY with the exception of the I^{NaP}. All intrinsic ionic currents (I^{j}) were modeled in a similar form:

$$I^{j} = g_{j} m^{M} h^{N} \left(V - E_{j}\right).$$

where g_{j} is the maximum conductance, m (activation) and h (inactivation) are the gating variables, V is the voltage of the compartment, and E_{j} is the reversal potential of the ionic current. The gating variable dynamics are described as follows:

$$\frac{dx}{dt} = -\frac{x - x_{\infty}}{\tau_{x}},$$

$$\tau_{x} = \frac{(1/(\alpha_{x} + \beta_{x}))}{Q_{T}},$$

$$x_{\infty} = \frac{\alpha_{x}}{(\alpha_{x} + \beta_{x})},$$

where $x = m$ or h, τ is the time constant, Q_{T} is the temperature related term, $Q_{T} = Q^{((T-23)/10)} = 2.9529$, with $Q = 2.3$ and $T = 36$.

In the thalamus, TCs and REs contained a single compartment with membrane potential dynamics given by:

$$C_{m} \frac{dV_{D}}{dt} = -I^{Na} - I^{K} - ACh_{gkl} I^{KL} - I^{T} - I^{h} - I^{L} - I^{syn},$$

where I^{Na} is the fast Na^{+} current, I^{K} is the fast K^{+} current, I^{KL} is the K^{+} leak current, I^{T} is the low-threshold Ca^{2+} current, I^{h} is the hyperpolarization-activated mixed cation current, I^{L} is the Cl^{-} leak current, and I^{syn} is the total synaptic current input to the neurons. The I^{h} current was

only expressed in TCs. The influence of histamine (HA) on I^h was implemented as a shift in the activation curve by HA_{gh} as described by:

$$m_\infty = \frac{1}{1 + exp\left(\frac{V + 75 + HA_{gh}}{5.5}\right)}.$$

Synaptic currents. The equations for our synaptic current models have been described in detail in our previous studies (Krishnan et al., 2016; Wei et al., 2018). To model the effects of ACh and GABA, we modified the standard equations as follows:

$$I_{syn}^{GABA} = \gamma_{GABA_A}\, g_{syn}\, [O](V - E_{syn}),$$

$$I_{syn}^{AMPA} = ACh_{AMPA}\, g_{syn}\, [O](V - E_{syn}),$$

where g_{syn} is the maximal conductance at the synapse, $[O]$ is the fraction of open channels, and E_{syn} is the channel reversal potential ($E_{GABA\text{-}A}$ = -70 mV, E_{AMPA} = 0 mV, and E_{NMDA} = 0 mv). The parameter γ_{GABA_A} modulated the GABA synaptic currents for IN-PY, RE-RE, and RE-TC connections. For INs γ_{GABA_A} was 0.22 and 0.44 for awake and N3 sleep, respectively, while for REs γ_{GABA_A} was 0.6 and 1.2. ACh_{AMPA} defined the influence of ACh levels on AMPA synaptic currents for PY-PY, TC-PY, and TC-IN. For PYs ACh_{AMPA} was 0.133 and 0.4332 for awake and N3 sleep, respectively, while for TCs ACh_{AMPA} was 0.6 and 1.2.

In addition to spike-triggered post-synaptic potentials (PSPs), spontaneous miniature PSPs (mPSPs) were implemented for both excitatory and inhibitory synapses within the cortex. The dynamics are similar to the typical PSPs described above, but the arrival times were governed by an inhomogeneous Poisson process where the next release time $t_{release}$ is given by:

$$t_{release} = (2/(1 + exp(-(t - t_0)/v)) - 1)/250\,,$$

where t_0 is the time of the last presynaptic spike, and υ was the mPSP frequency ($\upsilon_{mini(PY-PY)}^{AMPA} = 30$, $\upsilon_{mini(PY-IN)}^{AMPA} = 30$, and $\upsilon_{mini(IN-PY)}^{GABA} = 30$). The maximum conductances for mPSPs were $g_{mini(PY-PY)}^{AMPA} = 0.03\ \mu S$, $g_{mini(PY-IN)}^{AMPA} = 0.02\ \mu S$, and $g_{mini(IN-PY)}^{GABA} = 0.02\ \mu S$.

Finally, short-term synaptic depression was also implemented in AMPA synapses within the cortex. To model this phenomenon, the maximum synaptic conductance was multiplied by a depression variable ($D \leq 1$), which represents the amount of available "synaptic resources" as described in (Bazhenov et al., 2002). This short-term depression was modeled as follows:

$$D = 1 - \left(1 - D_i(1 - U)\right) exp\left(-\frac{t - t_i}{\tau}\right),$$

where D_i is the value of D immediately before the i_{th} event, $(t - t_i)$ is the time after the i_{th} event, $U = 0.073$ is the fraction of synaptic resources used per action potential, and $\tau = 700ms$ is time constant of recovery of synaptic resources.

Spike-timing-dependent plasticity. The potentiation and depression of AMPA synapses between PYs were governed by the following spike-timing-dependent plasticity (STDP) rule:

$$g_{AMPA} \leftarrow g_{AMPA} + g_{max}\, F(\Delta t),$$

$$F(\Delta t) = \begin{cases} A_+\, e^{-|\Delta t|/\tau_+}, if\ \Delta t > 0 \\ -A_-\, e^{-|\Delta t|/\tau_-}, if\ \Delta t < 0 \end{cases}$$

where g_{max} was the maximal conductance of g_{AMPA}, F was the STDP kernel, and □t was the relative timing of the pre- and post-synaptic spikes. The maximum potentiation/depression were set to $A_{+/-} = 0.002$, while the time constants were set to $t_{+/-} = 20$ ms. A_- was reduced to 0.001 during training to reflect the effects of changes in acetylcholine concentration during focused

attention on synaptic depression during task learning observed experimentally (Blokland, 1995; Shinoe et al., 2005; Sugisaki et al., 2016).

Sequence training and testing. Training and testing of memory sequences was performed similarly to our previous study (Wei et al., 2018). In brief, each sequence was comprised of the same 5 groups of 10 PYs (i.e PYs 200 - 249), with Sequence 1 (S1) ordered E(240-249), D(230-239), C(220-229), B(210-119), A(200-209), and Sequence 2 (S2) ordered A(200-209), B(210-219), C(220-229), D(230-239), E(240-249). Each training bout consisted of sequentially activating each group via a 10 ms direct current pulse with a 5 ms delay between group activations. Training bouts occurred every 1 s during the training period. This training structure was chosen to ensure strong interference between S1 and S2 according to our STDP rule. Test bouts occurred every 1 ms during testing periods, in which only the first group in each sequence was activated (E for S1; A for S2), and recall performance was measured based on the extent of pattern completion for the remainder of the sequence within a 350 ms window.

Data Analysis

All analyses were performed with standard MatLab and Python functions. Data are presented as mean □ standard error of the mean (SEM) unless otherwise stated. For each experiment a total of 6 simulations with different random seeds were used for statistical analysis.

Sequence performance measure. A detailed description of the performance measure used during testing can be found in (Wei et al., 2018) and the code is available in (https://github.com/o2gonzalez/sequencePerformanceAnalysis)(González, 2020). Briefly, the performance of the network on recalling a given sequence following activation of the first group

of that sequence was measured by the percent of successful sequence recalls. We first detected all spikes within the predefined 350 ms time window for all 5 groups of neurons in a sequence. The firing rate of each group was then smoothed by convolving the average instantaneous firing rate of the group's 10 neurons with a Gaussian kernel with window size of 50 ms. We then sorted the peaks of the smoothed firing rates during the 350 ms window to determine the ordering of group activations. Next, we applied a string match (SM) method to determine the similarity between the detected sequences and an ideal sequence (ie. A-B-C-D-E for S1). SM was calculated using the following equation:

$$SM = 2 * N - \sum_{i=1}^{N} |L(S_{test}, S_{sub}[i]) - i|,$$

where N is the sequence length of S_{test}, S_{test} is the test sequence generated by the network during testing, S_{sub} is a subset of the ideal sequence that only contains the same elements of S_{test}, and $L(S_{test}, S_{sub}[i])$ is the location of the element $S_{sub}[i]$ in sequence S_{test}. SM was then normalized by double the length of the ideal sequence. Finally, the performance was calculated as the percent of recalled sequences with $SM \geq Th = 0.8$, where Th is a threshold indicating that the recalled sequence must be at least 80% similar to the ideal sequence to be counted as a successful recall as previously done in (Wei et al., 2018).

Representation of firing rate space. To visualize dimensionality reduced trajectories during Up-states in firing rate space, we took single random seeds of a simulations in which only S1 or only S2 was trained prior to N3 sleep and detected Up-states in each. We then generated spike rasters of the PY activity in the trained region during each Up-state, and converted these into firing rates by taking a moving average of the spike rasters with a sliding window length of 10 ms, resulting

in a set of N-by-T_j matrices, where $N = 50$ was the number of PYs, and T_j was the duration of jth Up-state with $j = 1, \ldots N_{up}$. The firing rate matrices were then interpolated to be the same duration across each up state, and concatenated, resulting in an N-by-($T_{\max}$*N_{up}) matrix where T_{max} was the duration of longest Up-state across all data sets. Principal components analysis (PCA) was then performed on the firing rate data to reduce it from 50 to 2 dimensions. This linear PCA kernel was then applied to the data from all random seeds for a particular simulation paradigm, and the mean and standard deviation of PCs 1 and 2 were plotted for visualization.

Replay probability. To estimate the probability of S1 and S2 being replayed during a given Up-state, we took the interpolated firing rate data for each Up-state that was used to train the PCA kernel, unrolled the N-by-T_j matrices into (NxT_j)-dimensional column vectors, providing the observations to train a linear support vector machine (SVM), with labels determined by whether the Up-state came from the simulation with S1 or S2 training before N3 sleep, and scores were transformed to posterior probabilities. This SVM was then used to predict the posterior probabilities of S1 and S2 replay for each Up-state for a given random seed of a particular simulation paradigm. To compute the average posterior probabilities, we first interpolated the data so that each random seed had the same number of data points – specifically, the maximum number of Up-states in a single simulation from that paradigm.

Representation of synaptic weight space. In order to visualize the trajectories of the network through synaptic weight space, we trained a linear PCA kernel on the synaptic weight timeseries data of all synapses in the trained region from every random seed of each simulation paradigm discussed in the paper, except for the simulations involving over/under-training, or synaptic

resetting. The data was then transformed into PC space, with either 2 or 3 dimensions retained for plotting.

Performance contours in synaptic weight space. In order to construct the recall performance contours, we first estimated the locations of the potential attractor sites by computing the centroids of the final weight state configurations for the S1 training, S2 training, and the sequential training simulation paradigms in full-dimensional weight space, found the unique 2-dimensional planar subspace which these three points define, and then densely sampled weight state configurations from this subspace. These weight state configurations were then input into the network model to have recall performance evaluated. These sampled weight states were then projected into 2-dimensional PC space along with the corresponding mean performance values for S1 and S2 individually (Figure 3; red and blue), and S1&S2 jointly (Supp Figure 3; purple), and a mesh with contour gradients was computed for each set of recall performance measures.

Desnity of synaptic weight states. In order to compute the density of synaptic weight matrices, we adpated a measure previously conceived to compute the sparsity of neural population codes, operating over real-valued firing rates (Treves and Rolls, 1994),

$$a = \frac{\left(\sum_{i=1}^{n} \frac{r_i}{n}\right)^2}{\sum_{i=1}^{n} \left(\frac{r_i^2}{n}\right)}$$

where a is the density of the population code, r_i is the firing rate of the ith neuron, and n is the total number of neurons. The density here is maximal at 1 when all neurons have the same firing rate, $1/n$ when only one neuron responds, and zero when the population is silent.

To adapt this to compute the density of synaptic weight states, we first unrolled the synaptic weight matrix into a vector and then removed all non-existent synapses (i.e. zero-valued) from the vector so that we only computed over the number of existing synapses rather than potential synapses. We then computed

$$a = \frac{\left(\sum_{i=1}^{n} \frac{w_i}{n}\right)^2}{\sum_{i=1}^{n} \left(\frac{w_i^2}{n}\right)}$$

where a is the density of the weight state, w_i is the firing rate of the ith synapse, and n is the total number of synapses. Thus, the density here is maximal at 1 when all synapses have the same strength, and near $1/n$ when only one synapse is above its minimum strength value (note that existing synapses could not drop to zero strength).

ACKNOWLEDGEMENTS

This work was supported by NIH (1R01MH125557, 1RF1NS132913), NSF (2223839), and NIH (RF1NS132041)

References

Bazhenov A, Dewasurendra P, Krishnan G, Delanois E (2024) Sleep-Like Unsupervised Replay Improves Performance when Data are Limited or Unbalanced. In: AAAI-24.

Bazhenov M, Timofeev I, Steriade M, Sejnowski TJ (2002) Model of thalamocortical slow-wave sleep oscillations and transitions to activated states. The Journal of neuroscience : the official journal of the Society for Neuroscience 22:8691–8704.

Benton G, Maddox W, Lotfi S, Wilson AGG (2021) Loss Surface Simplexes for Mode Connecting Volumes and Fast Ensembling. In: Proceedings of the 38th International Conference on Machine Learning (Marina M, Tong Z, eds), pp 769—779. Proceedings of Machine Learning Research: PMLR.

Blokland A (1995) Acetylcholine: a neurotransmitter for learning and memory? Brain Res Brain Res Rev 21:285–300.

Born J (2010) Slow-wave sleep and the consolidation of long-term memory. World J Biol Psychiatry 11 Suppl 1:16–21.

Brod G, Lindenberger U, Wagner AD, Shing YL (2016) Knowledge Acquisition during Exam Preparation Improves Memory and Modulates Memory Formation. The Journal of neuroscience : the official journal of the Society for Neuroscience 36:8103–8111.

Delanois E, Ahuja A, Krishnan GP, Tadros T, McAuley J, Bazhenov M (2023) Improving robustness of convolutional networks through sleep-like replay. 2023 International Conference on Machine Learning and Applications (ICMLA):10.1109/ICMLA58977.52023.00043.

Dohare S, Hernandez-Garcia JF, Lan Q, Rahman P, Mahmood AR, Sutton RS (2024) Loss of plasticity in deep continual learning. Nature 632:768–774.

Eschenko O, Ramadan W, Mölle M, Born J, Sara SJ (2008) Sustained increase in hippocampal sharp-wave ripple activity during slow-wave sleep after learning. Learn Mem 15:222–228.

Evans BD, Stringer SM (2012) Transformation-invariant visual representations in self-organizing spiking neural networks. Front Comput Neurosci 6:46.

Flesch T, Balaguer J, Dekker R, Nili H, Summerfield C (2018) Comparing continual task learning in minds and machines. Proc Natl Acad Sci U S A 115:E10313–E10322.

French RM (1999) Catastrophic forgetting in connectionist networks. Trends Cogn Sci 3:128–135.

Golden R, Delanois JE, Sanda P, Bazhenov M (2022) Sleep prevents catastrophic forgetting in spiking neural networks by forming a joint synaptic weight representation. PLoS computational biology 18:e1010628.

Golden R, Saxena R, González O, Delanois J, Kilianski S, McNaughton B, Bazhenov M (2025) Interleaved Replay of Novel and Familiar Memory Traces During Slow-Wave Sleep Prevents Catastrophic Forgetting. bioRxiv 20250625661579.

Gonzalez OC, Sokolov Y, Krishnan GP, Delanois JE, Bazhenov M (2020) Can sleep protect memories from catastrophic forgetting? Elife 9.

González OC (2020) sequencePerformanceAnalysis. GitHub. https://github.com/o2gonzalez/sequencePerformanceAnalysis . 094c4be. In.

Hasselmo ME (2017) Avoiding Catastrophic Forgetting. Trends Cogn Sci 21:407–408.

Hayes TL, Krishnan GP, Bazhenov M, Siegelmann HT, Sejnowski TJ, Kanan C (2021) Replay in Deep Learning: Current Approaches and Missing Biological Elements. Neural computation 33:2908–2950.

Hennevin E, Hars B, Maho C, Bloch V (1995) Processing of learned information in paradoxical sleep: relevance for memory. Behav Brain Res 69:125–135.

Higgins I, Stringer S, Schnupp J (2017) Unsupervised learning of temporal features for word categorization in a spiking neural network model of the auditory brain. PloS one 12:e0180174.
Hong JY, Gallanter E, Muller-Oehring EM, Schulte T (2019) Phases of procedural learning and memory: characterisation with perceptual-motor sequence tasks. J Cogn Psychol (Hove) 31:543–558.
Isomura Y, Sirota A, Ozen S, Montgomery S, Mizuseki K, Henze DA, Buzsaki G (2006) Integration and segregation of activity in entorhinal-hippocampal subregions by neocortical slow oscillations. Neuron 52:871–882.
Javed K, White M (2019) Meta-Learning Representations for Continual Learning. In: arXiv e-prints, p arXiv:1905.12588.
Ji D, Wilson MA (2007) Coordinated memory replay in the visual cortex and hippocampus during sleep. Nature neuroscience 10:100–107.
Krishnan GP, Chauvette S, Shamie I, Soltani S, Timofeev I, Cash SS, Halgren E, Bazhenov M (2016) Cellular and neurochemical basis of sleep stages in the thalamocortical network. Elife 5.
Kudithipudi D, Aguilar-Simon M, Babb J, Bazhenov M, Blackiston D, Bongard J, Brna AP, Raja SC, Cheney N, Clune J, Daram A, Fusi S, Helfer P, Kay L, Ketz N, Kira Z, Kolouri S, Krichmar JL, Kriegman S, Levin M, Madireddy S, Manicka S, Marjaninejad A, McNaughton B, Miikkulainen R, Navratilova Z, Pandit T, Parker A, Pilly PK, Risi S, Sejnowski TJ, Soltoggio A, Soures N, Tolias AS, Urbina-Melendez D, Valero-Cuevas FJ, van de Ven GM, Vogelstein JT, Wang F, Weiss R, Yanguas-Gil A, Zou XY, Siegelmann H (2022) Biological underpinnings for lifelong learning machines. Nature Machine Intelligence 4:196–210.
Lemieux M, Chen J-Y, Lonjers P, Bazhenov M, Timofeev I (2014) The impact of cortical deafferentation on the neocortical slow oscillation. The Journal of Neuroscience 34:5689—5703.
Lewis PA, Durrant SJ (2011) Overlapping memory replay during sleep builds cognitive schemata. Trends Cogn Sci 15:343–351.
Lewis PA, Knoblich G, Poe G (2018) How Memory Replay in Sleep Boosts Creative Problem-Solving. Trends Cogn Sci 22:491–503.
McClelland JL, McNaughton BL, O'Reilly RC (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 102:419–457.
Mccloskey M, Cohen NJ (1989) CATASTROPHIC INTERFERENCE IN CONNECTIONIST NETWORKS: THE SEQUENTIAL LEARNING PROBLEM. The Psychology of Learning and Motivation 24:109–165.
Mednick SC, Cai DJ, Shuman T, Anagnostaras S, Wixted JT (2011) An opportunistic theory of cellular and systems consolidation. Trends in neurosciences 34:504–514.
Oudiette D, Antony JW, Creery JD, Paller KA (2013) The role of memory reactivation during wakefulness and sleep in determining which memories endure. The Journal of neuroscience : the official journal of the Society for Neuroscience 33:6672–6678.
Paller KA, Voss JL (2004) Memory reactivation and consolidation during sleep. Learn Mem 11:664–670.
Plihal W, Born J (1999) Effects of early and late nocturnal sleep on priming and spatial memory. Psychophysiology 36:571–582.
Rasch B, Born J (2013) About sleep's role in memory. Physiological reviews 93:681–766.
Sanda P, Malerba P, Jiang X, Krishnan GP, Gonzalez-Martinez J, Halgren E, Bazhenov M (2021) Bidirectional Interaction of Hippocampal Ripples and Cortical Slow Waves Leads to Coordinated Spiking Activity During NREM Sleep. Cereb Cortex 31:324–340.
Shinoe T, Matsui M, Taketo MM, Manabe T (2005) Modulation of synaptic plasticity by physiological activation of M1 muscarinic acetylcholine receptors in the mouse hippocampus. The Journal of neuroscience : the official journal of the Society for Neuroscience 25:11194–11200.
Steriade M, McCormick DA, Sejnowski TJ (1993a) Thalamocortical oscillations in the sleeping and aroused brain. Science 262:679–685.

Steriade M, Nuñez A, Amzica F (1993b) A novel slow (<1 Hz) oscillation of neocortical neurons *in vivo* : depolarizing and hyperpolarizing components. The Journal of neuroscience : the official journal of the Society for Neuroscience 13:3252–3265.

Stickgold R (2013) Parsing the role of sleep in memory processing. Curr Opin Neurobiol 23:847–853.

Stickgold R, Walker MP (2007) Sleep-dependent memory consolidation and reconsolidation. Sleep medicine 8:331–343.

Sugisaki E, Fukushima Y, Fujii S, Yamazaki Y, Aihara T (2016) The effect of coactivation of muscarinic and nicotinic acetylcholine receptors on LTD in the hippocampal CA1 network. Brain Res 1649:44–52.

Tadros T, Bazhenov M (2022) Role of Sleep in Formation of Relational Associative Memory. The Journal of neuroscience : the official journal of the Society for Neuroscience.

Tadros T, Krishnan GP, Ramyaa R, Bazhenov M (2020a) Biologically inspired sleep algorithm for increased generalization and adversarial robustness in deep neural networks. . In: International Conference on Learning Representations, p Paper1347.

Tadros T, Krishnan G, Ramyaa R, Bazhenov M (2020b) Biologically inspired sleep algorithm for reducing catastrophic forgetting in neural networks. . In: AAAI Conference on Artificial Intelligence, pp 13933–13934.

Tadros T, Krishnan GP, Ramyaa R, Bazhenov M (2022) Sleep-like unsupervised replay reduces catastrophic forgetting in artificial neural networks. Nat Commun 13:7742.

Timofeev I, Grenier F, Bazhenov M, Sejnowski TJ, Steriade M (2000) Origin of slow cortical oscillations in deafferented cortical slabs. Cer Cortex 10:1185–1199.

Treves A, Rolls ET (1994) Computational analysis of the role of the hippocampus in memory. Hippocampus 4:374–391.

Vanini G, Lydic R, Baghdoyan HA (2012) GABA-to-ACh ratio in basal forebrain and cerebral cortex varies significantly during sleep. Sleep 35:1325–1334.

Walker MP, Stickgold R (2004) Sleep-dependent learning and memory consolidation. Neuron 44:121–133.

Wei Y, Krishnan G, Bazhenov M (2016a) Synaptic Mechanisms of Memory Consolidation during Sleep Slow Oscillations. Journal of Neuroscience 36:4231–4247.

Wei Y, Krishnan GP, Bazhenov M (2016b) Synaptic Mechanisms of Memory Consolidation during Sleep Slow Oscillations. The Journal of neuroscience : the official journal of the Society for Neuroscience 36:4231–4247.

Wei Y, Krishnan GP, Komarov M, Bazhenov M (2018) Differential roles of sleep spindles and sleep slow oscillations in memory consolidation. PLoS computational biology 14:e1006322.

Wei Y, Krishnan GP, Marshall L, Martinetz T, Bazhenov M (2020) Stimulation Augments Spike Sequence Replay and Memory Consolidation during Slow-Wave Sleep. The Journal of neuroscience : the official journal of the Society for Neuroscience 40:811–824.

Wilson MA, McNaughton BL (1994) Reactivation of hippocampal ensemble memories during sleep. Science 265:676–679.

FIGURES

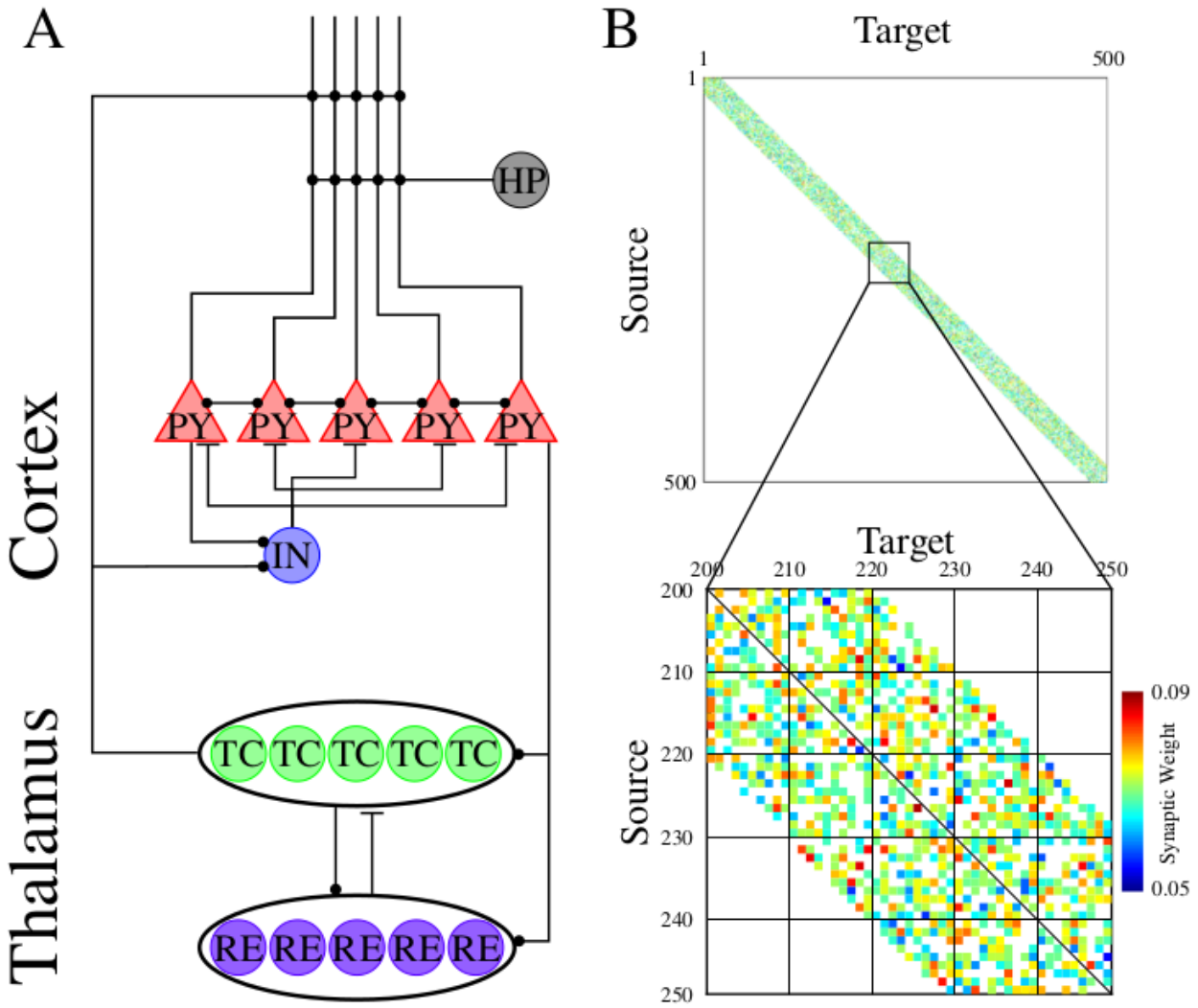


Fig 1. Network Architecture. (**A**) Basic network architecture (PY: excitatory pyramidal neurons; IN: inhibitory interneurons; TC: excitatory thalamocortical neurons; RE: inhibitory thalamic reticular neurons). Excitatory synapses are represented by lines terminating in a dot, while inhibitory synapses are represented by lines terminating in bars. Arrows indicate the direction of the connection. (**B**) Top panel shows the initial weighted synaptic matrix for the PYs. The color in this plot represents the strength of the AMPA connections between PY neurons, with white indicating the lack of synaptic connection. Bottom panel shows a zoom-in of the top panel for the subregion where training occurs.

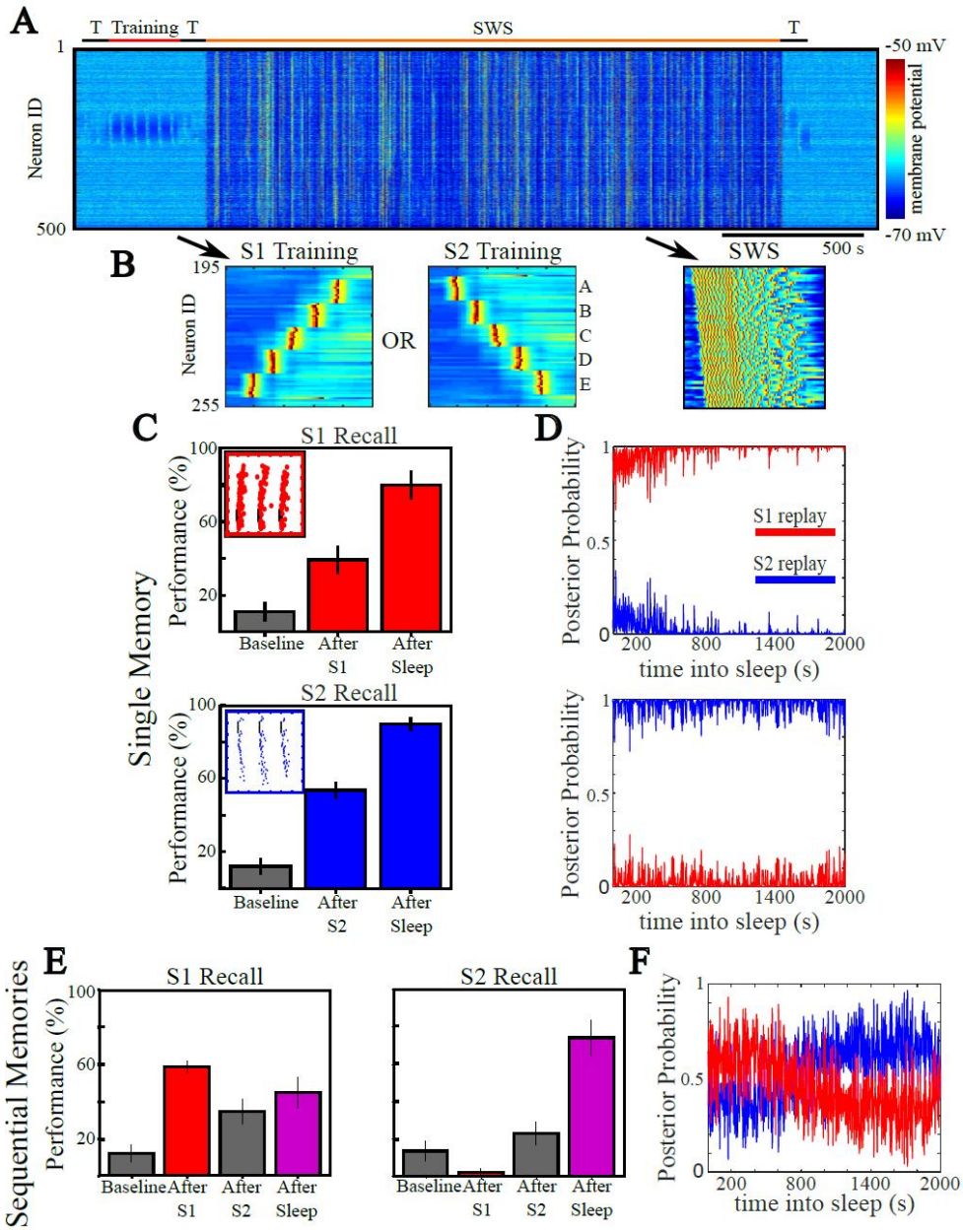


Fig 2. Sleep rescues interference induced by sequential training. (A) Network activity during an example simulation depicting all 500 PYs on the y-axis and the membrane potential (color scale) of each over time (x-axis). The network undergoes testing periods (T; black) at baseline, after Training (red), and after SWS (orange). **(B)** Examples of network activity during one bout of S1 (left) or S2 (middle) training, and during a single up state of SWS (right). All three subpanels depict the same network region. **(C)** Recall performance for S1 (top) or S2 (bottom) during all testing periods. Insets include spike rasters showing examples of pattern completion (red/blue dots) following cued (black dots) recall. Bars are colored according to which memories the network can pattern complete at that point in the simulation: none (gray), red (S1), blue (S2), or purple (S1&S2). **(D)** Average probability of replaying S1 (red) or S2 (blue) during an up state at a given point in sleep; top panel (S1 trained), bottom panel (S2 trained). Based on an SVM classifier trained on held out trials of sims where either S1 or S2 were trained before sleep. **(E)**

Same as (C) but for a simulation where S2 was trained sequentially after S1 and prior to sleep; S1 recall (left), S2 recall (right). **(F)** Same as (D) but for the sequential training simulation.

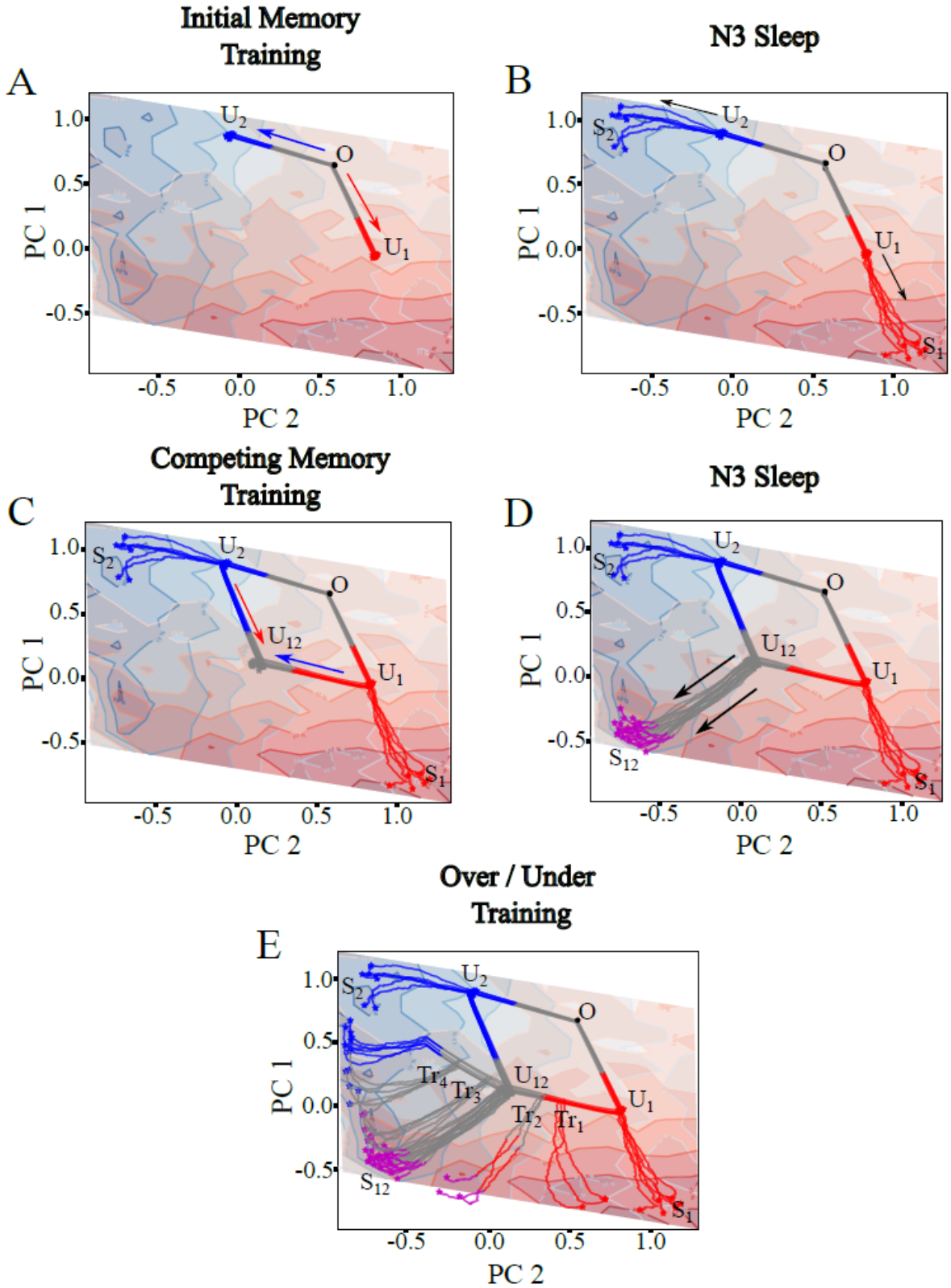


Fig 3. Synaptic performance landscape reveals multi-stability and fine-tuning. All panels show the evolution of the network through dimensionality reduced synaptic weight space. Contour lines and coloring correspond to single memory recall performance for S1 (red) and S2 (blue). Trajectories are colored according to which memories the network can pattern complete at that point in the simulation: none (gray), red (S1), blue (S2), or purple (S1&S2). **(A)** Evolution during S1 (red arrow) or S2 (blue arrow) training onto each memory manifold (red/blue trajectories). **(B)** SWS moves each network further along its current trajectory to putative single memory attractor states (S_1 and S_2). **(C)** Sequentially training the competing memory moves the network to a gray central region where neither memory can be recalled (U_{12}). **(D)** SWS moves the networks towards the putative joint memory attractor (purple trajectories; S_{12}) where both can be recalled. **(E)** Examples of under (Tr_1, Tr_2) and over (Tr_3, Tr_4) training S2 following S1 training reveal the necessity of fine-tuning the training durations for sleep to evolve the network to the joint memory manifold.

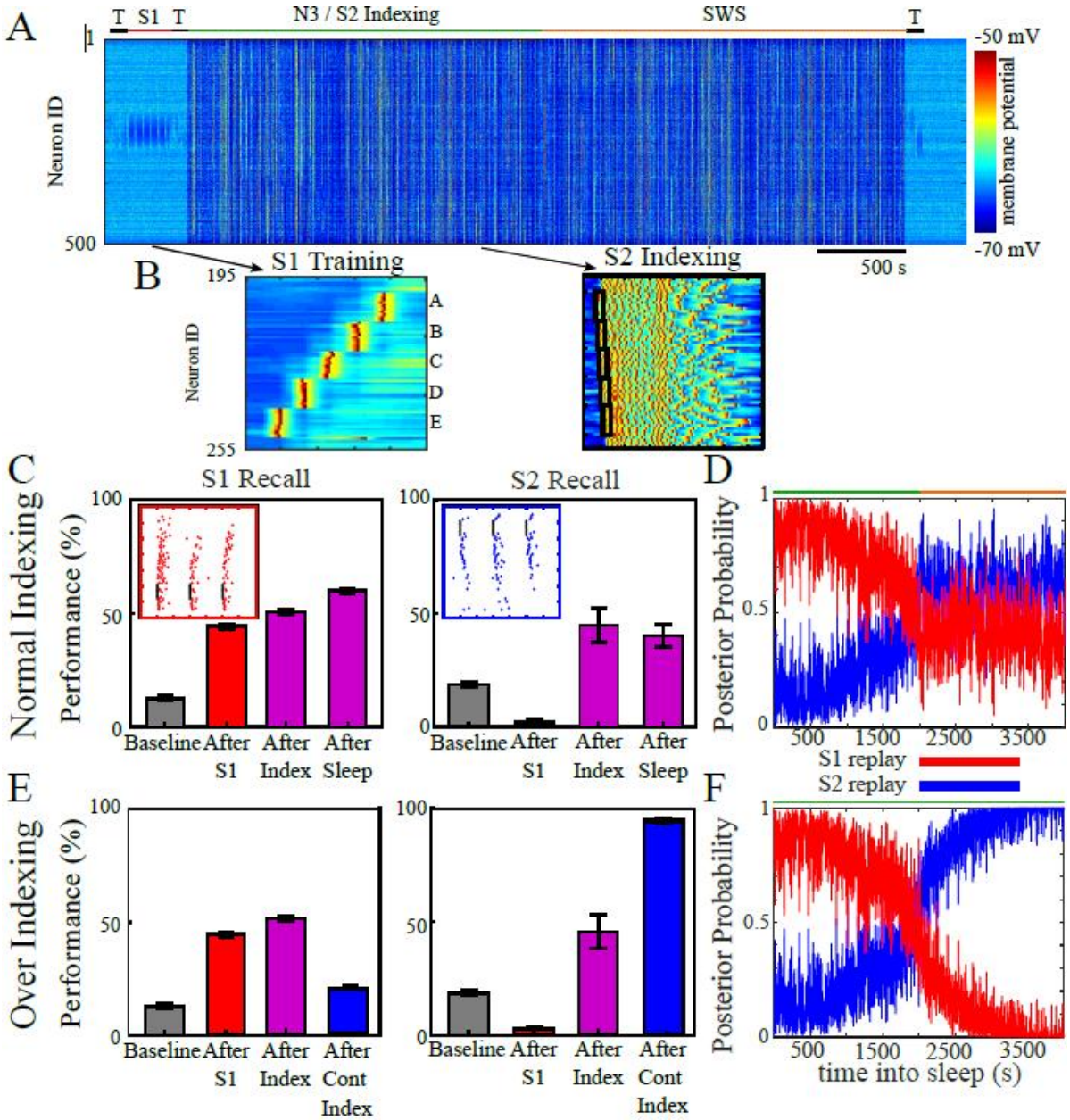


Fig 4. Hippocampal indexing during sleep induces consolidation without interference. (A) Network activity during the simulation in which the network undergoes periods of testing (T; black), training (S1; red), SWS with S2 Indexing (green), and SWS (orange). **(B)** Examples of network activity during one bout of S1 training (left), and during a single up state of SWS (right) where indexing is simulated at the beginning of the up state (black boxes). Both subpanels depict the same network region. **(C)** S1 (left) and S2 (right) recall performance show that the network consolidates S2 without interference to S1. **(D)** Average replay probability shows S2 replay probability slowly increases during the course of indexing (green) until it is roughly even with S1 replay probability by the time sleep without indexing (orange) begins. **(E)** Same as (C) but for simulations in which indexing continued for the entire duration of sleep. Only S2 could be recalled at the end of the simulation. **(F)** Same as (D) but for simulations in which indexing

continued for the entire duration of sleep. S2 replay probability was shown to increase to unity with continual indexing.

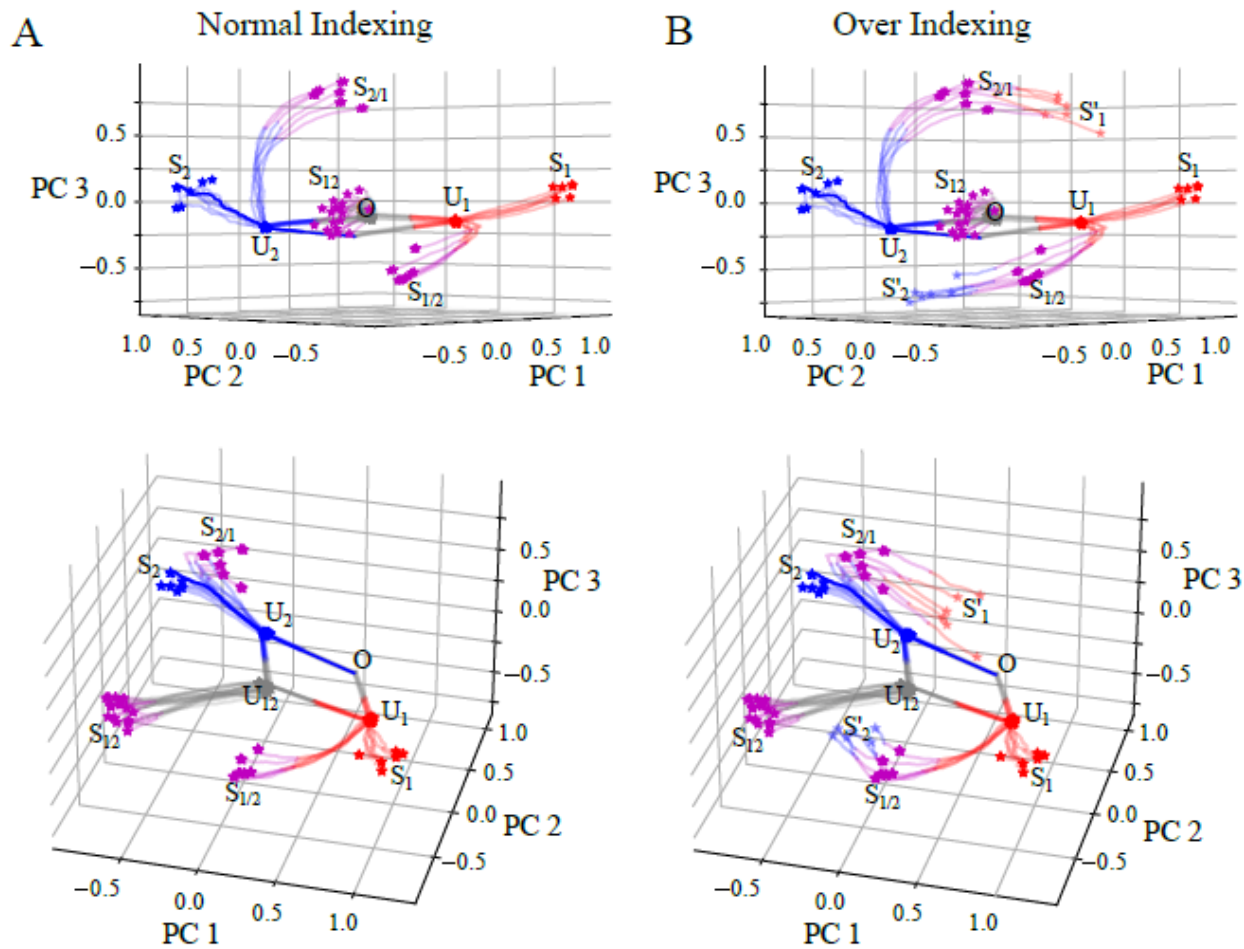


Fig 5. Indexing evolves the network along the current memory manifold toward the intersection. (A) Evolution of the networks during normal indexing (top) shows the network moves along its current memory manifold (i.e. S1 - red; S2 - blue), primarily in the PC3 dimension, until it reaches an intersection of the manifolds (purple) that is stable with subsequent sleep. **(B)** Evolution during over indexing (top) shows that if indexing is not halted, it pushes the network out of the intersection (purple) and back onto the single memory manifold which corresponds to the index. The bottom panels show the same as the top, but with the plot rotated 45 degrees about the PC3 axis.

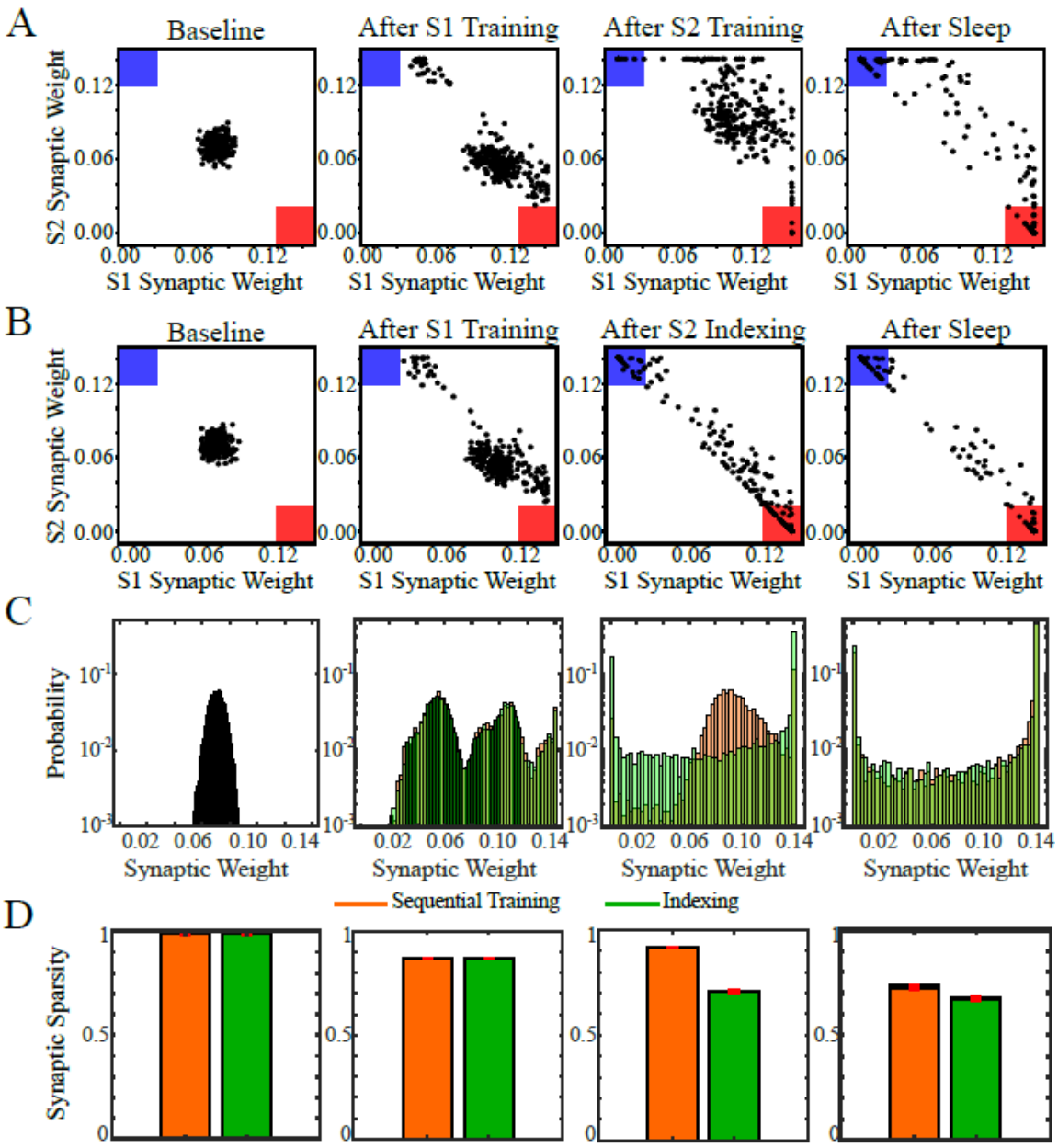


Fig 6. Indexing leads to sparser memory representations than sequential training and sleep. **(A-B)** Each panel shows scatter plots of synaptic weights between all PYs in the trained region that have bi-directional synapses between them. The weight in the S1 direction is on the x-axis, while that in the S2 direction is on the y-axis. Colored corners indicate regions where the synaptic pair has been strongly biased towards S1 (red) or S2 (blue) at the expense of the opposing memory. (A) Snapshots during a simulation with sequential training and sleep. (B) Snapshots during a simulation with normal indexing during sleep. **(C)** Distributions of synaptic weights at each time point from (A-B) for sequential training (orange) and normal indexing (green) with the y-axis on a log-scale. **(D)** Average synaptic density of the weight matrix across simulations with sequential training (orange) and normal indexing (green). Error bars show standard deviation.

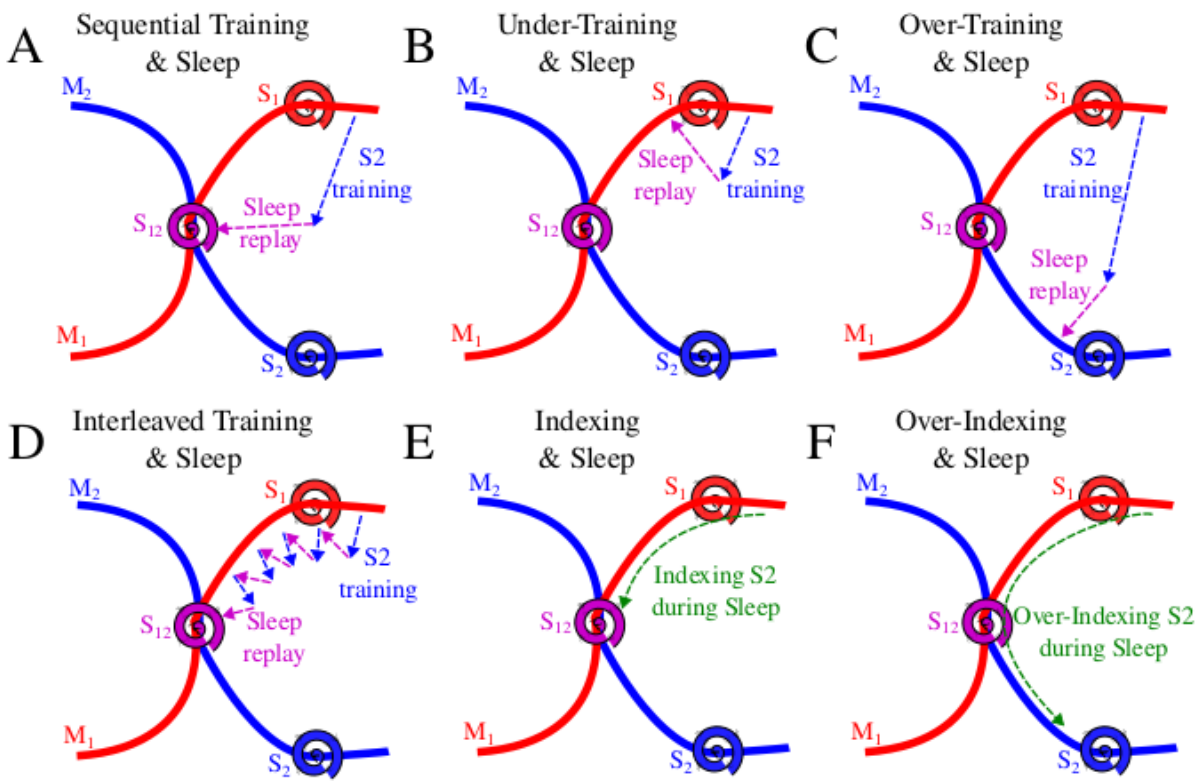
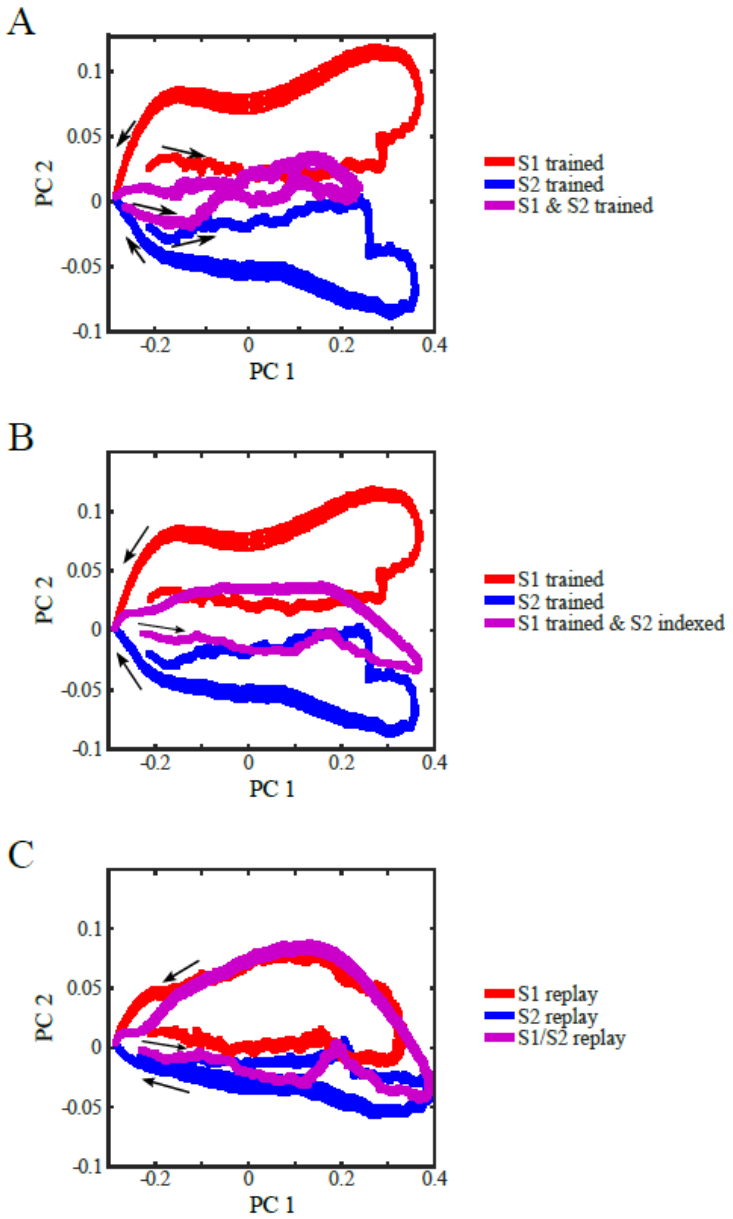
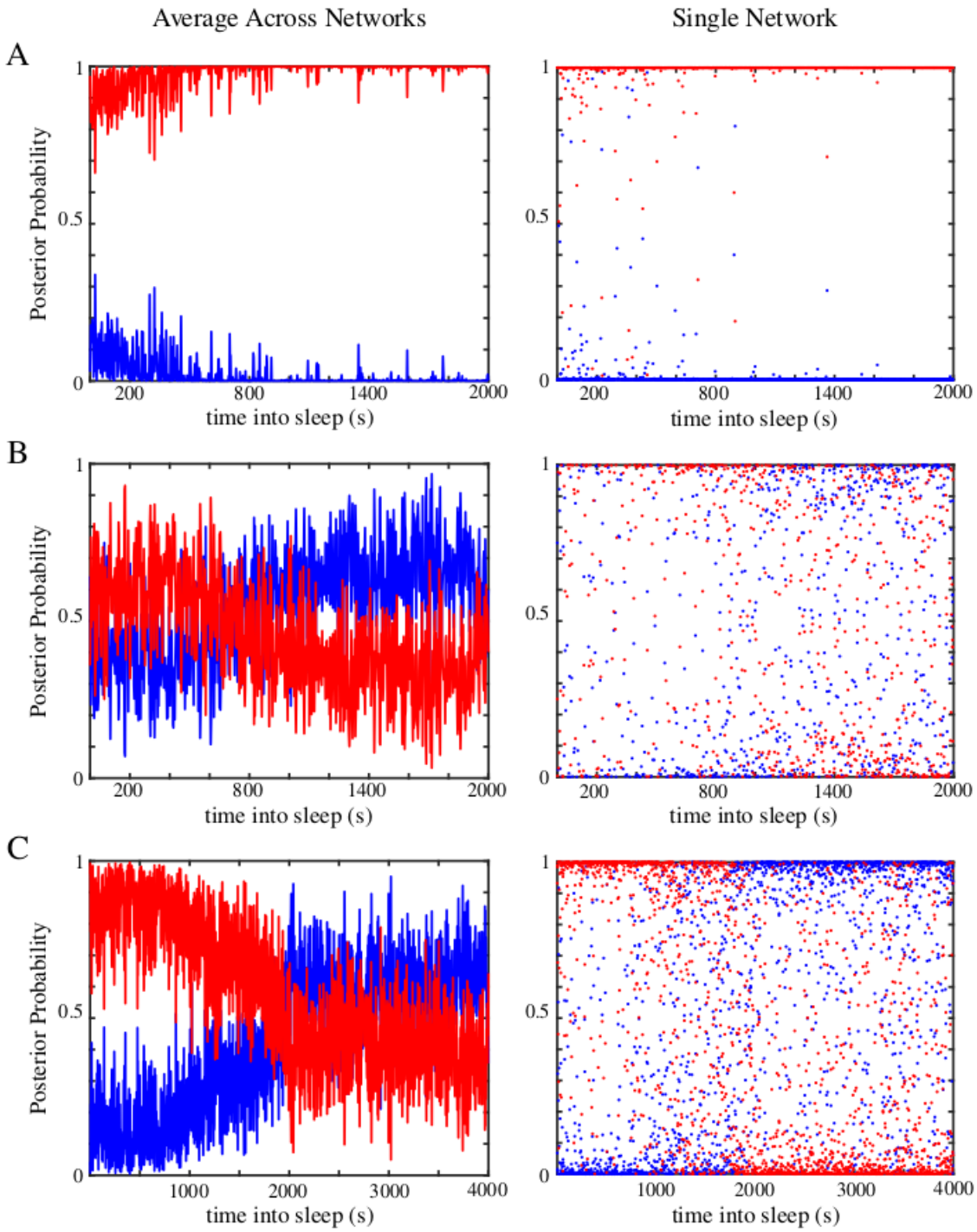


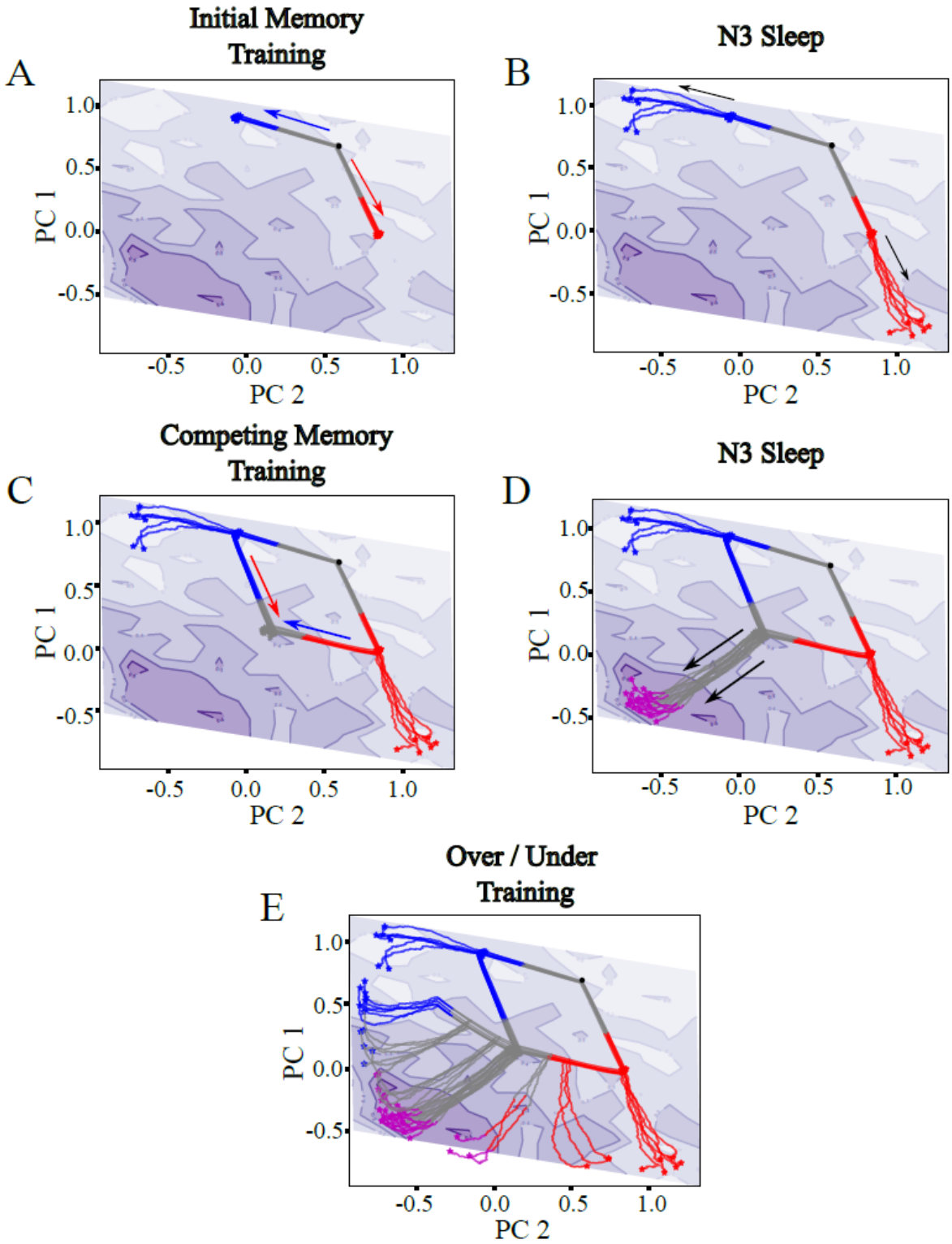
Fig 7. Systems consolidation prevents interference through interleaved replay within up states. Cartoon schematics illustrating the network dynamics across memory manifolds for **(A)** Sequential Training, **(B)** Under-Training, **(C)** Over-Training, **(D)** Interleaved Training, **(E)** Indexing, and **(F)** Over-Indexing. Spirals represent regions on manifolds which appear to be stable and attracting under sleep dynamics.



Supp Figure 1. The firing rate trajectories of each memory's replays during SWS are linearly separable in PC space. (A) Average firing rate trajectories in PC space during up states of a network trained on S1 (red), S2 (blue), or S1&S2 (purple) before sleep. **(B)** Average firing rate trajectories in PC space during up states after single memory training (S1 - red; S2 - blue) and during normal indexing (purple). **(C)** Average trajectories during normal indexing grouped according to the replay probability of each up state (>95% S1 = red; >95% S2 = blue; otherwise = purple).

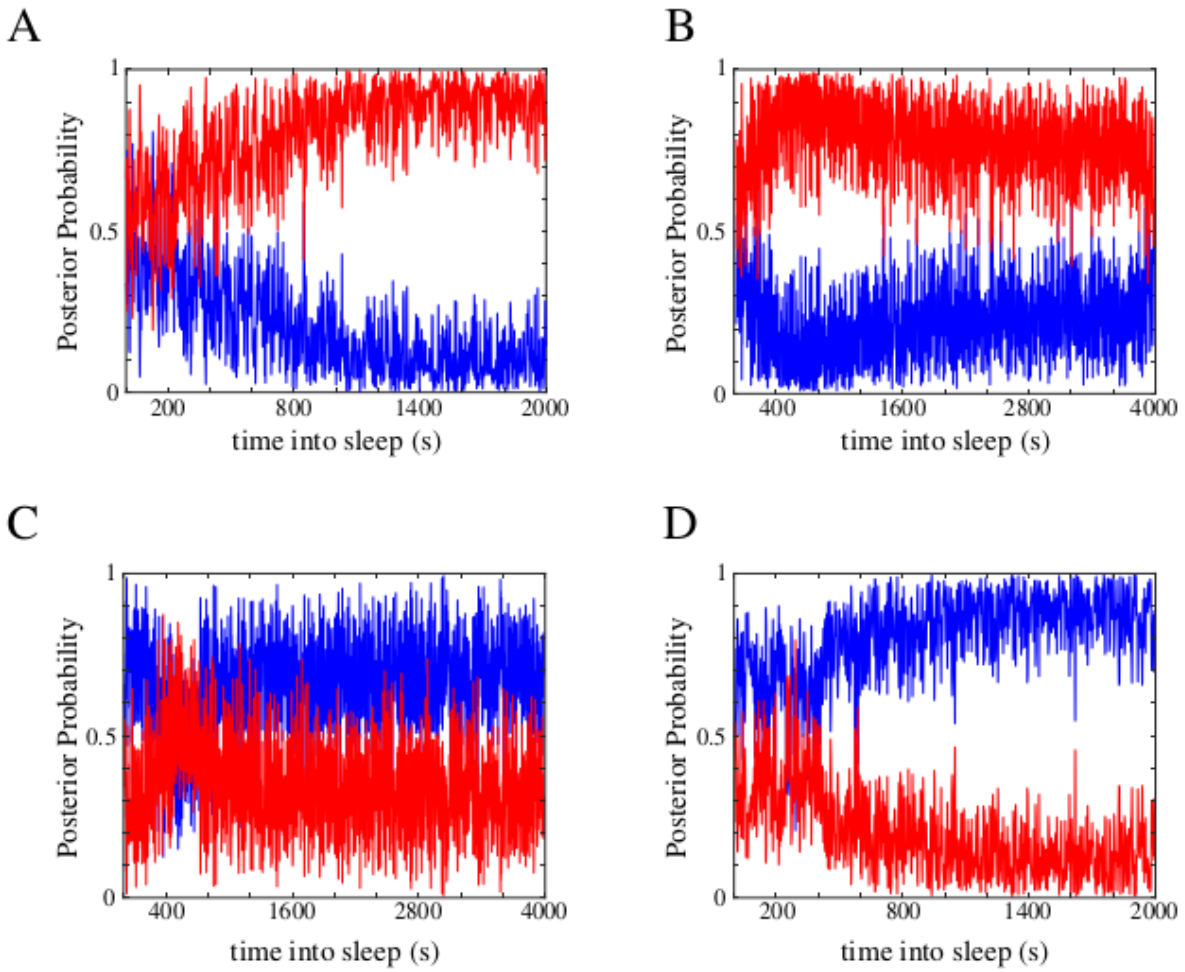


Supp Figure 2. Classification of memory replay on individual UP states is robust in a single network. Averaged (left) and single network (right) replay probabilities for **(A)** S1 training only; **(B)** sequential training; and **(C)** normal indexing. The single network plots highlight that the majority of individual Up-states are robustly classified as either S1 or S2 across all conditions.



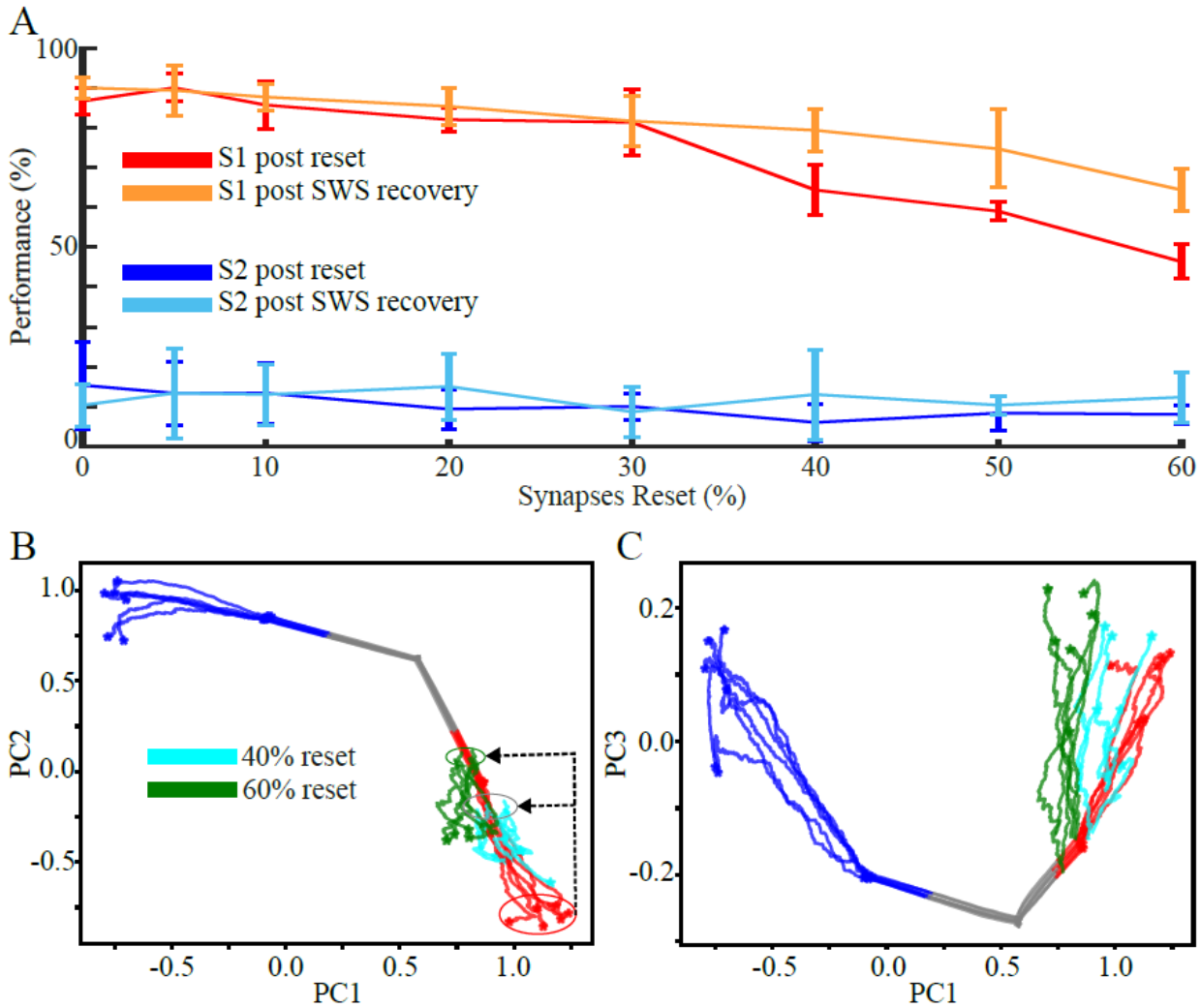
Supp Figure 3. Synaptic performance landscape using a joint-memory recall metric. All panels show the evolution of the network through dimensionality reduced synaptic weight space as in Fig 3, except now the background is colored according to a single contour plot which considers performance on both memories. Joint memory performance is the average of the

performance on each memory with the absolute difference between the two single memory performance scores subtracted as a penalty for favoring one memory over the other. Trajectories are colored according to which memories the network can pattern complete at that point in the simulation: none (gray), red (S1), blue (S2), or purple (S1&S2). **(A)** Evolution during S1 (red arrow) or S2 (blue arrow) training onto each memory manifold (red/blue trajectories). **(B)** SWS moves each network further along its current trajectory to putative single memory attractor states (S_1 and S_2). **(C)** Sequentially training the competing memory moves the network to a gray central region where neither memory can be recalled (U_{12}). **(D)** SWS moves the networks towards the putative joint memory attractor (purple trajectories; S_{12}) where both can be recalled. **(E)** Examples of under (Tr_1, Tr_2) and over (Tr_3, Tr_4) training S2 following S1 training reveal the necessity of fine-tuning the training durations for sleep to evolve the network to the joint memory manifold.

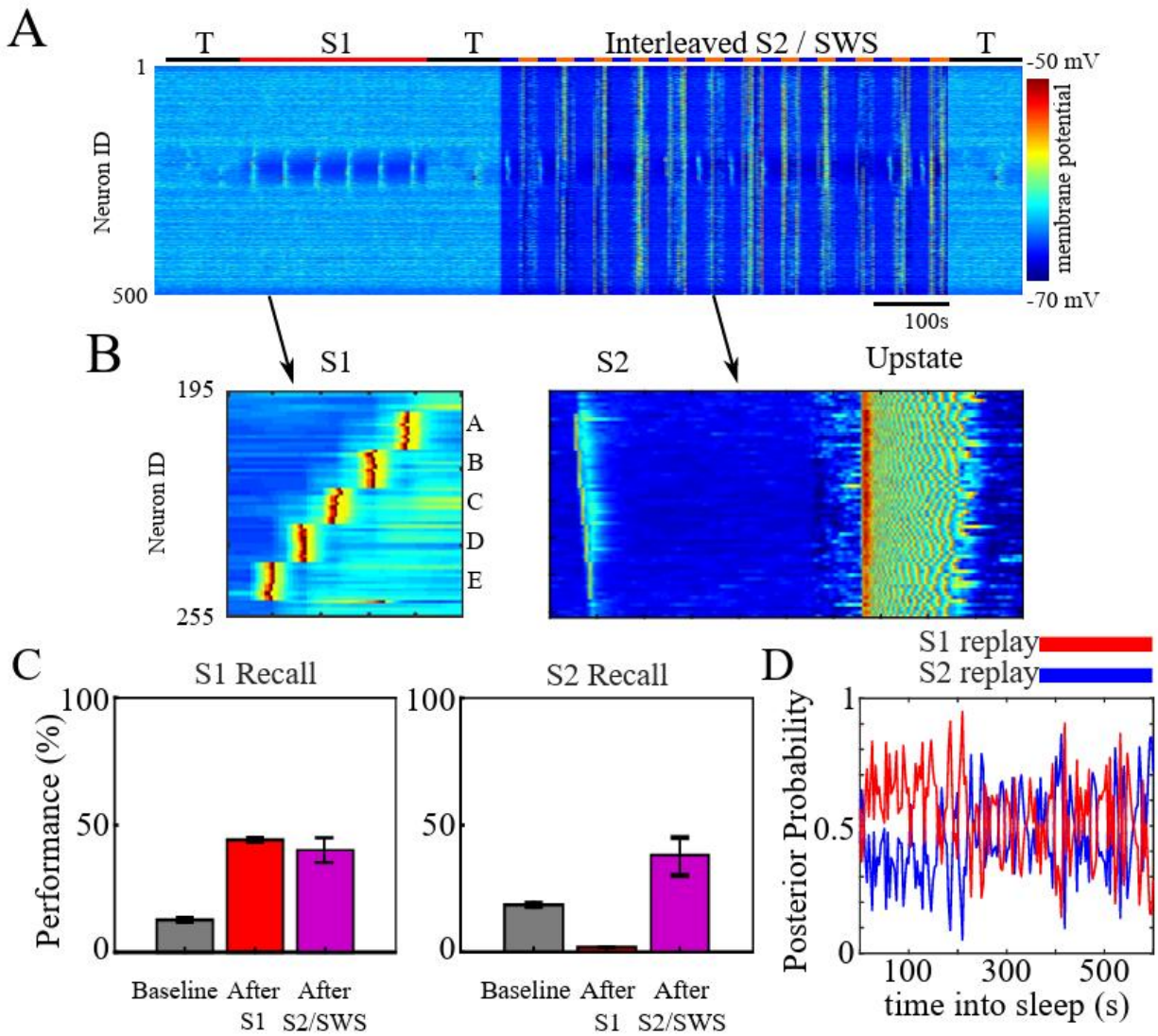


Supp Figure 4. Replay probabilities during sleep following under/over-training. (A) Following significant undertraining of S2 (Figure 3E; Tr_1) the network begins sleep in a mixed replay state and slowly converges to an S1 dominated replay state. **(B)** Following moderate undertraining of S2 (Figure 3E; Tr_2) the network diverges from a mixed replay state towards an

S1 dominated replay state before slowly relaxing towards a more mixed state. **(C)** Following moderate overtraining of S2 (Figure 3E; Tr_3) the network's replay state briefly oscillates about and then remains in a mixed replay state. **(D)** Following significant overtraining of S2 (Figure 3E; Tr_4) the network slowly converges towards an S2 dominated replay state from its initial mixed replay state.



Supp Figure 5. Network shows graceful degradation and recovery with SWS when weights are perturbed from putative single memory attractor states. (A) Line plot showing the recall performance of S1 and S2 after a random percentage of synapses reset to their initial value (S1, red; S2, blue) and after a SWS recovery period (S1, orange; S2; light blue). **(B)** Evolution of network through weight space after 40% (cyan) or 60% (green) of the synapses are reset. **(C)** Same as (C) but showing PCs 1 and 3 instead of 1 and 2.



Supp Figure 6. Interleaving S2 training with SWS following S1 training allows for recall of both old and new memories. (A) Network activity during the simulation in which the network undergoes periods of testing (T; black), training (S1; red), and 12 cycles of interleaved S2 training (blue) with SWS (orange), with each phase of the cycle lasting 25 s. **(B)** Examples of network activity during one bout of S1 training (left), and during a training to sleep transition with the last bout of an S2 training period and a single up state at the start of a SWS period (right). Both subpanels depict the same network region. **(C)** S1 (left) and S2 (right) recall performances show that the network consolidates S2 without interference to S1. **(D)** Average replay probability shows S2 (blue) replay probability begins below that of S1 (red) until roughly 250 s into interleaved S2/SWS.