UNIVERSITY OF CALIFORNIA SAN DIEGO

Sleep Rhythms and Consolidation Strategies: Advances in Modeling Life-Long Learning

A Dissertation submitted in partial satisfaction of the requirements
for the degree Doctor of Philosophy

in

Neurosciences with a Specialization in Computational Neurosciences

by

Ryan Golden

Committee in charge:

Professor Maxim Bazhenov, Chair
Professor Andrea Chiba
Professor Timothy Gentner
Professor Eric Halgren
Professor Terrence Sejnowski

2023

The Dissertation of Ryan Golden is approved, and it is acceptable in quality and form for publication on microfilm and electronically.

University of California San Diego

2023

DEDICATION

To all of those who cultivated my curiosity, all those who tolerated my tangents, and all those who convinced me of the beauty of chaos. But most of all to Jillian, who convinced me to wear a helmet.

TABLE OF CONTENTS

DISSERTATION APPROVAL PAGE iii

DEDICATION iv

TABLE OF CONTENTS v

LIST OF FIGURES vi

LIST OF TABLES viii

ACKNOWLEDGEMENTS ix

VITA xi

ABSTRACT OF THE DISSERTATION xii

Chapter 1 Sleep prevents catastrophic forgetting in spiking neural networks by forming a joint synaptic weight representation 1

Chapter 2 Hippocampal indexing alters the stability landscape of synaptic weight space allowing life-long learning 39

Chapter 3 Multielectrode cortical stimulation induces traveling waves 84

LIST OF FIGURES

Figure 1.1: Network architecture and foraging task structure ……4
Figure 1.2: Receptive fields of output and hidden layer neurons determine the agent behavior 5
Figure 1.3: Sleep prevents catastrophic forgetting during new task training ……8
Figure 1.4: Interleaved periods of new task training with sleep allows integrating synaptic information relevant to new task while preserving old task ……...10
Figure 1.5: Receptive fields following interleaved sleep and task 1 training reveal how the network can multiplex the complementary tasks ……12
Figure 1.6: Periods of sleep allow learning task 1 without interference with old task 2 through renormalization of task-relevant synapses ……...14
Figure 1.7: Periods of sleep push the network towards the intersection of task 1 and task 2 synaptic weight manifolds ……..……..16
Figure 1.8: (S1 Fig) Spike rasters showing network activity across various training regimes …….....…..……...32
Figure 1.9: (S2 Fig) Model displays graceful degradation in performance as a result of hidden layer dropout ……...…....……...33
Figure 1.10: (S3 Fig) Particle responsiveness metric (PRM) shows correspondence between type of training and particles preferred by the network ...……...……34
Figure 1.11: (S4 Fig) Effect of sleep to protect old memory does not depend on specific properties of noise applied during sleep phase ……...……35
Figure 1.12: (S5 Fig) Interleaving old and new task training allows integrating synaptic information relevant to new task while preserving old task information ……36
Figure 1.13: (S6 Fig) Freezing a fraction of task specific strong synapses preserves differing degrees of performance in a sequential learning paradigm ……38
Figure 2.1: Network architecture ……72
Figure 2.2: Sleep rescues interference induced by sequential training ……73
Figure 2.3: Synaptic performance landscape reveals multi-stability and fine-tuning ……..74
Figure 2.4: Hippocampal indexing during sleep induces consolidation without interference ……75
Figure 2.5: Indexing causes interleaved memory reactivation within individual up states ……76
Figure 2.6: Indexing evolves the network along the current memory manifold toward the intersection ……..77
Figure 2.7: Indexing leads to sparser memory representations than sequential training and sleep ……..78
Figure 2.8: Systems consolidation prevents interference through interleaved replay within up states ……...79
Figure 2.9: (S1) Classification of memory replay on individual up states is robust in a single network ……..80
Figure 2.10: (S2) Replay probabilities during sleep following under/over-training ……81
Figure 2.11: (S3) Interleaving S2 training with N3 sleep following S1 training allows for recall of both old and new memories ……82
Figure 3.1: Electric potential and activating function in the plane Y = 0 ……86
Figure 3.2: Representative reconstructions and averaged axonal density for

neuronal cell types modeled ……..88
Figure 3.3: Probability of spiking as a function of horizontal distance from the center of the electrode array for each cell type and cortical layer ……89
Figure 3.4: Microcircuit diagram of a single cortical column in modeled network ……90
Figure 3.5: Directed propagation of pyramidal activity in raster plot of microcircuit simulation trial ……..91
Figure 3.6: Voltage traces for layer II/III cells and pyramidal input conductances during simulation show how inhibition causes directionality of traveling wave ……92
Figure 3.7: Cell type-specific silencing indicates distinct roles of inhibitory interneurons in temporal and spatial dynamics of the traveling wave ……..93
Figure 3.8: Summary of interaction resulting in unidirectional propagation ……...94

LIST OF TABLES

Table 3.1: Summary of datasets with reconstructed cells ……85
Table 3.2: Structure of the network ……87
Table 3.3: Connectivity within the network ……87

ACKNOWLEDGEMENTS

I would like to acknowledge and express my gratitude to Professor Maxim Bazhenov for his continual support throughout my studies. He provided me with both the necessary room to refine my interests and mature as a scientist, as well as the soft landings I needed when I would begin becoming untethered.

I'd like to thank my committee for their valuable feedback as I struggled to define the scope of my thesis, as well as all of their advice in navigating academia and finding my niche. I would also like to thank Erin Gilbert for her infinite patience as my friends and I fumbled our ways through each and every administrative and organizational hurdle of graduate school these past eight years.

I'd also like to thank all members, past and present, of the Bazhenov Lab, as well as all of the friends I have made in the Neurosciences Graduate Program. You filled my time here with memories of collaboration, friendship – and the occasional bout of gleeful abandon.

I'd also like to thank my parents for all that they have contributed that has allowed my life to travel the courses it has, and my brother for instilling his passion for science in me. Lastly, I'd like to thank my partner, Jillian. Your enduring support has helped me to achieve my dearest dreams, and your love has taught me that it's alright if I don't.

Chapter 1, in full, is a reprint of the material as it appears in PLOS Computational Biology 18(11): e1010628, under the title "Sleep prevents catastrophic forgetting in spiking neural networks by forming a joint synaptic weight representation", Golden, Ryan; Delanois, J. Erik; Sanda, Pavel; Bazhenov, Maxim, PLOS, 2022. The dissertation author was the co-primary investigator and author of this paper, along with J. Erik Delanois.

Chapter 2, in full, is currently being prepared for submission for publication of the material. Gonzalez, Oscar C.; Golden, Ryan; Delanois, J. Erik; McNaughton, Bruce L.; Bazhenov, Maxim. The dissertation author was the co-primary investigator, along with Oscar C. Gonzalez, and the primary author of this paper.

Chapter 3, in full, is a reprint of the material as it appears in The Journal of Neuroscience, under the title “Multielectrode cortical stimulation selectively induces unidirectional wave propagation of excitatory neuronal activity in biophysical neural model”, April 2023; 43(14):2482-2496. Halgren, Alma S.; Siegel, Zarek; Golden, Ryan; Bazhenov, Maxim. The dissertation author was one of the primary investigators and authors of this paper, but also serving in a supervisory role as a mentor for the primary investigator, Alma S. Halgren.

VITA

2016 Bachelor of Science in Mathematics, Saint Ambrose University

2016 Bachelor of Science in Biology, Saint Ambrose University

2023 Doctor of Philosophy in Neurosciences with a Specialization in Computational Neurosciences, University of California San Diego

PUBLICATIONS

Golden R, Bazhenov M. "Thalamocortical Oscillations - Mechanisms and Impact" in S.M. Sherman and W.M. Usrey (ed.), The Cerebral Cortex and Thalamus, Oxford: Oxford University Press. 2023. pgs: 651-663.

Halgren A, Siegel Z, **Golden R**, Bazhenov M. (2023) "Multielectrode cortical stimulation selectively induces unidirectional wave propagation of excitatory neuronal activity in biophysical neural network". J. Neurosci.

Golden R*, Delanois JE*, Bazhenov M. (2022) "Sleep prevents catastrophic forgetting in spiking neural networks by forming joint synaptic weight representations". PLoS Comput. Biol.

Komarov M, Malerba P, **Golden R**, Nunez P, Halgren E, & Bazhenov M (2019). "Selective recruitment of cortical neurons by electrical stimulation". PloS Comput. Biol.

Golden R & Cho I (2017). "Matricial representations of certain finitely presented groups generated by order-2 generators and their applications". Am J Undergrad Res.

*indicates co-first author

FIELD OF STUDY

Major Field: Neurosciences with a Specialization in Computational Neuroscience
Professor Maxim Bazhenov

ABSTRACT OF THE DISSERTATION

Sleep Rhythms and Consolidation Strategies:
Advances in Modeling Life-Long Learning

by

Ryan Golden

Doctor of Philosophy in Neurosciences with a Specialization in Computational Neurosciences

University of California San Diego, 2023

Professor Maxim Bazhenov, Chair

This dissertation was an investigation into the computational roles of sleep rhythms in the consolidation of memory, and how these roles may be leveraged to the benefit of machine learning and medicine. In Chapter 1 we used an artificial spiking neural network model to validate that a consolidation strategy thought to be taken by the procedural memory system – incrementally learning a new skill by interleaving bouts of training with periods of sleep – can prevent catastrophic forgetting when faced with learning a novel task. In particular, we demonstrated that memory replay during sleep acted to keep the network's synaptic weight state near to previous memory manifolds as it learns the new task. In Chapter

2, we utilized a biophysical thalamocortical network model to further study this procedural memory consolidation strategy, as well as a declarative memory consolidation strategy – incrementally transferring a new memory to the cortex by hippocampal indexing during sleep. While both strategies were able to prevent catastrophic forgetting, we found that the procedural memory strategy suffers from fine-tuning and works best when training bouts are short and protracted in time. The declarative memory strategy does not suffer from this same fine-tuning problem, suggesting it may be engaged when training bouts are chunked rather than distributed in time. Moreover, our model suggests that the declarative memory consolidation strategy may simply be a compressed version of the procedural memory strategy, with the hippocampus generating simulated training samples to be indexed to the cortex during sleep. We anticipate that such a strategy will be useful in mitigating catastrophic forgetting in machine learning, as others in our lab have shown the procedural memory consolidation strategy to be. Finally, in Chapter 3, we made use of a two-phase biophysical-anatomical and dynamic-neuronal network in order to model the effects of electrical stimulation of the cortical surface and studies the circuit mechanisms behind how this could be used to induce directed traveling waves. We found that cortical surface stimulation differentially recruits distinct subtypes of inhibitory interneurons, which shape the oscillatory frequency and direction of the wave. In the future, we hope to develop this work further to model the induction of sleep rhythms with this network, and how this may be used to aid clinical treatment of memory and sleep disorders.

Chapter 1 Sleep prevents catastrophic forgetting in spiking neural networks by forming a joint synaptic weight representation

PLOS COMPUTATIONAL BIOLOGY

RESEARCH ARTICLE

Sleep prevents catastrophic forgetting in spiking neural networks by forming a joint synaptic weight representation

Ryan Golden[1,2☯], **Jean Erik Delanois**[2,3☯], **Pavel Sanda**[4], **Maxim Bazhenov**[1,2*]

1 Neurosciences Graduate Program, University of California, San Diego, La Jolla, California, United States of America, **2** Department of Medicine, University of California, San Diego, La Jolla, California, United States of America, **3** Department of Computer Science and Engineering, University of California, San Diego, La Jolla, California, United States of America, **4** Institute of Computer Science of the Czech Academy of Sciences, Prague, Czech Republic

☯ These authors contributed equally to this work.
* mbazhenov@ucsd.edu

OPEN ACCESS

Citation: 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 Comput Biol 18(11): e1010628. https://doi.org/10.1371/journal.pcbi.1010628

Editor: Daniel Bush, University College London, UNITED KINGDOM

Received: April 22, 2022

Accepted: October 3, 2022

Published: November 18, 2022

Peer Review History: PLOS recognizes the benefits of transparency in the peer review process; therefore, we enable the publication of all of the content of peer review and author responses alongside final, published articles. The editorial history of this article is available here: https://doi.org/10.1371/journal.pcbi.1010628

Data Availability Statement: All relevant data are within the paper and its Supporting Information files.

Abstract

Artificial neural networks overwrite previously learned tasks when trained sequentially, a phenomenon known as catastrophic forgetting. In contrast, the brain learns continuously, and typically learns best when new training is interleaved with periods of sleep for memory consolidation. Here we used spiking network to study mechanisms behind catastrophic forgetting and the role of sleep in preventing it. The network could be trained to learn a complex foraging task but exhibited catastrophic forgetting when trained sequentially on different tasks. In synaptic weight space, new task training moved the synaptic weight configuration away from the manifold representing old task leading to forgetting. Interleaving new task training with periods of off-line reactivation, mimicking biological sleep, mitigated catastrophic forgetting by constraining the network synaptic weight state to the previously learned manifold, while allowing the weight configuration to converge towards the intersection of the manifolds representing old and new tasks. The study reveals a possible strategy of synaptic weights dynamics the brain applies during sleep to prevent forgetting and optimize learning.

Author summary

Artificial neural networks can achieve superhuman performance in many domains. Despite these advances, these networks fail in sequential learning; they achieve optimal performance on newer tasks at the expense of performance on previously learned tasks. Humans and animals on the other hand have a remarkable ability to learn continuously and incorporate new data into their corpus of existing knowledge. Sleep has been hypothesized to play an important role in memory and learning by enabling spontaneous reactivation of previously learned memory patterns. Here we use a spiking neural network model, simulating sensory processing and reinforcement learning in animal brain, to demonstrate that interleaving new task training with sleep-like activity optimizes the

Funding: This study was supported by ONR (N00014-16-1-2829 to MB), Lifelong Learning Machines program from DARPA/MTO (HR0011-18-2-0021 to MB), NSF (EFRI BRAID 2223839 to MB), and NIH (1RF1MH117155 to MB; 1R01MH125557 to MB; 1R01NS109553 to MB). The funders had no role in study design, data collection and analysis, decision to publish, or preparation of the manuscript.

Competing interests: The authors have declared that no competing interests exist.

network's memory representation in synaptic weight space to prevent forgetting old memories. Sleep makes this possible by replaying old memory traces without the explicit usage of the old task data.

Introduction

Humans are capable of continuously learning to perform novel tasks throughout life without interfering with their ability to perform previous tasks. Conversely, while modern artificial neural networks (ANNs) are capable of learning to perform complicated tasks, ANNs have difficulty learning multiple tasks sequentially [1–3]. Sequential training commonly results in catastrophic forgetting, a phenomenon which occurs when training on the new task completely overwrites the synaptic weights learned during the previous task, leaving the ANN incapable of performing a previous task [1–4]. Attempts to solve catastrophic forgetting have drawn on insights from the study of neurobiological learning, leading to the growth of neuroscience-inspired artificial intelligence (AI) [5–8]. While proposed approaches are capable of mitigating catastrophic forgetting in certain circumstances, a general solution which can achieve human level performance for continual learning is still an open question [9].

Historically, an interleaved training paradigm, where multiple tasks are presented within a common training dataset, has been employed to circumvent the issue of catastrophic forgetting [4,10,11]. In fact, interleaved training was originally construed to be an approximation to what the brain may be doing during sleep to consolidate memories; spontaneously reactivating memories from multiple interfering tasks in an interleaved manner [11]. Unfortunately, explicit use of interleaved training, in contrast to memory consolidation during biological sleep, imposes the stringent constraint that the original training data be perpetually stored for later use and combined with new data to retrain the network [1,2,4,11]. Thus, the challenge is to understand how the biological brain enables memory reactivation during sleep without access to past training data.

Parallel to the growth of neuroscience-inspired ANNs, there has been increasing investigation of spiking neural networks (SNNs) which attempt to provide a more realistic model of brain functioning by taking into account the underlying neural dynamics and by using biologically plausible local learning rules [12–15]. A potential advantage of the SNNs, that was explored in our new study, is that local learning rules combined with spike-based communication allow previously learned memory traces to reactivate spontaneously and modify synaptic weights without interference during off-line processing–sleep. Indeed, a common hypothesis, supported by a vast range of neuroscience data, is that the consolidation of memories during sleep occurs through synaptic changes enabled by reactivation of the neuron ensembles engaged during learning [16–20]. It has been suggested that Rapid Eye Movement (REM) sleep supports the consolidation of non-declarative or procedural memories, while non-REM sleep supports the consolidation of declarative memories [16,21–23].

Here we used a multi-layer SNN with reinforcement learning to investigate whether interleaving periods of new task training with periods of sleep-like autonomous activity, can circumvent catastrophic forgetting. The network can be trained to learn one of two complementary complex foraging tasks involving pattern discrimination but exhibits catastrophic forgetting when trained on the tasks sequentially. Significantly, we show that catastrophic forgetting can be prevented by periodically interrupting reinforcement learning on a new task with sleep-like phases. From the perspective of synaptic weight space, while new task training alone moves the synaptic weight configuration away from the old task's manifold–a subspace of synaptic weight space that guarantees high performance on that task—and towards

the new task manifold, interleaving new task training with sleep replay allows the synaptic weights to stay near the old task manifold and still move towards its intersection with the manifold representing the new task, i.e., converge to the intersection of these manifolds. Our study predicts that sleep prevents catastrophic forgetting in the brain by forming joint synaptic weight representations suitable for storing multiple memories.

Results

Human and animal brains are complex and although there are many differences between species, critical common elements can still be identified from insects to humans. From an anatomic perspective, this includes largely the sequential processing of sensory information, from raw low level representations on the sensory periphery to high level representations deeper in the brain followed by decision making networks controlling the motor circuits. From a functional perspective, this includes local synaptic plasticity, combination of different plasticity rules and sleep-wake cycle that was shown to be critical for memory and learning in variety of species from insects [24–26] to vertebrates [16]. In this new study we model a basic brain neural circuit including many of these anatomical and functional elements. While our model is extremely simplified, it captures critical processing steps found, e.g., in insect olfactory system where odor information is sent from olfactory receptors to the mushroom bodies and then to the motor circuits. In vertebrates, visual information is sent from the retina to early visual cortex and then to decision making layers in associative cortices to drive motor output. Many of these steps are plastic, in particular decision making circuits utilize spike timing dependent plasticity (STDP) in insects [27] and vertebrates [28,29].

Fig 1A illustrates a feedforward spiking neural network (see also *Methods*: *Network Structure* for details) simulating the basic steps from sensory input to motor output. Excitatory synapses between the input (I) and hidden (H) layers were subjected to unsupervised learning (implemented as non-rewarded STDP) [28,29] while those between the H and output (O) layers were subjected to reinforcement learning (implemented using rewarded STDP) [30–33] (see *Methods*: *Synaptic plasticity* for details). Unsupervised plasticity allowed neurons in layer H to learn different particle patterns at various spatial locations of the input layer I, while rewarded STDP allowed the neurons in layer O to learn motor decisions based on the type of the particle patterns detected in the input layer [14]. While inspired by the processing steps of a biological brain, this structure also mimics basic elements of the feedforward artificial neural networks (ANNs), including convolutional layer (from I to H) and fully connected layer (from H to O) [34].

Complementary complex foraging tasks can be robustly learned

We trained the network on one of two complementary complex foraging tasks. In either task, the network learned to discriminate between rewarded and punished particle patterns in order to acquire as much reward as possible. We consider pattern discriminability (ratio of rewarded vs punished particles consumed) as a measure of performance, with chance performance being 0.5. All reported results are based on at least 10 trials with different random network initialization.

The paradigm for Task 1 is shown in Fig 1B. First, during an unsupervised learning period, all 4 types of 2-particle patterns (horizontal, vertical, positive diagonal, and negative diagonal) were present in the environment with equal densities. This was a period, equivalent to a developmental critical period in the brain (or training convolutional layers in ANN), when the network learned the environmental statistics and formed, in layer H, high level representations of all possible patterns found at the different visual field locations (see Fig 2 for details).

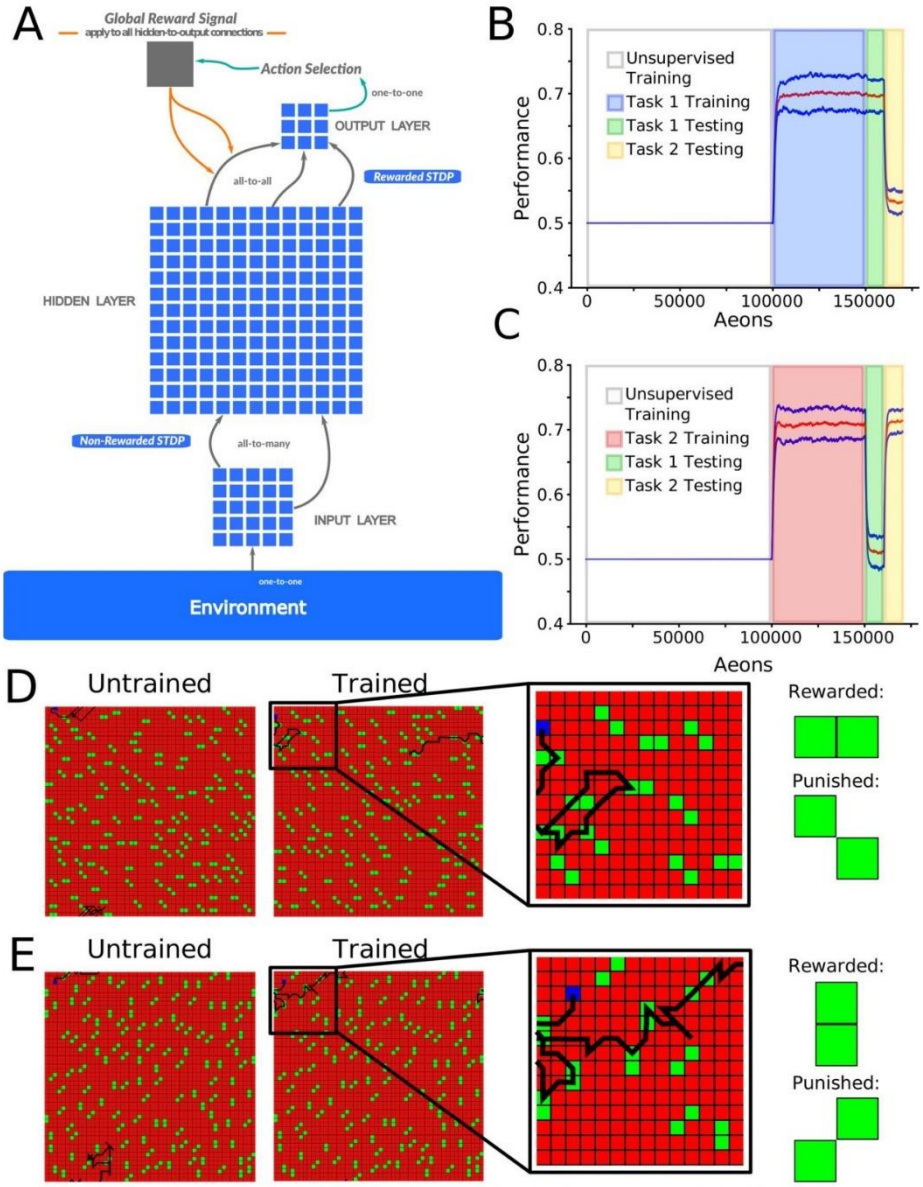


Fig 1. Network architecture and foraging task structure. (A) The network had three layers of neurons with a feed-forward connectivity scheme. Input from virtual environment was simulated as a set of excitatory inputs to the input layer neurons ("visual field"- 7x7 subspace of 50x50 environment) representing the position of food particles in an egocentric reference frame relative to the virtual agent. Each hidden layer neuron received an excitatory synapse from 9 randomly selected input layer neurons. Each output layer neuron received one excitatory and one inhibitory synapse from each hidden layer neuron. The most active neuron in the output layer (size 3x3) determined the direction of

movement. **(B)** Mean performance (redline) and standard deviation (blue lines) over time: unsupervised training (white), Task 1 training (blue), and Task 1 (green) and Task 2 (yellow) testing. The y-axis represents the agent's performance, or the probability of acquiring rewarded as opposed to punished particle patterns. The x-axis is time in aeons (1 aeon = 100 movement cycles). **(C)** The same as shown in (B) except now for: unsupervised training (white), Task 2 training (red), and Task 1 (green) and Task 2 (yellow) testing. **(D)** Examples of trajectories through the environment at the beginning (left) and at the end (middle-left) of training on Task 1, with a zoom in on the trajectory at the end of training (middle-right), and the values of the task-relevant food particles (right). **(E)**. The same as shown in (D) except for Task 2.

https://doi.org/10.1371/journal.pcbi.1010628.g001

Unsupervised training was followed by a reinforcement learning period, equivalent to task specific training in the brain (or training a specific set of classes in an ANN), during which the synapses between layers I and H were frozen while synapses from H to O were updated using a

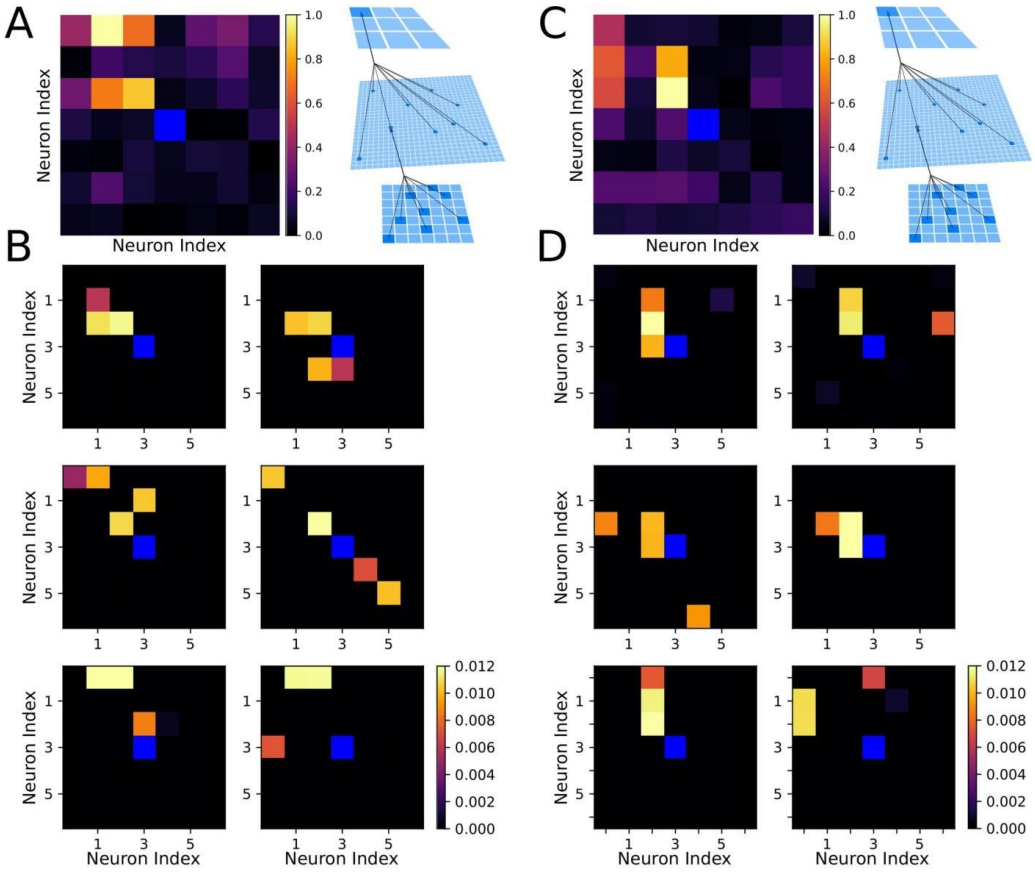


Fig 2. Receptive fields of output and hidden layer neurons determine the agent behavior. **(A)** Left, Receptive field of the output layer neuron controlling movement to the upper-left direction following training on Task 1. This neuron can be seen to selectively respond to horizontal orientations in the upper-left quadrant of the visual field. Right, Schematic of connections between layers. **(B)** Examples of receptive fields of hidden layer neurons which synapse strongly onto the output neuron from (A) after training on Task 1. **(C)** The same as shown in (A) except following training on Task 2. The upper-left decision neuron can be seen to selectively respond to vertical orientations in the upper-left quadrant of the visual field. **(D)** The same as shown in (B) except following training on Task 2.

https://doi.org/10.1371/journal.pcbi.1010628.g002

rewarded STDP rule. The reinforcement learning period was when the network learned to make decisions about which direction to move based on the visual input. For Task 1, horizontal patterns were rewarded and negative diagonal patterns were punished (Fig 1D). During both the rewarded training and the testing periods only 2 types of patterns were present in the environment (e.g. horizontal and negative diagonal for Task 1).

After training Task 1, mean performance across ten trials on Task 1 was 0.70 ± 0.02 while performance on the untrained Task 2 was 0.53 ± 0.02 (chance level). The naive agent moved randomly through the environment (Fig 1D, left), but after task training, moved to seek out horizontal patterns and largely avoid negative diagonal ones (Fig 1D, right). The complementary paradigm for Task 2 (vertical patterns are rewarded, and positive diagonal are punished) is shown in Fig 1C and 1E. These results demonstrate that the network is capable of learning and performing either one of the two complementary complex foraging tasks. The similarity between these tasks is evident in their definition (symmetrical particle orientations; Fig 1D and 1E), through the similar performances attained by the network on each task (Fig 1B and 1C), and through the similar levels of activity induced in the network when training each task (S1A and S1B Fig)

To understand how sensitive a trained network was to pruning, we employed a neuronal dropout procedure which progressively removes neurons from the hidden layer at random (S2 Fig). We found the network was able to keep performance steady on either task following training until around 70% of the hidden layer was pruned. Such high resiliency suggests the network utilizes a highly distributed coding strategy to develop its policy.

Next, to understand synaptic changes during training, we computed receptive fields of each neuron in layer O with respect to the inputs from layer I (see schematic in Fig 2A and 2C). This was done by first computing the receptive fields of all of the neurons in layer H with respect to I, then performing a weighted average where the weights were given by the synaptic strength from each neuron in layer H to the particular neuron in layer O. Fig 2A shows a representative example of the receptive field which developed after training on Task 1 for one specific neuron in layer O which controls movements to the upper-left direction. This neuron responded most robustly to bars of horizontal orientation (rewarded) in the upper-left quadrant of the visual field and, importantly, did not respond to bars of negative diagonal orientation (punished).

Fig 2B shows examples of receptive fields of six neurons in layer H which synapse strongly onto the upper-left neuron in layer O (the neuron shown in Fig 2A). These neurons formed high level representations of the input patterns, similar to the neurons in the primary visual system or later layers of a convolutional neural network [35–37]. The majority of these receptive fields revealed strong selection for the horizontal (i.e. rewarded) food particles in the upper-left quadrant of the visual field. As a particularly notable example, one of these layer H neurons (Fig 2B; middle-right) preferentially responded to negative diagonal (i.e. punished) food particles in the bottom-right quadrant of the visual field. Thus, spiking in this neuron caused the agent to move away from these punished food particles. Similar findings after training on Task 2 are shown in Fig 2C and 2D.

To further quantify the network's sensitivity to various particle types we developed a metric termed the Particle Responsiveness Metric (PRM) to gauge how specific particles influence activity of the output layer neurons (see the section Methods: Particle responsiveness metric for further details). Using PRM on all food particle orientations across ten trials, we found that following Task 1 training the network is drawn to horizontal particles (S3A Fig) while post Task 2 training vertical particles drive output layer activity (S3B Fig), thus quantitatively supporting the qualitative results displayed in Fig 2.

Sleep prevents catastrophic forgetting of the old task during new task training

We next tested whether the model exhibits catastrophic forgetting by training sequentially on Task 1 (old task) followed by Task 2 (new task) (Fig 3A). Following Task 2 training, mean performance across ten trials on Task 1 was down to no better than chance (0.52 ± 0.02), while performance on Task 2 improved to 0.69 ± 0.03 (Fig 3A and 3B). Thus, sequential training on a complementary task caused the network to undergo catastrophic forgetting of the task trained earlier, remembering only the most recent task.

Interleaved training was proposed as a solution for catastrophic forgetting [4,10,11]. In the next experiment, after training on Task 1, we simulated interleaved T1/T2 training (Interleaved$_{T1,T2}$) when we alternated short presentations of Task 1 and Task 2 every 100 movement cycles (Fig 3C). Sample network activity from this period can be seen to closely resemble single task training (S1C Fig). Following interleaved training, the network achieved a mean performance of 0.68 ±0.03 on Task 1 and a performance of 0.65 ± 0.04 on Task 2 across trials. Therefore, interleaved training allowed the network to learn new Task 2 without forgetting previously learned Task 1. However, while interleaved training made it possible to learn both tasks, it imposes the stringent constraint that all the original training data (in our case explicit access to the Task 1 environment) be stored for later use and combined with new data to retrain the network [1,2,4,11].

Sleep is believed to be an off-line processing period when recent memories are replayed to avoid damage from new learning. We previously showed that sleep replay improves memory in a thalamocortical network [38–40] and when a network was trained to learn interfering tasks sequentially, sleep prevented the old task memory from catastrophic forgetting [41]. Can we implement a sleep like phase to our model to protect an old task and still accomplish new task learning without explicit re-training of the old task? In vivo, activity of the neocortical neurons during REM sleep is low-synchronized and similar to baseline awake activity [42]. Therefore, to simulate REM sleep-like activity in the model, the rewarded STDP rule was replaced by unsupervised STDP, the input layer was silenced while hidden layer neurons were artificially stimulated by Poisson distributed spike trains in order to maintain spiking rates similar to that during task training (see *Methods*: *Simulated Sleep* for details). Sample network activity recorded during this sleep phase is visualized in the raster plots shown in S1D Fig.

Again, we first trained the network on Task 1. Next, we implemented a training phase consisted of alternating periods of training on Task 2 (new task) lasting 100 movement cycles and periods of "sleep" of the same duration (we will refer to this training phase as Interleaved$_{S,T2}$) (Fig 3E). Importantly, no training on Task 1 was performed at any time during Interleaved$_{S,T2}$. Following Interleaved$_{S,T2}$, the network achieved a mean performance across ten trials of 0.68 ± 0.05 on Task 2 and retained a performance of 0.70 ± 0.03 on Task 1 (Fig 3E and 3F), comparable to single Task 1 (0.70 ± 002) or Task 2 (0.69 ± 0.03) performances (Fig 1B and 1C) and exceeding those achieved through Interleaved$_{T1,T2}$ training (Fig 3C and 3D).

We interpret these results as follows (see below for detailed synaptic connectivity analysis). Each episode of new Task 2 training improves Task 2 performance but damages synaptic connectivity responsible for old Task 1. If continuous Task 2 training is long enough, the damage to Task 1 becomes irreversible. Having a sleep phase after a short period of Task 2 training enables spontaneous forward replay between hidden and output layers (H->O) that preferentially benefits the strongest synapses. Thus, if Task 1 synapses are still strong enough to maintain replay, they are replayed and weights are increased.

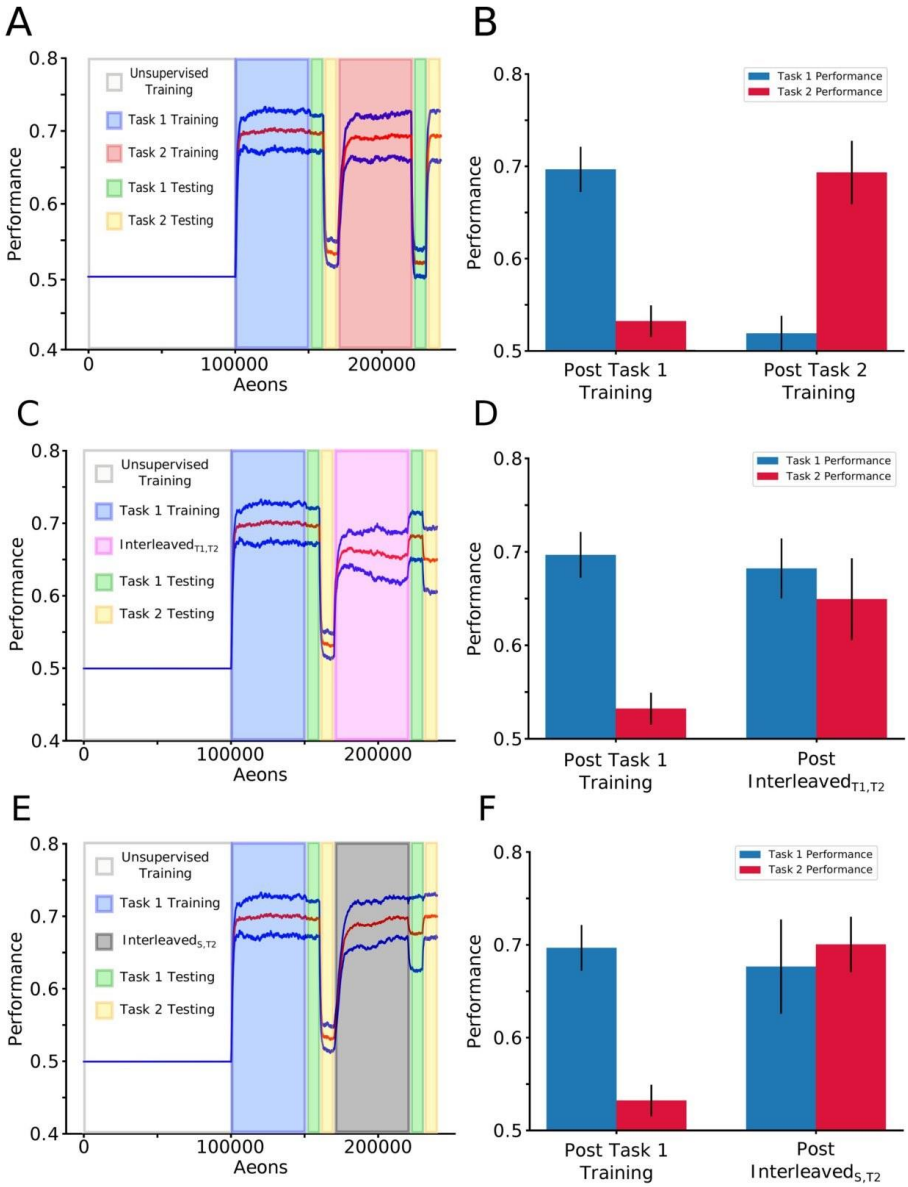


Fig 3. Sleep prevents catastrophic forgetting during new task training. (A) Mean performance (red line) and standard deviation (blue lines) over time: unsupervised training (white), Task 1 training (blue), Task 1/2 testing (green/yellow), Task 2 training (red), Task 1/2 testing (green/yellow). **(B)** Mean and standard deviation of performance during testing on Task 1 (blue) and Task 2 (red). Task 2 training after Task 1 training led to Task 1 forgetting. **(C)** Task paradigm similar to that shown in (A) but with Interleaved$_{T1,T2}$ training (pink) instead of Task 2 training. **(D)** Mean and standard deviation of performance during testing on Task 1 (blue) and Task 2 (red).

Interleaved$_{T1,T2}$ training allowed new Task 2 learning without forgetting old Task 1. **(E)** Task paradigm similar to that shown in (A) but with Interleaved$_{S,T2}$ training (gray) instead of Task 2 training. **(F)** Mean and standard deviation of performance during testing on Task 1 (blue) and Task 2 (red). Embedding sleep phases to the new Task 2 training protected old Task 1 memory.

https://doi.org/10.1371/journal.pcbi.1010628.g003

Sleep can protect synaptic configuration from previous training but does not provide training by itself

In simulations presented in Fig 3, during sleep phase, each hidden layer neuron was stimulated by noise, a Poisson distributed spike train, and we ensured that its firing rate during sleep would be close to the mean rate of that neuron firing across all the preceding training sessions. Therefore, intensity of the noise input during Interleaved$_{S,T2}$ was influenced by preceding Task 1 training and could also vary between H neurons. To eliminate the possibility that such input may provide direct Task 1 training during sleep, three additional experiments were conducted. First, we applied Interleaved$_{S,T1}$ phase to a completely naive network. Importantly, even though this network was never trained on Task 2, we used information about hidden layer neuron firing rates after Task 2 training from another experiment. In other words, we artificially took into account Task 2 firing rate data to design random input during sleep to check if this might be sufficient to improve the network performance on Task 2. We found that the network learns Task 1 but Task 2 performance remained at baseline (S4A and S4B Fig). In a second experiment, a similar period of Interleaved$_{S,T1}$ was applied following Task 1 training (S4C and S4D Fig) and we found that it maintained performance on Task 1 but again without any performance gain for Task 2.

In a third experiment, we repeated the sequence shown in Fig 3E, however, during the sleep phase, we provided each hidden layer neuron with a Poisson spike train input which was drawn (independently) from the same distribution, i.e., we used the same input firing rate for all hidden layer neurons determined by the mean firing of the entire hidden layer population as opposed to the private spiking history of individual H neurons in the Fig 3E and 3F experiments (termed Uniform-Noise Sleep (US)). The network's performance under this implementation of noise, Interleaved$_{US,T1}$, (S4E and S4F Fig) was similar to that from our original sleep implementation (see Fig 3E and 3F). Taken together, these results suggest that the properties of the input that drives firing during sleep are not essential to enable replay, any similar to awake random activity in layers H and O is sufficient to prevent forgetting.

Sleep replay protects critical synapses of the old tasks

To reveal synaptic weights dynamics during training and sleep, we next traced "task-relevant" synapses, i.e. synapses identified in the top 10% of the distribution following training on that specific task. We first trained Task 1, followed by Task 2 training (Fig 4A) and we identified "task-relevant" synapses after each task training. Next, we continued by training Task 1 again but we interleaved it with periods of sleep: T1->T2->Interleaved$_{S,T1}$. Sequential training of Task 2 after Task 1 led to forgetting of Task 1, but after Interleaved$_{S,T1}$ Task 1 was relearned while Task 2 was preserved (Fig 4A and 4B), as in the experiments in the previous section (Fig 3C). Importantly, this protocol allowed us to compare synaptic weights after Interleaved$_{S,T1}$ training with those identified as task-relevant after individual Task 1 and Task 2 training (Fig 4C). The structure in the distribution of Task 1-relevant synapses formed following Task 1 training (Fig 4C; top-left) was destroyed following Task 2 training (top-middle) but partially recovered following Interleaved$_{S,T1}$ training (top-right). The distribution structure of Task 2-relevant synapses following Task 2 training (bottom-middle) was not present following Task 1 training (bottom-left) and was partially retained following Interleaved$_{S,T1}$ training (bottom-

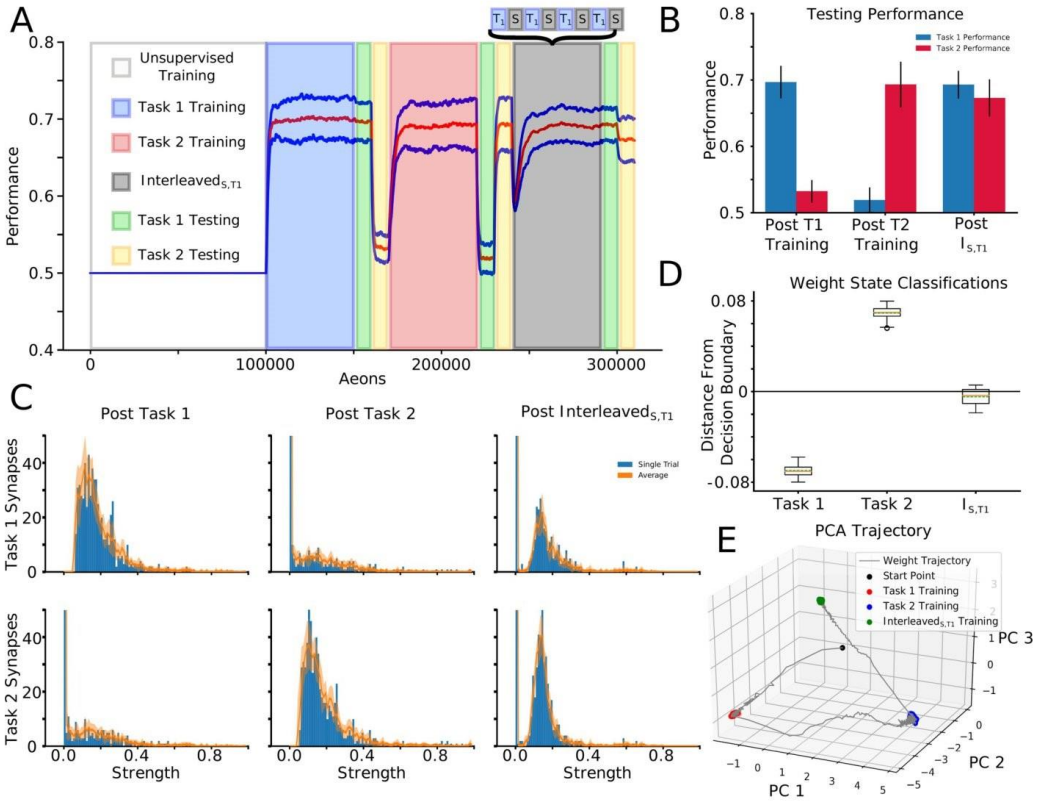


Fig 4. Interleaving periods of new task training with sleep allows integrating synaptic information relevant to new task while preserving old task information. **(A)** Mean performance (red line) and standard deviation (blue lines) over time: unsupervised training(white), Task 1 training (blue), Task 1/2 testing (green/yellow), Task 2 training (red), Task 1/2 testing (green/yellow), Interleaved$_{S,T1}$ training (grey), Task 1/2 testing (green/yellow). Note that performance for Task 2 remains high at the end despite no Task 2 training during Interleaved$_{S,T1}$. **(B)** Mean and standard deviation of performance during testing on Task 1 (blue) and Task 2 (red). **(C)** Distributions of task-relevant synaptic weights (blue bars–single trial, orange line / shaded region–mean / std across 10 trails). The distributional structure of Task 1-relevant synapses following Task 1 training (top-left) is destroyed following Task 2 training (top-middle), but partially recovered following Interleaved$_{S,T1}$ training (top-right). Similarly, the distributional structure of Task 2-relevant synapses following Task 2 training (bottom-middle), which was not present following Task 1 training (bottom-left), was partially preserved following Interleaved$_{S,T1}$ training (bottom-right). **(D)** Box plots with mean (dashed green line) and median (dashed orange line) of the distance to the decision boundary found by an SVM trained to classify Task 1 and Task 2 synaptic weight matrices for Task 1, Task 2, and Interleaved$_{S,T1}$ training across trials. Task 1 and Task 2 synaptic weight matrices had mean classification values of -0.069 and 0.069 respectively, while that of Interleaved$_{S,T1}$ training was -0.0047. **(E)** Trajectory of H to O layer synaptic weights through PC space. Synaptic weights which evolved during Interleaved$_{S,T1}$ training (green dots) clustered in a location of PC space intermediary between the clusters of synaptic weights which evolved during training on Task 1 (red dots) and Task 2 (blue dots).

https://doi.org/10.1371/journal.pcbi.1010628.g004

right). It should be noted that this qualitative pattern can be distinctly observed in a single trial (Fig 4C; Blue Bars), but also generalizes across trials (Fig 4C; Orange Line). Thus, sleep can preserve important synapses while incorporating new ones.

To better understand the effect of Interleaved$_{S,T1}$ training on the synaptic weights, we trained a support vector machine (SVM; see *Method*: *Support Vector Machine Training* for details) to classify the synaptic weight configurations between layers H and O according to whether they serve to perform Task 1 or Task 2 on every trial. Fig 4D shows that the SVMs

robustly and consistently classified the synaptic weight states after Task 1 and Task 2 training while those after $Interleaved_{S,T1}$ fell significantly closer to the decision boundary. This indicates that the synaptic weight matrices which result from $Interleaved_{S,T1}$ training are a mixture of Task 1 and Task 2 states. Using principal components analysis (PCA), we found that while synaptic weight matrices associated with Task 1 and Task 2 training cluster in distinct regions of PC space, $Interleaved_{S,T1}$ training pushes the synaptic weights to an intermediate location between Task 1and Task 2 (Fig 4E). Importantly, the smoothness of this trajectory to its steady state suggests that Task 2 information is never completely erased during this evolution. We take this as evidence that $Interleaved_{S,T1}$ training is capable of integrating synaptic information relevant to Task 1 while protecting Task 2 information.

This analysis applied during interleaved training of Task 1 and Task 2 ($Interleaved_{T1,T2}$), revealed similar results (S5 Fig), suggesting that $Interleaved_{S,T1}$ can enable similar synaptic weights dynamics as $Interleaved_{T1,T2}$ training, but without access to the old task data (old training environment).

Receptive fields of decision-making neurons after sleep represent multiple tasks

To confirm that the network had learned both tasks after $Interleaved_{S,T1}$ training, we visualized the receptive fields of decision-making neurons in layer O (Fig 5; see Fig 2 for comparison). Fig 5A shows the receptive field for the neuron in layer O which controls movement in the upper-left direction. This neuron responded to both horizontal (rewarded for Task 1) and vertical (rewarded for Task 2) orientations in the upper-left quadrant of the visual field. Although it initially appears that this layer O neuron may also be responsive to diagonal patterns in this region, analysis of the receptive fields of neurons in layer H (Fig 5B) revealed that these receptive fields are selective to either horizontal food particles (left six panels; rewarded for Task 1) or vertical food particles (right six panels; rewarded for Task 2) in the upper-left quadrant of the visual field. Other receptive fields were responsible for avoidance of punished particles for both tasks (see examples in Fig 5B, bottom-middle-right and bottom-middle-left). Thus, the network utilizes one of two distinct sets of layer H neurons, selective for either Task 1 or Task 2, depending on which food particles are present in the environment. To validate these qualitative results we inspected the PRM metrics for all food particle orientations across ten trials following $Interleaved_{S,T1}$ training. The comparatively high mean values for horizontal and vertical food particle orientations revealed the network's movement was significantly driven by these rewarded food particle orientations (horizontal and vertical), quantifying multitask memory integration into the network's synaptic weight matrix. (S3C Fig).

Periods of sleep allow for integration of a new task memory without interference through renormalization of task-relevant synapses

To visualize synaptic weight dynamics during $Interleaved_{S,T1}$ training, traces of all synapses projecting to a single representative layer O neuron were plotted (Fig 6A). As in Fig 4, we wanted to monitor task specific synapses, so we used the training paradigm of T1->T2->$Interleaved_{S,T1}$, and Task 1 and Task 2 relevant synapses were identified after each individual task training. At the onset of $Interleaved_{S,T1}$ training (i.e. 240,000 aeons), the network was only able to perform on Task 2, meaning the strong synapses in the network were specific to this task. These synapses were represented by a cluster ranging from ~0.08 to ~0.4; the rest of synapses grouped near 0. As $Interleaved_{S,T1}$ training progressed, Task 1 specific synapses moved to the strong cluster and some, presumably less important, Task 2 synapses moved to the weak

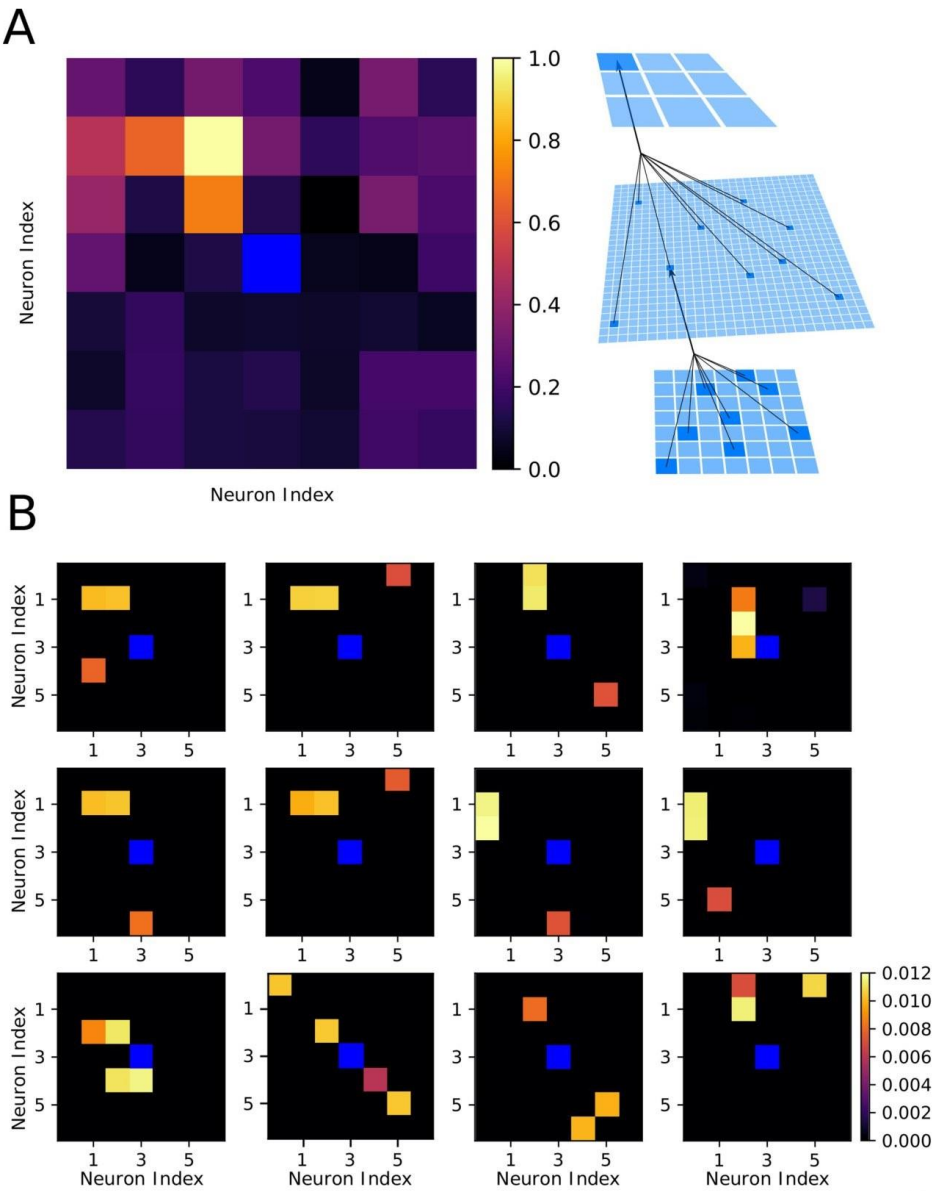


Fig 5. Receptive fields following interleaved Sleep and Task 1 training reveal how the network can multiplex the complementary tasks. (A) Left, Receptive field of the output layer neuron controlling movement to the upper-left direction following interleaved sleep and Task 1training. This neuron has a complex receptive field capable of responding to horizontal and vertical orientations in the upper-left quadrant of the visual field. Right, Schematic of the connectivity between layers. **(B)** Examples of receptive fields of hidden layer neurons which synapse strongly onto the output neuron from (A) after interleaved Sleep and Task 1 training. The majority of these neurons selectively respond to horizontal food

particles (left half) or vertical food particles (right half) in the upper-left quadrant of the visual field, promoting movement in that direction and acquisition of the rewarded patters.

https://doi.org/10.1371/journal.pcbi.1010628.g005

cluster. After a period of time the rate of transfer decreased and the total number of synapses in each group stabilized, showing that the network approached equilibrium (Fig 6B).

To visualize how sleep renormalizes task relevant synapses, we plotted two-dimensional weight distributions for T1->T2 (Fig 6C) and T2->Interleaved$_{S,T1}$ (Fig 6D) experiments (see *Methods*: *2-D Synaptic Weight Distributions* for details). To establish a baseline, in Fig 6C (left) the weight state at the end of Task 1 training (X-axis) (see overall timeline of this experiment in Fig 4A) was compared to itself (Y-axis). This formed a perfectly diagonal plot. The next comparison (Fig 6C, middle) was between the weight state after Task 1 training (X-axis) and a time early on Task 2 training (Y-axis). At that time, synapses were only able to modify their strength slightly, causing most points to lie close to the diagonal. As training on Task 2 continued, synapses moved far away from the diagonal (Fig 6C, right). Two trends were observed: (a) set of synapses that had a strength near zero following Task 1 training increased strength following Task 2 training (Fig 6D, right, red dots along Y-axis); (b) many strongly trained by Task 1 synapses were depressed down to zero (Fig 6C, right, red dots along X-axis). The latter illustrates the effect of catastrophic forgetting—complete overwriting of the synaptic weight matrix caused performance of Task 1 to return to baseline after training on Task 2.

Does sleep prevent overwriting of the synaptic weight matrix? Fig 6D plots used the weight state at the end of training Task 2 as a reference which is then compared to different times during Interleaved$_{S,T1}$ training. The first two plots (Fig 6D, left/middle) are similar to those in Fig 6C. However, after continuing Interleaved$_{S,T1}$ training (Fig 6D, right) many synapses that were strong following Task 2 training were not depressed to zero but rather were pushed to an intermediate strength (note cluster of points parallel to X-axis). Thus, Interleaved$_{S,T1}$ training preserved strong synapses from a previously learned task while also introducing new strong synapses to perform the new task.

Can we prevent old task forgetting simply by freezing a fraction of old task-relevant synapses to prevent their damage by new training? We found that freezing 1% of Task 1-relevant weights allowed Task 2 to be learned, but was not capable of preserving Task 1 (S6A Fig). Freezing 5% of Task 1-relevant weights (S6B Fig) resulted in modest performance on both tasks, but significantly below that seen after Interleaved$_{S,T2}$ (see Fig 3F). Finally, freezing 10% of Task 1-relevant weights (S6C Fig) was capable of fully preserving Task 1 performance, but prevented Task 2 from being learned.

Thus, in all cases, some degree of retroactive or prospective interference was observed highlighting the fact that the sleep-like phase performs a significantly more sophisticated modification to the weight matrix than simply freezing (or amplifying) task relevant synapses. Sleep is capable of intelligently selecting which certain strong synapses to maintain in addition to which weak synapses should be strengthened. Indeed, the sleep phase results in a large cluster of weights being renormalized around an intermediate value of synaptic strength in the network. This may also explain why we observed somewhat better overall performance (combined performance on both tasks) after sleep compare with interleaved training (see Fig 3). Indeed, interleaved training requires repetitive activation of the entire memory pattern, so if different memory patterns compete for synaptic resources then each phase of interleaved training will enhance one memory trace but damage the others. This is in contrast to spontaneous replay during sleep when only task specific subsets of neurons and synapses may be involved in each replay episode. It is worth mentioning that freezing a fraction of synaptic weights that are most relevant to old tasks (however, implemented in more complex form) is

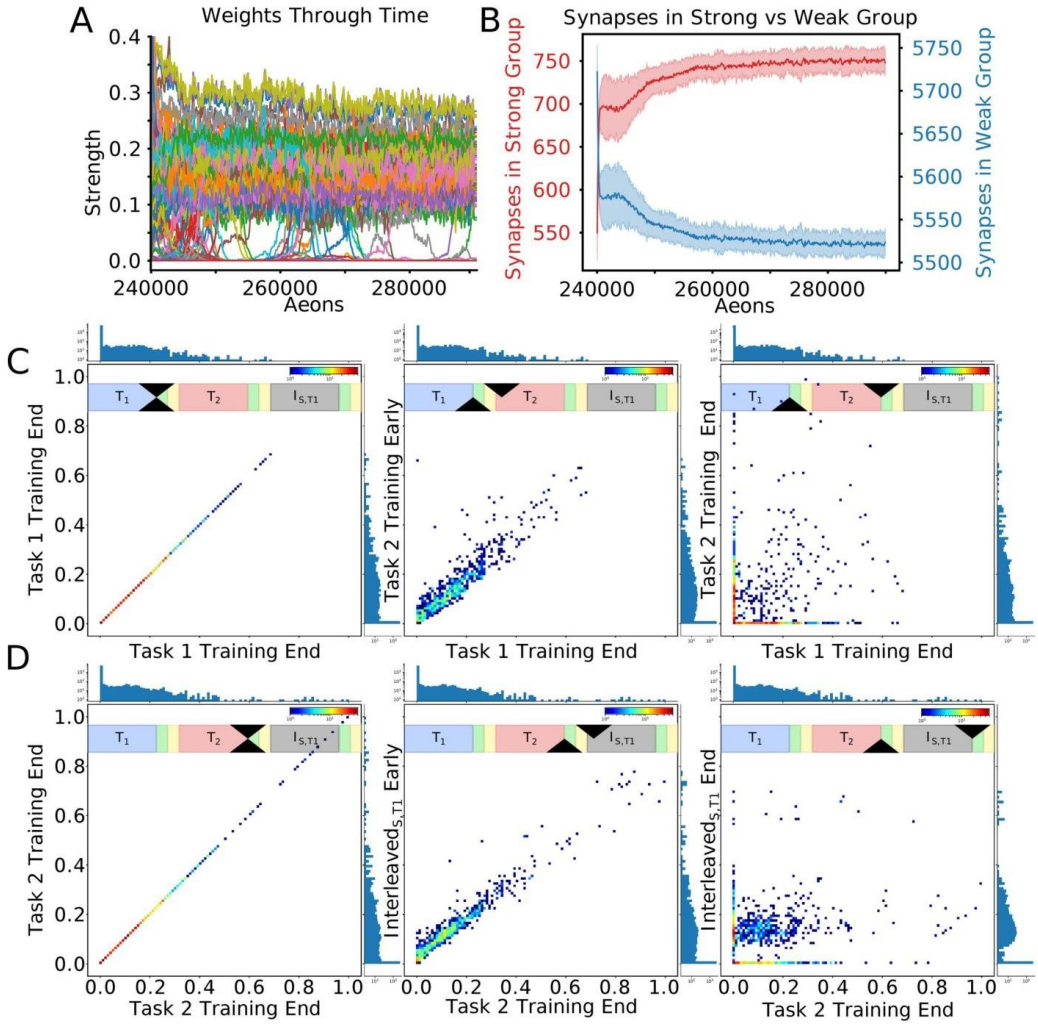


Fig 6. Periods of sleep allow learning Task 1 without interference with old Task 2 through renormalization of task-relevant synapses. **(A)** Dynamics of all incoming synapses to a single output layer neuron during $\text{Interleaved}_{S,T1}$ training shows the synapses separate into two clusters. The network was trained in the following order: T1->T2->$\text{Interleaved}_{S,T1}$. **(B)** Number of synapses in the strong (red) and weak (blue) clusters during $\text{Interleaved}_{S,T1}$. **(C)** Two-dimensional histograms illustrating synaptic weights dynamics. For each plot, the x-axis represents synaptic weight after Task 1 training and the y-axis represents the synaptic weight at a different point in time (Scale bar: brown—50 synapses/bin, blue—1 synapse/bin). One-dimensional projections along top and right sides show the global distribution of synapses at the time slices for a given plot. **(D)** Same as (C) except the x-axis refers to the end of Task 2 training. Note, that after a full period of $\text{Interleaved}_{S,T1}$ training (right), weak synapses were recruited to support Task 1 (red cluster along the y-axis) and many Task 2 specific synapses remained moderately strong (blue cluster along x-axis).

https://doi.org/10.1371/journal.pcbi.1010628.g006

one of the approaches in machine learning to avoid catastrophic forgetting–Elastic Weight Consolidation [7].

Periods of interleaved sleep and new task training push the network weight state towards the intersection of Task 1 and Task 2 synaptic weights configuration manifolds

Can many distinct synaptic weight configurations support a given task, or is each task supported by a unique synaptic connectivity matrix? Our previous analysis suggests that each task can be served by at least two different configurations–one unique for that task (Task 1 or Task 2) and another one that supports both Task 1 and Task 2. To further explore this question and to identify possible task-specific solution manifolds (M_{T1} and M_{T2}) and their intersection ($M_{T1\cap T2}$) in synaptic weights space, we used multiple trials of Task 1 and Task 2 training to sample the manifolds (Fig 7A). Here, red/blue dots indicate an exclusive high degree of performance on Task 1/2 respectively, while cyan and greed dots indicate states where the network is able to perform on both tasks simultaneously. Since this analysis was generated utilizing a wide variety of simulation paradigms with many corresponding trials differing in randomness, we believe it allows us to draw generalized conclusions. We therefore interpret these results as evidence that synaptic weight space includes a manifold, M_{T1}, where different configurations of weights (red, green, cyan dots) all allow for Task 1 to perform well. This manifold intersects with another one, M_{T2}, where different weights configurations (blue, green, cyan dots) are all suitable for Task 2. Fig 7B and 7C show 2D dimensionality reductions to PCA space, and include trajectories in addition to end states. One can see that PC 1 seems to capture the extent to which a synaptic weight configuration is associated with Task 1 (positive values) or Task 2 (negative values), while PC 2 and PC 3 capture the variance in synaptic weight configurations associated with Task 1 and Task 2, respectively. Note, the trajectories through this space (red/blue lines) during $\text{Interleaved}_{\text{T1,T2}}$ and $\text{Interleaved}_{\text{S,T1/T2}}$ training would also belong to the respective task manifolds as performance on the old tasks was never lost in these training scenarios.

We next calculated the distance from the current synaptic weight configurations to M_{T1} (Fig 7D), M_{T2} (Fig 7E), and $M_{T1\cap T2}$ (Fig 7F; see *Methods*: *Distance from Solution Manifolds* for details) during different training protocols. Fig 7D and 7E show that while Sequential (T1->T2 or T2->T1) training causes synaptic weight configurations to diverge quickly from its initial solution manifold (i.e. M_{T1} or M_{T2}) and to remain far (suggesting quick forgetting of the original task), both $\text{Interleaved}_{\text{T1,T2}}$ and $\text{Interleaved}_{\text{S,T1/T2}}$ training cause synaptic weight configurations to stay relatively close to the initial solution manifold as a new task was learned. (Note, that we certainly under sampled M_{T1} and M_{T2}, which may explains initial distance increase.) Importantly, Fig 7F shows that both $\text{Interleaved}_{\text{T1,T2}}$ and $\text{Interleaved}_{\text{S,T1/T2}}$ training cause synaptic weight configurations to smoothly converge towards $M_{T1\cap T2}$, while Sequential training avoids this intersection entirely.

In Fig 7G we present a schematic depiction of these results. The task-specific manifolds, M_{T1} and M_{T2}, are depicted in 3D as two volumes whose boundaries are defined by two orthogonal elliptic paraboloids with opposite orientation. The ellipsoidal intersection approximates the volume comprising $M_{T1\cap T2}$. Fig 7H and 7I depict a cartoon of trajectories taken by the network in this space following Task 2 and Task 1 training, respectively. Sequential training causes the network to jump directly from one task-specific solution manifold to the other, resulting in catastrophic forgetting. In contrast, interleaving new task training with sleep ($\text{Interleaved}_{\text{S,T1/T2}}$) prevents catastrophic forgetting by keeping the network close to the old task solution manifold as it converges towards $M_{T1\cap T2}$ –a region capable of supporting both tasks simultaneously.

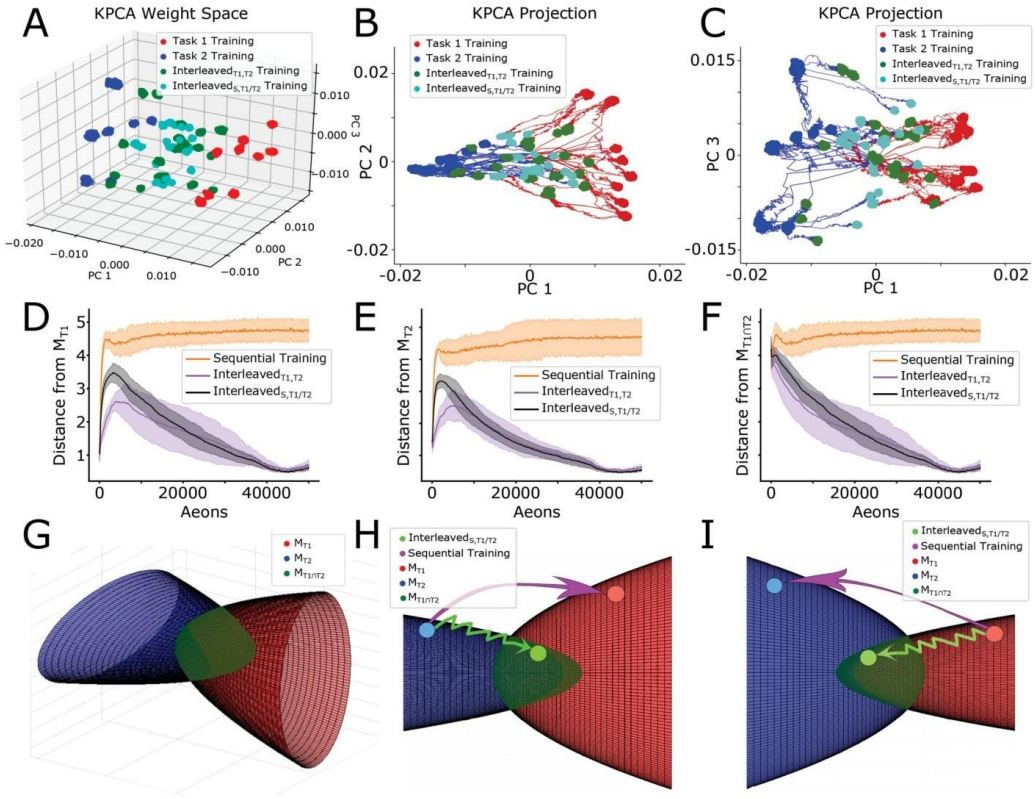


Fig 7. Periods of sleep push the network towards the intersection of Task 1 and Task 2 synaptic weight manifolds. (A-C) Low-dimensional visualizations of the synaptic weight configurations of 10 networks obtained through kPCA for 3-dimensions (A) and 2-dimensions (B-C). Synaptic weight configurations taken from the last fifth of Task 1 (red dots), Task 2 (blue dots), Interleaved$_{T1,T2}$ (green dots), and Interleaved$_{S,T1/T2}$ (cyan dots) training are shown. Trajectories resulting from Interleaved$_{T1,T2}$ and Interleaved$_{S,T1/T2}$ training following Task 1 (Task 2) training are shown in red (blue). **(D-F)** Average (solid lines) and standard deviation (shaded regions) of the Euclidean distances between the current synaptic weight configuration and M_{T1} (D), M_{T2} (E), and $M_{T1\cap T2}$ (F) during Sequential (orange), Interleaved$_{T1,T2}$ (purple), and Interleaved$_{S,T1/T2}$ (black) training. **(G)** Cartoon illustration of the task-specific point-sets shown in (A-C) as solution manifolds M_{T1} (red) and M_{T2} (blue). M_{T1} and M_{T2} can be thought of as two volumes with boundaries defined by the interiors of oppositely oriented elliptic paraboloids which intersect orthogonally defining an approximately ellipsoidal volume near the origin ($M_{T1\cap T2}$; dark green). **(H, I)** Sequential training (pink arrow) causes the network to jump from one solution manifold to the other while avoiding $M_{T1\cap T2}$, while Interleaved$_{S,T1/T2}$ training (light green arrow) keep the network close to the initial solution manifold as it converges towards $M_{T1\cap T2}$.

https://doi.org/10.1371/journal.pcbi.1010628.g007

Discussion

We report that a multi-layer spiking neural network utilizing reinforcement learning exhibits catastrophic forgetting upon sequential training of two complementary complex foraging tasks, however the problem is mitigated if the network is allowed, during new task training, to undergo intervening periods of spontaneous reactivation which are equivalent to the periods of sleep in a biological brain. Old task was spontaneously replayed during sleep, therefore interleaving new task training with sleep was effectively equivalent to explicit interleaved training of the old and new tasks without the need to store and train on previous task data or environments. At the synaptic level, training a new task alone led to complete overwriting of

synaptic weights responsible for the old task. In contrast, interleaving periods of reinforcement learning on a new task with periods of unsupervised plasticity during sleep preserved critical old task synapses to avoid forgetting and enhanced synapses relevant for a new task to allow new task learning. Thus, in synaptic weight space, the network weight configuration was pushed towards the intersection of the manifolds representing synaptic weight configurations associated with individual tasks—an optimal compromise for performing both tasks.

The critical role that sleep plays in learning and memory is supported by a vast, interdisciplinary literature spanning both psychology and neuroscience [16,22,43–45]. Specifically, it has been suggested that REM sleep supports the consolidation of non-declarative or procedural memories while non-REM sleep supports the consolidation of declarative memories [16,21,22]. In particular, REM sleep has been shown to be important for the consolidation of memories of hippocampus-independent tasks involving perceptual pattern separation, such as the texture discrimination task [16,46]. Despite the difference in the cellular and network dynamics during these two stages of sleep [16,22], both are thought to contribute to memory consolidation through repeated reactivation, or replay, of specific memory traces acquired during learning [16,21,39,44,47–49]. These studies suggest that through replay, sleep can support the process of off-line memory consolidation to circumvent the problem of catastrophic forgetting.

From mechanistic perspective, the sleep phase in our model protects old memories by enabling spontaneous reactivation of neurons and changing synapses responsible for previously learned tasks. We previously reported that in the thalamocortical model a sleep phase may enable replay of spike sequences learned in awake to improve post-sleep performance [38–40] and to protect old memories from catastrophic forgetting [41]. Here we found, however, that a single episode of new task training using reinforcement learning could quickly erase old memories to the point that they cannot be recovered by subsequent sleep. The solution was similar to how the brain slowly learns procedural (hippocampal-independent) memories [16,21,22,46,50]. Each episode of new task training improves new task performance only slightly but also damages slightly synaptic connectivity responsible for the older task. Subsequent sleep phases enable replay that preferentially benefits the strongest synapses, such as those from old memory traces, to allow them to recover.

We found that multiple distinct configurations of synaptic weights can support each task, suggesting the existence of task specific solution manifolds in synaptic weight space. Sequential training of new tasks makes the network to jump from one solution manifold to another, enabling memory for the most recent task but erasing memories of the previous tasks. Interleaving new task training with sleep phases enables the system to evolve towards intersection of these manifolds where synaptic weight configurations can support multiple tasks (a similar idea was recently proposed in the machine learning literature to minimize catastrophic interference by learning representations that accelerate future learning [51]). From this point of view having multiple episodes of new task training interleaved with multiple sleep episodes allows gradual convergence to the intersection of the manifolds representing old and new tasks, while staying close to the old task manifold. In contrast, a single long episode of new task learning would push the network far away from the old task manifold making it impossible to recover by subsequent sleep.

Although classical interleaved training of the old and new tasks showed similar performance results in our model as interleaving new task training with sleep, we believe the latter to be superior on the following theoretical grounds. Classical interleaved training will necessarily cause the system to oscillate about the optimal location in synaptic weight space which can support both tasks because each training cycle uses a cost function specific to only a single task. While this can be ameliorated with a learning rate decay schedule, the system is never actually optimizing for the desired dual-task state. Sleep, on the other hand, can support not

only replays of the old task, but also support replays which are a mixture of both tasks [41,52,53]. Thus, through unsupervised plasticity during sleep replay, the system is able to perform approximate optimization for the desired dual-task (or multi-task) state.

Our results are in line with a large body of literature suggesting that interleaved training is capable of mitigating catastrophic forgetting in ANNs [4,10,11] and SNNs [12,13], which led to a number of replay-like algorithms involving storing a subset of previous veridical inputs and mixing them with more recent inputs to update the networks (reviewed in [9]). The novel contribution from our study is that the data intensive process of storing old data and using them for retraining can be avoided in SNN by implementing periods of noise-induced spontaneous reactivation during new task training; similar to how brains undergo offline consolidation periods during sleep resulting in reduced retroactive interference to previously learned tasks [16,50]. Indeed, we recently successfully implemented a similar approach in feedforward ANNs, where sleep-like phase prevented catastrophic forgetting and improved generalization and adversarial robustness [54–56]. And our results are in line with previous work done in humans showing that perceptual learning tasks are subject to retroactive interference by competing memories without an intervening period of REM sleep [21,46]. Moreover, performance on visual discrimination tasks in particular have been shown to steadily improve over successive nights of sleep [46], consistent with our findings that interleaving multiple periods of sleep with novel task learning leads to optimal performance on each task.

In comparing our modeling results to those found in the literature on biological learning, it is important to note an important difference in the "baseline" state of an animal undergoing an experimental training condition versus a neural network model. In our model, and indeed in all neural network models, the system begins as a "blank slate" without knowledge of any previous learning or competing demands. In contrast, animals under experimental training paradigms have a wealth of experiences which would serve as priors to bias the subsequent learning during training, leading potentially to proactive interference. Moreover, training is typically conducted across multiple days, with intervening periods during which the animal will be subject to an array of various task-irrelevant stimuli and organismal demands possibly leading to retroactive interference. Both of these ensure that the baseline state of the animal entering a given training session is far from that of the "blank slate" a neural network model enters with, as well as that recently learned memories may start degrading quickly in the brain while the network weights remain unchanged post training (unless new task is explicitly trained). Due to this stark differences, we focus our attention on the interference phenomena which follow training on an initial task as opposed to initial learning. Viewed from this perspective, initial task training in our network can serve a similar role to the prior personal history of an animal subject.

While our model represents a dramatic simplification of a living system, we believe that it captures some important elements of how animal and human brains interact with the external world. The primary visual system is believed to employ a sequence of processing steps when visual information is increasingly represented by neurons encoding higher level features [35–37]. In insects, complex patterns of olfactory receptors activation by odors are encoded by sparse patterns of the mushroom body Kenyon cells firing [57–59]. This processing step is also similar to the function performed by convolutional layers of an ANN [34] and it was reduced to very simple convolution from the input to hidden layer in our model. Subsequently, in the vertebrate brain, associative areas and motor cortex are trained to make decisions based on reward signals released by neuro modulatory centers [10,60–62]. In insects, Kenyon cells make plastic (subject to rewarded STDP) projections to the lobes [27,63]. This was reduced in our model to synaptic projections from the hidden to output (decision making) layer implementing rewarded STDP to learn a task [30–32]. While NREM sleep in vertebrates is characterized

by complex patterns of synchronized neuronal activity [16], REM sleep is characterized by low-synchronized firing [42], similar to activity during sleep-like phase in our model and paradoxical sleep with similar properties has been reported in honeybee and fruit fly [64–66].

Our study predicts synaptic level mechanisms of how sleep-based memory reactivation can protect old memory traces during training of a new interfering memory task. It suggests that, at least for procedural memories that are directly encoded to the cortical network connectivity during new training, multiple episodes of training interleaved with periods of sleep provide necessary mechanisms to prevent forgetting old memories. Interleaving new task training with sleep enables the connectivity matrix to evolve towards the joint synaptic weight configuration, representing the intersection of manifolds supporting individual tasks. Sleep makes this possible by replaying old memory traces without explicit usage of the old training data.

Methods

Environment

Foraging behavior took place in a virtual environment consisting of a 50x50 grid with randomly distributed "food" particles. Each particle was two pixels in length and could be classified into one of four types depending on its orientation: vertical, horizontal, positively sloped diagonal, or negatively sloped diagonal. During the initial unsupervised training period, the particles are distributed at random with the constraints that each of the four types are equally represented and no two particles can be directly adjacent. During training and testing periods only the task-relevant particles were present. When a particle was acquired as a result of the virtual agent moving, it was removed from its current location (simulating consumption) and randomly assigned to a new location on the grid, again with the constraint that it not be directly adjacent to another particle. This ensures a continuously changing environment with a constant particle density. The density of particles in the environment was set to 10%. The virtual agent can see a 7x7 grid of squares (the "visual field") centered on its current location and it could move to any adjacent square, including diagonally, for a total of eight directions.

Network structure

The network was composed of 842 spiking reduced (map-based) model neurons (see *Methods*: *Map-based neuron model* below) [67,68], arranged into three feed-forward layers to mimic a basic biological circuit: a 7x7 input layer (I), a 28x28 hidden layer (H), and a 3x3 output layer (O) with a nonfunctional center neuron (Fig 1). Input to the network was simulated as a set of suprathreshold inputs to the neurons in layer I, equivalent to the lower levels of the visual system, which represent the position of particles in an egocentric reference frame relative to the virtual agent (positioned in the center of the 7x7 visual field). The most active neuron in layer O, playing the role of biological motor cortex, determined the direction of the subsequent movement. Each neuron in layer H, which can be loosely defined as higher levels of the visual system or associative cortex, received excitatory synapses from 9 randomly selected neurons inlayer I. These connections initially had random strengths drawn from a normal distribution. Each neuron in layer H connected to every neuron in layer O with both an excitatory (*Wij*) and an inhibitory (*WIij*) synapse. This provided an all-to-all connectivity pattern between these two layers and accomplished a balanced feed-forward inhibition [69] found in many biological structures [69–74]. Initially, all these connections had uniform strengths and the responses in layer O were due to the random synaptic variability. Random variability was a property of all synaptic interactions between neurons and was implemented as variability in the magnitude of the individual synaptic events.

Policy

Simulation time was divided up into epochs of 600 timesteps, each roughly equivalent to 300 ms. At the start of each epoch the virtual agent received input corresponding to locations of nearby particles within the 7x7 "visual field". Thus 48 of the 49 neurons in layer I received input from a unique location relative to the virtual agent. At the end of the epoch the virtual agent made a single move based on the activity in layer O. If the virtual agent moved to a grid location with a "food" particle present, the particle was removed and assigned to a randomly selected new location.

Each epoch was of sufficient duration for the network to receive inputs, propagate activity forward, produce outputs, and return to a resting state. Neurons in layer I which represent locations in the visual field containing particles received a brief pulse of excitatory stimulation sufficient to trigger a spike; this stimulation was applied at the start of each movement cycle (epoch). At the end of each epoch the virtual agent moved according to the activity which has occurred in layer O. Each simulation consisted of millions of these movement cycles / epochs, therefore a unit of time was introduced termed aeon (1 aeon = 100 epochs) for concise reporting.

The activity in layer O controlled the direction of the virtual agent's movement. Each of the neurons in layer O mapped onto a specific direction (i.e. one of the eight adjacent locations or the current location). The neuron in layer O which spiked the greatest number of times during the first half of the epoch defined the direction of movement for that epoch. If there was a tie, the direction was chosen at random from the set of tied directions. If no neurons in layer O spiked, the virtual agent continued in the direction it had moved during the previous epoch.

There was a 1% chance on every move that the virtual agent would ignore the activity inlayer O and instead move in a random direction. Moreover, for every movement cycle that passed without the virtual agent acquiring a particle, this probability was increased by 1%. The random variability promoted exploration vs exploitation dynamics and essentially prevented the virtual agent from getting stuck in movement patterns corresponding to infinite loops. While biological systems could utilize various different mechanisms to achieve the same goal, the method we implemented was efficient and effective for the scope of our study.

Neuron models

For all neurons we used spiking model identical to the model used in in [14,15] that can be described by the following set of difference equations [68,75,76]:

$$V_{n+1} = f_\alpha(V_n, I_n + \beta_n),$$

$$I_{n+1} = I_n - \mu(V_n + 1) + \mu\sigma + \mu\sigma_n,$$

where Vn is the membrane potential, In is a slow dynamical variable describing the effects of slow conductances, and n is a discrete time-step (0.5 ms). Slow temporal evolution of In was achieved by using small values of the parameter $\mu << 1$. Input variables β_n and σ_n were used to incorporate external current $I^{ext}n$ (e.g. background synaptic input): $\beta_n = \beta^e I^{ext}{}_n$, $\sigma_n = \sigma^e I^{ext}{}_n$. Parameter values were set to $\sigma = 0.06$, $\beta^e = 0.133$, $\sigma^e = 1$, and $\mu = 0.0005$. The nonlinearity $f\alpha$ (V_n, I_n) was defined in the form of the piece-wise continuous function:

$$f_\alpha(V_n, I_n) = \begin{cases} \alpha(1 - V_n)^{-1} + I_n, & V_n \leq 0 \\ \alpha + I_n, & 0 < V_n < \alpha + I_n \& V_{n-1} \leq 0 \\ -1 & \alpha + I_n \leq V_n \text{ or } V_{n-1} > 0, \end{cases}$$

where $\alpha = 3.65$. This model is very computationally efficient, and, despite its intrinsic low dimensionality, produces a rich repertoire of dynamics capable of mimicking the dynamics of Hodgkin-Huxley type neurons both at the single neuron level and in the context of network dynamics [68,75,77].

To model the synaptic interactions, we used the following piece-wise difference equation:

$$g_{n+1}^{syn} = \gamma g_n^{syn} + \begin{cases} (1 - R + 2XR)g_{syn}/W_j, & spike_{pre} \\ 0, & \text{otherwise,} \end{cases}$$

$$I_n^{syn} = -g_n^{syn}(V_n^{post} - V_{rp}).$$

Here *gsyn* is the strength of the synaptic coupling, modulated by the target rate *Wj* of receiving neuron *j*. Indices *pre* and *post* stand for the pre- and post-synaptic variables, respectively. The first condition, *spikepre*, is satisfied when the pre-synaptic spikes are generated. Parameter γ controls the relaxation rate of synaptic current after a presynaptic spike is received ($0 \leq \gamma < 1$). The parameter *R* is the coefficient of variability in synaptic release. The standard value of *R* is 0.12. *X* is a random variable sampled from a uniform distribution with range [0, 1]. Parameter *Vrp* defines the reversal potential and, therefore, the type of synapse (i.e. excitatory or inhibitory). The term (1-*R*+2*XR*) introduces a variability in synaptic release such that the effect of any synaptic interaction has an amplitude that is pulled from a uniform distribution with range [1-*R*,1+*R*] multiplied by the average value of the synapse.

Synaptic plasticity

Synaptic plasticity closely followed the rules introduced in [14,15]. A rewarded STDP rule [30–33] was operated on synapses between layers H and O while a standard STDP rule operated on synapses between layers I and H. A spike in a post-synaptic neuron that directly followed a spike in a pre-synaptic neuron created a *pre before post* event while the converse created a *post before pre* event. Each new post-synaptic (pre-synaptic) spike was compared to all pre-synaptic (post-synaptic) spikes with a time window of 120 iterations.

The value of an STDP event (trace) was calculated using the following equation [28,29]:

$$p = \frac{-|t_r - t_p|}{T_c},$$

$$tr_k = Ke^p$$

where t_r and t_p are the times at which the pre- and post-synaptic spike events occurred respectively, *Tc* is the time constant and is set to 40 ms, and *K* is maximum value of the trace tr_k and is set to -0.04 for a *post before pre* event and 0.04 for a *pre before post* event.

A trace was immediately applied to synapse between neurons in layers I and H. However, for synapses between neurons in layers H and O the traces were stored for 6 epochs after its creation before being erased. During storage, a trace had an effect whenever there was a rewarding or punishing event. In such a case, the synaptic weights are updated as follows:

$$W_{ij} \leftarrow W_{ij} \prod_k^{traces} \left(1 + \frac{W_{i0}}{W_i} * \Delta_k\right),$$

$$\Delta_k = S_{rp}\left(\frac{tr_k}{t - t_k + c}\right)\frac{Sum_{tr}}{Avg_{tr}},$$

$$Sum_{tr} = \sum_{k}^{traces} \frac{tr_k}{t - t_k + c},$$

$$Avg_{tr} \leftarrow (1 - \delta) Avg_{tr} + \delta Sum_{tr},$$

where t is the current timestep, S_{rp} is a scaling factor for reward/punishment, trk is the magnitude of the trace, tk is the time of the trace event, c is a constant (= 1 epoch) used for decreasing sensitivity to very recent spikes, $W_i = \Sigma_j W_{ij}$ is the total synaptic strength of all connections from the neuron i in layer H to all neurons in layer O, W_{i0} is a constant that is set to the initial value(*target value*) of Wi at the beginning of the simulation. The term W_{i0}/W_i helped to keep the output weight sum close to the initial target value. The effect of these rules was that neurons with lower total output strength could increase their output strength more easily.

The network was rewarded when the virtual agent moved to a location which contained a particle from a "food" pattern (horizontal in Task 1, vertical in Task 2) and S_{rp} = 1, and received a punishment of S_{rp} = -0.001 when it moved to a location with a particle from a neutral pattern(negative/positive diagonal in Task 1/2). A small punishment of S_{rp} = -0.0001 was applied if the agent moved to a location without a particle present to help the virtual agent learn to acquire "food" as rapidly as possible. During periods of sleep the network received a constant reward of S_{rp} = 0.5 on each movement cycle.

To ensure that neurons in layer O maintained a relatively constant long-term firing rate, the model incorporated homeostatic synaptic scaling which was applied every epoch. Each timestep, the total strength of synaptic inputs $W_j = \Sigma_i W_{ij}$ to a given neuron in layer O was set equal to the target synaptic input W_{j0} –a slow variable which varied over many epochs depending on the activity of the given neuron in layer O–which was updated according to:

$$W_{j0} \leftarrow \begin{cases} W_{j0}(1 + D_{tar}) \text{ spike rate} < \text{target rate} \\ W_{j0}(1 - D_{tar}) \text{ spike rate} > \text{target rate} \end{cases}$$

To ensure that the net synaptic input W_j to any neuron was unaffected by plasticity events at the individual synapses at distinct timesteps and equal to W_{j0}, we implemented a scaling process akin to heterosynaptic plasticity which occurs after each STDP event. When any excitatory synapse of neuron in layer O changed in strength, all other excitatory synapses received by that neuron were updated according to:

$$W_{ij} \leftarrow W_{ij} \frac{W_{j0}}{\sum_i W_{ij}}$$

Additionally, all inhibitory synapses were modified via a similar heterosynaptic update rule following each STDP event where the strength of every outgoing inhibitory weight from a given neuron was set to the negative mean of all outgoing excitatory synapses of that same neuron. More rigorously:

$$WI_{ij} \leftarrow -\frac{1}{|j|} \sum_j W_{ij}$$

Simulated sleep

To simulate the sleep phase, we inactive the sensory receptors (i.e. the input layer of network), cut off all sensory signals (i.e. remove all particles from the environment), and decouple output

layer activity from motor control (i.e. the output layer can spike but no longer causes the agent to move). We also change the learning rule between the hidden and output layer from rewarded to unsupervised STDP (see *Methods*: *Synaptic Plasticity* for details) as there is no way to evaluate decision-making without sensory input or motor output.

To simulate the spontaneous activity observed during REM sleep, we provided noise to each neuron in the hidden layer in a way which ensured that the spiking statistics of each neuron was conserved across awake and sleep phases. To determine these spiking rates, we recorded average spiking rates of neurons in the hidden layer H during preceding training of both Task 1 and Task 2; these task specific spiking rates were then averaged to generate target spiking rates for hidden layer neurons. $\text{Interleaved}_{\text{S,T1}}$ training consisted of alternating intervals of this sleep phase and training on Task 1, with each interval lasting 100 movement cycles (although no movement occurred).

Support vector machine training

A support vector machine with a radial basis function kernel was trained to classify synaptic weight configurations as being related to Task 1 or Task2. Labeled training data were obtained by taking the excitatory synaptic weight matrices between the hidden and output layers from the last fifth of the Task 1 and Task 2 training phases (i.e. after performance had appeared to asymptote). These synaptic weight matrices were then flattened into column vectors, and the column vectors were concatenated to form a training data matrix of size *number of features* x *number of samples*. The number of features was equal to the total number of excitatory synapses between the hidden and output layer– 6272 dimensions. We then used this support vector machine to classify held out synaptic weight configurations from Task 1 and Task 2 training, as well as ones which resulted from $\text{Interleaved}_{\text{T1,T2}}$ and $\text{Interleaved}_{\text{S,T1}}$ training.

2-D synaptic weight distributions (Fig 6)

First for each synapse we found how its synaptic strength changes between two slices in time, where the given synapse's strength at time slice 1 is the point's X-value and strength at time slice 2 is its Y-value. Then we binned this space and counted synapses in each bin to make two dimensional histograms where blue color corresponds to a single synapse found in a bin and brown corresponds to the max of 50 synapses. These two-dimensional histograms assist in visualizing the movement of all synapses between the two slices in time that are specified by the timelines at the top of each plot. Conceptually, it is important to note that if a synapse does not change in strength between time slice 1 and time slice 2, then point the synapse corresponds to in this space will lie on the diagonal of the plot since the X-value will match the Y-value. If a great change in the synapse's strength has occurred between time slice 1 and time slice 2, then the synapse's corresponding point will lie far from the diagonal since the X-value will be distant from the Y-value. The points on the X-(Y-) axis represent synapses that lost (gained) all synaptic strength between time slice 1 and time slice 2.

Distance from solution manifolds (Fig 7)

Each of the two solution manifolds (i.e. Task 1 and Task 2 specific manifolds) were defined by the point-sets in synaptic weight space which were capable of supporting robust performance on that particular task, namely the sets M_{T1} and M_{T2}. This included the synaptic weight states from the last fifth of training on a particular task(i.e. after performance on that task appeared to asymptote) and all of the synaptic weight states from the last fifth of both $\text{Interleaved}_{\text{T1,T2}}$ and $\text{Interleaved}_{\text{S,T1/T2}}$ training. The intersection of the two solution manifolds (i.e. the point-set $M_{T1 \cap T2}$) was defined solely by the synaptic weight states from the last fifth of both

$\text{Interleaved}_{\text{T1,T2}}$ and $\text{Interleaved}_{\text{S,T1}}$ training. As the network evolved along its trajectory in synaptic weight space, the distance from the current point in synaptic weight space, *pt*, to the two solution manifolds and their intersection were computed as follows:

$$d^n(p_t, M_\tau) = \min_{x \in M\tau}(d^n(p_t, x)).$$

Here, d^n is the n-dimensional Euclidean-distance function, where n is the dimensionality of synaptic weight space (i.e. $n = 6272$ here), M_τ is the point-set specific to the manifold or intersection in question (i.e. either M_{T1}, M_{T2}, or $M_{T1 \cap T2}$), and x is a particular element of the point-set M_τ.

Particle responsiveness metric (PRM)

The particle responsiveness metric (PRM) developed to quantify how responsive the network's weight matrix is to specific food particle orientations thereby allowing the quality of the receptive field for a given task to be determined was defined as follows:

$$\text{PRM}(\text{Particle Type}) = \sum_{\forall O \in Output} \text{grand}(\text{DirectionMask}(\text{O}) \odot \sum_{\forall H \in Hidden} W_{H \to O} * \sum_{\forall \text{P} \in \text{Particle}Masks} (W_H \odot P) * \text{grand}(W_H \odot P)^2)$$

Here *Output* is the set of all output layer neurons, *O; Hidden* is the set of all hidden layer neurons, *H; ParticleMasks* is the set of masks, *P*, representing all possible locations of a single instance of a *ParticleType* in the input field (e.g., horizontal bars would be a set of masks with a single horizontal bar placed in all possible locations in the visual field; each particle mask *P* consists of a *7 x 7* matrix of zeros with ones being placed in locations that correspond to current food pixels). W_H is a *7x7* synaptic weights matrix of a given hidden layer neuron *H;* $\odot$ gives Hadamard (or element-wise) product of two matrixes, *grand*(*A*) is a grand sum of all the elements of a matrix A (*grand*(*A*) = $e^T Ae$, where *e* is all-ones vector*). DirectionMask*(*O*) takes in an output layer neuron, *O*, and returns a matrix that represents the direction of motion with respect to the input field. For example, when the neuron that directs the critter to move up and to the left is supplied as input, the function returns a *7 x 7* matrix of zeros with the top left *3 x 3* submatrix being ones. $W_{H \to O}$ simply returns the synapse strength from the source *(H)* to destination *(O)* neuron.

Although this is seemingly an intricate metric, it captures many desired features of the network's connectivity and responses to food particles present in the visual field. Conceptually, this metric is similar to the method used for developing the receptive fields of output layer neurons with respect to the input field (Figs 2 and 5). PRM builds upon this qualitative visualization, allowing us to numerically assess how specific particles influence output layer neurons to spike when present in the portion of the visual field that corresponds to the direction of motion for that neuron. The intuitions of the metric are as follows: $W_H \odot P$ develops a notion of how well the current hidden neuron's (*H*) connections to the input layer overlaps with the current food particle (*P*) placed at specific location. The resulting matrix is then multiplied by $\text{grand}(W_H \odot P)^2$, which emphasizes contribution of the *H* neurons receiving input from adjacent pixels in correct orientation (i.e., sensitive to the food particles) vs those receiving input from random pixels. Indeed, when a hidden layer neuron *H* overlaps strongly with a food particle *P*, the chances of spiking are significantly increased, thus this nonlinear term captures the high impact overlapping receptive fields and food particles has on output layer activity. $W_{H \to O}$ captures how strongly the current output layer neuron *O* is listening to the current hidden layer neuron *H*.

These described pieces are multiplied together to form a weighted input receptive field of the output layer neuron with respect to a specific hidden layer neuron and food particle type / location. The sum of these terms for all hidden layer neurons and food particle locations is taken for a single output layer neuron, achieving a global view of all hidden layer neurons and food particle types / locations influencing the current output layer neuron. The *grand*(*A*) operation between the *DirectionMask*(*O*) and the previously described summed term is then taken to see how much the summed weighted receptive fields overlap with the corresponding direction of movement for output neuron *O*. This process is repeated for all output layer neurons to get a global quantification of how the current food particle influences activity in the direction of motion for all output layer neurons. When this metric is calculated for a given network state across food particle types we can observe what food particles impact output layer activity and drive the critter to move, highlighting what particle orientations the network is attracted to.

Supporting information

S1 Fig. Spike rasters showing network activity across various training regimes. (A-D) Representative spike rasters from various training regimes. The vertical axis specifies a unique neuron in the network while time in epochs is shown horizontally. Here a single dot represents a specific neuron spiking at a given time while the color of the dot dictates what layer that neuron belongs to (green, blue, red corresponding to input, hidden, and output layers respectively). Panels A, B, C, D correspond to sample activity from Task 1 training, Task 2 training, $I_{T1,T2}$ training and $I_{S,T1}$ training respectively. Note, in panel D activity is taken during a period of sleep when the hidden layer is spontaneously activated. Thus, there are hidden (blue) and output (red) layer spikes while the input (green) layer is completely silent.
(EPS)

S2 Fig. Model displays graceful degradation in performance as a result of hidden layer dropout. (A) Mean performance (red line) and standard deviation (blue lines) over time: unsupervised training (white), Task 1 training (blue), Task 1 testing (green). Hidden layer neurons are randomly removed during testing period. Gradient bar above Task 1 testing (green) displays the number of hidden layer neurons over time starting at 784 and decreasing down to 0. The testing performance remains high until ~25% of neurons are left, after which it starts to drop. This highlights the formation of a distributed synaptic structure between hidden and output layer neurons developed during training, ensuring output layer activity is not dictated by a select few hidden layer neurons. **(B)** Same as in (A) but for Task 2.
(EPS)

S3 Fig. Particle responsiveness metric (PRM) shows correspondence between type of training and particles preferred by the network. (A-D) Mean and standard deviation (blue bars and black lines respectively) of the PRM for various types of training and particle orientations across ten trials. The title of each plot reflects the most recently trained stage, the vertical axis corresponds to the value of the PRM while the horizontal axis identifies the particle type (bold labels indicate ideal particles the network would be attracted to following the corresponding training). It can be seen that the metric indicates the network is most responsive to the corresponding ideal particle types following a specific training regime e.g. Post Task 1 the network is most responsive to horizontal particles (A), Post Task 2 the network is most responsive to vertical particles (B), Post $I_{S,T1}$ the network is most responsive to horizontal and vertical particles (C), Post $I_{T1,T2}$ the network is most responsive to horizontal and vertical particles (D).
(EPS)

S4 Fig. Effect of sleep to protect old memory does not depend on specific properties of noise applied during sleep phase. **(A)** Mean performance (red line) and standard deviation (blue lines) over time: unsupervised training (white), Interleaved$_{S,T1}$ (grey), Task 1/2 testing (green/yellow). **(B)** Mean and standard deviation of performance during testing on Task 1 (blue) and Task 2 (red). Following Interleaved$_{S,T1}$, mean performance on Task 1 was 0.60 ± 0.03 while Task 2 was 0.49 ± 0.05. (In all experiments, 0.5 represents chance performance.) Note that periods of Task 1 training interleaved with sleep do not lead to increase in performance on untrained Task 2, even when Task 2 data from another experiment were used to set up mean firing rates of the random input during sleep. **(C)** Same as in (A) but the sequence of training was: unsupervised training (white), Task 1 training (blue), Task 1/2 testing (green/yellow), Interleaved$_{S,T1}$ (grey), Task 1/2 testing (green/yellow). **(D)** Mean and standard deviation of performance during testing on Task 1 (blue) and Task 2 (red) after Task 1 training and after Interleaved$_{S,T1}$. Following Task 1 training, mean performance on Task 1 was 0.70 ± 0.02 while Task 2 was 0.53 ± 0.02. Post Interleaved$_{S,T1}$ training, mean performance on Task 1 was 0.71 ± 0.02 and Task 2 was 0.51 ± 0.02. Task 1 performance remained high after Interleaved$_{S,T1}$ but no improvement on Task 2 was observed. **(E)** Mean performance (red line) and standard deviation (blue lines) over time: unsupervised training (white), Task 1 training (blue), Task 1/2 testing (green/yellow), Interleaved$_{US,T2}$ (burnt orange), Task 1/2 testing (green/yellow). **(F)** Mean and standard deviation of performance during testing on Task 1 (blue) and Task 2 (red). Following Task 1 training, mean performance on Task 1 was 0.70 ± 0.02 while Task 2 was 0.53 ± 0.02. Post Interleaved$_{US,T2}$ training, mean performance on Task 1 was 0.67 ± 0.05 and Task 2 was 0.69 ± 0.03.
(EPS)

S5 Fig. Interleaving old and new task training allows integrating synaptic information relevant to new task while preserving old task information. **(A)** Mean performance (red line) and standard deviation (blue lines) over time: unsupervised training (white), Task 1 training (blue), Task 1/2 testing (green/yellow), Task 2 training (red), Task 1/2 testing (green/yellow), Interleaved$_{T1,T2}$ training (purple), Task 1/2 testing (green/yellow). **(B)** Mean and standard deviation of performance during testing on Task 1 (blue) and Task 2 (red). Following Task 1training, mean performance on Task 1 was 0.69 ± 0.02 while Task 2 was 0.53 ± 0.02. Conversely, following Task 2 training, mean performance on Task 1 was 0.52 ± 0.02 while Task2 was 0.69 ± 0.04. Following Interleaved$_{T1,T2}$ training, mean performance on Task 1 was 0.65 ±0.03 while Task 2 was 0.67 ± 0.04. **(C)** Distributions of task-relevant synaptic weights (blue bars–single trial, orange line / shaded region–mean / std across 10 trails. The distributional structure of Task 1-relevant synapses following Task 1 training (top-left) is destroyed following Task 2 training (top-middle), but partially recovered following. Interleaved$_{T1,T2}$ training (top-right). Similarly, the distributional structure of Task 2-relevantsynapses following Task 2 training (bottom-middle), which was not present following Task 1training (bottom-left), was partially preserved following Interleaved$_{T1,T2}$ training (bottom-right).**(D)** Box plots with mean (dashed green line) and median (dashed orange line) of the distance to the decision boundary found by an SVM trained to classify Task 1 and Task 2 synaptic weight matrices for Task 1, Task 2, and Interleaved$_{T1,T2}$ training across trials. Task 1 and Task 2synaptic weight matrices had mean classification values of -0.069 and 0.069 respectively, while that of Interleaved$_{T1,T2}$ training was 0.016. **(E)** Trajectory of H to O layer synaptic weights through PC space. Synaptic weights which evolved during Interleaved$_{T1,T2}$ training (green dots)clustered in a location of PC space intermediary between the clusters of synaptic weights which evolved during training on Task 1 (red dots) and Task 2 (blue dots).
(EPS)

S6 Fig. Freezing a fraction of task specific strong synapses preserves differing degrees of performance in a sequential learning paradigm. (A-C) Mean and standard deviation of performance during testing on Task 1 (blue) and Task 2 (red). Left, Performance after Task 1 training. Right, Performance after Task 2 training when a fraction of the strongest (after Task 1 training) synapses remained frozen– 1% (A), 5% (B), 10% (C). In all cases, after Task 1 training, Task 1 performance was 0.70 ± 0.02 and Task 2 performance was 0.53 ± 0.02. (A) Freezing the top 1% of Task 1 synapses resulted in a Task 1 performance of 0.54 ± 0.02 and Task 2 performance of 0.68 ± 0.03. (B) Freezing the top 5% of Task 1 synapses resulted in a Task 1 performance of 0.65 ± 0.02 and Task 2 performance of 0.61 ± 0.01. (C) Freezing the top 10% of Task 1 synapses resulted in a Task 1 performance of 0.70 ± 0.03 and Task 2 performance of 0.53 ± 0.03. Freezing the top 1% of Task 1 synapses was not sufficient to maintain Task 1 performance, thus enabling Task 2 relevant synapses to dominate the network; however, freezing the top 10% of Task 1 synapses fully retains Task 1 performance preventing Task 2 to be learned.
(EPS)

Author Contributions

Conceptualization: Ryan Golden, Pavel Sanda, Maxim Bazhenov.

Data curation: Jean Erik Delanois.

Formal analysis: Ryan Golden, Jean Erik Delanois, Maxim Bazhenov.

Funding acquisition: Maxim Bazhenov.

Investigation: Jean Erik Delanois.

Methodology: Ryan Golden, Jean Erik Delanois, Pavel Sanda, Maxim Bazhenov.

Project administration: Maxim Bazhenov.

Resources: Maxim Bazhenov.

Software: Jean Erik Delanois.

Supervision: Pavel Sanda, Maxim Bazhenov.

Visualization: Jean Erik Delanois.

Writing – original draft: Ryan Golden, Jean Erik Delanois, Pavel Sanda, Maxim Bazhenov.

Writing – review & editing: Ryan Golden, Jean Erik Delanois, Pavel Sanda, Maxim Bazhenov.

References

1. French RM. Catastrophic forgetting in connectionist networks. Trends Cogn Sci. 1999; 3(4):128–35. https://doi.org/10.1016/s1364-6613(99)01294-2 PMID: 10322466
2. Mccloskey M, Cohen NJ. CATASTROPHIC INTERFERENCE IN CONNECTIONIST NETWORKS: THE SEQUENTIAL LEARNING PROBLEM. The Psychology of Learning and Motivation. 1989; 24:109–65.
3. Ratcliff R. Connectionist models of recognition memory: constraints imposed by learning and forgetting functions. Psychol Rev. 1990; 97(2):285–308. https://doi.org/10.1037/0033-295x.97.2.285 PMID: 2186426
4. Hasselmo ME. Avoiding Catastrophic Forgetting. Trends Cogn Sci. 2017; 21(6):407–8. https://doi.org/10.1016/j.tics.2017.04.001 PMID: 28442279
5. Hassabis D, Kumaran D, Summerfield C, Botvinick M. Neuroscience-Inspired Artificial Intelligence. Neuron. 2017; 95(2):245–58. https://doi.org/10.1016/j.neuron.2017.06.011 PMID: 28728020

6. Kemker R, Abitino A, McClure M, Kanan C. Measuring Catastrophic Forgetting in Neural Networks. arXiv:170802072 [Internet]. 2017.
7. Kirkpatrick J, Pascanu R, Rabinowitz N, Veness J, Desjardins G, Rusu AA, et al. Overcoming catastrophic forgetting in neural networks. Proceedings of the national academy of sciences. 2017; 114 (13):3521–6. https://doi.org/10.1073/pnas.1611835114 PMID: 28292907
8. Kemker R, Kanan C. Fearnet: Brain-inspired model for incremental learning. arXiv:171110563. 2017.
9. Hayes TL, Krishnan GP, Bazhenov M, Siegelmann HT, Sejnowski TJ, Kanan C. Replay in Deep Learning: Current Approaches and Missing Biological Elements. Neural computation. 2021; 33(11):2908–50. https://doi.org/10.1162/neco_a_01433 PMID: 34474476
10. Flesch T, Balaguer J, Dekker R, Nili H, Summerfield C. Comparing continual task learning in minds and machines. Proc Natl Acad Sci U S A. 2018; 115(44):E10313–E22. https://doi.org/10.1073/pnas.1800755115 PMID: 30322916
11. McClelland JL, McNaughton BL, O'Reilly RC. Why there are complementary learning systems in the hippocampus and neocortex: insights from the successes and failures of connectionist models of learning and memory. Psychol Rev. 1995; 102(3):419–57. https://doi.org/10.1037/0033-295X.102.3.419 PMID: 7624455
12. Evans BD, Stringer SM. Transformation-invariant visual representations in self-organizing spiking neural networks. Front Comput Neurosci. 2012; 6:46. https://doi.org/10.3389/fncom.2012.00046 PMID: 22848199
13. Higgins I, Stringer S, Schnupp J. Unsupervised learning of temporal features for word categorization in a spiking neural network model of the auditory brain. PLoS One. 2017; 12(8):e0180174. https://doi.org/10.1371/journal.pone.0180174 PMID: 28797034
14. Sanda P, Skorheim S, Bazhenov M. Multi-layer network utilizing rewarded spike time dependent plasticity to learn a foraging task. PLoS Comput Biol. 2017; 13(9):e1005705. https://doi.org/10.1371/journal.pcbi.1005705 PMID: 28961245
15. Skorheim S, Lonjers P, Bazhenov M. A spiking network model of decision making employing rewarded STDP. PLoS One. 2014; 9(3):e90821. https://doi.org/10.1371/journal.pone.0090821 PMID: 24632858
16. Rasch B, Born J. About sleep's role in memory. Physiological reviews. 2013; 93(2):681–766. https://doi.org/10.1152/physrev.00032.2012 PMID: 23589831
17. Ji D, Wilson MA. Coordinated memory replay in the visual cortex and hippocampus during sleep. Nat Neurosci. 2007; 10(1):100–7. https://doi.org/10.1038/nn1825 PMID: 17173043
18. Euston DR, Tatsuno M, McNaughton BL. Fast-forward playback of recent memory sequences in prefrontal cortex during sleep. Science. 2007; 318(5853):1147–50. https://doi.org/10.1126/science.1148979 PMID: 18006749
19. Peyrache A, Khamassi M, Benchenane K, Wiener SI, Battaglia FP. Replay of rule-learning related neural patterns in the prefrontal cortex during sleep. Nat Neurosci. 2009; 12(7):919–26. https://doi.org/10.1038/nn.2337 PMID: 19483687
20. Barnes DC, Wilson DA. Slow-wave sleep-imposed replay modulates both strength and precision of memory. J Neurosci. 2014; 34(15):5134–42. https://doi.org/10.1523/JNEUROSCI.5274-13.2014 PMID: 24719093
21. Mednick SC, Cai DJ, Shuman T, Anagnostaras S, Wixted JT. An opportunistic theory of cellular and systems consolidation. Trends Neurosci. 2011; 34(10):504–14. https://doi.org/10.1016/j.tins.2011.06.003 PMID: 21742389
22. Stickgold R. Parsing the role of sleep in memory processing. Curr Opin Neurobiol. 2013; 23(5):847–53. https://doi.org/10.1016/j.conb.2013.04.002 PMID: 23618558
23. Ramanathan DS, Gulati T, Ganguly K. Sleep-Dependent Reactivation of Ensembles in Motor Cortex Promotes Skill Consolidation. PLOS Biology. 2015; 13(9):e1002263. https://doi.org/10.1371/journal.pbio.1002263 PMID: 26382320
24. Zwaka H, Bartels R, Gora J, Franck V, Culo A, Gotsch M, et al. Context odor presentation during sleep enhances memory in honeybees. Curr Biol. 2015; 25(21):2869–74. https://doi.org/10.1016/j.cub.2015.09.069 PMID: 26592345
25. Melnattur K, Kirszenblat L, Morgan E, Militchin V, Sakran B, English D, et al. A conserved role for sleep in supporting Spatial Learning in Drosophila. Sleep. 2021; 44(3). https://doi.org/10.1093/sleep/zsaa197 PMID: 32959053
26. Donlea JM, Thimgan MS, Suzuki Y, Gottschalk L, Shaw PJ. Inducing sleep by remote control facilitates memory consolidation in Drosophila. Science. 2011; 332(6037):1571–6. https://doi.org/10.1126/science.1202249 PMID: 21700877

27. Cassenaer S, Laurent G. Hebbian STDP in mushroom bodies facilitates the synchronous flow of olfactory information in locusts. Nature. 2007; 448(7154):709–13. https://doi.org/10.1038/nature05973 PMID: 17581587

28. Bi GQ, Poo MM. Synaptic modifications in cultured hippocampal neurons: dependence on spike timing, synaptic strength, and postsynaptic cell type. J Neurosci. 1998; 18(24):10464–72. https://doi.org/10.1523/JNEUROSCI.18-24-10464.1998 PMID: 9852584

29. Markram H, Lubke J, Frotscher M, Sakmann B. Regulation of synaptic efficacy by coincidence of postsynaptic APs and EPSPs. Science. 1997; 275(5297):213–5. https://doi.org/10.1126/science.275.5297.213 PMID: 8985014

30. Farries MA, Fairhall AL. Reinforcement learning with modulated spike timing dependent synaptic plasticity. J Neurophysiol. 2007; 98(6):3648–65. https://doi.org/10.1152/jn.00364.2007 PMID: 17928565

31. Florian RV. Reinforcement learning through modulation of spike-timing-dependent synaptic plasticity. Neural Comput. 2007; 19(6):1468–502. https://doi.org/10.1162/neco.2007.19.6.1468 PMID: 17444757

32. Izhikevich EM. Solving the distal reward problem through linkage of STDP and dopamine signaling. Cereb Cortex. 2007; 17(10):2443–52. https://doi.org/10.1093/cercor/bhl152 PMID: 17220510

33. Legenstein R, Pecevski D, Maass W. A learning theory for reward-modulated spike-timing-dependent plasticity with application to biofeedback. PLoS Comput Biol. 2008; 4(10):e1000180. https://doi.org/10.1371/journal.pcbi.1000180 PMID: 18846203

34. LeCun Y, Bengio Y, Hinton G. Deep learning. Nature. 2015; 521(7553):436–44. https://doi.org/10.1038/nature14539 PMID: 26017442

35. Cadieu CF, Hong H, Yamins DL, Pinto N, Ardila D, Solomon EA, et al. Deep neural networks rival the representation of primate IT cortex for core visual object recognition. PLoS Comput Biol. 2014; 10(12):e1003963. https://doi.org/10.1371/journal.pcbi.1003963 PMID: 25521294

36. Yamins DL, DiCarlo JJ. Using goal-driven deep learning models to understand sensory cortex. Nat Neurosci. 2016; 19(3):356–65. https://doi.org/10.1038/nn.4244 PMID: 26906502

37. Yamins DL, Hong H, Cadieu CF, Solomon EA, Seibert D, DiCarlo JJ. Performance-optimized hierarchical models predict neural responses in higher visual cortex. Proc Natl Acad Sci U S A. 2014; 111(23):8619–24. https://doi.org/10.1073/pnas.1403112111 PMID: 24812127

38. Wei Y, Krishnan G, Bazhenov M. Synaptic Mechanisms of Memory Consolidation during Sleep Slow Oscillations. Journal of Neuroscience. 2016; 36(15):4231–47. https://doi.org/10.1523/JNEUROSCI.3648-15.2016 PMID: 27076422

39. Wei Y, Krishnan GP, Komarov M, Bazhenov M. Differential roles of sleep spindles and sleep slow oscillations in memory consolidation. PLoS Comput Biol. 2018; 14(7):e1006322. https://doi.org/10.1371/journal.pcbi.1006322 PMID: 29985966

40. Wei Y, Krishnan GP, Marshall L, Martinetz T, Bazhenov M. Stimulation Augments Spike Sequence Replay and Memory Consolidation during Slow-Wave Sleep. The Journal of neuroscience: the official journal of the Society for Neuroscience. 2020; 40(4):811–24. https://doi.org/10.1523/JNEUROSCI.1427-19.2019 PMID: 31792151

41. Gonzalez OC, Sokolov Y, Krishnan GP, Delanois JE, Bazhenov M. Can sleep protect memories from catastrophic forgetting? Elife. 2020;9. https://doi.org/10.7554/eLife.51005 PMID: 32748786

42. Peever J, Fuller PM. The Biology of REM Sleep. Curr Biol. 2017; 27(22):R1237–R48. https://doi.org/10.1016/j.cub.2017.10.026 PMID: 29161567

43. Oudiette D, Antony JW, Creery JD, Paller KA. The role of memory reactivation during wakefulness and sleep in determining which memories endure. J Neurosci. 2013; 33(15):6672–8. https://doi.org/10.1523/JNEUROSCI.5497-12.2013 PMID: 23575863

44. Paller KA, Voss JL. Memory reactivation and consolidation during sleep. Learn Mem. 2004; 11(6):664–70. https://doi.org/10.1101/lm.75704 PMID: 15576883

45. Walker MP, Stickgold R. Sleep-dependent learning and memory consolidation. Neuron. 2004; 44(1):121–33. https://doi.org/10.1016/j.neuron.2004.08.031 PMID: 15450165

46. Stickgold R, James L, Hobson JA. Visual discrimination learning requires sleep after training. Nat Neurosci. 2000; 3(12):1237–8. https://doi.org/10.1038/81756 PMID: 11100141

47. Hennevin E, Hars B, Maho C, Bloch V. Processing of learned information in paradoxical sleep: relevance for memory. Behav Brain Res. 1995; 69(1–2):125–35. https://doi.org/10.1016/0166-4328(95)00013-j PMID: 7546303

48. Lewis PA, Knoblich G, Poe G. How Memory Replay in Sleep Boosts Creative Problem-Solving. Trends Cogn Sci. 2018; 22(6):491–503. https://doi.org/10.1016/j.tics.2018.03.009 PMID: 29776467

49. Oudiette D, Paller KA. Upgrading the sleeping brain with targeted memory reactivation. Trends Cogn Sci. 2013; 17(3):142–9. https://doi.org/10.1016/j.tics.2013.01.006 PMID: 23433937

50. McDevitt EA, Duggan KA, Mednick SC. REM sleep rescues learning from interference. Neurobiol Learn Mem. 2015; 122:51–62. https://doi.org/10.1016/j.nlm.2014.11.015 PMID: 25498222

51. Javed K, White M. Meta-Learning Representations for Continual Learning. arXiv e-prints [Internet]. 2019 May 01, 2019:[arXiv:1905.12588 p.]. Available from: https://ui.adsabs.harvard.edu/abs/2019arXiv190512588J.

52. Roumis DK, Frank LM. Hippocampal sharp-wave ripples in waking and sleeping states. Curr Opin Neurobiol. 2015; 35:6–12. https://doi.org/10.1016/j.conb.2015.05.001 PMID: 26011627

53. Swanson RA, Levenstein D, McClain K, Tingley D, Buzsáki G. Variable specificity of memory trace reactivation during hippocampal sharp wave ripples. Current Opinion in Behavioral Sciences. 2020; 32:126–35. https://doi.org/10.1016/j.cobeha.2020.02.008 PMID: 36034494

54. Krishnan GP, Tadros T, Ramyaa R, Bazhenov M. Biologically inspired sleep algorithm for artificial neural networks. arXiv. 2019:1908.02240v1.

55. Tadros T, Krishnan G, Ramyaa R, Bazhenov M. Biologically inspired sleep algorithm for reducing catastrophic forgetting in neural networks. AAAI Conference on Artificial Intelligence 2020. p. 13933–4.

56. Tadros T, Krishnan GP, Ramyaa R, Bazhenov M. Biologically inspired sleep algorithm for increased generalization and adversarial robustness in deep neural networks. International Conference on Learning Representations [Internet]. 2019.

57. Laurent G. Olfactory network dynamics and the coding of multidimensional signals. Nat Rev Neurosci. 2002; 3(11):884–95. https://doi.org/10.1038/nrn964 PMID: 12415296

58. Assisi C, Stopfer M, Laurent G, Bazhenov M. Adaptive regulation of sparseness by feedforward inhibition. Nature neuroscience. 2007; 10(9):1176–84. https://doi.org/10.1038/nn1947 PMID: 17660812

59. Perez-Orive J, Bazhenov M, Laurent G. Intrinsic and circuit properties favor coincidence detection for decoding oscillatory input. The Journal of neuroscience: the official journal of the Society for Neuroscience. 2004; 24(26):6037–47.

60. Schultz W. Dopamine reward prediction-error signalling: a two-component response. Nat Rev Neurosci. 2016; 17(3):183–95. https://doi.org/10.1038/nrn.2015.26 PMID: 26865020

61. Schultz W. Dopamine reward prediction error coding. Dialogues Clin Neurosci. 2016; 18(1):23–32. https://doi.org/10.31887/DCNS.2016.18.1/wschultz PMID: 27069377

62. Schultz W, Dayan P, Montague PR. A neural substrate of prediction and reward. Science. 1997; 275 (5306):1593–9. https://doi.org/10.1126/science.275.5306.1593 PMID: 9054347

63. Cassenaer S, Laurent G. Conditional modulation of spike-timing-dependent plasticity for olfactory learning. Nature. 2012; 482(7383):47–52. https://doi.org/10.1038/nature10776 PMID: 22278062

64. Tainton-Heap LAL, Kirszenblat LC, Notaras ET, Grabowska MJ, Jeans R, Feng K, et al. A Paradoxical Kind of Sleep in Drosophila melanogaster. Curr Biol. 2021; 31(3):578–90 e6. https://doi.org/10.1016/j.cub.2020.10.081 PMID: 33238155

65. Kaiser W, Steiner-Kaiser J. Neuronal correlates of sleep, wakefulness and arousal in a diurnal insect. Nature. 1983; 301(5902):707–9. https://doi.org/10.1038/301707a0 PMID: 6828153

66. Sauer S, Kinkelin M, Herrmann E, Kaiser W. The dynamics of sleep-like behaviour in honey bees. Journal of comparative physiology A, Neuroethology, sensory, neural, and behavioral physiology. 2003; 189 (8):599–607. https://doi.org/10.1007/s00359-003-0436-9 PMID: 12861424

67. Rulkov NF, Bazhenov M. Oscillations and synchrony in large-scale cortical network models. J Biol Phys. 2008; 34(3–4):279–99. https://doi.org/10.1007/s10867-008-9079-y PMID: 19669478

68. Rulkov NF, Timofeev I, Bazhenov M. Oscillations in large-scale cortical networks: map-based model. J Comput Neurosci. 2004; 17(2):203–23. https://doi.org/10.1023/B:JCNS.0000037683.55688.7e PMID: 15306740

69. Bazhenov M, Stopfer M. Forward and back: motifs of inhibition in olfactory processing. Neuron. 2010; 67(3):357–8. https://doi.org/10.1016/j.neuron.2010.07.023 PMID: 20696373

70. Bruno RM. Synchrony in sensation. Curr Opin Neurobiol. 2011; 21(5):701–8. https://doi.org/10.1016/j.conb.2011.06.003 PMID: 21723114

71. Dong H, Shao Z, Nerbonne JM, Burkhalter A. Differential depression of inhibitory synaptic responses in feedforward and feedback circuits between different areas of mouse visual cortex. J Comp Neurol. 2004; 475(3):361–73. https://doi.org/10.1002/cne.20164 PMID: 15221951

72. Pouille F, Scanziani M. Enforcement of temporal fidelity in pyramidal cells by somatic feed-forward inhibition. Science. 2001; 293(5532):1159–63. https://doi.org/10.1126/science.1060342 PMID: 11498596

73. Shao Z, Burkhalter A. Different balance of excitation and inhibition in forward and feedback circuits of rat visual cortex. Journal of Neuroscience. 1996; 16(22):7353–65. https://doi.org/10.1523/JNEUROSCI.16-22-07353.1996 PMID: 8929442

74. Silberberg G. Polysynaptic subcircuits in the neocortex: spatial and temporal diversity. Curr Opin Neurobiol. 2008; 18(3):332–7. https://doi.org/10.1016/j.conb.2008.08.009 PMID: 18801433

75. Bazhenov M, Rulkov NF, Fellous JM, Timofeev I. Role of network dynamics in shaping spike timing reliability. Phys Rev E Stat Nonlin Soft Matter Phys. 2005; 72(4 Pt 1):041903. https://doi.org/10.1103/PhysRevE.72.041903 PMID: 16383416

76. Rulkov NF. Modeling of spiking-bursting neural behavior using two-dimensional map. Phys Rev E Stat Nonlin Soft Matter Phys. 2002; 65(4 Pt 1):041922. https://doi.org/10.1103/PhysRevE.65.041922 PMID: 12005888

77. Komarov M, Krishnan G, Chauvette S, Rulkov N, Timofeev I, Bazhenov M. New class of reduced computationally efficient neuronal models for large-scale simulations of brain dynamics. J Comput Neurosci. 2018; 44(1):1–24. https://doi.org/10.1007/s10827-017-0663-7 PMID: 29230640

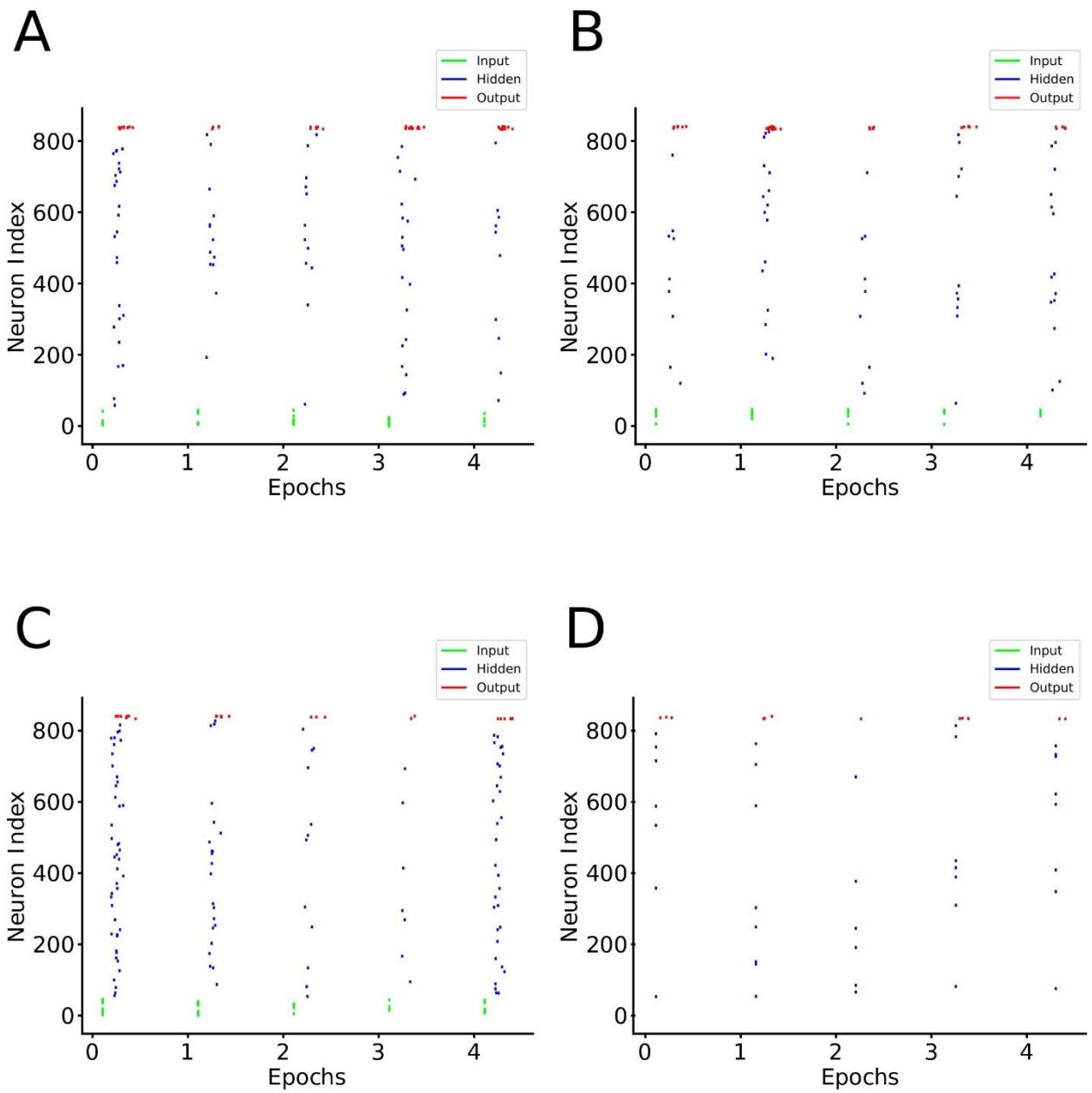


Figure 1.8. (S1 Fig) Spike rasters showing network activity across various training regimes. **(A-D)** Representative spike rasters from various training regimes. The vertical axis specifies a unique neuron in the network while time in epochs is shown horizontally. Here a single dot represents a specific neuron spiking at a given time while the color of the dot dictates what layer that neuron belongs to (green, blue, red corresponding to input, hidden, and output layers respectively). Panels A, B, C, D correspond to sample activity from Task 1 training, Task 2 training, $I_{T1,T2}$ training and $I_{S,T1}$ training respectively. Note, in panel D activity is taken during a period of sleep when the hidden layer is spontaneously activated. Thus, there are hidden (blue) and output (red) layer spikes while the input (green) layer is completely silent.

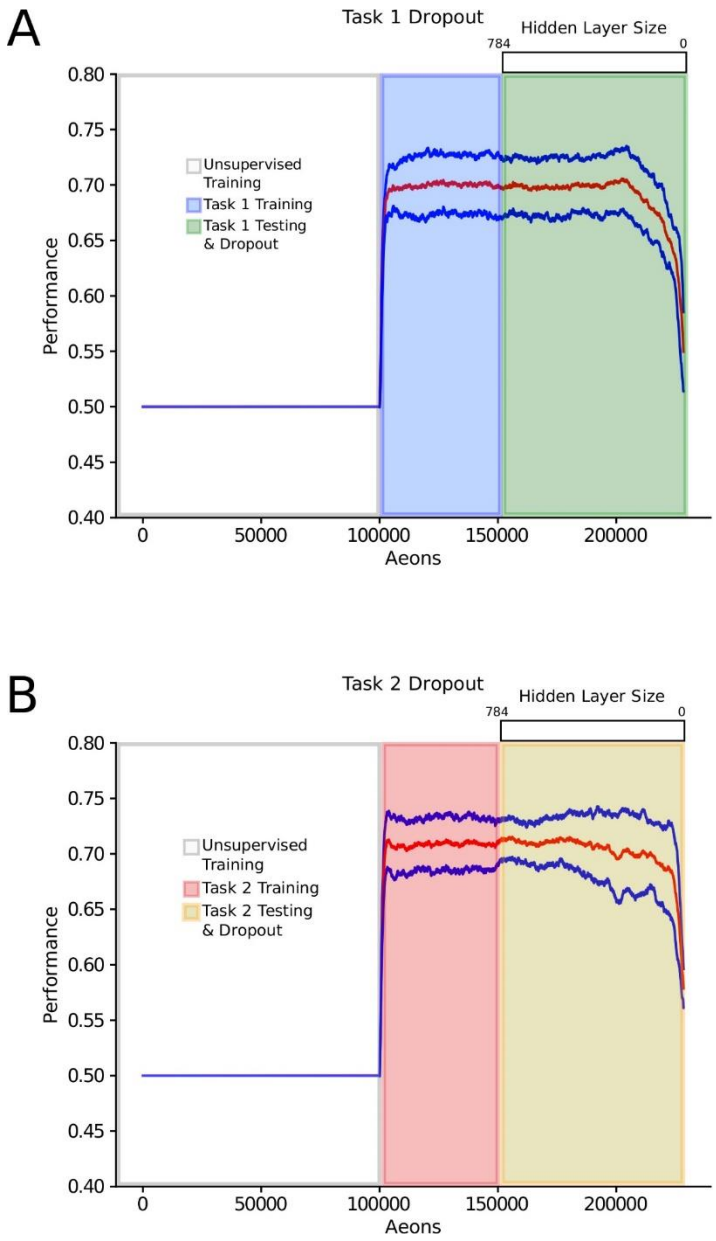


Figure 1.9. (S2 Fig) Model displays graceful degradation in performance as a result of hidden layer dropout.
(A) Mean performance (red line) and standard deviation (blue lines) over time: unsupervised training (white), Task 1 training (blue), Task 1 testing (green). Hidden layer neurons are randomly removed during testing period. Gradient bar above Task 1 testing (green) displays the number of hidden layer neurons over time starting at 784 and decreasing down to 0. The testing performance remains high until ~25% of neurons are left, after which it starts to drop. This highlights the formation of a distributed synaptic structure between hidden and output layer neurons developed during training, ensuring output layer activity is not dictated by a select few hidden layer neurons. **(B)** Same as in (A) but for Task 2.

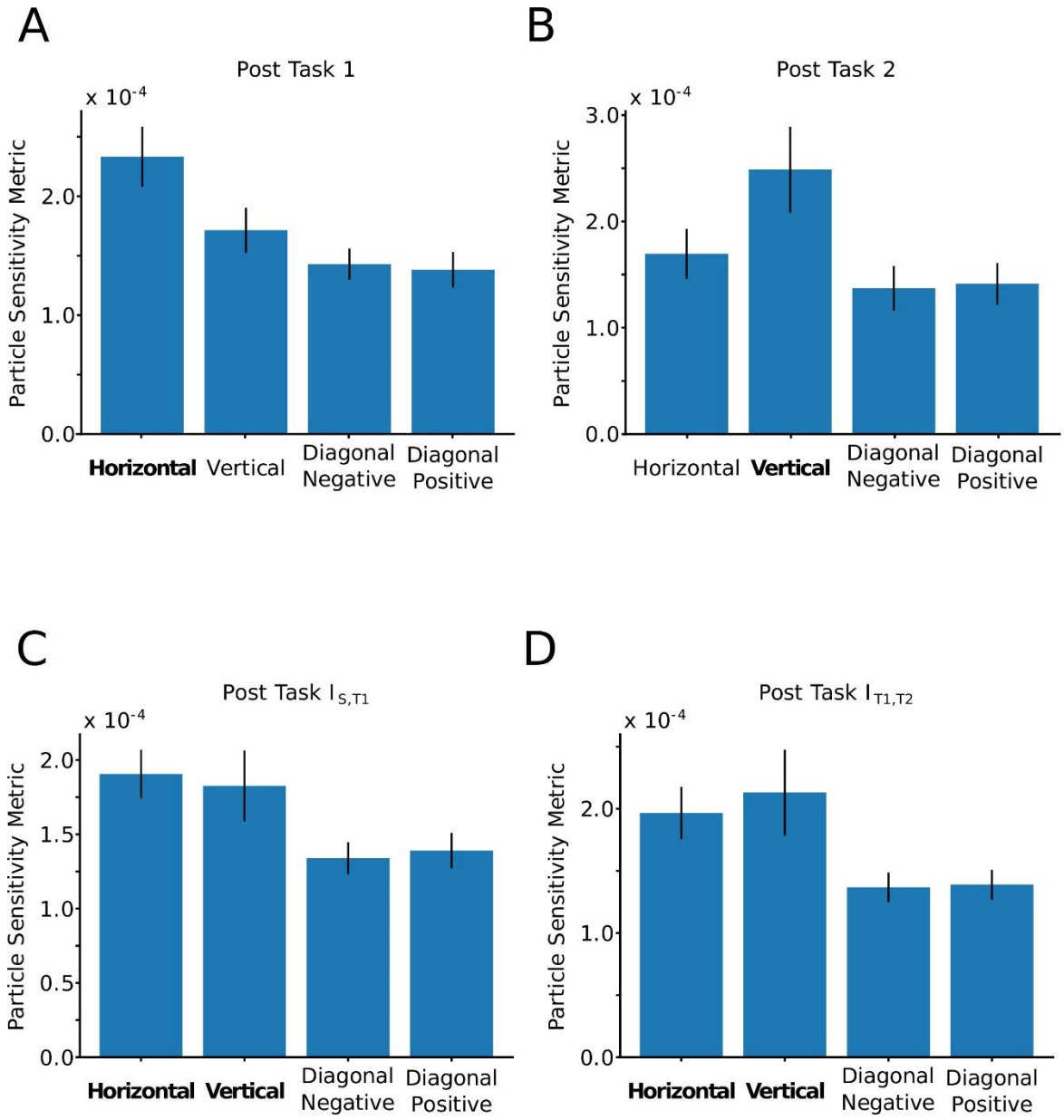


Figure 1.10. (S3 Fig) Particle responsiveness metric (PRM) shows correspondence between type of training and particles preferred by the network.

(A-D) Mean and standard deviation (blue bars and black lines respectively) of the PRM for various types of training and particle orientations across ten trials. The title of each plot reflects the most recently trained stage, the vertical axis corresponds to the value of the PRM while the horizontal axis identifies the particle type (bold labels indicate ideal particles the network would be attracted to following the corresponding training). It can be seen that the metric indicates the network is most responsive to the corresponding ideal particle types following a specific training regime e.g. Post Task 1 the network is most responsive to horizontal particles (A), Post Task 2 the network is most responsive to vertical particles (B), Post $I_{S,T1}$ the network is most responsive to horizontal and vertical particles (C), Post $I_{T1,T2}$ the network is most responsive to horizontal and vertical particles (D).

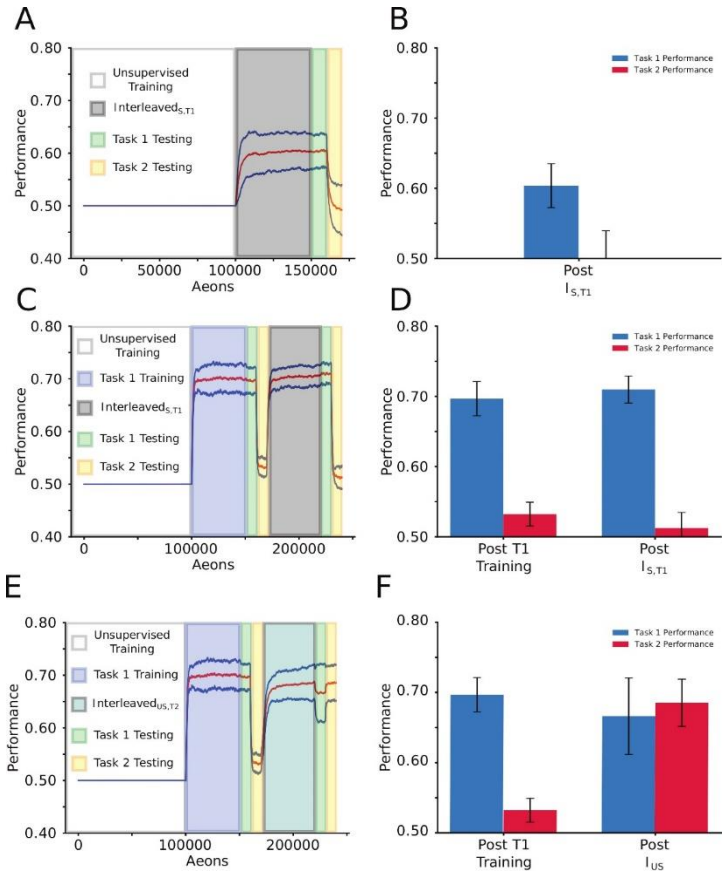


Figure 1.11. (S4 Fig) Effect of sleep to protect old memory does not depend on specific properties of noise applied during sleep phase.

(A) Mean performance (red line) and standard deviation (blue lines) over time: unsupervised training (white), Interleaved$_{S,T1}$ (grey), Task 1/2 testing (green/yellow). **(B)** Mean and standard deviation of performance during testing on Task 1 (blue) and Task 2 (red). Following Interleaved$_{S,T1}$, mean performance on Task 1 was 0.60 ± 0.03 while Task 2 was 0.49 ± 0.05. (In all experiments, 0.5 represents chance performance.) Note that periods of Task 1 training interleaved with sleep do not lead to increase in performance on untrained Task 2, even when Task 2 data from another experiment were used to set up mean firing rates of the random input during sleep. **(C)** Same as in (A) but the sequence of training was: unsupervised training (white), Task 1 training (blue), Task 1/2 testing (green/yellow), Interleaved$_{S,T1}$ (grey), Task 1/2 testing (green/yellow). **(D)** Mean and standard deviation of performance during testing on Task 1 (blue) and Task 2 (red) after Task 1 training and after Interleaved$_{S,T1}$. Following Task 1 training, mean performance on Task 1 was 0.70 ± 0.02 while Task 2 was 0.53 ± 0.02. Post Interleaved$_{S,T1}$ training, mean performance on Task 1 was 0.71 ± 0.02 and Task 2 was 0.51 ± 0.02. Task 1 performance remained high after Interleaved$_{S,T1}$ but no improvement on Task 2 was observed. **(E)** Mean performance (red line) and standard deviation (blue lines) over time: unsupervised training (white), Task 1 training (blue), Task 1/2 testing (green/yellow), Interleaved$_{US,T2}$ (burnt orange), Task 1/2 testing (green/yellow). **(F)** Mean and standard deviation of performance during testing on Task 1 (blue) and Task 2 (red). Following Task 1 training, mean performance on Task 1 was 0.70 ± 0.02 while Task 2 was 0.53 ± 0.02. Post Interleaved$_{US,T2}$ training, mean performance on Task 1 was 0.67 ± 0.05 and Task 2 was 0.69 ± 0.03.

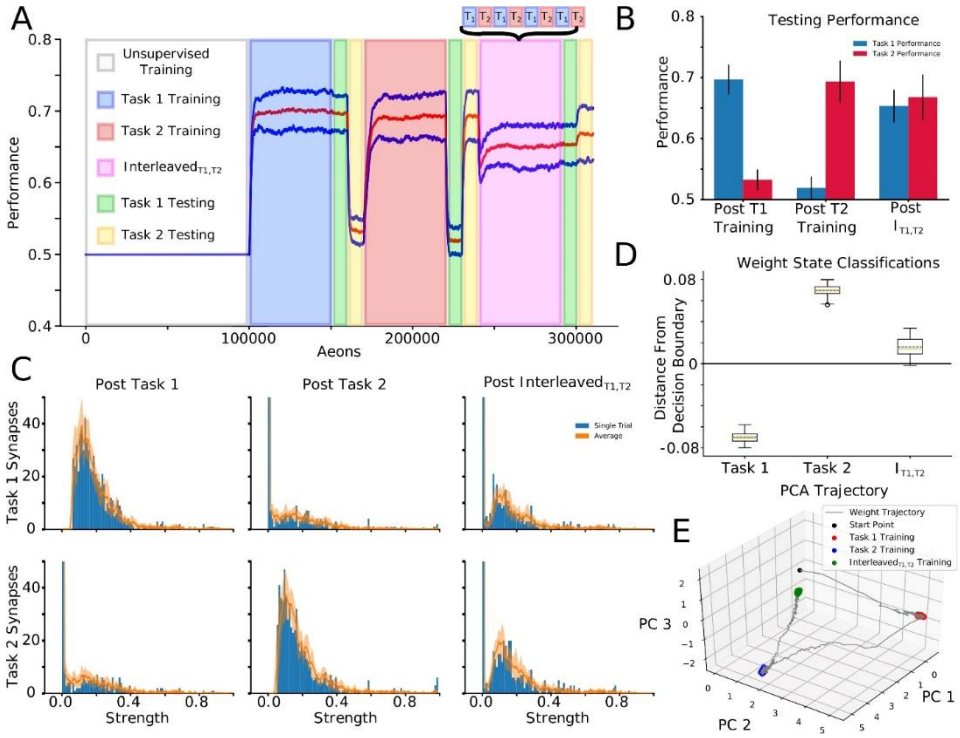


Figure 1.12. (S5 Fig) Interleaving old and new task training allows integrating synaptic information relevant to new task while preserving old task information.

(A) Mean performance (red line) and standard deviation (blue lines) over time: unsupervised training (white), Task 1 training(blue), Task 1/2 testing (green/yellow), Task 2 training (red), Task 1/2 testing (green/yellow), Interleaved$_{T1,T2}$ training (purple), Task 1/2 testing (green/yellow). **(B)** Mean and standard deviation of performance during testing on Task 1 (blue) and Task 2 (red). Following Task 1training, mean performance on Task 1 was 0.69 ± 0.02 while Task 2 was 0.53 ± 0.02. Conversely, following Task 2 training, mean performance on Task 1 was 0.52 ± 0.02 while Task2 was 0.69 ± 0.04. Following Interleaved$_{T1,T2}$ training, mean performance on Task 1 was 0.65 ±0.03 while Task 2 was 0.67 ± 0.04. **(C)** Distributions of task-relevant synaptic weights (blue bars–single trial, orange line / shaded region–mean / std across 10 trails. The distributional structure of Task 1-relevant synapses following Task 1 training (top-left) is destroyed following Task 2 training (top-middle), but partially recovered following. Interleaved$_{T1,T2}$ training (top-right). Similarly, the distributional structure of Task 2-relevantsynapses following Task 2 training (bottom-middle), which was not present following Task 1training (bottom-left), was partially preserved following Interleaved$_{T1,T2}$ training (bottom-right).**(D)** Box plots with mean (dashed green line) and median (dashed orange line) of the distance to the decision boundary found by an SVM trained to classify Task 1 and Task 2 synaptic weight matrices for Task 1, Task 2, and Interleaved$_{T1,T2}$ training across trials. Task 1 and Task 2synaptic weight matrices had mean classification values of -0.069 and 0.069 respectively, while that of Interleaved$_{T1,T2}$ training was 0.016. **(E)** Trajectory of H to O layer synaptic weights through PC space. Synaptic weights which evolved during Interleaved$_{T1,T2}$ training (green dots)clustered in a location of PC space intermediary between the clusters of synaptic weights which evolved during training on Task 1 (red dots) and Task 2 (blue dots).

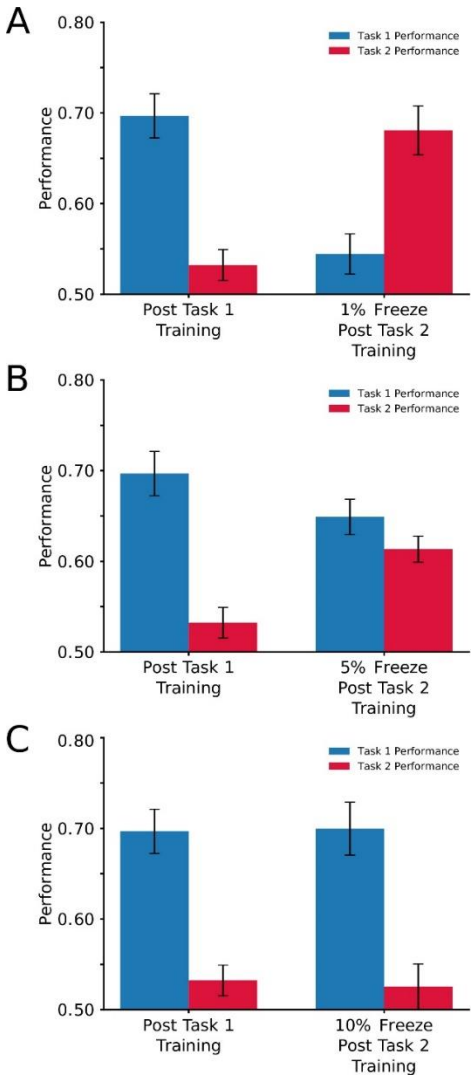


Figure 1.13. (S6 Fig) Freezing a fraction of task specific strong synapses preserves differing degrees of performance in a sequential learning paradigm.
(A-C) Mean and standard deviation of performance during testing on Task 1 (blue) and Task 2 (red). Left, Performance after Task 1 training. Right, Performance after Task 2 training when a fraction of the strongest (after Task 1 training) synapses remained frozen– 1% (A), 5% (B), 10% (C). In all cases, after Task 1 training, Task 1 performance was 0.70 ± 0.02 and Task 2 performance was 0.53 ± 0.02. (A) Freezing the top 1% of Task 1 synapses resulted in a Task 1 performance of 0.54 ± 0.02 and Task 2 performance of 0.68 ± 0.03. (B) Freezing the top 5% of Task 1 synapses resulted in a Task 1 performance of 0.65 ± 0.02 and Task 2 performance of 0.61 ± 0.01. (C) Freezing the top 10% of Task 1 synapses resulted in a Task 1 performance of 0.70 ± 0.03 and Task 2 performance of 0.53 ± 0.03. Freezing the top 1% of Task 1 synapses was not sufficient to maintain Task 1 performance, thus enabling Task 2 relevant synapses to dominate the network; however, freezing the top 10% of Task 1 synapses fully retains Task 1 performance preventing Task 2 to be learned.

Chapter 1, in full, is a reprint of the material as it appears in PLOS Computational Biology 18(11): e1010628, under the title “Sleep prevents catastrophic forgetting in spiking neural networks by forming a joint synaptic weight representation”, Golden, Ryan; Delanois, J. Erik; Sanda, Pavel; Bazhenov, Maxim, PLOS, 2022. The dissertation author was the co-primary investigator and author of this paper, along with J. Erik Delanois.

Chapter 2 Hippocampal indexing alters the stability landscape of synaptic weight space allowing life-long learning

ABSTRACT

Systems consolidation theory posits that the hippocampus rapidly encodes new information during wakeful behavior. This hippocampal memory trace is subsequently assimilated into the cortex during sleep. This powerful idea explains fundamental learning principles, but the process by which sleep modulates synaptic weight space to intricately integrate new memories into the existing knowledge pool remains unknown. In this study, we employed a biophysically-realistic thalamocortical network model to assess the stability landscape of synaptic weight dynamics during task training and subsequent sleep. Our findings indicate that a cortical network synaptic weight space comprises a subspace, a memory manifold, where various weight configurations yield high performance for a given task. After training, sleep acts to propel the system further along this memory-specific manifold, thus improving performance. When training for a new, competing memory occurs, the system may move away from the established memory manifold toward the new task manifold, potentially leading to 'catastrophic forgetting.' This issue is mitigated by employing a dual cortico-hippocampal memory system. Offline memory consolidation, involving mapping newly established hippocampal memory traces to the cortex, guides the system along the old memory manifold toward its intersection with the new memory manifold, thereby circumventing the risk of forgetting the old task. Our study presents a novel theory regarding the role of sleep in memory consolidation, offering a convenient 'geometric' framework for understanding the dynamics of the synaptic weight space induced by sleep and predicting the usefulness of dual-memory system in preventing catastrophic forgetting and facilitating robust memory consolidation.

SIGNIFICANCE STATEMENT

The ability to store, process, and retrieve information is arguably the foundation of intelligent behavior. Sleep extracts invariant features from learned information, leading to the generation of explicit knowledge and insight. Despite a wealth of facts, our fundamental understanding of how memories are encoded in brain networks is very modest. Here, we propose a novel framework for understanding how memories are encoded in synaptic weight space and how sleep dynamics alter the synaptic landscape. Our approach explains why new task learning may lead to memory interference and how sleep enables continual learning. The results advance our knowledge of how the brain solves some fundamental problems in life-long learning.

INTRODUCTION

Continual learning is a foundation of human intelligence. To survive under constantly changing environmental conditions, human and animal brains must continuously encode and assimilate new memories to appropriately guide behavior. Under such circumstances, minimizing memory interference becomes a priority. Not only we can learn without interference, but we learn better when related information was learned in the past and new learning can improve on what we learned before. The difficulty of performing such learning tasks is illustrated by the on-going attempts to achieve scalable continual learning in artificial neural networks (ANNs) without suffering severe retroactive interference known as catastrophic forgetting (Mccloskey and Cohen 1989, McClelland, McNaughton et al. 1995, French 1999, Hayes, Krishnan et al. 2021).

Sleep has been hypothesized to play an important role in memory consolidation and generalization of knowledge in biological systems (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, McCormick et al. 1993). 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, Krishnan et al. 2016). Using biophysical models of thalamocortical network implementing sleep-wake transition (Krishnan, Chauvette et al. 2016), we previously showed that replay of recently learned memories along with relevant old memories can improve learning and enables the network to form orthogonal memory representations to enable coexistence of competing memories within overlapping populations of neurons (Wei, Krishnan et al. 2016, Wei, Krishnan et al. 2018, Gonzalez, Sokolov et al. 2020, Wei, Krishnan et al. 2020). Recent work also revealed that implementing sleep-like processing in ANNs can mitigate catastrophic forgetting and improve generalization (Tadros, Krishnan et al. 2020, Tadros, Krishnan et al. 2020, Tadros and Bazhenov 2022, Tadros, Krishnan et al. 2022, Delanois, Ahuja et al. 2023).

While the progress has been made by incorporating ideas of how the brain functions during sleep in biophysical and artificial networks, we still lack understanding of the fundamental principles governing sleep-induced dynamics in the synaptic weights space. There are little doubts that both awake learning and sleep consolidation change synaptic weight landscape, however, the extreme dimensionality of the synaptic weight space hinders progress in understanding the principles behind learning- and consolidation-induced synaptic weights dynamics. It further limits our ability to apply neuroscience principles to artificial intelligence, in which continual learning remains an unsolved problem (Hayes, Krishnan et al. 2021, Kudithipudi, Aguilar-Simon et al. 2022).

In this new study, we applied biophysical models of the thalamocortical system capable of task learning in awake and memory replay during sleep to develop the concept of a "memory manifold" – a subspace of the synaptic weight space that describes a set of synaptic weights which allow for strong cued recall of specific memories. We then applied this theory to explain how selective memory replay during slow-wave sleep allows formation of new memory traces without catastrophic forgetting of the old memories. Starting from procedural (hippocampus-independent) memories, we expanded idea of memory manifolds to declarative memories and we demonstrated that the dual hippocampo-cortical memory system, as proposed in Systems Consolidation Theory (Wilson and McNaughton 1994, Rasch and Born 2013), may provide an optimal mechanism of a new memory training as the system never leaves the vicinity of the old tasks memory manifolds.

RESULTS

Network Model

The network model utilized throughout the study was built upon thalamocortical models previously used in earlier work (Krishnan, Chauvette et al. 2016, Gonzalez, Sokolov et al. 2020). 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 synapses were set probabilistically within a local radius with the initial weight values Gaussian distributed (Figure 1B) but allowed to vary according to local spike-

timing-dependent plasticity (STDP) rules. In the model, transitions between awake and non-rapid-eye-movement stage 3 (N3) sleep were simulated by changing cellular and synaptic parameters to mimic the effects of the distinct neuromodulatory tone of each brain state (Krishnan, Chauvette et al. 2016). See **Methods** for more details about the network model.

Differential replay of sequence memories during sleep improves recall performance

An example of an experimental simulation paradigm consisting of testing, training, and N3 sleep stages is shown in Figure 2A. During the awake state, the network was trained with one of two sequence memories (S1 or S2), represented by five sequentially ordered, cell groups of ten neurons (EDCBA or ABCDA, respectively; see Figure 2B, left and middle). These sequences were chosen to elicit maximal interference according to our STDP rules. Training proceeded by simulating DC current injections to activate each cell group, with a small delay between groups allowing for STDP to strengthen connections. During N3 sleep, the network exhibited spontaneous slow oscillations (<1 Hz) 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 during the awake state by activating the first cell group and testing for pattern completion of the trained sequence. (See Figure 2C insets). Recall performance was found to reliably increase for S1 (Figure 2C, top) and S2 (bottom) following both training and 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. The later result suggested that the network was reactivating the trained sequence during sleep.

In order to verify differential neural reactivation that depends on the trained memory sequence, we used PCA to perform a dimensionality reduction on the neural firing rate data during Up states in each of the two memory sequence conditions (Figure 2D; see **Methods** for details). It can be seen that the Up state trajectories following training S1 (red) and S2 (blue) are separable in the first two principal components, indicating robust, reliable, differential reactivation. We took advantage of this differential reactivation to train a linear SVM to predict the probability of S1 or S2 replay on each Up state by using a single trial of each sequence memory as training data (see **Methods** for details). Figure 2E shows the replay probability for each memory during each Up state of sleep averaged across trials (for an example of the replay probabilities for a single random seed, see Supp. Figure 1A).

Sleep can rescue both interference effects induced by sequentially training sequence memories

Next, we simulated the case of two, sequentially trained sequence memories (S1 and S2) followed by a period of sleep. Under this simulation paradigm, the network was shown to exhibit a retroactive interference effect on S1 recall (Figure 2F, left) that can be seen after S2 training. Additionally, a prospective interference effect was exhibited on S2 recall (Figure 2F, right) that can be seen by comparison to the single memory case (Figure 2C, bottom) after S2 training. Figure 2G shows that the average replay probability for each sequence oscillated about 0.5, indicating that each memory was replayed approximately evenly during sleep. Importantly, for any given trial simulation, replay for each Up state was nearly always robustly classified as S1 or S2 (Supp. Figure 1B).

Interference can be characterized by the network falling off its "memory manifolds"

To visualize the synaptic changes which occur from training and memory consolidation during sleep, we used PCA to perform dimensionality reduction on the synaptic weight data between PY-PY cells. Additionally, we sampled a subset of synaptic weight space and tested recall performance to generate contour plots in PC space which serve as the synaptic performance landscape – indicating the approximate recall performance on S1 (red) and S2 (blue) independently (Figure 3, left), and jointly on S1 & S2 (purple; Figure 3, right) for a given synaptic weight state. See the **Methods** for details on the PCA implementation and the synaptic performance landscape.

Figure 3A shows that initial training of S1 (red arrow) and S2 (blue arrow) push the network in the performance landscape from its initial location (black dot) to regions of greater recall performance for the respective memory. The coloring of the trajectories was determined by the memories that could be robustly recalled: gray – neither; red – S1 only; blue – S2 only; purple – both. We refer to regions in synaptic weight space with robust memory recall as either single or joint "memory manifolds". Figure 3B shows that subsequent N3 sleep pushes the networks further along their current memory manifolds in the same direction as training. However, if instead of sleep, competing memory training occurred (Figure 3C), the networks were pushed off each single memory manifold (red/blue) into a region of ignorance (gray) where neither sequence could be recalled. From this region of ignorance, subsequent sleep (Figure 3D) was able to move the network onto the joint memory manifold. It should be noted that the terminal locations of all three memory manifolds displayed stable dynamics, such that further sleep resulted negligible movement synaptic weight space.

Sleep moves the network towards stable regions of memory manifolds

Finally, Figure 3E shows examples of under/overtraining of S2 (blue) following initial S1 (red) training. The first undertraining example halted S2 training before the network fell off the S1 memory manifold, and subsequent sleep pushed the network towards the stable region of the S1 manifold. Supp. Figure 2A 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 halted S2 training just after the network fell off the S1 manifold. In this case, subsequent sleep pushed the network through a region of the S1 manifold before moving onto the joint memory manifold. Supp. Figure 2B shows that replay initially becomes biased towards S1 before slowly returning towards more equal replay. Similarly, the first overtraining example halts the network just after it passes the intersection in the region of ignorance, and subsequent sleep pushed it towards the joint memory manifold. Supp. Figure 2C shows the average replay probabilities stay more balanced in this case. However, in another overtraining example, the network was trained on S2 until it reached the S2 manifold, where subsequent sleep pushed the network towards its stable region. Here the average replay probabilities are initial balanced before S2 comes to dominate by the end of sleep (Supp. Figure 2D). Taken together, these under/overtraining dynamics reveal that sleep tends to move the network towards stable regions of memory manifolds.

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 stable region of the joint memory manifold, it relied on the network passing through a region of ignorance 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 immediately following sleep. Given this disparity, we hypothesized that simulating the effects of hippocampal indexing to guide cortical replay during sleep might allow the network to encode the competing memory with minimal interference to the initial memory. 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 N3 sleep and hippocampal indexing was induced by detecting the onset of each Up State (i.e. the Down-to-Up transition) and applying DC input the network to sequentially activate each cell group in S2 with a 5 ms delay (Figure 4B, right). Finally, the network was left in the sleep state without hippocampal indexing being applied before a final test. Hippocampal indexing was chosen to only last for the first half of sleep 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, Cai et al. 2011, Rasch and Born 2013).

Figure 4C shows the recall performance for S1 (left) steadily increased throughout all stages of the simulation, while indexing causes a jump in S2 recall (right) that is maintained following subsequent sleep. The average replay probabilities (Figure 4D) show that the network became increasingly likely to replay S2 during the course of indexing. Moreover, halting indexing when the memories had an approximately equal chance of being replayed allowed the network to maintain continued replay of each during subsequent sleep. However, if indexing were allowed to continue throughout the entirety of sleep, catastrophic retroactive interference occurred (Figure 4E, left) at the expense of an extremely robust competing memory (right).

Figure 4F illustrates that this occurs by the competing memory coming to completely dominate replay well before the end of sleep.

Finally, we investigated how these Up state trajectories appeared in neural activity space by utilizing the PCA representation employed previously (Figure 2D; see **Methods** for details). Figure 4G shows the average Up state trajectory for the network during indexing (purple) fall within an intermediate region of those of Up states after only training S1 (red) or S2 (blue) on the left. On the right, Up state trajectories during indexing were sorted as follows: $p(S1 \geq 0.95)$ – S1 replay (red); $p(S2 \geq 0.95)$ – S2 replay (blue); $p(0.05 < S1 < 0.95)$ – S1/S2 replay (purple). This was based off the densities obtained from integrating the single trial replay probabilities for indexing over time (Supp. Figure 1C); most of the replay probabilities are very near to 0 or 1, with a more sparse and uniform distribution between $p=0.05$ and $p=0.95$. From this sorting of Up state trajectories, it is clear that S1 and S2 replays which occur during indexing are separable in neural activity space. Moreover, the way the S1/S2 replay trajectories first overlap the S2 replays (i.e. the sequence which is indexed at the beginning of each Up state) before jumping over to overlap the S1 replays (i.e. the sequence which was initially trained in the network) suggests that interleaved replay of the two memories may allow the network to stay on its initial memory manifold as it moves towards the joint memory manifold. This is further suggested by the fact that S1/S2 replays become increasingly common throughout the second half of indexing (see Figure 2 – Supplement 1C). Figure 4H quantifies the trajectory overlaps just described (Figure 4G, right).

Hippocampal indexing causes interleaved memory replay within individual Up states

Based off the above observation (Figure 4G, right), we quantified the amount of synaptic reactivation in the S1 and S2 direction during the indexed phase of each up state and the post-index phase. The indexed phase was defined as the first 50 ms following the detection of the Down-to-Up transition during which the simulated index was applied. This was done by counting the number of Up states each synapse experienced a net change towards potentiation. Figure 5A displays this reactivation count (color scale) for the top 30% of synapses during the indexed (left) and post-indexed (right) phases during the first 1000 s of indexed sleep. During the indexed phase more reactivation occurred among synapses that facilitate S2 recall (blue template), while during the post-index phase more reactivation occurred among synapses that facilitate S1 recall (red template).

Figure 5B summarizes this further, displaying the proportion of the sum total of reactivations in the S1 (red line) and S2 (blue line) templates compared to the sum of both templates, averaged over 1000 s intervals of indexed sleep. During the indexed phase (left) greater reactivation of S2 facilitating synapses occurs throughout the entirety of sleep. However, during the post-index phase S1 reactivation is initially greater, before declining and being overtaken by S2 reactivation around the middle of sleep, after which S2 reactivation dominates. Therefore, our model predicts that indexing causes interleaved memory replay within individual Up states, with the recent, indexed memory being replayed first, followed by replay of older memory traces in the cortical network.

Hippocampal indexing moves the network to its joint memory manifold without falling off its initial single memory manifold

Examining 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 6 displays two different perspective of a three-dimensional principal component representation of the paths of the networks through synaptic weight space, with the trajectory coloring encoding the same information as in Figure 3. In the case of normal indexing (left) the network begins at the red (resp. blue) corner of the rhombus after training S1 (resp. S2) and moves negatively (resp. positively) in the third principal component dimension towards a newly discovered stable region of the joint memory manifold. Significantly, this happens without the network falling off the S1 (resp. S2) manifold. In the case of over indexing (right), the networks are pushed past the new stable regions and continue onto the opposite single memory manifolds from which they began.

Hippocampal indexing keeps synaptic weights lower and results in sparser stable solutions

Next, we investigated how the synaptic weight dynamics differed between the cases on sequential training and indexing. Figures 7A & B plot the synaptic weight values for all pairs of bi-directionally connected neurons in these cases. This was done because pairs without bi-directional connectivity are not subject to competition under the STDP rules used here. To make this more concrete, take a pair of neurons with one in cell group A and the other in group B. In this case, the S1 synaptic weight would be the strength of the synapse going from B->A (since S1 is defined as EDCBA), while the S2 synaptic weight would be the strength of the synapse from A->B. The colored regions of the plots identify regions where the bi-directional pair had

prioritized either the S1 (red) or S2 (blue) synapse at the expense of the opposing synapse becoming effectively disconnected.

For both simulation paradigms, the weight values had a Gaussian distribution at baseline (Figures 7A&B, left), and the bulk of the density moved towards the red corner as a result of S1 training (Figures 7A&B, middle-left). After this point the two simulations paradigms diverge. Both sequential training and indexing resulted in densities that were more symmetric about the line y=x, indicating a roughly equal distribution of synaptic resources between the two memories. However, S2 training (Figure 7A, middle-right) resulted in the density moving much further into top-right corner of the plot compared to S2 indexing (Figure 7B, middle-right), which kept the density more localized around the line y=-x. This indicates that sequential training tends to result in many bi-directional pairs which are strongly activating but in both directions; this ambiguity prevents them from contributing to differential memory encoding. Under both simulation paradigms, sleep resulted in bi-directional pairs being pushed further into either the red or blue corners by amplifying any S1/S2 encoding bias which was already present in the pair (Figure 7A&B, right). In the case of sequential training (Figure 7A, right), the reduction of ambiguous bi-directional pairs can also be observed by the void that appears in the top-right corner of the plot.

Figure 7C shows the synaptic weight distributions obtained by sequential training (purple) and indexing (green) paradigms just after the simulation phase in the corresponding columns panels A and B, respectively. It can be seen that after sleep, the distribution converged on by indexing appears to have more weak synapses and less strong synapses when compared to that found after sequential training.

To better quantify this, we computed the sparsity of a synaptic weight filtration at each time of the time points discussed above. Briefly, a synaptic weight filtration is generated from the data by taking the synaptic weight matrix and repeatedly binarizing it according to a dense set of thresholds which cover all possible weight values a synapse can take in simulation. This results in a stack of binary matrices which preserve all of the information from the original floating-point populated synaptic weight matrix, but allows computing binary matrix properties, such as sparsity (i.e., the ratio of zero-valued entries to the total number of entries) on the weight data (see **Methods** for more details).

Figure 7D plots the average sparsity of the filtrations across trials (solid line), and the standard error (shaded region), for the sequential training (purple) and indexing (green) paradigms. At baseline (Figure 7D, left), both paradigms resulted in a sharp sigmoidal curve characterizing the sparsity of the filtration stacks. This makes sense since, at baseline, the weight values are Gaussian distributed around an initial value (see Figure 7C, left). In the plot, we see that when we use the initial weight value as a threshold, approximately half of the weights get zeroed out upon binarization, and most weight values are very near to the initial value. Further evidence of the underlying Gaussian structure can be shown by choosing a threshold slightly left of the initial weight and observing that there are no weights in the network which have a lower value than this threshold (i.e. sparsity = 0). Alternatively, choosing a threshold slightly to the right results in a sparsity of 1, indicating that all synaptic weight values are below this threshold.

As a result of S1 training (Figure 7D, middle-left), the sparsity curves of both filtration curves undergo a deformation – primarily to the right of the initial weight value with less deformation to the left. This indicates that many synaptic weights have increased, with some at maximum strength (i.e., presence of non-unity sparsity values near the max weight threshold)

while another subset has decreased, but not to the extent of the minimum strength (i.e., lack of non-zero sparsity values near the min weight threshold). Following subsequent S2 training or S2 indexing, the sparsity filtration curves (Figure 7D, middle-right) change differentially. The sparsity curve after S2 training (purple) indicates that <10% of synaptic weights are below the initial weight value, while that of S2 indexing (green) has this number at >30%. Moreover, the asymmetric sigmoid shape of the S2 training (purple) curve indicates that the weight distribution in biased towards strong values, while the logit shape of S2 indexing (green) suggests the weight values are distributed bimodally at the boundaries (see Figure 7C, middle-right).

Finally, further sleep has the effect of deforming the S2 training (purple) curve from an asymmetric sigmoid into a logit shape, and further flattening the logit shape of the S2 indexing curve (green; Figure 7D, right). While both of these sparsity curves correspond to weight values which are bimodally distributed at the boundaries, the fact that the S2 indexing curve is significantly larger than that of S2 training for all threshold values at the end of the simulation indicates that indexing finds regions on the joint memory manifold which are significantly sparser, and thus, more resource efficient, than sequential training.

DISCUSSION

Using a biophysically realistic thalamocortical network model capable of task learning in the awake state and consolidation during sleep, we characterized the neural network dynamics in synaptic weight space during both sleep and wake. Figure 8 summarizes these dynamics with illustrative schematics. When the initial training was sufficient to bring the system into the vicinity of the task-specific memory manifold (a task-specific subspace in synaptic weight space), sleep replay induced a convergence dynamic towards the memory attractor. Learning a

new task introduced a transition away from the old task-specific manifolds and towards the new task manifold (Figure 8; top). With subsequent sleep it was possible to transition to joint-task memory manifold (Figure 8A), but this was found to require a fine-tuned training duration and was more likely to induce a failure to consolidate the new memory (i.e. proactive interference; Figure 8B), or catastrophic forgetting (i.e. retroactive interference; Figure 8C). In this scenario, slow task training interleaved with periods of sleep (Supp. Fig 3), as in procedural (hippocampus-independent) learning, was necessary to prevent damage to the old tasks (Figure 8D). Fast learning of a new task could be accomplished by utilizing a complementary learning systems approach, involving a fast-learning hippocampus and a slow-learning cortex (Figure 8E), as is the case with declarative memories. In such a case, each slow wave included hippocampus-dependent replay of a new memory during the initial slow-wave phase and intrinsically driven cortical replay of old memories during the later slow-wave phase. This dynamic allowed the system to remain near the old task manifold while converging toward its intersection with the new task manifold, providing optimal learning dynamics so long as indexing did not persist too long into sleep (Figure 8F).

It is interesting to note that our model suggests the complementary learning systems approach can be seen as a maximally compressed version of the strategy taken for procedural memory consolidation, with the indexed and post-indexed phases corresponding to the training and sleep phases, respectively. Importantly, each indexed phase only receives a single sample from the hippocampus, and each post-indexed phase to the spontaneous replay during a single Up-state. Both are at the minimum limit of what could conceivably be labeled training and sleep periods.

Humans and animals have a remarkable ability to learn continuously, incorporate new data into their corpus of existing knowledge, and generalize episodic memories beyond a single experience. In contrast, artificial neural networks (ANNs) suffer from "catastrophic forgetting" whereby they achieve optimal performance on newer tasks at the expense of performance on previously learned tasks (Mccloskey and Cohen 1989, McClelland, McNaughton et al. 1995, French 1999, Hayes, Krishnan et al. 2021). ANNs have poor generalization properties when tested on datasets with even small deviations from the training distribution such as non-Gaussian data noise (Geirhos, Temme et al. 2018), which makes ANN predictions unreliable in “real-life” scenarios. This dichotomy between learning new tasks and the ability to retain and generalize knowledge across all tasks in mammals and ANNs has given rise to the stability-plasticity dilemma (French 1999, Abraham and Robins 2005, Mermillod, Bugaiska et al. 2013). On the one hand, a network must be plastic such that the parameters in the network can change in order to accurately represent and respond to new tasks. On the other hand, a network must be stable such that it maintains knowledge of older tasks. Although deep neural networks (LeCun, Bengio et al. 2015) can achieve supra-human level of performance on tasks ranging from complex games to image recognition, they lie at a sub-optimal point on the stability-plasticity spectrum.

ANNs have long been known to be able overcome catastrophic forgetting under varying degrees of data-intensive interleaved training strategies (McClelland, McNaughton et al. 1995, Hasselmo 2017, Saxena, Shobe et al. 2022), and it has recently been shown that the same applies to biophysical (Gonzalez, Sokolov et al. 2020) and artificial spiking neural networks (Golden, Delanois et al. 2022). Moreover, the procedural memory consolidation strategy of interleaved training and sleep has now been shown to mitigate catastrophic forgetting in ANNs (Tadros, Krishnan et al. 2020, Tadros, Krishnan et al. 2022), artificial spiking networks (Golden, Delanois

et al. 2022), and biophysical spiking networks (Supp. Figure 3). Our results with biophysical spiking networks suggest that it may be advantageous to adapt the indexing strategy described here to ANNs.

The critical role that sleep plays in learning and memory is supported by a vast, interdisciplinary literature spanning both psychology and neuroscience (Paller and Voss 2004, Walker and Stickgold 2004, Oudiette, Antony et al. 2013, Rasch and Born 2013, Stickgold 2013). Specifically, it has been suggested that REM sleep supports the consolidation of non-declarative or procedural memories, while non-REM sleep supports the consolidation of declarative memories (Mednick, Cai et al. 2011, Rasch and Born 2013, Stickgold 2013). In particular, REM sleep has been shown to be important for the consolidation of memories of tasks involving perceptual pattern separation, such as the texture discrimination task (Stickgold, James et al. 2000, Rasch and Born 2013). Despite the difference in the cellular and network dynamics during these two stages of sleep (Rasch and Born 2013, Stickgold 2013), both are thought to contribute to memory consolidation through repeated reactivation, or replay, of specific memory traces acquired during learning (Hennevin, Hars et al. 1995, Paller and Voss 2004, Mednick, Cai et al. 2011, Oudiette, Antony et al. 2013, Rasch and Born 2013, Lewis, Knoblich et al. 2018, Wei, Krishnan et al. 2018).

During NREM sleep, the features of the neocortical SO to repeatedly reset networks during the Down phase has led to the hypothesis that the neocortical SO provides a global temporal frame within the cortex and between brain regions for offline memory processing and reactivation (Isomura, Sirota et al. 2006, Ji and Wilson 2007, Rasch, Buchel et al. 2007, Molle, Eschenko et al. 2009, Wierzynski, Lubenov et al. 2009). The key element of the consolidation stage during NREM sleep is cortical replay triggered by hippocampal SWR events (Peyrache,

Khamassi et al. 2009). In rodents, temporally ordered firing sequences related to a recent experience are replayed in both hippocampus and neocortex synchronously (Ji and Wilson 2007, Mehta 2007) during SO. Such sequence replay has been proposed to be a neural substrate of memory consolidation (Barnes and Wilson 2014) and is believed to result in synaptic changes in the neocortex responsible for integration of memory representations (Schwindel and McNaughton 2011).

What are the underlying mechanisms that support continual learning in biological systems? What is the basis for robust learning that is resilient against potential interference from new experiences? Building upon our recent work (Wei, Krishnan et al. 2016, Wei, Krishnan et al. 2018, Gonzalez, Sokolov et al. 2020, Wei, Krishnan et al. 2020), here we proposed and tested using biophysical model a hypothesis that: (a) The same memory can be represented by multiple different configurations of synaptic weights, forming a "memory manifold" in the space of all synaptic weights, i.e., any point on this manifold would allow successful retrieval of a memory; (b) New task training moves the synaptic weight configuration away from the manifold representing old tasks potentially leading to forgetting. (c) Biological sleep allows simultaneous replay of old and new memory traces, and thus mitigates catastrophic forgetting by pushing the synaptic weight configuration towards the intersection of the solution manifolds representing multiple tasks. (d) Complementary memory systems including a fast learning hippocampus and a slow learning cortex, provides an optimal mechanism of a new memory training as the system never leaves the vicinity of the old tasks' memory manifolds.

METHODS

Thalamocortical network model

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

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

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, Chauvette 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, Lydic 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, Krishnan 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:

$$I^j = g_j m^M h^N (V - E_j).$$

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, Chauvette et al. 2016, Wei, Krishnan 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:

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

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, Timofeev et al. 2002). This short-term depression was modeled as follows:

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

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 Δt 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 $\tau_{+/-} = 20$ 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, Matsui et al. 2005, Sugisaki, Fukushima et al. 2016).

Sequence training and testing. Training and testing of memory sequences was performed similarly to our previous study (Wei, Krishnan 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, Krishnan et al. 2018) and the code is available at 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, Krishnan 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 j^{th} 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 seeds 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. 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, left; red and blue), and S1&S2 jointly (Figure 3, right; purple), and a mesh with contour gradients was computed for each set of recall performance measures.

Sparsity of synaptic weight filtrations. In order to compute the sparsity of synaptic weight matrices, we first transformed these floating-point valued matrices into a stack of binary matrices using the following information-preserving filtration. We first compiled a dense set of synaptic weight values to act as thresholds, in particular, generated the set by starting at the mean initial weight value, and then incrementing/decrementing by the minimum potentiation/depression value permitted by our simulation until we reached the maximum/minimum weight value. Each threshold was then used to generate a binarized synaptic weight matrix by setting all entries less than the threshold equal to zero, and all those greater than or equal to it to 1. The sparsity of these binarized matrices was then computed, and the sparsity at each threshold value was averaged across random seeds.

ACKNOWLEDGEMENTS

REFERENCES

Abraham, W. C. and A. Robins (2005). "Memory retention--the synaptic stability versus plasticity dilemma." Trends Neurosci **28**(2): 73-78.

Barnes, D. C. and D. A. Wilson (2014). "Slow-wave sleep-imposed replay modulates both strength and precision of memory." J Neurosci **34**(15): 5134-5142.

Bazhenov, M., I. Timofeev, M. Steriade and T. J. Sejnowski (2002). "Model of thalamocortical slow-wave sleep oscillations and transitions to activated states." J Neurosci **22**(19): 8691-8704.

Blokland, A. (1995). "Acetylcholine: a neurotransmitter for learning and memory?" Brain Res Brain Res Rev **21**(3): 285-300.

Delanois, E., A. Ahuja, G. P. Krishnan, T. Tadros, J. McAuley and M. Bazhenov (2023). "Improving robustness of convolutional networks through sleep-like replay." ICMLA: in press.

French, R. M. (1999). "Catastrophic forgetting in connectionist networks." Trends Cogn Sci **3**(4): 128-135.

Geirhos, R., C. R. Temme, J. Rauber, H. H. Schütt, M. Bethge and F. A. Wichmann (2018). Generalisation in humans and deep neural networks. Advances in Neural Information Processing Systems.

Golden, R., J. E. Delanois, P. Sanda and M. Bazhenov (2022). "Sleep prevents catastrophic forgetting in spiking neural networks by forming a joint synaptic weight representation." PLoS Comput Biol **18**(11): e1010628.

González, O. C. (2020). sequencePerformanceAnalysis. GitHub. https://github.com/o2gonzalez/sequencePerformanceAnalysis . 094c4be.

Gonzalez, O. C., Y. Sokolov, G. P. Krishnan, J. E. Delanois and M. Bazhenov (2020). "Can sleep protect memories from catastrophic forgetting?" Elife **9**.

Hasselmo, M. E. (2017). "Avoiding Catastrophic Forgetting." Trends Cogn Sci **21**(6): 407-408.

Hayes, T. L., G. P. Krishnan, M. Bazhenov, H. T. Siegelmann, T. J. Sejnowski and K. C. (2021). "Replay in Deep Learning: Current Approaches and Missing Biological Elements." Neural Computations (in press): arXiv:2104.04132.

Hayes, T. L., G. P. Krishnan, M. Bazhenov, H. T. Siegelmann, T. J. Sejnowski and C. Kanan (2021). "Replay in Deep Learning: Current Approaches and Missing Biological Elements." Neural Comput **33**(11): 2908-2950.

Hennevin, E., B. Hars, C. Maho and V. Bloch (1995). "Processing of learned information in paradoxical sleep: relevance for memory." Behav Brain Res **69**(1-2): 125-135.

Isomura, Y., A. Sirota, S. Ozen, S. Montgomery, K. Mizuseki, D. A. Henze and G. Buzsaki (2006). "Integration and segregation of activity in entorhinal-hippocampal subregions by neocortical slow oscillations." Neuron **52**(5): 871-882.

Ji, D. and M. A. Wilson (2007). "Coordinated memory replay in the visual cortex and hippocampus during sleep." Nature neuroscience **10**(1): 100-107.

Krishnan, G. P., S. Chauvette, I. Shamie, S. Soltani, I. Timofeev, S. S. Cash, E. Halgren and M. Bazhenov (2016). "Cellular and neurochemical basis of sleep stages in the thalamocortical network." Elife **5**.

Kudithipudi, D., M. Aguilar-Simon, J. Babb, M. Bazhenov, D. Blackiston, J. Bongard, A. P. Brna, S. C. Raja, N. Cheney, J. Clune, A. Daram, S. Fusi, P. Helfer, L. Kay, N. Ketz, Z. Kira, S. Kolouri, J. L. Krichmar, S. Kriegman, M. Levin, S. Madireddy, S. Manicka, A. Marjaninejad, B. McNaughton, R. Miikkulainen, Z. Navratilova, T. Pandit, A. Parker, P. K. Pilly, S. Risi, T. J. Sejnowski, A. Soltoggio, N. Soures, A. S. Tolias, D. Urbina-Melendez, F. J. Valero-Cuevas, G. M. van de Ven, J. T. Vogelstein, F. Wang, R. Weiss, A. Yanguas-Gil, X. Y. Zou and H. Siegelmann (2022). "Biological underpinnings for lifelong learning machines." Nature Machine Intelligence **4**(3): 196-210.

LeCun, Y., Y. Bengio and G. Hinton (2015). "Deep learning." Nature **521**(7553): 436-444.

Lewis, P. A. and S. J. Durrant (2011). "Overlapping memory replay during sleep builds cognitive schemata." Trends Cogn Sci **15**(8): 343-351.

Lewis, P. A., G. Knoblich and G. Poe (2018). "How Memory Replay in Sleep Boosts Creative Problem-Solving." Trends Cogn Sci **22**(6): 491-503.

McClelland, J. L., B. L. McNaughton and R. C. O'Reilly (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." Psychol Rev **102**(3): 419-457.

Mccloskey, M. and N. J. Cohen (1989). "CATASTROPHIC INTERFERENCE IN CONNECTIONIST NETWORKS: THE SEQUENTIAL LEARNING PROBLEM." The Psychology of Learning and Motivation **24**: 109-165.

Mednick, S. C., D. J. Cai, T. Shuman, S. Anagnostaras and J. T. Wixted (2011). "An opportunistic theory of cellular and systems consolidation." Trends in neurosciences **34**(10): 504-514.

Mehta, M. R. (2007). "Cortico-hippocampal interaction during up-down states and memory consolidation." Nat Neurosci **10**(1): 13-15.

Mermillod, M., A. Bugaiska and P. Bonin (2013). "The stability-plasticity dilemma: investigating the continuum from catastrophic forgetting to age-limited learning effects." Front Psychol **4**: 504.

Molle, M., O. Eschenko, S. Gais, S. J. Sara and J. Born (2009). "The influence of learning on sleep slow oscillations and associated spindles and ripples in humans and rats." Eur J Neurosci **29**(5): 1071-1081.

Oudiette, D., J. W. Antony, J. D. Creery and K. A. Paller (2013). "The role of memory reactivation during wakefulness and sleep in determining which memories endure." J Neurosci **33**(15): 6672-6678.

Paller, K. A. and J. L. Voss (2004). "Memory reactivation and consolidation during sleep." Learn Mem **11**(6): 664-670.

Peyrache, A., M. Khamassi, K. Benchenane, S. I. Wiener and F. P. Battaglia (2009). "Replay of rule-learning related neural patterns in the prefrontal cortex during sleep." Nature neuroscience **12**(7): 919-926.

Plihal, W. and J. Born (1999). "Effects of early and late nocturnal sleep on priming and spatial memory." Psychophysiology **36**(5): 571-582.

Rasch, B. and J. Born (2013). "About sleep's role in memory." Physiological reviews **93**(2): 681-766.

Rasch, B., C. Buchel, S. Gais and J. Born (2007). "Odor cues during slow-wave sleep prompt declarative memory consolidation." Science **315**(5817): 1426-1429.

Saxena, R., J. L. Shobe and B. L. McNaughton (2022). "Learning in deep neural networks and brains with similarity-weighted interleaved learning." Proc Natl Acad Sci U S A **119**(27): e2115229119.

Schwindel, C. D. and B. L. McNaughton (2011). "Hippocampal-cortical interactions and the dynamics of memory trace reactivation." Prog Brain Res **193**: 163-177.

Shinoe, T., M. Matsui, M. M. Taketo and T. Manabe (2005). "Modulation of synaptic plasticity by physiological activation of M1 muscarinic acetylcholine receptors in the mouse hippocampus." J Neurosci **25**(48): 11194-11200.

Steriade, M., D. A. McCormick and T. J. Sejnowski (1993). "Thalamocortical oscillations in the sleeping and aroused brain." Science **262**(5134): 679-685.

Stickgold, R. (2013). "Parsing the role of sleep in memory processing." Curr Opin Neurobiol **23**(5): 847-853.

Stickgold, R., L. James and J. A. Hobson (2000). "Visual discrimination learning requires sleep after training." Nat Neurosci **3**(12): 1237-1238.

Stickgold, R. and M. P. Walker (2007). "Sleep-dependent memory consolidation and reconsolidation." Sleep medicine **8**(4): 331-343.

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

Tadros, T. and M. Bazhenov (2022). "Role of Sleep in Formation of Relational Associative Memory." J Neurosci.

Tadros, T., G. Krishnan, R. Ramyaa and M. Bazhenov (2020) "Biologically inspired sleep algorithm for reducing catastrophic forgetting in neural networks. ." AAAI Conference on Artificial Intelligence **34 (10)**, 13933-13934.

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

Tadros, T., G. P. Krishnan, R. Ramyaa and M. Bazhenov (2022). "Sleep-like unsupervised replay reduces catastrophic forgetting in artificial neural networks." Nat Commun **13**(1): 7742.

Vanini, G., R. Lydic and H. A. Baghdoyan (2012). "GABA-to-ACh ratio in basal forebrain and cerebral cortex varies significantly during sleep." Sleep **35**(10): 1325-1334.

Walker, M. P. and R. Stickgold (2004). "Sleep-dependent learning and memory consolidation." Neuron **44**(1): 121-133.

Wei, Y., G. Krishnan and M. Bazhenov (2016). "Synaptic Mechanisms of Memory Consolidation during Sleep Slow Oscillations." Journal of Neuroscience **36**(15): 4231-4247.

Wei, Y., G. P. Krishnan, M. Komarov and M. Bazhenov (2018). "Differential roles of sleep spindles and sleep slow oscillations in memory consolidation." PLoS Comput Biol **14**(7): e1006322.

Wei, Y., G. P. Krishnan, L. Marshall, T. Martinetz and M. Bazhenov (2020). "Stimulation Augments Spike Sequence Replay and Memory Consolidation during Slow-Wave Sleep." J Neurosci **40**(4): 811-824.

Wierzynski, C. M., E. V. Lubenov, M. Gu and A. G. Siapas (2009). "State-dependent spike-timing relationships between hippocampal and prefrontal circuits during sleep." Neuron **61**(4): 587-596.

Wilson, M. A. and B. L. McNaughton (1994). "Reactivation of hippocampal ensemble memories during sleep." Science **265**(5172): 676-679.

FIGURES

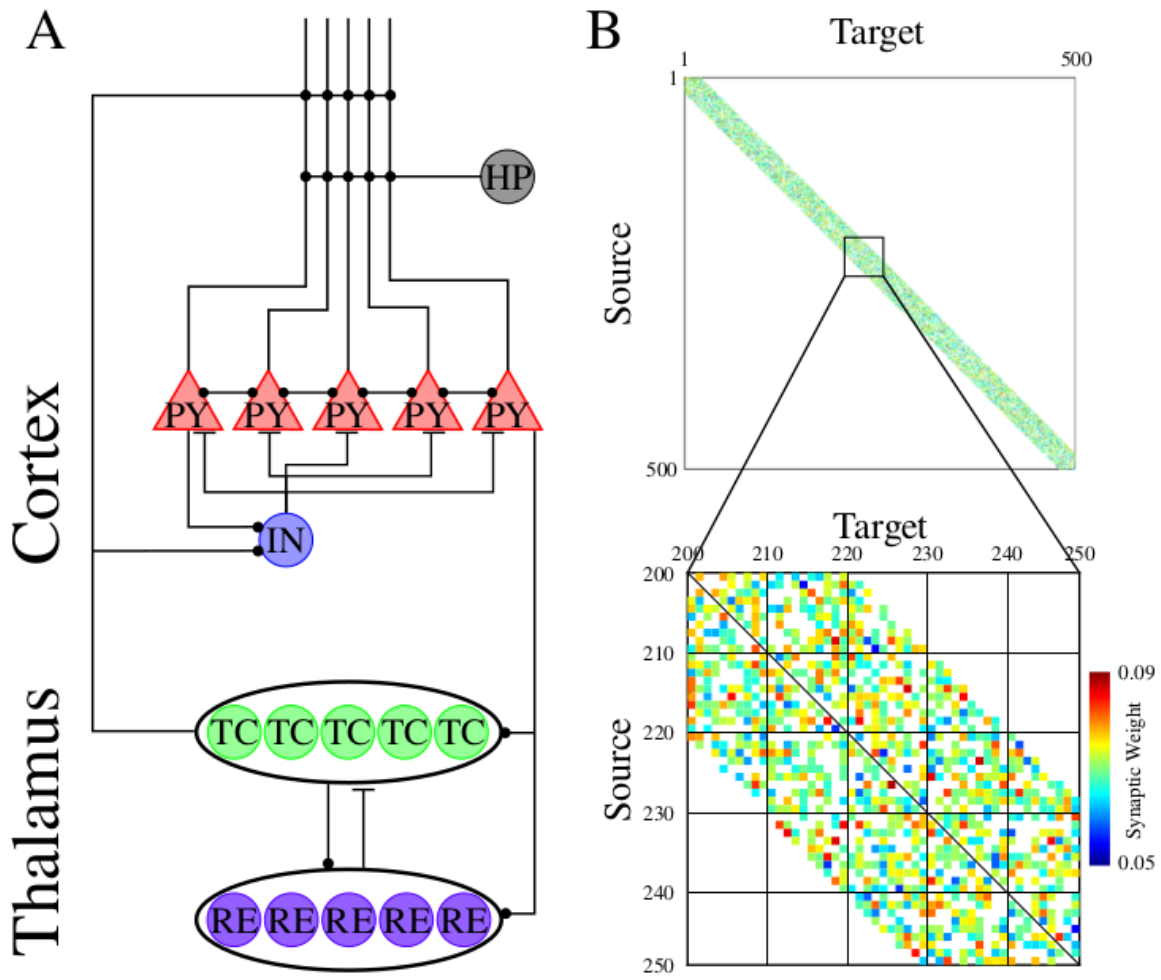


Figure 2.1. Network Architecture. (**A**) Basic network architecture (PY: excitatory pyramidal neurons; IN: inhibitory interneurons; TC: excitatory thalamocortical neurons; RE: inhibitory thalamic reticular neurons). Excitatory synapses are represented by lines terminating in a dot, while inhibitory 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.

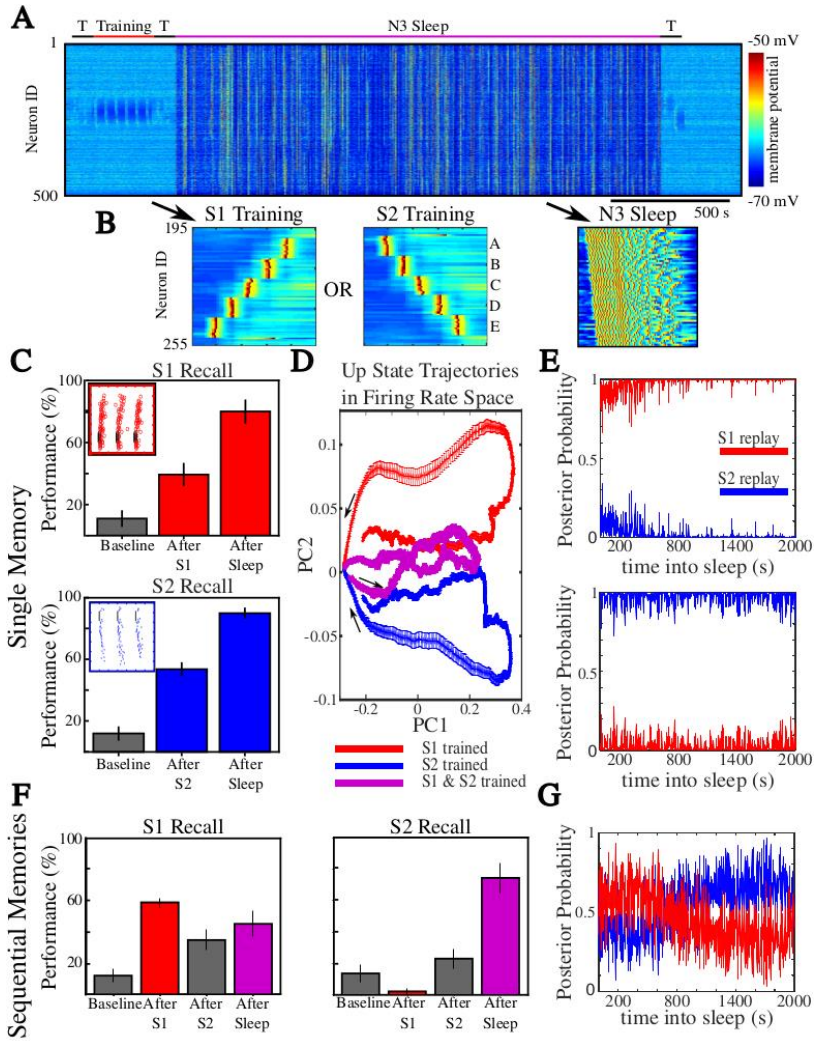


Figure 2.2. Sleep rescues interference induced by sequential training. (A) Network activity during an example simulation depicting all 500 PYs on the y-axis and the membrane potential (color scale) of each over time (x-axis). The network undergoes testing periods (T; black) at baseline, after Training (red), and after N3 Sleep (purple). **(B)** Examples of network activity during one bout of S1 (left) or S2 (middle) training, and during a single up state of N3 Sleep (right). **(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 firing rate trajectory in PC space during up states of a network trained on S1 (red), S2 (blue), or S1&S2 (purple) before sleep. **(E)** 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. **(F)** Same as (C) but for a simulation where S2 was trained sequentially after S1 and prior to sleep; S1 recall (left), S2 recall (right). **(G)** Same as (F) but for the sequential training simulation.

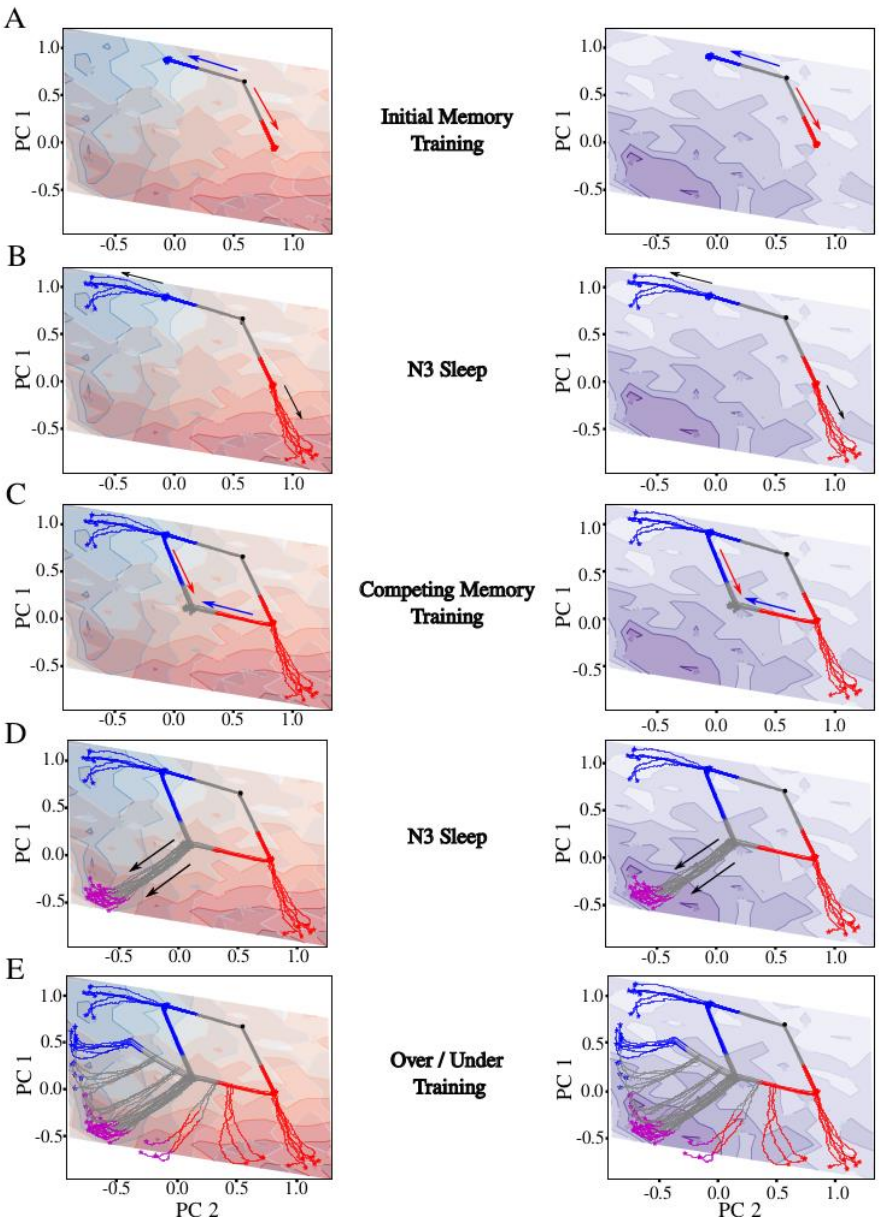


Figure 2.3. Synaptic performance landscape reveals multi-stability and fine-tuning. All panels show the evolution of the network through dimensionality reduced synaptic weight space. Contour lines and coloring correspond to single memory recall performance (left panels) for S1 (red) and S2 (blue), and to joint memory recall performance (right panels; purple). 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)** N3 sleep moves each network further along its current memory manifold. **(C)** Sequential training by the competing memory moves the network to a gray central region where neither memory can be recalled. **(D)** N3 sleep moves the networks onto the joint memory manifold (purple trajectories) where both can be recalled. **(E)** Examples of under/over-training S2 following S1 training reveals the necessity of fine-tuning the training durations for sleep to evolve the network to the joint memory manifold.

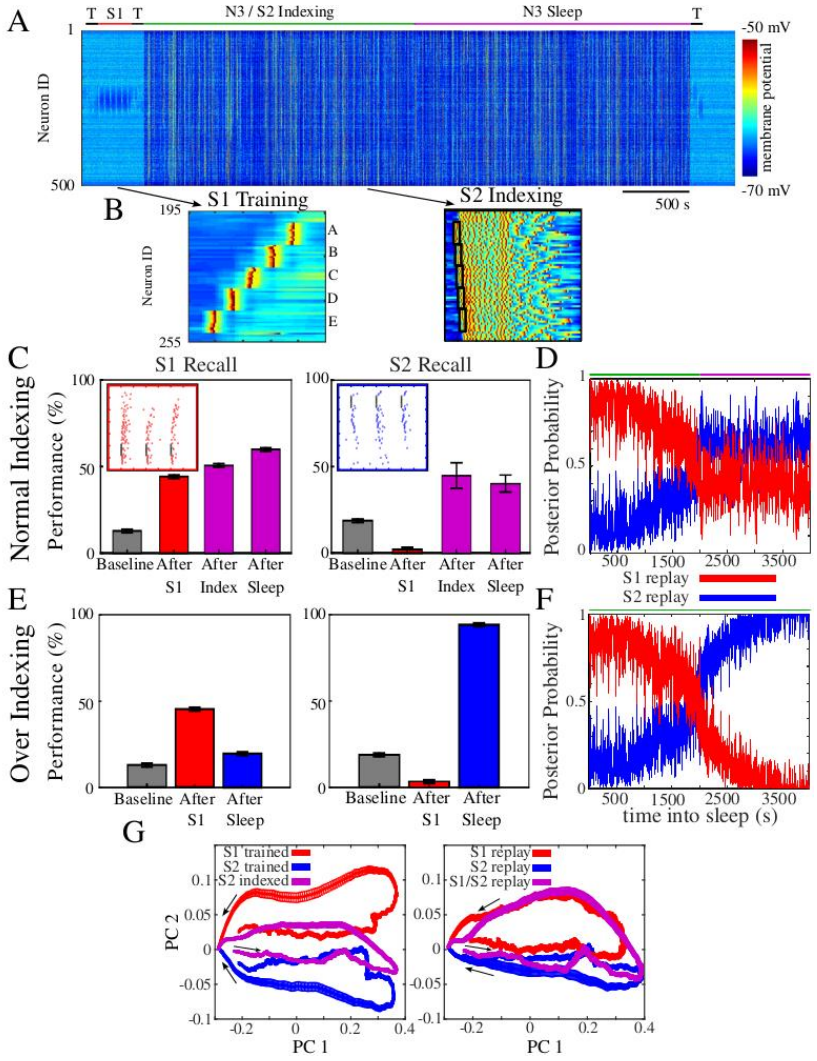


Figure 2.4. Hippocampal indexing during sleep induces consolidation without interference. **(A)** Network activity during the simulation in which the network undergoes periods of testing (T; black), training (S1; red), N3 Sleep with S2 Indexing (green), and N3 sleep (purple). **(B)** Examples of network activity during one bout of S1 training (left), and during a single up state of N3 Sleep (right) where indexing is simulated at the beginning of the up state (black boxes). **(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 (purple) 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. **(G)** 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. **(G)** Left panel shows the average firing rate trajectory in PC space during up states after single memory training (S1 - red; S2 - blue) and during normal indexing (purple). Right panel shows average trajectories during normal indexing grouped according to the replay probability of each up state (>95% S1 = red; >95% S2 = blue; otherwise = purple).

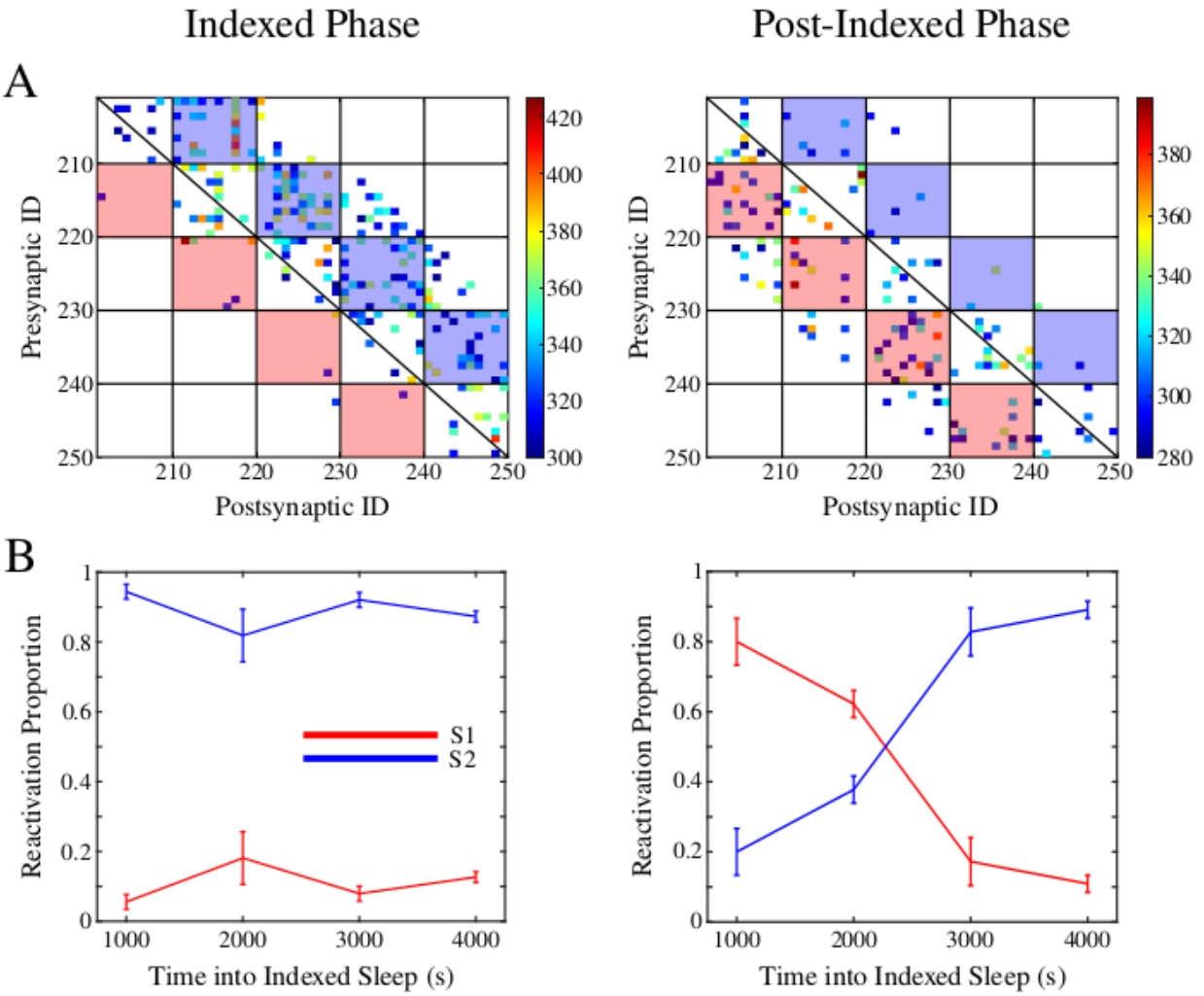


Figure 2.5. Indexing causes interleaved memory reactivation within individual up states. **(A)** Synaptic reactivation during the indexed (left) and post-indexed (right) phases for the top 30% of reactivated synapses. The Y-/X-axis correspond to the pre/postsynaptic neuron IDs, and the color scale indicates the number of Up-states a particular synapse experienced a net potentiation event during the referenced phase. Red and blue squares indicate synapses considered relevant for S1 and S2 respectively. **(B)** The relative proportions of S1 (red) and S2 (blue) relevant synaptic reactivations during the indexed (left) and post-indexed (right) phases of sleep, averaged over 1000 s in non-overlapping windows. Error bars depict standard deviation.

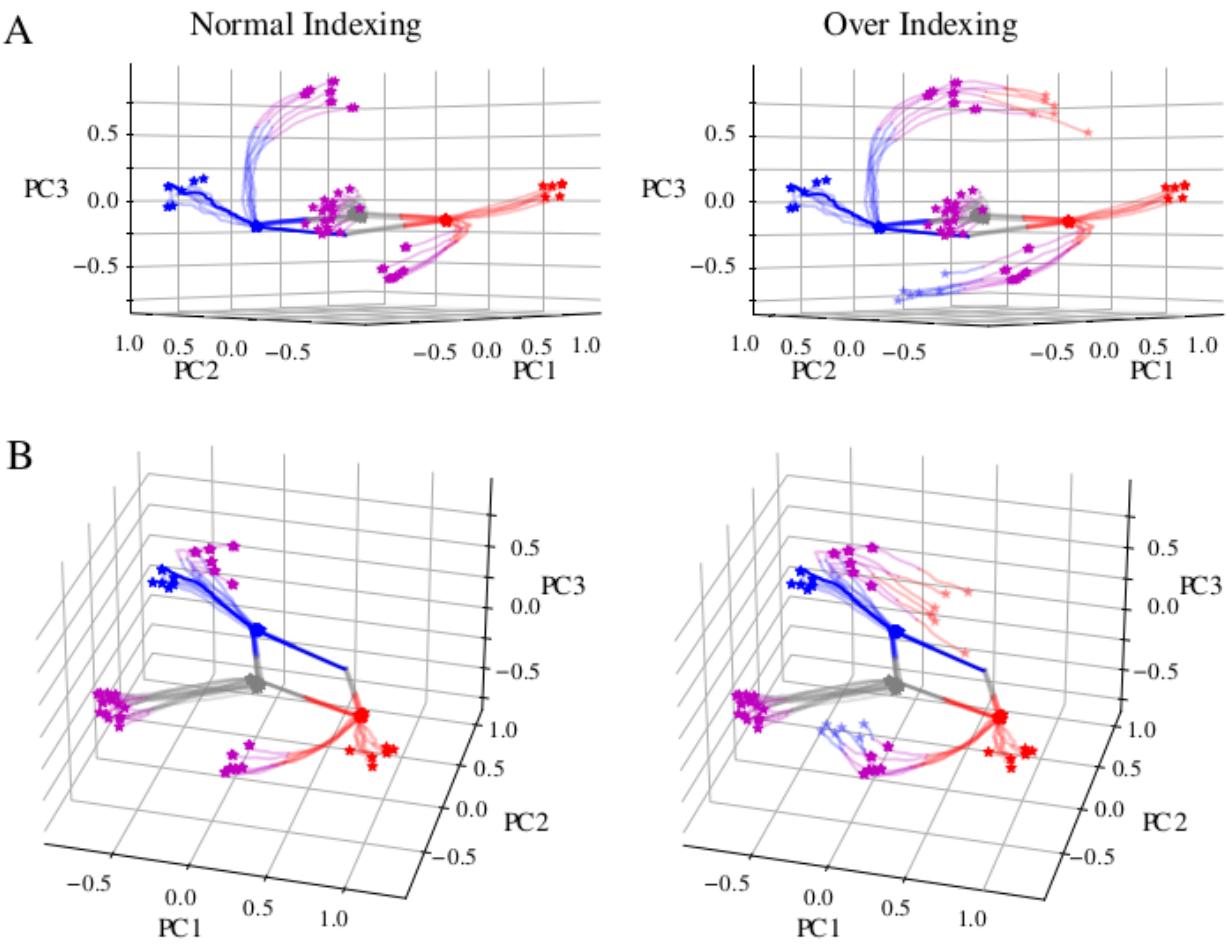


Figure 2.6. Indexing evolves the network along the current memory manifold toward the intersection. (A) Evolution of the networks during normal indexing (left) shows the network moves along its current memory manifold (i.e. S1 - red; S2 - blue), primarily in the PC3 dimension, until it reaches an intersection of the manifolds (purple) that is stable with subsequent sleep. Evolution during over indexing (right) 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. **(B)** Same as (A) but with the plot rotated 45 degrees about the PC3 axis.

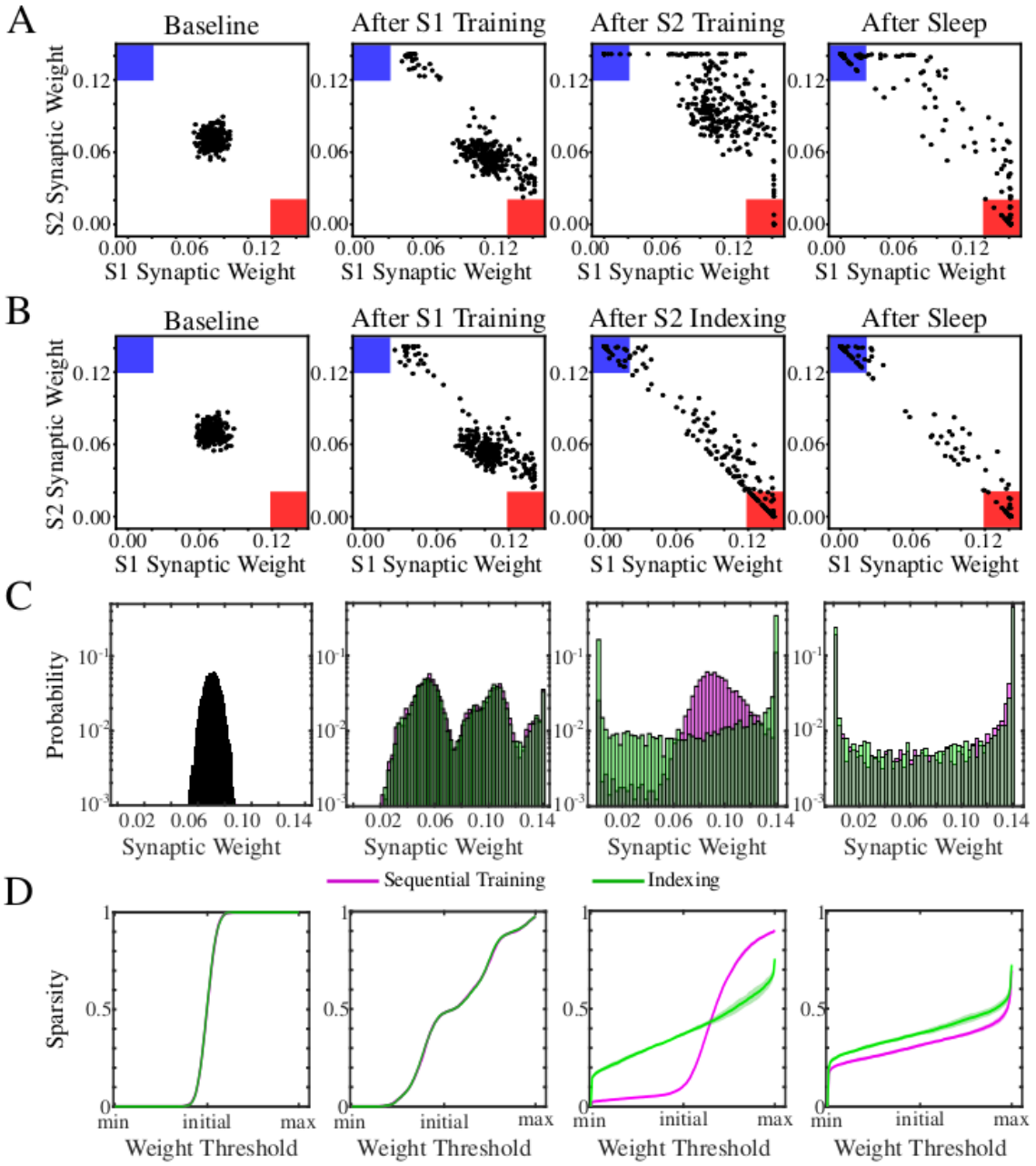


Figure 2.7. Indexing leads to sparser memory representations than sequential training and sleep. (A-B) Each panel shows scatter plots of synaptic weights between all PYs in the trained region that have bi-directional synapses between them. The weight in the S1 direction is on the x-axis, while that in 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 (purple) and normal indexing (green) with the y-axis on a log-scale. **(D)** Average sparsity across a dense set of synaptic weight thresholds for simulations with sequential training (purple) and normal indexing (green).

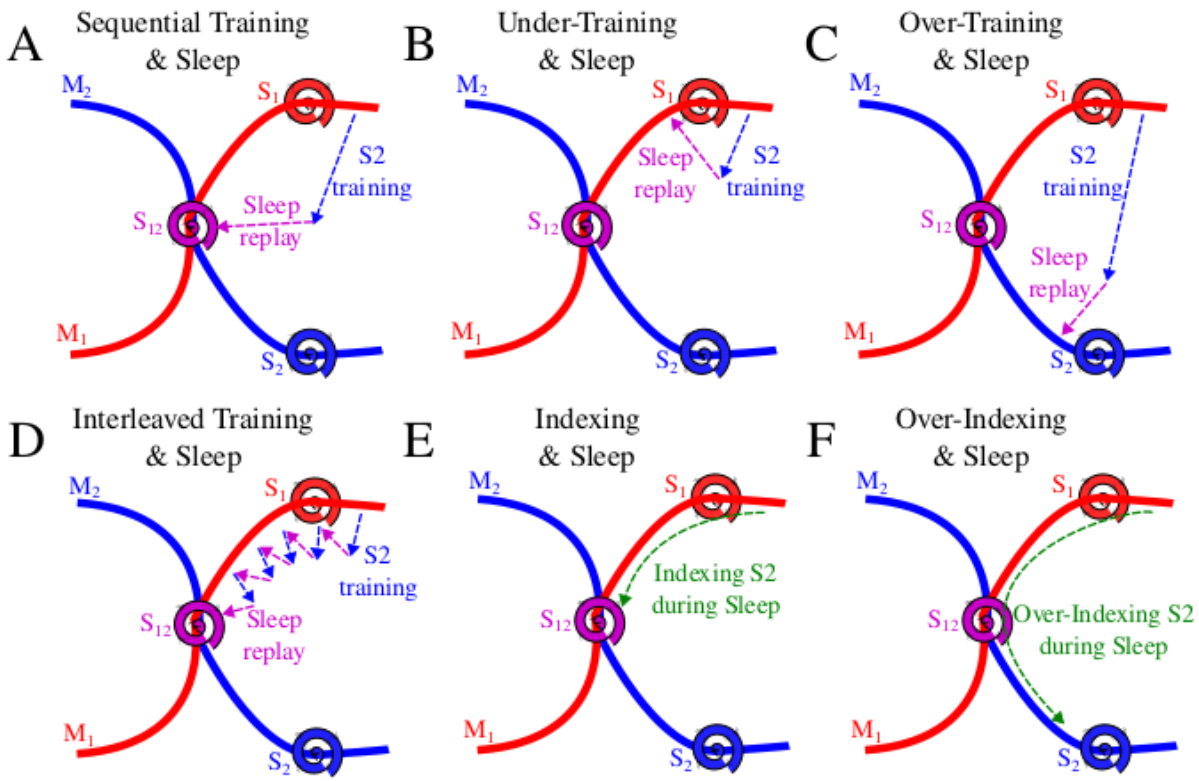


Figure 2.8. Systems consolidation prevents interference through interleaved replay within up states. Cartoon schematics illustrating the network dynamics across memory manifolds for **(A)** Sequential Training, **(B)** Under-Training, **(C)** Over-Training, **(D)** Interleaved Training, **(E)** Indexing, and **(F)** Over-Indexing. Spirals represent regions on manifolds which appear to be stable and attracting under sleep dynamics.

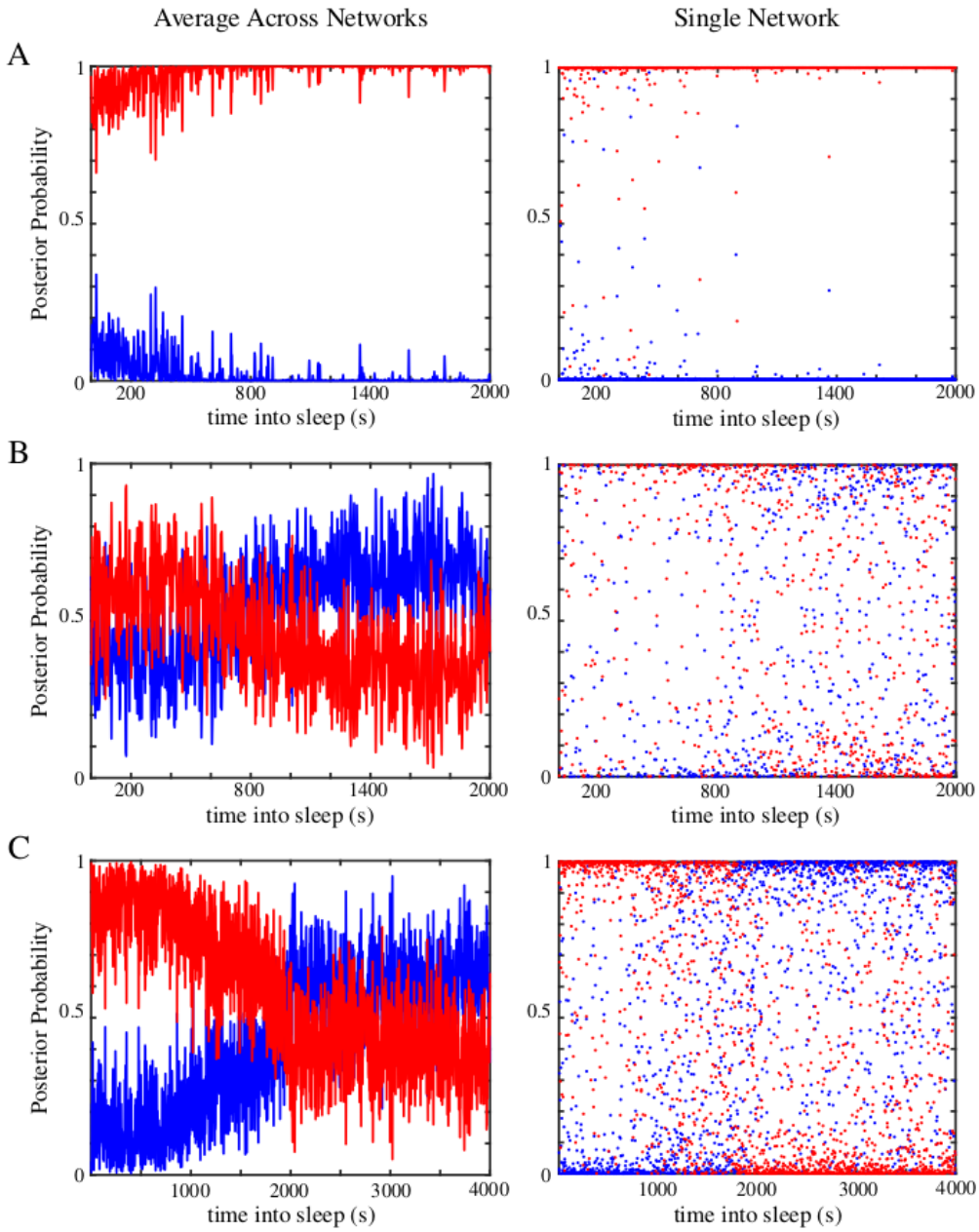


Figure 2.9. (S1) Classification of memory replay on individual UP states is robust in a single network. Averaged (left) and single network (right) replay probabilities for **(A)** S1 training only; **(B)** sequential training; and **(C)** normal indexing. The single network plots highlight that the majority of individual Up-states are robustly classified as either S1 or S2 across all conditions.

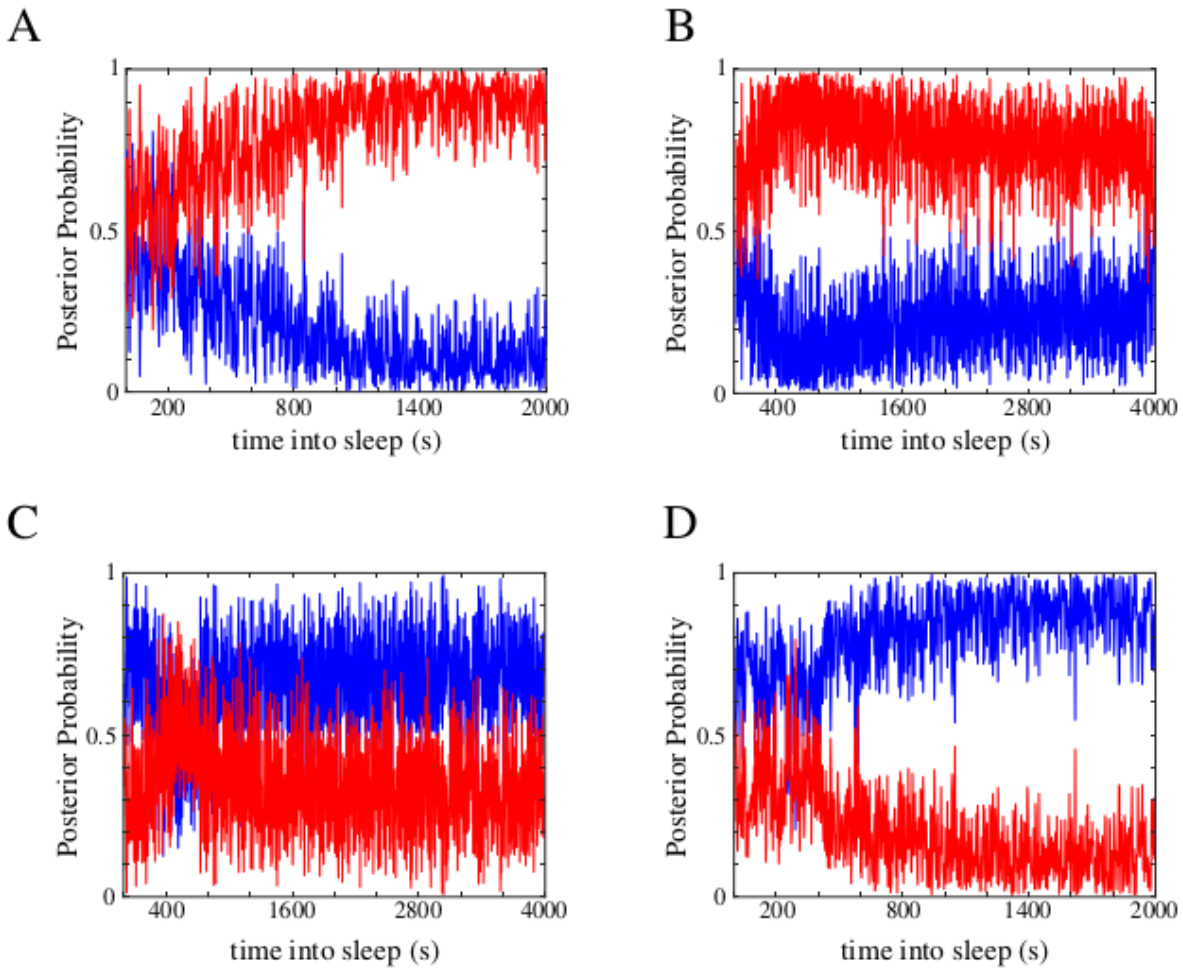


Figure 2.10. (S2) Replay probabilities during sleep following under/over-training. (A) Following significant undertraining of S2 the network begins sleep in a mixed replay state and slowly converges to an S1 dominated replay state. **(B)** Following moderate undertraining of S2 the network diverges from a mixed replay state towards an S1 dominated replay state before slowly relaxing towards a more mixed state. **(C)** Following moderate overtraining of S2 the network's replay state briefly oscillates about and then remains in a mixed replay state. **(D)** Following significant overtraining of S2 the network slowly converges towards an S2 dominated replay state from its initial mixed replay state.

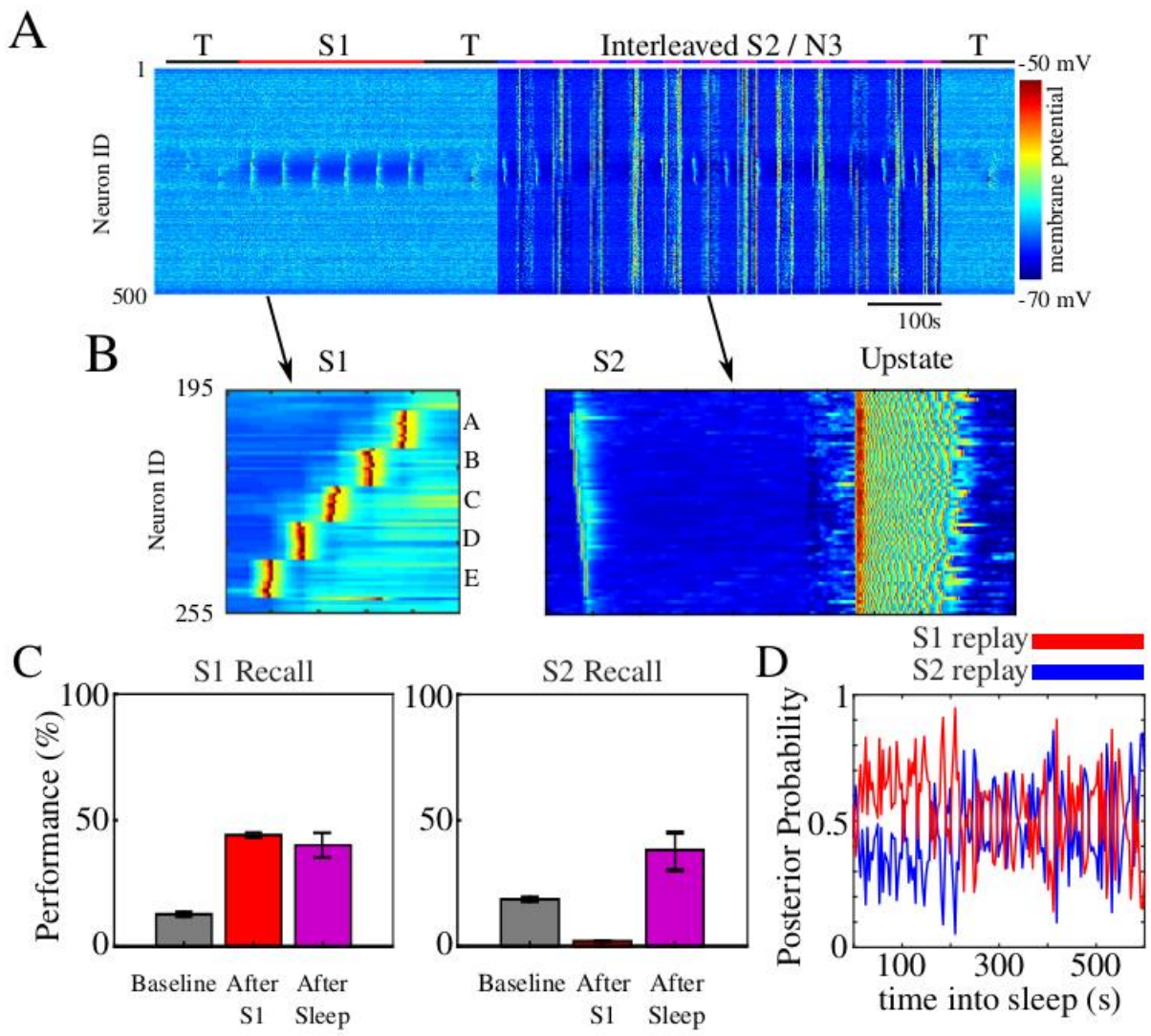


Figure 2.11. (S3) Interleaving S2 training with N3 sleep following S1 training allows for recall of both old and new memories. (A) Network activity during the simulation in which the network undergoes periods of testing (T; black), training (S1; red), and 12 cycles of interleaved S2 training (blue) with N3 Sleep (purple), 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 an N3 sleep period (right). **(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/N3 sleep.

Chapter 2, in full, is currently being prepared for submission for publication of the material. Gonzalez, Oscar C.; Golden, Ryan; Delanois, J. Erik; Bazhenov, Maxim. The dissertation author was the co-primary investigator, along with Oscar C. Gonzalez, and the primary author of this paper.

Systems/Circuits

Multielectrode Cortical Stimulation Selectively Induces Unidirectional Wave Propagation of Excitatory Neuronal Activity in Biophysical Neural Model

Alma S. Halgren,[1,2] **Zarek Siegel,**[1,3] **Ryan Golden,**[1,3] **and Maxim Bazhenov**[1,3]

[1]Department of Medicine, University of California - San Diego, La Jolla, California 92093-7374, [2]Department of Integrative Biology, University of California - Berkeley, Berkeley, California 94720, and [3]Neurosciences Graduate Program, University of California - San Diego, La Jolla, California 92093-7374

Cortical stimulation is emerging as an experimental tool in basic research and a promising therapy for a range of neuropsychiatric conditions. As multielectrode arrays enter clinical practice, the possibility of using spatiotemporal patterns of electrical stimulation to induce desired physiological patterns has become theoretically possible, but in practice can only be implemented by trial-and-error because of a lack of predictive models. Experimental evidence increasingly establishes traveling waves as fundamental to cortical information-processing, but we lack an understanding of how to control wave properties despite rapidly improving technologies. This study uses a hybrid biophysical-anatomical and neural-computational model to predict and understand how a simple pattern of cortical surface stimulation could induce directional traveling waves via asymmetric activation of inhibitory interneurons. We found that pyramidal cells and basket cells are highly activated by the anodal electrode and minimally activated by the cathodal electrodes, while Martinotti cells are moderately activated by both electrodes but exhibit a slight preference for cathodal stimulation. Network model simulations found that this asymmetrical activation results in a traveling wave in superficial excitatory cells that propagates unidirectionally away from the electrode array. Our study reveals how asymmetric electrical stimulation can easily facilitate traveling waves by relying on two distinct types of inhibitory interneuron activity to shape and sustain the spatiotemporal dynamics of endogenous local circuit mechanisms.

Key words: **biophysical model; cortical simulation; multielectrode arrays; network model; stimulation**

Significance Statement

Electrical brain stimulation is becoming increasingly useful to probe the workings of brain and to treat a variety of neuropsychiatric disorders. However, stimulation is currently performed in a trial-and-error fashion as there are no methods to predict how different electrode arrangements and stimulation paradigms will affect brain functioning. In this study, we demonstrate a hybrid modeling approach, which makes experimentally testable predictions that bridge the gap between the microscale effects of multielectrode stimulation and the resultant circuit dynamics at the mesoscale. Our results show how custom stimulation paradigms can induce predictable, persistent changes in brain activity, which has the potential to restore normal brain function and become a powerful therapy for neurological and psychiatric conditions.

Introduction

Brain stimulation is widely used in both experimental and clinical settings. In basic research, it is used to probe neural function by disrupting or hyperactivating local brain processing (E. Halgren et al., 1978; Salzman et al., 1990; Tehovnik et al., 2002). In clinical settings, direct manipulation of activity via stimulation has also been shown to be effective in the treatment of several neurological and psychiatric disorders. Deep brain stimulation (DBS) has been successful in the treatment of movement disorders, such as Parkinson's disease (Blumenfeld and Bronte-Stewart, 2015; Baizabal-Carvallo and Alonso-Juarez, 2016; Papageorgiou et al., 2017), depression (Mayberg et al., 2005; Schlaepfer et al., 2008), and obsessive-compulsive disorder (Abelson et al., 2005; B. D. Greenberg et al., 2006). Superficial cortical stimulation is an effective therapy for epilepsy (Nagaraj et al., 2015) and stroke patients (Hummel and Cohen, 2006). Increasingly, electrical

Received Sep. 1, 2021; revised Dec. 27, 2022; accepted Jan. 13, 2023.

Author contributions: A.S.H., Z.S., R.G., and M.B. designed research; A.S.H. and Z.S. performed research; A.S.H., Z.S., and R.G. analyzed data; A.S.H. wrote the first draft of the paper; A.S.H., R.G., and M.B. edited the paper; A.S.H., Z.S., and M.B. wrote the paper.

This work was supported by National Institutes of Health, National Institute of Neurological Disorders and Stroke Grant 1R01NS109553 and National Institutes of Health, National Institute of Mental Health Grant 1RF1MH117155.

The authors declare no competing financial interests.

Correspondence should be addressed to Maxim Bazhenov at mbazhenov@ucsd.edu.

https://doi.org/10.1523/JNEUROSCI.1784-21.2023

stimulation has also shown promise in both the restoration and enhancement of critical cognitive functions, such as memory. DBS has been shown experimentally to enhance memory encoding when applied during learning (Suthana et al., 2012), and closed-loop stimulation protocols have proven to be effective during periods of poor memory encoding as well as during memory recall (Ezzyat et al., 2018; Kahana et al., 2018; Kucewicz et al., 2018a,b).

While brain stimulation is sometimes conceptualized as disrupting pathological activity to restore normal activity, increasingly the explicit goal is to directly generate normal activity. Experimental evidence supports traveling waves as critical to normal brain activity. These propagating waves are fundamental to brain information-processing as they coordinate neural behavior across all spatial scales, from within-layer to whole-brain interactions, as well as across temporal scales, from tens to hundreds of milliseconds. By mediating communication across multiple brain areas, propagating activity putatively performs a variety of cognitive functions, such as the processing of visual stimuli or long-term memory consolidation (Muller et al., 2018). For example, sleep spindles traveling across the cortical surface at multiple scales have been hypothesized to synchronize convergent co-firing of neurons, resulting in spike timing-dependent plasticity and consequent memory consolidation (Dickey et al., 2021). Similarly, the alpha rhythm which modulates visual processing appears to be a traveling wave from association to primary areas (M. Halgren et al., 2019). Accordingly, the ability to predict and control traveling waves has far-reaching implications for improving and controlling cognitive function.

Currently, there is no method for reliably generating directional traveling waves with electrical stimulation. Past efforts to develop new paradigms of stimulation which reinstate particular brain activity states have largely depended on trial and error. Recently, we described a method for modeling the effects of cortical stimulation that enables one to predict the consequences of stimulation *in silico*, and thereby develop stimulation protocols that achieve desired results *in vivo* (Komarov et al., 2019). This earlier study was limited to the effects of stimulation of a single electrode and thus did not evoke directional propagation. In this new study, we describe an initial attempt to model a multielectrode stimulation paradigm that produces unidirectional traveling waves in the cortex. With multielectrode arrays increasingly entering clinical practice (Ha et al., 2017), our model harnesses the additional complexity and control during stimulation that multielectrode protocols allow for.

This modeling approach includes two phases. In the first phase, a biophysical model is used to predict spiking probability in response to a spatially-varied electric field potential in reconstructed rat somatosensory cortical neurons obtained from www.neuromorpho.org (Ascoli et al., 2007). We found that the hyperpolarization or depolarization of individual neurons varied according to cell type and cortical depth, and also varied with respect to the polarity of the applied electric field. The diversity in excitation responses underlies the propagating wave activity that we observed in the second phase of the model. In this second phase, we constructed a Hodgkin–Huxley model of a rat somatosensory cortical network composed of multiple interconnected cortical columns, each containing a circuit of inhibitory and excitatory cells connected within and across cortical layers. Approximating stimulation effects using the activation probabilities calculated in Phase 1, we found that fast inhibitory activity, coupled with excitatory cells' preference for anodal stimulation, resulted in a unidirectional, propagating wave of activity. Importantly, we found that the temporal component of the wave (the timing of cell firing within the excitatory-inhibitory feedback loop) resulted from an interaction between pyramidal and basket inhibitory cells, suggesting that a brief pulse of suprathreshold amplitude is sufficient to facilitate oscillatory activity thought to be endogenous to cortical columnar circuits. The spatial component (asymmetry) of the wave, however, was shown to depend on the unique activation profile of Martinotti inhibitory cells under this stimulation paradigm relative to the pyramidal and basket cells. These results suggest a simple multielectrode pattern for evoking traveling waves and provide testable predictions for experimental confirmation and parameter optimization.

Table 1. Summary of datasets with reconstructed cells[a]

Cell type	No. of reconstructions	Reference	Strain (age)
Pyramidal cells (Layer II/III)	21	Tehovnik et al., 2006	Wistar (P20-P25)
Pyramidal cells (Layer IV)	11	Traub et al., 1994	Wistar (P19-P21)
Pyramidal cells (Layer Va)	14	Wang et al., 2002	Wistar (P20-P21)
Basket cells (Layer II/III)	96	Wester and Contreras, 2012	Wistar (P13-P15)
Basket cells (Layer IV)	82	Wester and Contreras, 2012	Wistar (P13-P15)
Basket cells (Layer V)	57	Wester and Contreras, 2012	Wistar (P13-P15)
Martinotti cells (Layer II/III)	13	Douglas and Martin, 1991	Wistar (P13-P16)
Martinotti cells (Layer V)	7	Douglas and Martin, 1991	Wistar (P13-P16)
Horizontal interneurons (Layer I)	59	Wang et al., 2004	Wistar (P13-P16)
Descending interneurons (Layer I)	29	Wang et al., 2004	Wistar (P13-P16)
Small interneurons (Layer I)	27	Wang et al., 2004	Wistar (P13-P16)

[a]Cell types (from www.neuromorpho.org) that were used in the biophysical component of the model.

Materials and Methods

Cell reconstruction selection. All neuronal cell reconstructions were chosen from publicly available datasets on www.neuromorpho.org (Wang et al., 2002, 2004; Staiger et al., 2004; Schubert et al., 2006; Ascoli et al., 2007; Muralidhar et al., 2013). The types of cells and their respective datasets are listed in Table 1. We used multiple cell reconstructions for each cell type to account for anatomical diversity. We used experimental measurements to approximate the cutoff depths for each layer (see Fig. 1*a*) (Markram et al., 2004; Wang et al., 2004).

Calculating the electric field potential generated by the electrode array. The electrode array modeled in this study was composed of three square electrodes (each 150 μm × 150 μm) placed linearly on the surface of the cortex. Two electrodes had negative current (−75 μA each) and one electrode had positive current (150 μA) (see Fig. 1*a*), and stimulation was applied for 200 μs. This was done to adhere to the constraints of clinical applications of electrical stimulation which require a net neutral current to be delivered to the tissue. These values are in accordance with common experimental parameters (Ha et al., 2017). Assuming that the current sources are homogeneous square electrodes, we calculated the electric field potential of each electrode as follows:

$$\Phi(X,Y,Z) = \frac{\rho_e I}{4\pi A^2} \iint_{-A/2}^{A/2} \frac{\mathrm{d}x\mathrm{d}y}{\sqrt{(X-x)^2+(Y-y)^2+Z^2}} \quad (1)$$

Here I is net current, ρ_e is extracellular resistivity, and A is the length of the square electrode edge (see Fig. 1*a*). In this study, $A = 150\,\mu$m and net current I is either −75 or 150 μA. The derivation of this formula can be found in our previous work (Komarov et al., 2019). We summed all three electric field potentials at each point in space to determine the overall electric field potential.

Estimating the activating function. This paper used a similar approach to that of Komarov et al. (2019) to estimate the firing probability for each neuronal reconstruction. The basis of this approach is the activating function, a formula derived from cable theory that describes the effective transmembrane current that arises because of extracellular

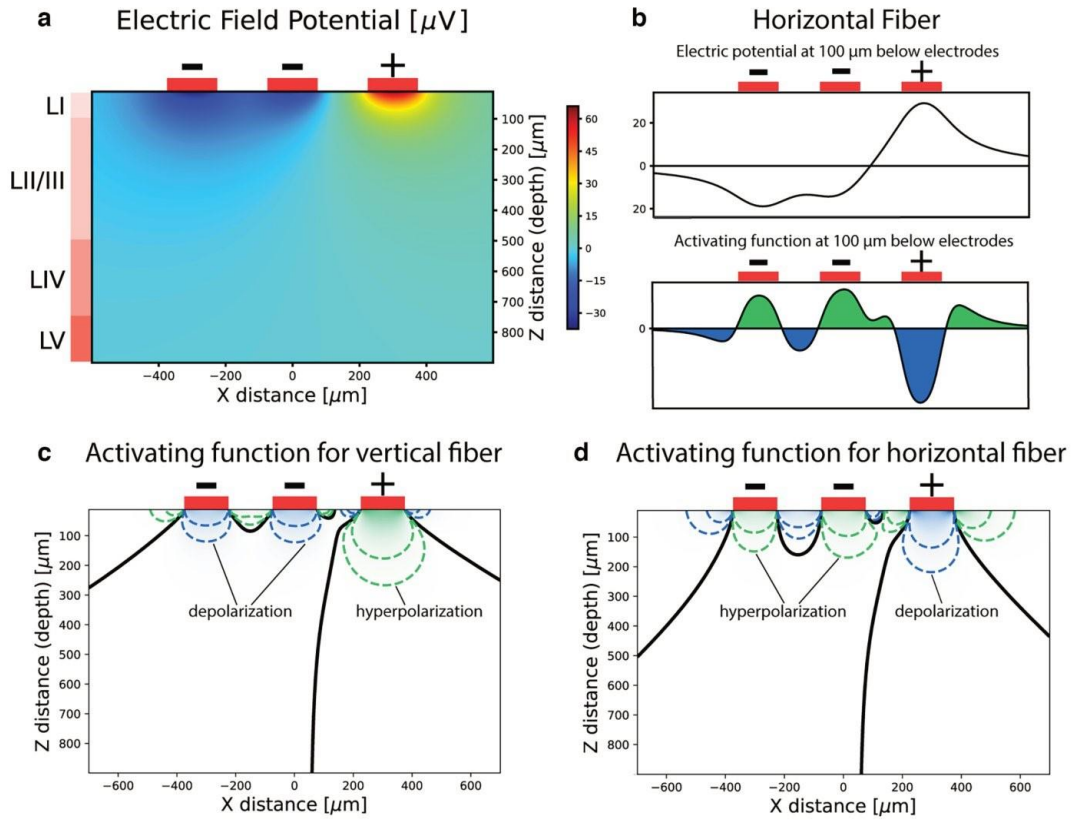


Figure 1. Electric potential and activating function in the plane $Y = 0$. ***a***, Schematic of the electrode configuration and electric potential in the plane $Y = 0$. Each electrode is square-shaped and measures 150 μm × 150 μm, and the spacing between each electrode is also 150 μm. The two leftmost electrodes are negative (cathodal) and deliver negative 75 μA of current each, while the rightmost electrode is positive (anodal) and delivers 150 μA of current, and the stimulation period is 200 μs. Layer depths are approximated from experimental measurements of rat cortex (Markram et al., 2004; Wang et al., 2004; Defelipe et al., 2012) as follows: layer I is 0-100 μm, layer II/III is 100-500 μm, layer IV is 500-750 μm, and layer V is 750-900 μm. ***b***, Axial potential along a horizontal axonal fiber located 100 μm below the cortical surface. The activating function is the second spatial derivative of the electric potential. ***c***, ***d***, As a result of ***b***, horizontal fibers are activated by the anode and hyperpolarized by the cathodes (***d***). The opposite is true for vertical fibers (***c***). The vast majority of direct stimulation occurs in layers I and II/III because of the decay of electric potential with depth.

electrical stimulation. This was used to calculate the transmembrane voltage in each axonal segment of every cell reconstruction. We then applied a threshold of activation to determine whether an axonal response was triggered. This threshold was drawn from *in vivo* experiments which define the threshold injected current I required to induce a threshold effective current f at the axonal initial segment located at a distance d from the electrode (Douglas and Martin, 1991; R. J. Greenberg et al., 1999). In previous work (Komarov et al., 2019), we simulated one such experiment by computing the activation current f at the axon initial segment of a layer II/III pyramidal cell while varying the stimulation current I and distance d. This simulation used a 200 μs-duration stimulating pulse, which is typical of similar *in vivo* experiments (Douglas and Martin, 1991; R. J. Greenberg et al., 1999). The value $f = f_{th} = 3\,\text{pA}/\mu\text{m}^2$ fully replicated the experimentally-observed current-distance relationship across varying stimulation currents and distances. Thus, this threshold value f_{th} was used to determine whether each axonal segment was activated by induced transmembrane current.

Since we compute the activating function in each small compartment composing an entire axonal arborization, jitter in the edges of the anatomical reconstruction could introduce numeric noise in our calculation. To minimize this issue, we estimate the direction of each axonal component (the mini segment forming a compartment in the reconstruction) using the position in space of neighboring compartments up to 10 μm away. The estimated direction is then crucial to computing the activating function, which, by definition, is calculated along the axonal element direction.

An important note is that this protocol neglects the effects of axonal branching at adjacent segments. Instead, it acts as though the axon continued, unbranched, in a direction given by the sum of the two orientation vectors of the bifurcated segments. However, given that we base our activation probabilities (see Computing the average activation probability per cell type below) on the total length of activated segments within a reconstruction, and then average across all reconstructions of that cell type, we believe that this impact is negligible.

Computing the average activation probability per cell type. To determine whether a given neuronal reconstruction would be activated by applied current, we used the activating function to calculate how many axonal segments had above-threshold transmembrane current values that could initiate an axonal action potential (assumed at nodes of Ranvier); these above-threshold axonal segments are collectively called the trigger area. The activating function threshold was set to $f_{th} = 3\,\text{pA}/\mu\text{m}^2$ for myelinated axons and to $f_{th} = 60\,\text{pA}/\mu\text{m}^2$ for unmyelinated

axons, since unmyelinated segments are significantly less excitable as they have fewer sodium channels (Cogan et al., 2016). In this model, we assumed pyramidal and basket cells are myelinated and Martinotti and layer I interneurons are unmyelinated based on experimental data (Thomson et al., 2002; Wang et al., 2004; Tomassy et al., 2014). For each cell reconstruction at a given point in space (i.e., a given coordinate in the *x*-*z* plane; see Fig. 1*a*), we found the probability of firing as outlined in prior work (Komarov et al., 2019). Given that variation in cell position and orientation within each cortical layer are present in nature, we additionally average across neuronal rotations and vertical shifts. Thus, we computed the average activation probability for each cell type/cortical layer pairing as follows: for each cell reconstruction, we first positioned the soma at a point along the *x* axis within its cortical layer. We then performed four rotations (0°, 90°, 180°, and 270°) about the vertical axis of the cell reconstruction in three-dimensional (3D) space, and at every rotation we calculated the likelihood of activation by computing the activating function across the axonal arbor, as outlined above. We then averaged across these four probabilities and set the result as the activation probability for said point. After this, we incrementally shifted the soma of the reconstruction vertically within its cortical layer and found the mean activation probability at each point. We averaged across all vertical shifts within the cortical layer to obtain an approximate spiking probability for this cell reconstruction at this point along the *x* axis. Along the *x* axis, we computed the activation probability of all cell reconstructions and then averaged across all cell reconstructions within a given cell type and cortical layer (see Fig. 3).

Computational model of the cortical circuit. The network model is composed of 11 interconnected cortical columns, where each column contains layer I interneurons and pyramidal, basket, and Martinotti cells from layers II-V (same as the biophysical analysis). The number of cells within each column is outlined in Table 2. This balance of excitatory to inhibitory cell types approximates the true cell composition of the rat somatosensory cortex, where pyramidal cells are the primary excitatory cells and basket and Martinotti cells comprise the majority of inhibition within and across layers (Markram et al., 2004; Wang et al., 2004). Cells were constructed to only spike if receiving synaptic input or electrical stimulation. Each cell behaves according to Hodgkin–Huxley dynamics, with a handful of parameters differentiating excitatory and inhibitory cells. Basket cells were modeled as fast-spiking cells, while all other cell types were modeled as regular-spiking cells with spike rate adaptation. Inhibitory cells fired more quickly than excitatory cells in response to activation because all interneurons were modeled as having a lower leak current than excitatory cells (Santos et al., 2012). Some additional parameters were as follows: a fast Na^+-K^+ spike generating mechanism (all cells), a high-threshold activated Ca^{2+} current (for pyramidal cells), and a slow calcium-dependent potassium (AHP) current (for regular-spiking cells).

The network architecture and function mirror that of Komarov et al. (2019) with the following exceptions: first, our network contains multiple columns while that in Komarov et al. (2019) only contains one; second, the initial activation probabilities in our network are derived from the first phase of our model (see Fig. 3) instead of from the probabilities calculated in prior work (Komarov et al., 2019); and third, our model contains slightly different cells (e.g., layer I cells) than those included previously (Komarov et al., 2019). The Hodgkin–Huxley equations that govern the dynamics of our model can be found in our prior work (Komarov et al., 2019).

Initially, the network runs without stimulating input for 200 ms to simulate preexisting activity. Then the network is stimulated and runs for an additional 500 ms. To simulate electrical stimulation, we used the binary term I_i^{ext} to inject above-threshold current into a subset of randomly chosen neurons within each cell type/cortical layer pairing such that the fraction of neurons induced to spike corresponds to the activation probabilities calculated in the biophysical phase of the model (see Fig. 3). For neurons in columns 1-3 and 9-11, the activation probabilities were set to zero. This was because the electric field generated by the electrode design was effectively null at locations this far from the stimulating electrodes. The term $\eta\xi_i(t)$ models spontaneous background activity as a white noise process (ξ) with SD η. All model parameters are listed in Komarov et al. (2019, their Table S2) (unless specified in the description of simulations), and the network structure and connectivity are described in Table 2 and Table 3, respectively. Cells were synaptically coupled by excitatory (AMPA) and inhibitory ($GABA_A$) connections. The strength and probability of connections between layers and cell types were set according to a canonical cortical circuit (Thomson et al., 2002).

Average network activity. To quantify network behavior across 50 simulations per current strength, we averaged the percentage of spiking across all cell types at each cortical column (see Fig. 5*c*).

Table 2. Structure of the network[a]

Cortical cell type	Layer	Cells/column	Total cells
IN	I	12	132
PY	II/III	100	1100
BC	II/III	100	1100
MC	II/III	24	264
PY	IV	12	132
BC	IV	12	132
MC	IV	12	132
PY	V	12	132
BC	V	12	132
MC	V	12	132
Total		308	3388

[a]The layer, cells per column, and total cells per cell type used in the network model.

Table 3. Connectivity within the network[a]

Presynaptic		Postsynaptic					
Type	Layer	Type	Layer	Cross-column	Type	Strength	Probability
PY	II/III	PY	II/III	True	AMPA	0.4	0.1
PY	IV	PY	IV	True	AMPA	0.4	0.1
PY	V	PY	V	True	AMPA	0.4	0.1
PY	II/III	PY	II/III	False	AMPA	0.75	0.1
PY	II/III	BC	II/III	False	AMPA	0.75	0.1
PY	II/III	MC	II/III	False	AMPA	0.75	0.1
PY	II/III	PY	IV	False	AMPA	0.25	0.05
PY	IV	PY	II/III	False	AMPA	1.5	0.05
PY	IV	PY	IV	False	AMPA	0.75	0.1
PY	IV	BC	IV	False	AMPA	0.75	0.1
PY	IV	MC	IV	False	AMPA	0.75	0.1
PY	IV	BC	II/III	False	AMPA	1.5	0.05
PY	IV	MC	II/III	False	AMPA	1.5	0.05
PY	V	PY	V	False	AMPA	0.75	0.1
PY	V	PY	II/III	False	AMPA	0.5	0.05
PY	V	BC	V	False	AMPA	0.75	0.1
PY	V	MC	V	False	AMPA	0.75	0.1
MC	II/III	PY	II/III	False	GABA	0.75	0.05
MC	II/III	PY	IV	False	GABA	0.75	0.05
MC	II/III	PY	V	False	GABA	0.75	0.05
MC	IV	PY	II/III	False	GABA	0.75	0.05
MC	IV	PY	IV	False	GABA	0.75	0.05
MC	IV	PY	V	False	GABA	0.75	0.05
MC	V	PY	II/III	False	GABA	0.75	0.05
MC	V	PY	IV	False	GABA	0.75	0.05
MC	V	PY	V	False	GABA	0.75	0.05
BC	II/III	PY	II/III	False	GABA	1.5	0.25
BC	IV	PY	IV	False	GABA	1.5	0.25
BC	V	PY	V	False	GABA	1.5	0.25
MC	V	PY	IV	False	GABA	1	0.2
MC	V	PY	V	False	GABA	1.5	0.25
IN	I	PY	II/III	False	GABA	0.3	0.3
IN	I	PY	IV	False	GABA	0.3	0.3
IN	I	PY	V	False	GABA	0.3	0.3
PY	II/III	IN	I	False	AMPA	0.3	0.3

[a]Type, strength, and probability of connections between all cell types and layers. These values were estimated from experimental data of anatomic connectivity across slices of the rat cortex (Thomson et al., 2002).

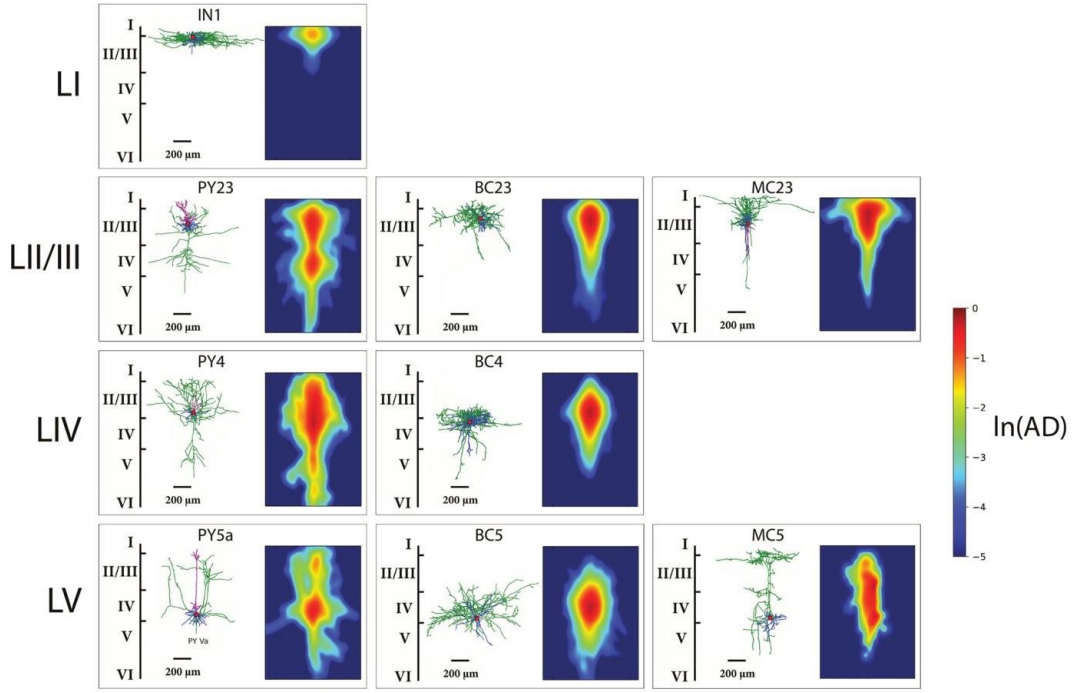


Figure 2. Representative reconstructions and averaged axonal density for neuronal cell types modeled. For each neuronal cell type used in the cortical microcircuit model, we plot both a single representative anatomical reconstruction as well as an averaged axonal density heatmap for all the reconstructions of that type. Cells are arranged by layer and type; the first row represents layer I inhibitory neurons. The representative layer I interneuron is a horizontal cell, but we average across small, descending, and horizontal layer I interneurons in both the axonal heatmap as well as in our analyses. The following rows represent pyramidal, basket, and Martinotti cells across layers II/III, IV, and V (layer Va for pyramidal cells). We did not have neuronal reconstructions for layer IV Martinotti cells and averaged spiking probability results from layers II/III and layer V Martinotti cells for subsequent analyses. In the axonal density (AD) plots, color represents the averaged density computed using all available reconstructions for the given cell type. The color scale is logarithmic for visual clarity. AD gives a sense of the general orientation and density of axon branches for each cell type, which is key to understanding subsequently computed activation probabilities.

Results

Building upon existing *in silico* models that simulated the effects of cortical surface stimulation from a single electrode (Komarov et al., 2019), in this study we sought to explore the spatial dynamics of stimulation by modeling an asymmetrical three-electrode configuration applied to the rat somatosensory cortex (Fig. 1*a*). This configuration was initially chosen to break the symmetry between the anodal and cathodal currents and pursued further because of the wave propagation observed in the network model as a result of the electrode choice. The paper is organized as follows. We first calculated the electric field potential created by the system of three electrodes: two cathodes (at -75 μA each) and a single anode (at 150 μA). Next, we estimated the activation probability for each cell type/cortical layer pairing by computing the activating function in biophysical reconstructions of axonal arbors. We then constructed a cortical microcircuit model with Hodgkin–Huxley dynamics to model the network effects of stimulation based on the previously-calculated spiking probabilities.

Cell activation results from a combination of morphology (cell type) and depth within the column

The applied electric field potential generated by the system of three electrodes (assuming homogeneous tissue) is shown in Figure 1*a*. To estimate the probability of specific cell types being activated by stimulation, we simulated the various cell types based on 3D morphological reconstructions of neurons derived from electron microscopy available from www.neuromorpho.org (Ascoli et al., 2007). The excitatory cells we considered were pyramidal cells across layers II-V, while the inhibitory neurons included basket cells and Martinotti cells across layers II-V in addition to layer I interneurons (Table 1). Example reconstructions as well as average axon density plots per cell type/cortical layer (Fig. 2) demonstrate the significant differences in axonal arborization and density among the different cell types, as well as between cells of the same type based in different layers.

The hyperpolarization or depolarization of a neuronal fiber within a constant electric field can be modeled with one-dimensional cable theory in conjunction with the activating function. The activating function (for details, see Materials and Methods) (Komarov et al., 2019) computes the net transmembrane current generated by external stimulation (while ignoring preexisting synaptic currents). According to one-dimensional cable theory, the activating function is the second-order spatial derivative of the electric potential along the neuronal fiber. The case of a perfectly horizontal fiber is shown in Figure 1*b*. Through this, we can draw relationships between the orientation and excitation of a fiber in response to a given stimulation polarity. Indeed, horizontal fibers were depolarized by anodal stimulation and hyperpolarized

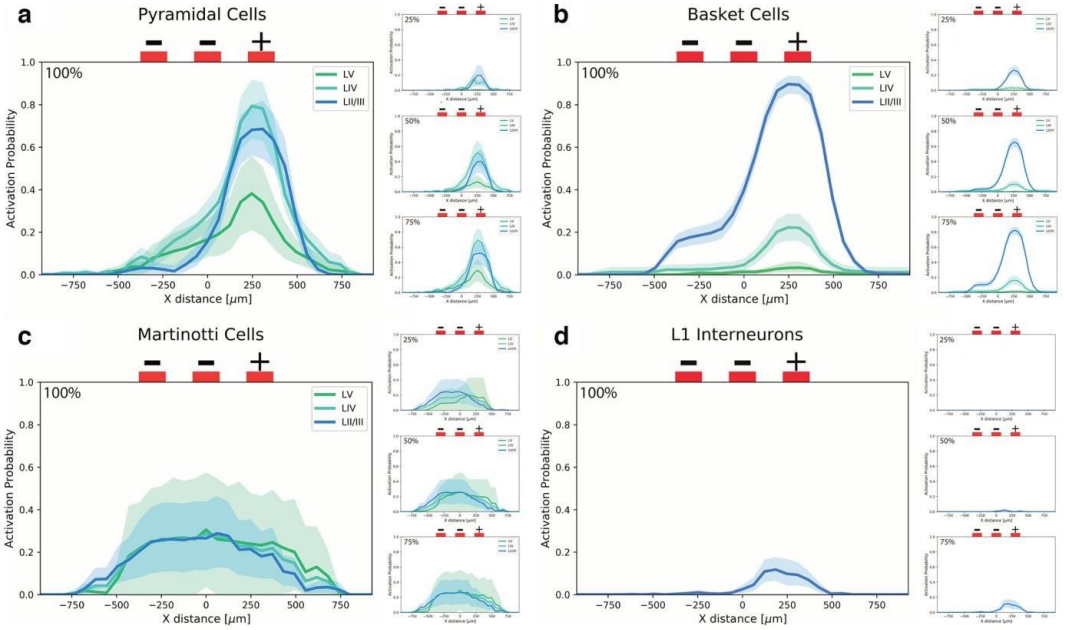


Figure 3. Probability of spiking as a function of horizontal distance from the center of the electrode array for each cell type and cortical layer. Average (solid line) cell spiking probability and 95% confidence intervals (shaded region) for each reconstruction were calculated for soma locations across the entire *X-Z* plane of the corresponding cortical layer by averaging spiking probability across rotations and vertical shifts of all cell reconstructions. Activation probabilities were calculated across four stimulating current strengths, with the maximum anodal stimulation current set to 150 μA and the maximum cathodal stimulation current set to -75 μA per electrode over a 200 μs stimulation period. Right, Insets, Activation profiles for 25%, 50%, and 75% stimulation strength. Pyramidal cells (excitatory, ***a***) and basket cells (inhibitory, ***b***) are highly activated by anodal stimulation and are minimally activated by cathodal stimulation because of their myelination and horizontally oriented axonal arbors. Martinotti cells (inhibitory, ***c***) are activated by all electrodes with a slight preference for the cathodes but lack myelination and thus show less activation overall. Layer I interneurons (***d***) are also unmyelinated and are minimally excited by stimulation.

by cathodal stimulation; in contrast, vertical fibers were hyperpolarized by anodal stimulation and depolarized by cathodal stimulation (Fig. 1*d* and Fig. 1*c*, respectively). While each neuron has unique axonal fibers that span 3D space, these maps of activation and suppression zones for orthogonal axonal orientations give us insight into how each cell type will behave across the stimulated space given its average axon density and orientation (Fig. 2).

We next calculated spiking probability in response to the applied electric field potential for each cell type/cortical layer pairing by averaging across the activating function results of their respective cell reconstructions; each cell reconstruction was shuffled by rotating and shifting along the vertical axis, and multiple reconstructions were considered for each cell type (for details, see Materials and Methods) (Komarov et al., 2019). This calculation compares the overall excitability of each reconstruction to an experimentally derived threshold (f_{th} = 3 pA/μm^2) to determine the probability of spiking. This threshold was set 20 times higher for unmyelinated cell types (Martinotti cells and layer I interneurons) compared to myelinated cell types (pyramidal and basket cells) since unmyelinated fibers are relatively unexcitable and lack nodes of Ranvier (Markram et al., 2004; Wang et al., 2004; Defelipe et al., 2012). The results of these calculations are shown in Figure 3.

To explore the parameter space of the model, we calculated activation probabilities at 25%, 50%, and 75% of the maximum stimulation current (150 μA for the anode and -75 μA for each cathode), which are displayed in Figure 3. The activation probabilities scale upward with increasing applied current for all cell types except Martinotti. At the weakest applied current, at which the absolute values of current amplitudes are $<$50 μA, layer II/III Martinotti cells exhibit a slight preference for cathodal stimulation, layer V Martinotti cells exhibit a slight preference for anodal stimulation, and layer IV cells show no strong preference. However, at all currents above the weakest, Martinotti cells across all layers display a slight preference for cathodal stimulation, and the activation probabilities appear to have reached a plateau; that is, increasing the applied current increases activation probabilities for all cell types except Martinotti. The average axonal density heatmaps in Figure 2 as well as the presence or absence of myelination explain the variation of activation responses across cell types and cortical layers.

Across all layers, pyramidal cells were strongly activated by the anode and minimally activated by the cathodes, with layer IV showing the greatest, layer V the least, and layer II/III an intermediate probability of activation (Fig. 3*a*). As shown in Figure 2, all pyramidal cells vertically span the cortical layers regardless of soma position. However, layer II/III and layer IV pyramidal cells exhibit significant horizontal axonal density close to the cortical surface and thus responded more strongly to stimulation overall (and to cathodal stimulation in particular), whereas the bulk of layer Va pyramidal axons lie in deeper layers and lack the superficial axonal density to be adequately stimulated above threshold. Pyramidal cells display a strong overall response to stimulation due to their myelinated axons.

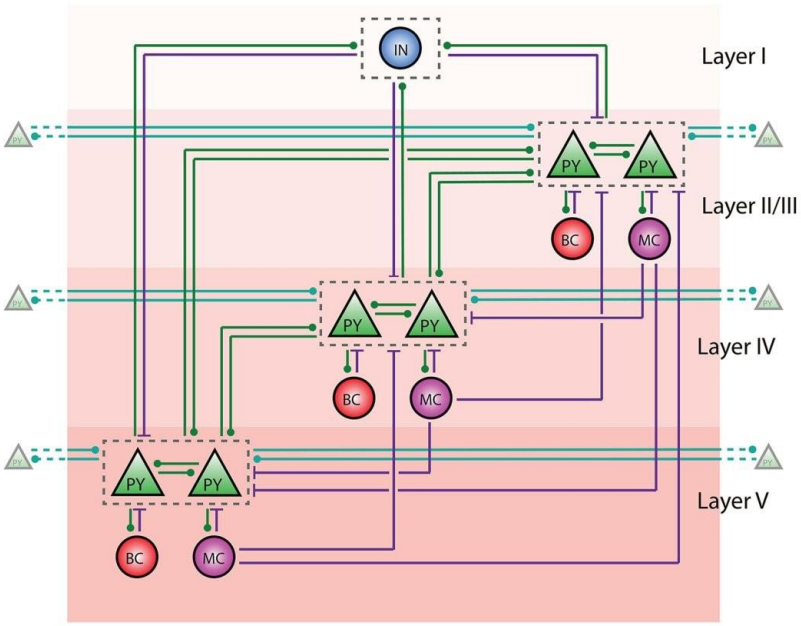


Figure 4. Microcircuit diagram of a single cortical column in modeled network. This depiction of a single cortical column details the cell types across cortical layers and the synaptic connections between them in our microcircuit network model. Circular labels (IN, LI interneurons; BC, basket cells; MC, Martinotti cells) represent inhibitory neurons. Triangular labels (PY, pyramidal cells) represent excitatory neurons. Green connections represent excitatory synapses, with the circular end indicating the postsynaptic cell and the unlabeled end indicating the presynaptic cell. Purple connections represent inhibitory synapses, with the perpendicular line indicating the postsynaptic end. Teal-colored connections are also excitatory but represent connections from pyramidal cells to others in adjacent columns (cross-columnar synapses). The gray box that surrounds the pyramidal cells within each layer includes all synaptic connections to all pyramidal cells, not just the ones closest to the synapse in the figure. The only cross-column synapses present are within-layer pyramidal-pyramidal excitatory connections to adjacent columns. Layer I includes only inhibitory interneurons, which inhibit pyramidal cells in all three deeper layers and are reciprocally excited by the same cells. Each of the deeper layers contain pyramidal, basket, and Martinotti cells. Within a cortical column, pyramidal cells reciprocally excite other pyramidal cells in their same layer as well as across cortical layers. Basket cells act as local interneurons as they only inhibit and are excited by pyramidal cells within their own layer. In contrast, while Martinotti cells are only excited by pyramidal cells within their own layer, they universally inhibit pyramidal cells across all layers. In this model, inhibitory cells receive only excitatory synaptic inputs. The number of neurons in each column is listed in Table 2 and the probability and strength of each synaptic connection is listed in Table 3.

Basket cells also exhibited a strong preference for anodal stimulation and little activation underneath the cathodes (Fig. 3*b*). However, their responses were significantly more tiered according to cortical layer compared with pyramidal cells because basket cell arborization is localized within the same layer as the soma (Fig. 2). Their preference for anodal stimulation is due to their largely horizontal axonal arbor that stretches out within each layer. Basket cells were the only myelinated inhibitory cell type in our model and therefore demonstrated a significantly stronger spiking response overall relative to Martinotti or layer I interneurons.

Martinotti cells across all layers are moderately activated by both anodal and cathodal stimulation but showed a slight preference for the latter (Fig. 3*c*). This is because all Martinotti cells make universal connections with pyramidal cells via layer I (Fig. 2); therefore, the majority of their arborizations lie in vertical axonal fibers connecting the soma to layer I, with additional density spread out horizontally across layer I. However, they exhibited a dampened stimulation response overall because of their unmyelinated axons.

Last, since layer I axon fibers are unmyelinated and stay localized to layer I (resulting in mainly horizontal arborization), layer I interneurons displayed a slight preference for anodal stimulation but little activation overall (Fig. 3*d*).

Cortical microcircuit model shows directional propagation when stimulated with three electrode array

In the previous section, we estimated the activation probabilities of isolated neurons within an applied electric field. To understand how stimulation affects the dynamics between neurons and ultimately the overall dynamics of the cortex, we constructed and stimulated a network model of the cortex using simplified neuron models and previously-calculated activation probabilities. Each cortical column was modeled as a canonical microcircuit (Douglas et al., 1989; Thomson et al., 2002; da Costa and Martin, 2010; Defelipe et al., 2012) containing the same cell types and cortical layers as the biophysical analysis above. A schematic of the cell types across cortical layers and their synaptic connections is shown in Figure 4.

The only cross-column synapses present are pyramidal-pyramidal excitatory connections to adjacent columns. Layer I includes only inhibitory interneurons, which inhibit pyramidal cells in all three deeper layers and are reciprocally excited by the same cells. Each of the deeper layers contain pyramidal, basket, and Martinotti cells. Within a cortical column, pyramidal cells reciprocally excite other pyramidal cells in their same layer as well as across cortical layers. Basket cells act as local interneurons as they only inhibit and are excited by pyramidal cells within their own layer. In contrast, while Martinotti cells are

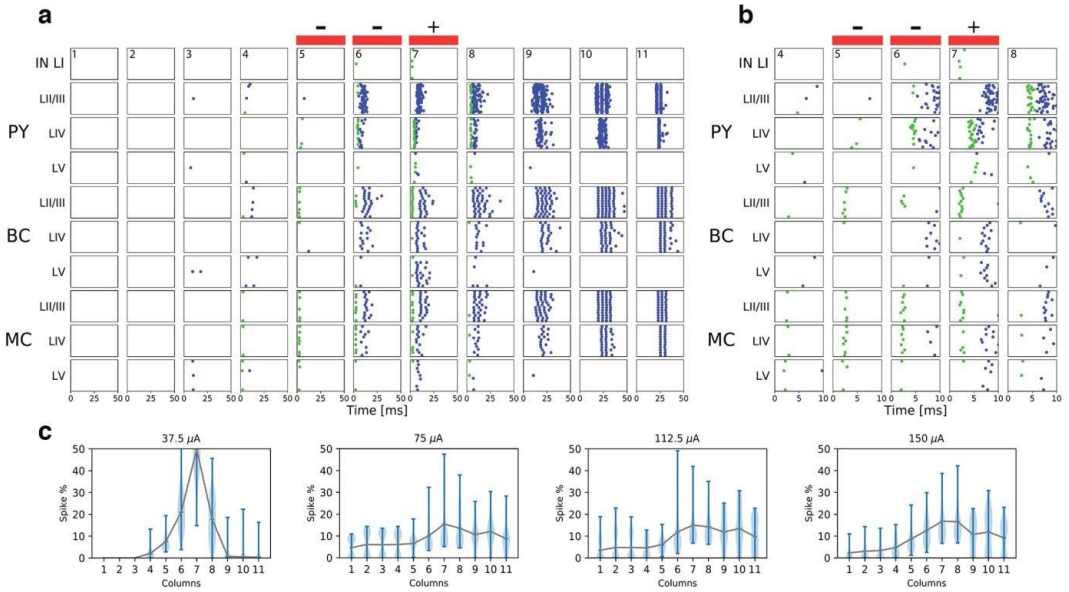


Figure 5. Directed propagation of pyramidal activity in raster plot of microcircuit simulation trial. A raster plot displaying network behavior during and after stimulation in one trial of the microcircuit simulation across all cortical columns at maximum applied current. Each cell within the microcircuit has its own coordinate on the *y* axis. Each dot is an action potential. Green dots indicate spikes that are directly triggered by electrical stimulation (occurs during first 5 ms). Blue dots indicate spikes triggered via synaptic input. ***a***, The first 50 ms of the simulation (the network is silent beyond this period). ***b***, Zoom in to the first 10 ms in the five central columns. ***c***, The percentage of spikes (all cell types) per cortical column, then averaged across 50 simulations at each applied current value (25%, 50%, 75%, and 100% applied current, with the anodal current listed, from left to right, respectively).

only excited by pyramidal cells within their own layer, they universally inhibit pyramidal cells across all layers. In this model, inhibitory cells receive only excitatory synaptic inputs. The number of neurons in each column is listed in Table 2 and the connectivity within the network is described in Table 3.

Following a brief stimulation period in which the activation probabilities from the biophysical calculations were applied to the model, the network was allowed to run without any external input for 500 ms, during which time it behaved according to synaptic interactions between neurons.

To test the robustness of network behavior, we conducted 50 simulations for four different values of stimulation current (25%, 50%, 75%, and 100% current, respectively) and calculated the percentage of total spikes in each cortical column averaged across these simulations (Fig. 5*c*). At the weakest applied current, spiking is highest underneath the anode and tapers off on either side as the stimulating current is not strong enough to induce propagating activity in either direction (Fig. 5*c*, leftmost plot). At ≥50% current, however, activity propagates unidirectionally as a traveling wave to the right. The following simulation example and subsequent explanations and analyses focus on network stimulation with the maximum current applied as it corresponds to strong asymmetric network behavior within physiological current bounds.

A raster plot and voltage and conductance traces of one microcircuit trial at maximum applied current are shown in Figures 5 and 6. The trial shown is one example of the general behavior of the microcircuit in the majority of trials at maximum applied current (Fig. 5*c*, rightmost plot) in which spiking activity, particularly from LII/III pyramidal cells, propagates to the rightward columns but not past the leftmost electrode. Given this unique spiking activity and the biological importance of LII/III pyramidal cells in mediating communication across cortical regions, we chose to focus our analyses on the network behavior of LII/III pyramidal cells.

Directionality of the stimulation-triggered wave can be explained by network inhibition

The activation probability curves in Figure 3 provide intuition into the network behavior during and immediately after the stimulation period (0-5 ms; Fig. 5*b*). Let us first examine the column underneath the anodal electrode (Fig. 5*b*, column 7). While both layer II/III and layer IV pyramidal cells were predicted to be highly activated underneath the anodal electrode (Fig. 3), only layer IV pyramidal cells were directly activated. Although all pyramidal cells were inhibited by moderate Martinotti activity, only layer II/III pyramidal cells were locally inhibited by strong synchronous layer II/III basket cell activity while other layers were not because basket cell response drops off with increasing cortical depth and because basket cells are only inhibit pyramidal cells within their own layer. Following stimulation, layer IV pyramidal cells excited layer II/III pyramidal cells and triggered a cluster of layer II/III activity. There was negligible layer V excitation during stimulation and none following because of their low excitation probabilities and relatively small neuronal population.

Network behavior underneath the cathodal electrodes contrasted sharply with anodal stimulation response and underpinned unidirectional excitatory propagation (Fig. 5*b*, columns 5 and 6). Although layers II/III and IV pyramidal cells were still moderately activated by cathodal stimulation (Fig. 3*a*), very few cells were pushed above threshold because of strong inhibition. Martinotti

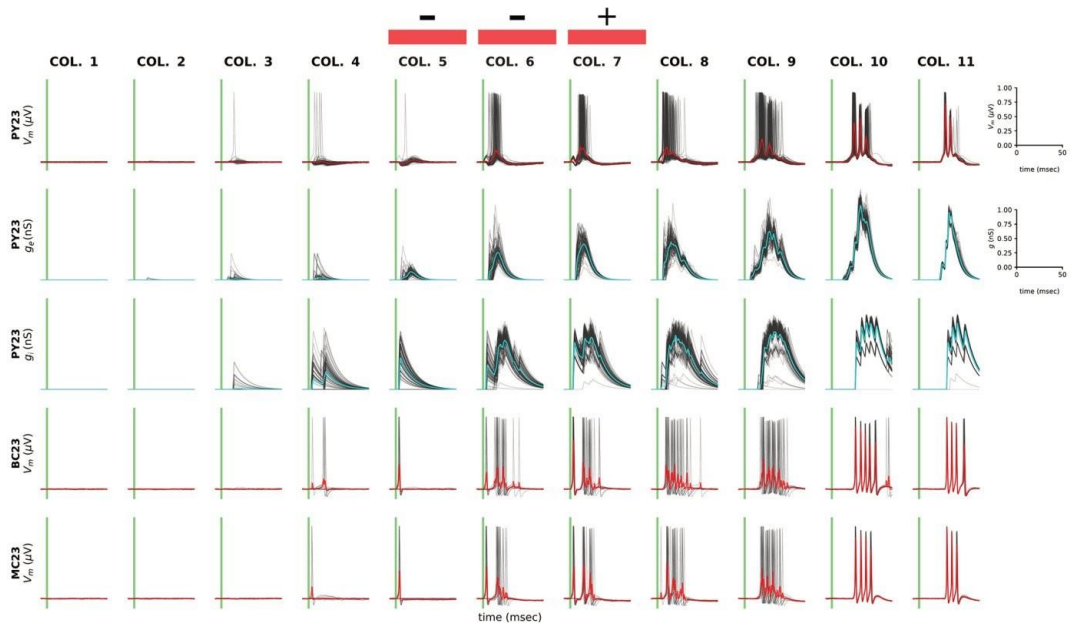


Figure 6. Voltage traces for layer II/III cells and pyramidal input conductances during simulation show how inhibition causes directionality of traveling wave. Each of the subplots in the grid contain data for the collection of layer II/III cells specified by type and column. The *x* axis for each subplot is time in milliseconds and is restricted to the interval from 5 ms before stimulation to 50 ms after. Region in light green represents stimulation period. Subplot columns, labeled at top, indicate the model columns. The first three rows represent data from pyramidal cells, with the first row representing the voltage trace, the second the total excitatory conductance (the sum of all incoming pyramidal connections), and the third the total inhibitory conductance (sum of inhibitory connections from layer I interneurons, basket cells, and Martinotti cells). The fourth and fifth rows represent the voltage traces of layer II/III basket and Martinotti cells, respectively. In each subplot, there are light gray traces representing each of the individual cells in the group, and their average (within the same column, layer, and subtype) in red. Sharp increases in the light gray trace indicate action potentials (spikes). Darker gray regions correspond to times when large numbers of cells spiked.

cells showed a preference for cathodal stimulation (according to Fig. 3) and thus fired early in the stimulation period, inhibiting pyramidal cells across all cortical depths (since Martinotti cells universally synapsed to all pyramidal cells in the network). This strong inhibitory force coupled with moderate superficial basket cell activity silenced almost all pyramidal activity across cortical layers.

In the column directly to the left of the electrode array (Fig. 5*b*, column 4), there was negligible activity across all cell types and cortical layers. Not only did the electric field potential drop off significantly at this distance, but pyramidal and basket cells were already minimally activated by the cathodal electrodes, and Martinotti cells were only moderately activated by the cathodal electrodes because of their lack of myelination. In the absence of stimulating electric field potential or activating input from neighboring cortical columns, the leftmost three columns exhibited no spiking activity at all (Fig. 5*a*, columns 1-3). Hence, the excitatory pyramidal activity present underneath the electrodes did not propagate leftward past the cathodal electrodes. This activity, however, did travel rightward past the electrode array, growing stronger and more synchronous as it propagated.

On the other side of the array, in the cortical column directly to the right of the anodal electrode (Fig. 5*b*, column 8), there was moderate direct activation of pyramidal cells and little direct activation of inhibitory cells. This follows from Figure 3, which depicts pyramidal cells continuing to be activated by the anodal electrode. While basket cells were also moderately activated by anodal stimulation, their joint inhibition with Martinotti cells was not enough to counter pyramidal stimulation response. This allowed for dense clusters of excitatory activity in pyramidal cells following stimulation.

In the second column to the right of the electrode array (Fig. 5*a*, column 9), we see a dense cluster of highly synchronized pyramidal layer II/III activity that was slightly delayed from the activity in the column to its left. Although pyramidal cells were no longer directly stimulated in this cortical column, this activity resulted from the rightward cross-columnar propagation of excitatory signaling. Remarkably, this substantial, synchronized pyramidal activity grew more and more synchronous as the wave propagated rightward across cortical columns. This unidirectional propagation of excitatory signaling is an exceptional product of asymmetrical stimulation (Fig. 5*a*, columns 9-11).

Following the stimulation period and initial clusters of activity that die down at ~10 ms after stimulation, there were a handful of waves of activity that ping-ponged between excitatory and inhibitory cells in columns with pyramidal excitation (Fig. 5*a*, columns 6-11). In these columns, pyramidal activity activated both Martinotti and basket cells, which in turn inhibited pyramidal activity. There were a few iterations of this negative feedback loop over the course of a few milliseconds before pyramidal spiking was halted entirely.

Analysis of synaptic currents reveals mechanism of asymmetrical spiking activity

Next, we analyzed the voltage traces of individual neurons and synaptic dynamics to explain the causes of pyramidal asymmetrical spiking activity (Fig. 6).

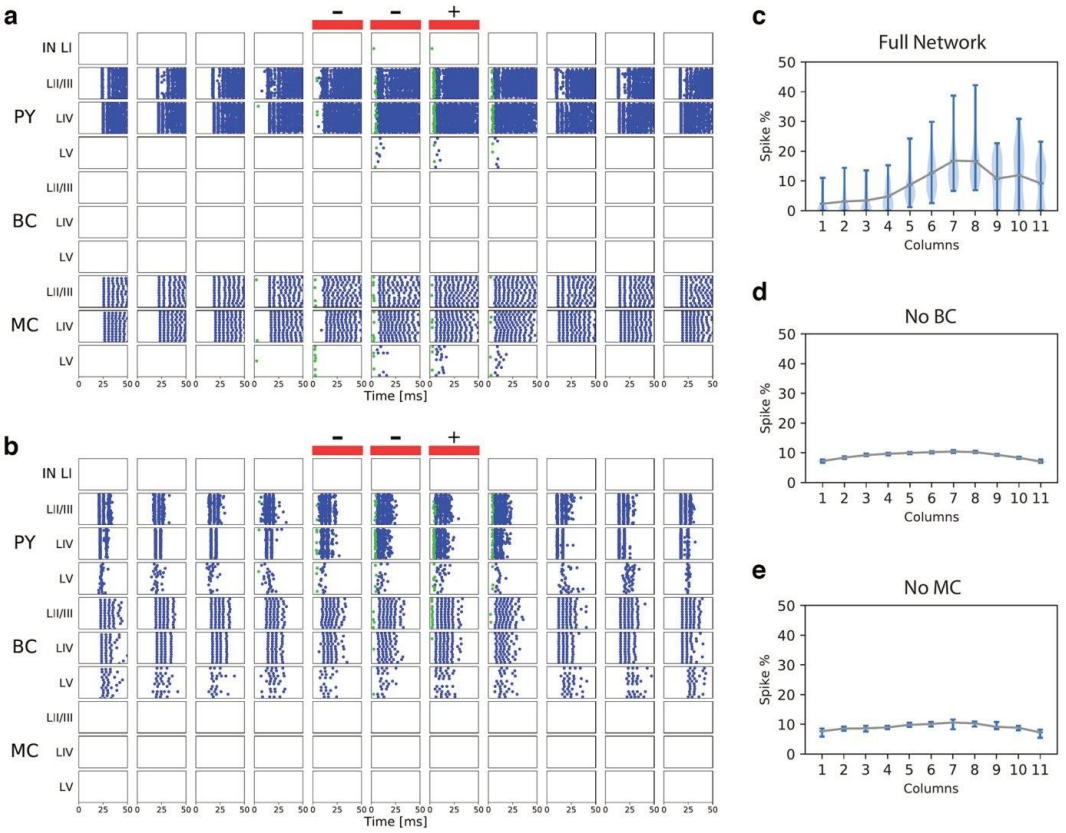


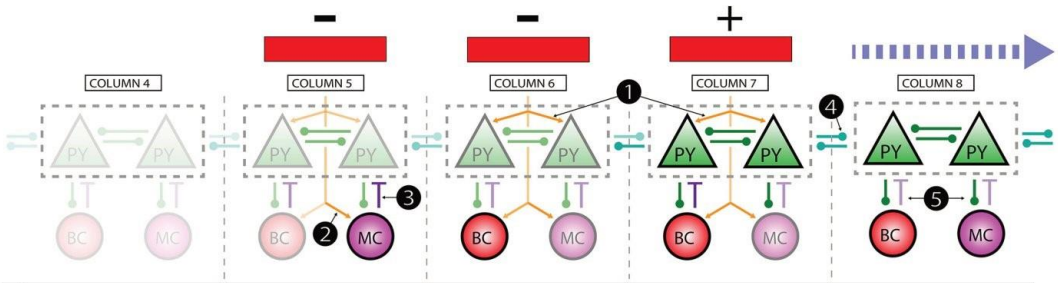
Figure 7. Cell type-specific silencing indicates distinct roles of inhibitory interneurons in temporal and spatial dynamics of the traveling wave. Raster plots displaying network activity in the same experimental setup as that shown in Figure 5*a*, except with either Basket cells (***a***) or Martinotti cells (***b***) silenced. ***c–e***, The average spike percentage per column (relative to the total number of spikes in the simulation) averaged across 20 simulations for the full network model (***c***), with silenced Basket cells (***d***), or with silenced Martinotti cells (***e***).

Voltage traces revealed that layer II/III pyramidal cells were initially depolarized underneath the anodal electrode (column 7), as they were highly activated by anodal stimulation. However, a large inhibitory conductance immediately following the stimulation period (the first g_i peak) dampened any stimulation-induced depolarization and hyperpolarized these pyramidal cells. Inhibitory conductance then fell and excitatory conductance rose as layer II/III pyramidal cells gained excitatory input from neighboring layers and columns, bringing layer II/III pyramidal cells to threshold and triggering action potentials. The second peak in inhibitory conductance midway through the pyramidal action potential was caused by pyramidal cell input into the inhibitory cells, which initiated a negative feedback loop that quickly subsided as pyramidal cells were silenced by inhibition. This effect occurred in all columns with substantial pyramidal activity.

Although there were similar inhibitory and excitatory conductance dynamics in the column underneath the central cathodal electrode (column 6), the excitatory conductance was of a smaller magnitude overall, and there were fewer action potentials because pyramidal cells were less excited by cathodal stimulation. Underneath the leftmost electrode (column 5), moderate inhibitory conductance outweighed negligible excitatory conductance, leading to minimal pyramidal activity. There was little excitation in column 4 or any of the other leftward columns (not shown). On the other side of the electrode array, in column 8, high initial excitatory conductance and minimal inhibitory conductance resulted in strong initial pyramidal spiking. Pyramidal action potentials became more and more synchronous as they traveled rightward, as evidenced by increasingly overlapped voltage traces. Together, these observations reveal the mechanisms of activity propagation to the right but not to the left in these stimulation settings.

Cell type-specific silencing reveals distinct roles of inhibitory interneurons in shaping and sustaining traveling waves

To better understand the role of basket and Martinotti cells in shaping and sustaining traveling waves, we performed cell type-specific silencing experiments using the previously described network model paradigm in Figure 5. Silencing basket cells resulted in persistent activity rapidly spreading bidirectionally from the central column similar to what may be observed during a seizure-like event (Fig. 7*a*). Compared with the control network in Figure 5*a*, without basket cells, the pyramidal cell activity within a given column never halts once initiated and appears more akin to a standing wave which spreads bidirectionally throughout the network rather than propagates. Alternatively, silencing Martinotti



PY	No direct activation + no excitatory input + small inhibitory input = little spiking	Low direct activation + low excitatory input + moderate inhibition = low spiking	Direct activation + very little inhibition = strong spiking	Direct activation + excitatory input outweighs inhibition = strong spiking	Direct activation + high excitatory input + very low inhibition = strong spiking
BC	No direct activation + no excitatory input = little spiking	Low direct activation + low excitatory input = low spiking	Direct activation + moderate excitatory input = strong spiking	Direct activation + excitatory input = strong spiking	Strong excitatory input = strong spiking in response to PYs
MC	Small activation + no excitatory input = little spiking	High direct activation = strong initial spiking	Direct activation + moderate excitatory input = moderate spiking	Direct activation + excitatory input = moderate spiking	Excitatory input = strong spiking in response to PYs

Figure 8. Summary of interactions resulting in unidirectional propagation. Key interactions are depicted graphically above and summarized in text in the table below. Columns 4-8 of the total 11 are included, indicated by labels at top. Cell and synaptic labels are as in Figure 4. Orange arrows coming from the electrodes and pointing toward cells represent direct electrical stimulation, as opposed to synaptic inputs, indicated by the green, purple, and teal lines. The opacity of the cell labels approximately depicts the degree of spiking activity, with more transparent cells showing little or no spiking activity, and more opaque ones corresponding to more actively spiking cells. The opacity of lines indicates the strength of the input to cells, either synaptic or from the stimulating electrodes. For example, more opaque orange arrows indicate that the cell being pointed to was strongly stimulated directly. The light blue dashed arrow at the top right indicates the initiation of synchronous, unidirectional propagation of activity to the direct right of the electrodes. Key events in initiation of this unidirectional propagation are indicated by the white numbers in black circles: ① Direct activation under (+) [column 7] and central (–) [column 6] induces strong spiking in PYs, BCs, and MCs. ② Under left (–) [column 5] direct activation induces strong spiking in MCs, but little in PYs and BCs. ③ PYs in column 5 further inhibited by MCs leads to little activity to left of electrode. ④ Strong PY activity in columns 6 and 7 propagates rightward via cross-column PY-PY synapses, overcoming moderate inhibition. ⑤ PY activity causes spiking in MCs and BCs as it propagates rightward; feedback between PYs and inhibitory cells causes increasingly synchronous spiking. Layer I interneurons have been omitted, as they did not contribute significantly to the propagation described here.

cells preserved the propagating nature of the wave across columns (i.e., activity in each column self-terminated after 20-25 ms). However, this propagation was still bidirectional (Fig. 7*b*). Thus, these experiments demonstrate that both types of inhibitory cells may be necessary for the spatial asymmetry of the wave, while suggesting that basket cells may be more important for wave propagation by more effectively halting excitatory activity once initiated.

Discussion

In this work, we predict that an asymmetrical cortical stimulation protocol using a combination of anodal and cathodal electrodes may trigger propagating excitatory activity that shows strong directional preference. Our model had two steps: we first constructed a biophysical model to predict activation probabilities across cell types in response to an asymmetrically-applied electric field potential, and then incorporated these probabilities into a cortical microcircuit to model the network effects of stimulation. We found that pyramidal cells and basket cells are highly activated by the anodal electrode and minimally activated by the cathodal electrodes because of their myelination and horizontal axonal arbors, while layer I interneurons are only moderately activated by the anodal electrode despite their horizontal axonal arbors due to lack of myelination. Martinotti cells also exhibit moderate activation due to lack on myelination, but show a slight preference for cathodal stimulation due to their predominately vertically-oriented axonal arbors. Network model simulations revealed that this asymmetrical activation results in a traveling wave in superficial excitatory cells that propagates away from the electrode array, past the anodal electrode, and into adjacent cortical columns, but does not propagate in the opposite direction past the leftmost cathodal electrodes (Fig. 8).

We found that distinct classes of inhibitory cells are the cause of separable components of the unidirectional propagation. Basket cells were necessary for the wave to propagate (as opposed to spread as a standing wave), but both basket and Martinotti cells were needed for asymmetrical spatial propagation. While activity directly under the central column is defined by the stimulation protocol, as the activity propagates laterally through horizontal excitatory connections, the increasing synchrony of cell firing is likely mediated by a PING type mechanism (Whittington et al., 2000) (Martinotti cells were not necessary for this aspect of the activity, as shown in Figure 7*b*).

Importantly, only the spatial component of the traveling wave is dependent on the particular electrode design of the stimulation paradigm; the temporal component is a manifestation of endogenous cortical columnar circuitry. This allows for endogenous oscillatory activity to be spatially guided through the network, without inducing artificial frequencies as a result of stimulation. These two classes of inhibitory cells could be deactivated optogenetically in rodents during stimulation to test the predictions of the model. Electrical stimulation paradigms that rely on current steering to guide activity along particular trajectories are widely used both experimentally and clinically and have demonstrated robust efficacy at the macroscopic network level, but are still poorly understood at the mesoscopic circuit level. Ultimately, we believe that this model is a first step toward understanding the circuit mechanisms which are engaged during these stimulation practices.

Traveling waves in the brain

Multielectrode recordings in human and animal subjects have demonstrated the ubiquity of traveling waves in cognitive

function (Muller et al., 2018). They relay information across a range of distances and thereby coordinate fundamental processes such as memory, perception, language, orientation, executive functions, and more across distant brain regions (Rubino et al., 2006; Wu et al., 2008; Muller et al., 2018; Salimpour and Anderson, 2019). Recordings have also shown that propagating activity present in the human cortex is often directional, traveling from one point to another. The ability to generate directional propagation via stimulation would allow for unique precision in and control over induced activity. This has widespread implications for the restoration or enhancement of high-level brain function, particularly because many neuropsychiatric disorders are marked by abnormal or absent propagating activity.

In the model, propagating activity was confined to supragranular cortical layers. This may not be a limitation if the goal is to reproduce natural waves, because spontaneous traveling waves in the human cortex have also been found to be largely confined to upper layers, including the alpha rhythm during waking (M. Halgren et al., 2019), and spindles and slow oscillation during sleep (Cash et al., 2009; M. Halgren et al., 2018).

Traveling waves have long been studied with a variety of computational models, and numerous mechanisms have been proposed to explain how activity propagates across neuronal networks (Ermentrout and Kleinfeld, 2001; Breakspear, 2017). Cortical propagating waves that are triggered specifically by electrical stimulation have been recorded in mammalian cortical slices (Kim et al., 1995; Wu et al., 2001; Wester and Contreras, 2012), as well as in non-human mammals (Contreras et al., 1997; Xu et al., 2007; Stieger et al., 2020) and simulated in non-mammalian computational models (Chen et al., 2008), but remain understudied in human subjects. This previous experimental work, both *in vivo* and *in silico*, has yielded scarce evidence of asymmetrical traveling wave propagation or reliable wave generation analogous to that reported here. To the extent that previous work has focused on traveling wave propagation initiated by stimulation (Alekseichuk et al., 2019), these studies examined stimulation through the skull and meninges, and are thus not directly comparable to this model of intracranial stimulation of the cortical surface. Many computational models exist that model the effects of stimulation on the brain, including some that have constructed Hodgkin–Huxley microcircuits (Douglas et al., 1995; Haeusler and Maass, 2007) and modeled cortical surface stimulation (Anderson et al., 2009). However, few have modeled multielectrode or asymmetrical stimulation or reproduced traveling waves using surface electrodes. Many existing stimulation models have focused on stimulation of a particular nerve (Raspopovic et al., 2011; Helmers et al., 2012) or an isolated cell type (Traub et al., 1994; R. J. Greenberg et al., 1999), as opposed to the functioning cortical microcircuit presented here, which can be more readily adapted to other cortical regions by adjusting the parameters and neuronal reconstructions used. In addition, stimulation has more often been simulated in these models by a simple application of suprathreshold current or a uniform electric field (Radman et al., 2007, 2009). Thus, by combining the two-phase biophysical model with asymmetrical, multipolar surface stimulation, our approach synthesizes existing achievements into a single coherent, clinically-adaptable model that uniquely sheds light on the generation of propagating wave activity.

Clinical relevance of our findings

Brain stimulation is becoming increasingly common in clinical and experimental settings, especially using multielectrode arrays (Lewis et al., 2015). As such, it is pressing that we develop accurate models of the effects that multielectrode stimulation has on neural activity. While sometimes the explicit goal of stimulation may be to disrupt aberrant activity to restore normal functioning, increasingly the goal is to induce the desired brain activity directly via stimulation, as our work demonstrates.

Changes in neural plasticity result from patterned activity, with the particular changes in connectivity contingent on the specific timing and order of activity (Bennett and Bair, 2015). Stimulation protocols that induce neural activity which continues past the stimulus duration are more likely to alter cellular and synaptic properties in favor of the induced activity, in contrast to stimulation protocols that briefly activate broad swaths of cells without triggering existing activity patterns. Thus, initiating propagating waves within tailored spatiotemporal constraints is a promising way to retrain neural networks and enhance or silence brain functions in a targeted way.

The generation of traveling waves may serve as a promising therapy for a variety of neurological disorders. For example, it has been previously suggested that triggering propagating activity in perilesional areas where waves are otherwise aberrant or absent may be an effective therapy for post-stroke aphasia (Beuter et al., 2020). While the ultimate clinical applications of this technique are uncertain, stimulation-induced traveling waves may have the potential to offset inhibition in cortical spreading depression (Liebetanz et al., 2006; Santos et al., 2012), reduce the risk of seizure while determining which brain tissue to remove from epilepsy patients (Nagaraj et al., 2015), and enhance memory formation when applied during learning or recall periods (Suthana et al., 2012; Batterink et al., 2016; Ezzyat et al., 2018; Kucewicz et al., 2018a,b), as consistency in traveling wave direction is positively correlated with working memory efficiency (Zhang et al., 2018).

Limitations

In this work, we modeled a single, short stimulating pulse. However, clinical stimulation is most often composed of longer pulses or pulse trains and is usually performed with bipolar electrodes delivering biphasic pulses to prevent damaging Faradic currents (Merrill et al., 2005; Cogan et al., 2016). These stimulation paradigms modulate properties over time that are not accounted for in the current model, such as underlying dendritic and axonal dynamics as well as synaptic interactions. Thus, our biophysical approach may be expanded in future studies to incorporate these steady-state properties through alternative modeling approaches, such as the cylinder model (Rall, 1962; Tranchina and Nicholson, 1986) or the multicompartmental model (Berzhanskaya et al., 2013). However, we chose to model the activation probability of the axonal instead of the dendritic arbor in this work because experimental evidence shows that the nodes of Ranvier, followed by the axon hillock, are the most excitable neuronal elements by far via direct stimulation (Gustafsson and Jankowska, 1976; Swadlow, 1992; Rattay, 1999; Tehovnik et al., 2006) as they both have a high concentration of sodium channels (Catterall, 1981). In contrast, direct stimulation of the dendritic arbor generates transmembrane currents that propagate to the axon hillock, but these effects are strongly attenuated and delayed, and are negligible compared with direct stimulation of the nodes of Ranvier and axon hillock.

Consistent with the findings that the axon is the most likely site of action potential initiation under electrical stimulation, our approach has focused specifically on estimating this probability while neglecting other aspects of stimulation which may alter

subsequent network activity. In particular, effects of stimulation on nonlinear, often calcium-mediated, properties of the dendrites and axon terminals would be expected to significantly outlast the duration of stimulation. Both of these locations can directly influence synaptic efficacy, or even decorrelate synaptic release from action potential initiation (Katz and Miledi, 1967), and therefore could substantially alter subsequent network dynamics in neural tissue.

Moreover, previous modeling studies have indicated that the most depolarized neural element is not always the site of action potential initiation (McIntyre and Grill, 1999). This was particularly found to be the case when the electrode was positioned near the cell body, which resulted in maximal depolarization in the dendrites or soma, but with action potential initiation taking place in the axon or at the initial segment. For the present study, this breakdown in the assumptions of the activating function approach is mitigated by the more proximal relationship of the axons than the soma to the electrode and the short pulse duration, as such conditions have been found to show greater correspondence between the site of maximal depolarization and action potential initiation, both of which typically occurred in axonal segments (Rattay and Aberham, 1993; McIntyre and Grill, 1999). In future studies which consider longer stimulus durations or DBS paradigms where the electrode may be more proximal to the soma than axonal – conditions that are particularly pertinent to clinical applications – an active cable theory model would need to be used to properly account for action potential initiation.

In the microcircuit phase of the model, the connectivity between different cell types follows a canonical microcircuit model. While this approach characterizes the main signal pathways and feedback loops present within cortical columns (Douglas et al., 1989; da Costa and Martin, 2010; Defelipe et al., 2012), finer details are not modeled, such as descending projections to inhibitory cells from excitatory cells (Thomson et al., 2002) or the contribution of less common interneuron cell types. The cells within the cortical microcircuit model could be extended from single-compartment to multicompartment neurons (Bonjean et al., 2012) to distinguish tuft versus soma-targeting interneurons, which may further differentiate the inhibitory power of interneuron cell types (Markram et al., 2004). This phase of the model may be further expanded from a 2D plane to a 3D circuit in the volume of cortex underneath the electrodes to understand how activity spreads across space.

While moving to multicompartment neurons with active properties would alleviate many of the limitations of our approach discussed above, it also presents unique difficulties. Such models are vastly higher-dimensional than the passive cable and point-neuron models considered here and are difficult to properly constrain because of the lack of experimental data on the distribution of passive and active ion channels within different cell types necessary for data-driven parameterization. Without such constraints, these high-dimensional models are liable to be finely tuned within their vast parameter space to be able to exhibit nearly any desired activity and run the risk of diverging from biologically realistic parameter regimens and decoupling the modeled activity from plausible cellular mechanisms. Given that our goal in this study was to shed light on the cellular and circuit mechanisms underlying electrical stimulation and current steering rather than provide robust statistical predictions of the results of this particular electrical stimulation design, we opted to use models which, although known to be incomplete, are capable of more robustly constrained parameterization.

Alternative modeling approaches

In this modeling work, we have focused on estimating transient, cell type-specific responses to electrical stimulation, and subsequently incorporated these estimates into a biophysical network model of the canonical cortical column. In this sense, we directly model certain effects of electrical stimulation at the microscopic, cellular level, and then import these findings into a mesoscopic, cortical circuit model. Previous work has approached the problem of modeling the effects of electrical stimulation in diverse ways at multiple scales. At the microscopic level, McIntyre et al. (2004a) used a multicompartmental cable theory model with active properties to study the effect of DBS on the cellular properties of single thalamocortical relay neurons (McIntyre et al., 2004a). More recently, such an approach was used to study single-cell responses to intracortical and uniform electric field stimulation for a variety of cell types obtained from human and rat cortical neuron reconstructions (Aberra et al., 2018). At more macroscopic levels, researchers have used finite-element methods to model epidural electrical stimulation of the motor cortex which can account for cortical folding (Wongsarnpigoon and Grill, 2008, 2012; Aberra et al., 2018), and have incorporated data from human diffusion tensor MRI to estimate the volume of tissue activated by DBS in the subthalamic nucleus (McIntyre et al., 2004b).

Generalization of this approach

While this work modeled a specific stimulation protocol for a particular and still simplified cortical network architecture, the approach is generalizable to a variety of basic science and clinical applications. This versatility comes from the modular structure of the model, which completely decouples the biophysical-anatomical model from the dynamic-neuronal network model. This makes it possible to use the same activation probabilities for a variety of network models so long as they contain analogs of the initial cell types. Additionally, circumventing the computational complexity of simulating high-dimensional compartmental models facilitates the widespread investigation of much larger networks than studied in this paper. Indeed, recent empirical studies have collected an enormous amount of new data regarding cell properties and local- and long-range connectivity, but current modeling efforts have yet to take advantage of these data. It is still not feasible to simulate large-scale network models, including different brain structures (Sanda et al., 2021) and/or multiple cortical regions and long-range connectivity (Rosen et al., 2019), which would be built on anatomically realistic cell reconstructions and include multiple layers and different cell types. This severely limits how new anatomical and functional data are used in the model design, and we suggest that the hybrid approach we present here may help to partially overcome these limitations.

In conclusion, this work models an asymmetrical stimulation paradigm that could be implemented to initiate unidirectional traveling waves in the cortex. A biophysical model is integrated with a computational network model to predict the behavior of single neurons as well as the cortical network dynamics resulting from multielectrode stimulation. This model provides hypotheses and stimulation paradigms which can be verified experimentally. It expands on the capabilities of our hybrid modeling approach to show how it can be deployed to probe the relationship between the microscale effects of electrical stimulation and the mesoscale consequences at the level of circuit dynamics. These results demonstrate how complex stimulation protocols

could be harnessed to generate persistent changes in activity with the potential to restore normal brain function in neurological and psychiatric conditions.

References

Abelson JL, Curtis GC, Sagher O, Albucher RC, Harrigan M, Taylor SF, Martis B, Giordani B (2005) Deep brain stimulation for refractory obsessive-compulsive disorder. Biol Psychiatry 57:510–516.

Aberra AS, Peterchev AV, Grill WM (2018) Biophysically realistic neuron models for simulation of cortical stimulation. J Neural Eng 15:066023.

Alekseichuk I, Falchier AY, Linn G, Xu T, Milham MP, Schroeder CE, Opitz A (2019) Electric field dynamics in the brain during multi-electrode transcranial electric stimulation. Nat Commun 10:2573.

Anderson WS, Kudela P, Weinberg S, Bergey GK, Franaszczuk PJ (2009) Phase-dependent stimulation effects on bursting activity in a neural network cortical simulation. Epilepsy Res 84:42–55.

Ascoli GA, Donohue DE, Halavi M (2007) NeuroMorpho.Org: a central resource for neuronal morphologies. J Neurosci 27:9247–9251.

Baizabal-Carvallo JF, Alonso-Juarez M (2016) Low-frequency deep brain stimulation for movement disorders. Parkinsonism Relat Disord 31:14–22.

Batterink LJ, Creery JD, Paller KA (2016) Phase of spontaneous slow oscillations during sleep influences memory-related processing of auditory cues. J Neurosci 36:1401–1409.

Bennett JE, Bair W (2015) Refinement and pattern formation in neural circuits by the interaction of traveling waves with spike timing-dependent plasticity. PLoS Comput Biol 11:e1004422.

Berzhanskaya J, Chernyy N, Gluckman BJ, Schiff SJ, Ascoli GA (2013) Modulation of hippocampal rhythms by subthreshold electric fields and network topology. J Comput Neurosci 34:369–389.

Beuter A, Balossier A, Vassal F, Hemm S, Volpert V (2020) Cortical stimulation in aphasia following ischemic stroke: toward model-guided electrical neuromodulation. Biol Cybern 114:5–21.

Blumenfeld Z, Bronte-Stewart H (2015) High frequency deep brain stimulation and neural rhythms in Parkinson's disease. Neuropsychol Rev 25:384–397.

Bonjean M, Baker T, Bazhenov M, Cash S, Halgren E, Sejnowski T (2012) Interactions between core and matrix thalamocortical projections in human sleep spindle synchronization. J Neurosci 32:5250–5263.

Breakspear M (2017) Dynamic models of large-scale brain activity. Nat Neurosci 20:340–352.

Cash SS, Halgren E, Dehghani N, Rossetti AO, Thesen T, Wang C, Devinsky O, Kuzniecky R, Doyle W, Madsen JR, Bromfield E, Eross L, Halasz P, Karmos G, Csercsa R, Wittner L, Ulbert I (2009) The human K-complex represents an isolated cortical down-state. Science 324:1084–1087.

Catterall WA (1981) Localization of sodium channels in cultured neural cells. J Neurosci 1:777–783.

Chen E, Stiefel KM, Sejnowski TJ, Bullock TH (2008) Model of traveling waves in a coral nerve network. J Comp Physiol A Neuroethol Sens Neural Behav Physiol 194:195–200.

Cogan SF, Ludwig KA, Welle CG, Takmakov P (2016) Tissue damage thresholds during therapeutic electrical stimulation. J Neural Eng 13:021001.

Contreras D, Destexhe A, Sejnowski TJ, Steriade M (1997) Spatiotemporal patterns of spindle oscillations in cortex and thalamus. J Neurosci 17:1179–1196.

da Costa NM, Martin KAC (2010) Whose cortical column would that be? Front Neuroanat 4:16.

Defelipe J, Markram H, Rockland KS (2012) The neocortical column. Front Neuroanat 6:22.

Dickey CW, Sargsyan A, Madsen JR, Eskandar EN, Cash SS, Halgren E (2021) Travelling spindles create necessary conditions for spike-timing-dependent plasticity in humans. Nat Commun 12:1027.

Douglas RJ, Martin KAC (1991) A functional microcircuit for cat visual cortex. J Physiol 440:735–769.

Douglas RJ, Martin KAC, Whitteridge D (1989) A canonical microcircuit for neocortex. Neural Comput 1:480–488.

Douglas RJ, Koch C, Mahowald M, Martin KAC, Suarez HH (1995) Recurrent excitation in neocortical circuits. Science 269:981–985.

Ermentrout GB, Kleinfeld D (2001) Traveling electrical waves in cortex: insights from phase dynamics and speculation on a computational role. Neuron 29:33–44.

Ezzyat Y, et al. (2018) Closed-loop stimulation of temporal cortex rescues functional networks and improves memory. Nat Commun 9:365.

Greenberg BD, Malone DA, Friehs GM, Rezai AR, Kubu CS, Malloy PF, Salloway SP, Okun MS, Goodman WK, Rasmussen SA (2006) Three-year outcomes in deep brain stimulation for highly resistant obsessive-compulsive disorder. Neuropsychopharmacology 31:2384–2393.

Greenberg RJ, Velte TJ, Humayun MS, Scarlatis GN, de Juan E Jr (1999) A computational model of electrical stimulation of the retinal ganglion cell. IEEE Trans Biomed Eng 46:505–514.

Gustafsson B, Jankowska E (1976) Direct and indirect activation of nerve cells by electrical pulses applied extracellularly. J Physiol 258:33–61.

Ha S, Akinin A, Park AJ, Kim C, Wang H, Maier C, Mercier P, Cauwenberghs G (2017) Silicon-integrated high-density electrocortical interfaces. Proc IEEE 105:11–33.

Haeusler S, Maass W (2007) A statistical analysis of information-processing properties of lamina-specific cortical microcircuit models. Cereb Cortex 17:149–162.

Halgren E, Walter RD, Cherlow DG, Crandall PH (1978) Mental phenomena evoked by electrical stimulation of the human hippocampal formation and amygdala. Brain 101:83–115.

Halgren M, Fabo D, Ulbert I, Madsen JR, Eross L, Doyle WK, Devinsky O, Schomer D, Cash SS, Halgren E (2018) Superficial slow rhythms integrate cortical processing in humans. Sci Rep 8:2055.

Halgren M, Ulbert I, Bastuji H, Fabo D, Eross L, Rey M, Devinsky O, Doyle WK, Mak-McCully R, Halgren E, Wittner L, Chauvel P, Heit G, Eskandar E, Mandell A, Cash SS (2019) The generation and propagation of the human alpha rhythm. Proc Natl Acad Sci USA 116:23772–23782.

Helmers SL, Begnaud J, Cowley A, Corwin HM, Edwards JC, Holder DL, Kostov H, Larsson PG, Levisohn PM, De Menezes MS, Stefan H, Labiner DM (2012) Application of a computational model of vagus nerve stimulation. Acta Neurol Scand 126:336–343.

Hummel FC, Cohen LG (2006) Non-invasive brain stimulation: a new strategy to improve neurorehabilitation after stroke? Lancet Neurol 5:708–712.

Kahana MJ, Aggarwal EV, Phan TD (2018) The variability puzzle in human memory. J Exp Psychol Learn Mem Cogn 44:1857–1863.

Katz B, Miledi R (1967) A study of synaptic transmission in the absence of nerve impulses. J Physiol 192:407–436.

Kim U, Bal T, McCormick DA (1995) Spindle waves are propagating synchronized oscillations in the ferret LGNd in vitro. J Neurophysiol 74:1301–1323.

Komarov M, Malerba P, Golden R, Nunez P, Halgren E, Bazhenov M (2019) Selective recruitment of cortical neurons by electrical stimulation. PLoS Comput Biol 15:e1007277.

Kucewicz MT, Berry BM, Kremen V, Miller LR, Khadjevand F, Ezzyat Y, Stein JM, Wanda P, Sperling MR, Gorniak R, Davis KA, Jobst BC, Gross RE, Lega B, Stead SM, Rizzuto DS, Kahana MJ, Worrell GA (2018a) Electrical stimulation modulates high gamma activity and human memory performance. eNeuro 5:ENEURO.0369-17.2018.

Kucewicz MT, Berry BM, Miller LR, Khadjevand F, Ezzyat Y, Stein JM, Kremen V, Brinkmann BH, Wanda P, Sperling MR, Gorniak R, Davis KA, Jobst BC, Gross RE, Lega B, Van Gompel J, Stead SM, Rizzuto DS, Kahana MJ, Worrell GA (2018b) Evidence for verbal memory enhancement with electrical brain stimulation in the lateral temporal cortex. Brain 141:971–978.

Lewis PM, Ackland HM, Lowery AJ, Rosenfeld JV (2015) Restoration of vision in blind individuals using bionic devices: a review with a focus on cortical visual prostheses. Brain Res 1595:51–73.

Liebetanz D, Fregni F, Monte-Silva KK, Oliveira MB, Amancio-dos-Santos A, Nitsche MA, Guedes RC (2006) After-effects of transcranial direct current stimulation (tDCS) on cortical spreading depression. Neurosci Lett 398:85–90.

Markram H, Toledo-Rodriguez M, Wang Y, Gupta A, Silberberg G, Wu C (2004) Interneurons of the neocortical inhibitory system. Nat Rev Neurosci 5:793–807.

Mayberg HS, Lozano AM, Voon V, McNeely HE, Seminowicz D, Hamani C, Schwalb JM, Kennedy SH (2005) Deep brain stimulation for treatment-resistant depression. Neuron 45:651–660.

McIntyre CC, Grill WM (1999) Excitation of central nervous system neurons by nonuniform electric fields. Biophys J 76:878–888.

McIntyre CC, Grill WM, Sherman DL, Thakor NV (2004a) Cellular effects of deep brain stimulation: model-based analysis of activation and inhibition. J Neurophysiol 91:1457–1469.
McIntyre CC, Mori S, Sherman DL, Thakor NV, Vitek JL (2004b) Electric field and stimulating influence generated by deep brain stimulation of the subthalamic nucleus. Clin Neurophysiol 115:589–595.
Merrill DR, Bikson M, Jefferys JG (2005) Electrical stimulation of excitable tissue: design of efficacious and safe protocols. J Neurosci Methods 141:171–198.
Muller L, Chavane F, Reynolds J, Sejnowski TJ (2018) Cortical travelling waves: mechanisms and computational principles. Nat Rev Neurosci 19:255–268.
Muralidhar S, Wang Y, Markram H (2013) Synaptic and cellular organization of layer 1 of the developing rat somatosensory cortex. Front Neuroanat 7:52.
Nagaraj V, Lee ST, Krook-Magnuson E, Soltesz I, Benquet P, Irazoqui PP, Netoff TI (2015) Future of seizure prediction and intervention: closing the loop. J Clin Neurophysiol 32:194–206.
Papageorgiou PN, Deschner J, Papageorgiou SN (2017) Effectiveness and adverse effects of deep brain stimulation: umbrella review of meta-analyses. J Neurol Surg A Cent Eur Neurosurg 78:180–190.
Radman T, Su Y, An JH, Parra LC, Bikson M (2007) Spike timing amplifies the effect of electric fields on neurons: implications for endogenous field effects. J Neurosci 27:3030–3036.
Radman T, Ramos RL, Brumberg JC, Bikson M (2009) Role of cortical cell type and morphology in subthreshold and suprathreshold uniform electric field stimulation in vitro. Brain Stimul 2:215–228.e3.
Rall W (1962) Electrophysiology of a dendritic neuron model. Biophys J 2:145–167.
Raspopovic S, Capogrosso M, Micera S (2011) A computational model for the stimulation of rat sciatic nerve using a transverse intrafascicular multichannel electrode. IEEE Trans Neural Syst Rehabil Eng 19:333–344.
Rattay F (1999) The basic mechanism for the electrical stimulation of the nervous system. Neuroscience 89:335–346.
Rattay F, Aberham M (1993) Modeling axon membranes for functional electrical stimulation. IEEE Trans Biomed Eng 40:1201–1209.
Rosen BQ, Krishnan GP, Sanda P, Komarov M, Sejnowski T, Rulkov N, Ulbert I, Eross L, Madsen J, Devinsky O, Doyle W, Fabo D, Cash S, Bazhenov M, Halgren E (2019) Simulating human sleep spindle MEG and EEG from ion channel and circuit level dynamics. J Neurosci Methods 316:46–57.
Rubino D, Robbins KA, Hatsopoulos NG (2006) Propagating waves mediate information transfer in the motor cortex. Nat Neurosci 9:1549–1557.
Salimpour Y, Anderson WS (2019) Cross-frequency coupling based neuromodulation for treating neurological disorders. Front Neurosci 13:125.
Salzman CD, Britten KH, Newsome WT (1990) Cortical microstimulation influences perceptual judgements of motion direction. Nature 346:174–177.
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.
Santos E, Sanchez-Porras R, Dohmen C, Hertle D, Unterberg AW, Sakowitz OW (2012) Spreading depolarizations in a case of migraine-related stroke. Cephalalgia 32:433–436.
Schlaepfer TE, Cohen MX, Frick C, Kosel M, Brodesser D, Axmacher N, Joe AY, Kreft M, Lenartz D, Sturm V (2008) Deep brain stimulation to reward circuitry alleviates anhedonia in refractory major depression. Neuropsychopharmacology 33:368–377.
Schubert D, Kotter R, Luhmann HJ, Staiger JF (2006) Morphology, electrophysiology and functional input connectivity of pyramidal neurons characterizes a genuine layer Va in the primary somatosensory cortex. Cereb Cortex 16:223–236.
Staiger JF, Flagmeyer I, Schubert D, Zilles K, Kotter R, Luhmann HJ (2004) Functional diversity of layer IV spiny neurons in rat somatosensory cortex: quantitative morphology of electrophysiologically characterized and biocytin labeled cells. Cereb Cortex 14:690–701.
Stieger KC, Eles JR, Ludwig KA, Kozai TD (2020) In vivo microstimulation with cathodic and anodic asymmetric waveforms modulates spatiotemporal calcium dynamics in cortical neuropil and pyramidal neurons of male mice. J Neurosci Res 98:2072–2095.
Suthana N, Haneef Z, Stern J, Mukamel R, Behnke E, Knowlton B, Fried I (2012) Memory enhancement and deep-brain stimulation of the entorhinal area. N Engl J Med 366:502–510.
Swadlow HA (1992) Monitoring the excitability of neocortical efferent neurons to direct activation by extracellular current pulses. J Neurophysiol 68:605–619.
Tehovnik EJ, Slocum WM, Schiller PH (2002) Differential effects of laminar stimulation of V1 cortex on target selection by macaque monkeys. Eur J Neurosci 16:751–760.
Tehovnik EJ, Tolias AS, Sultan F, Slocum WM, Logothetis NK (2006) Direct and indirect activation of cortical neurons by electrical microstimulation. J Neurophysiol 96:512–521.
Thomson AM, West DC, Wang Y, Bannister AP (2002) Synaptic connections and small circuits involving excitatory and inhibitory neurons in layers 2-5 of adult rat and cat neocortex: triple intracellular recordings and biocytin labelling in vitro. Cereb Cortex 12:936–953.
Tomassy GS, Berger DR, Chen HH, Kasthuri N, Hayworth KJ, Vercelli A, Seung HS, Lichtman JW, Arlotta P (2014) Distinct profiles of myelin distribution along single axons of pyramidal neurons in the neocortex. Science 344:319–324.
Tranchina D, Nicholson C (1986) A model for the polarization of neurons by extrinsically applied electric fields. Biophys J 50:1139–1156.
Traub RD, Jefferys JG, Miles R, Whittington MA, Toth K (1994) A branching dendritic model of a rodent CA3 pyramidal neurone. J Physiol 481:79–95.
Wang Y, Gupta A, Toledo-Rodriguez M, Wu CZ, Markram H (2002) Anatomical, physiological, molecular and circuit properties of nest basket cells in the developing somatosensory cortex. Cereb Cortex 12:395–410.
Wang Y, Toledo-Rodriguez M, Gupta A, Wu C, Silberberg G, Luo J, Markram H (2004) Anatomical, physiological and molecular properties of Martinotti cells in the somatosensory cortex of the juvenile rat. J Physiol 561:65–90.
Wester JC, Contreras D (2012) Columnar interactions determine horizontal propagation of recurrent network activity in neocortex. J Neurosci 32:5454–5471.
Whittington MA, Traub RD, Kopell N, Ermentrout B, Buhl EH (2000) Inhibition-based rhythms: experimental and mathematical observations on network dynamics. Int J Psychophysiol 38:315–336.
Wongsarnpigoon A, Grill WM (2008) Computational modeling of epidural cortical stimulation. J Neural Eng 5:443–454.
Wongsarnpigoon A, Grill WM (2012) Computer-based model of epidural motor cortex stimulation: effects of electrode position and geometry on activation of cortical neurons. Clin Neurophysiol 123:160–172.
Wu JY, Guan L, Bai L, Yang Q (2001) Spatiotemporal properties of an evoked population activity in rat sensory cortical slices. J Neurophysiol 86:2461–2474.
Wu JY, Xiaoying H, Chuan Z (2008) Propagating waves of activity in the neocortex: what they are, what they do. Neuroscientist 14:487–502.
Xu W, Huang X, Takagaki K, Wu JY (2007) Compression and reflection of visually evoked cortical waves. Neuron 55:119–129.
Zhang H, Watrous AJ, Patel A, Jacobs J (2018) Theta and alpha oscillations are traveling waves in the human neocortex. Neuron 98:1269–1281.e4.

Chapter 3, in full, is a reprint of the material as it appears in The Journal of Neuroscience, under the title “Multielectrode cortical stimulation selectively induces unidirectional wave propagation of excitatory neuronal activity in biophysical neural model”, April 2023; 43(14):2482-2496. Halgren, Alma S.; Siegel, Zarek; Golden, Ryan; Bazhenov, Maxim. The dissertation author was one of the primary investigators and authors of this paper, but also serving in a supervisory role as a mentor for the primary investigator, Alma S. Halgren.