PNAS

Supplementary Information for
A model of autonomous interactions between hippocampus and neocortex driving sleep-dependent memory consolidation

Dhairyya Singh[1], Kenneth A. Norman[2], & Anna C. Schapiro[1]

[1]Department of Psychology, University of Pennsylvania
[2]Department of Psychology and Princeton Neuroscience Institute, Princeton University

Authors for correspondence: Dhairyya Singh, Anna C. Schapiro
Email: dsin@sas.upenn.edu, aschapir@sas.upenn.edu

This PDF file includes:

Supplementary Information Text

Supplementary Methods

Emergent. The model, implemented in the Emergent framework (1), consists of units with rate code activations (using the NXX1 sigmoidal activation function, which approximates average firing rate in neuronal populations as a function of excitatory input), organized into layers and connected by learnable weights. Feedforward feedback (FFFB) inhibition simulates inhibitory network dynamics as implemented by inhibitory interneurons in biological neural networks (2). These are Emergent defaults.

Model architecture. The model architecture for both simulations consisted of the four C-HORSE hidden layers representing the DG, CA3, pCA1, and dCA1 subfields of the hippocampus (3), and one neocortical hidden layer (see Supplementary Material for parameter details). Previously C-HORSE has been employed to simulate hippocampal contributions to statistical learning, associative inference, and category learning (3–5) and comes from a lineage of hippocampal models of episodic memory (6–8). We adopted the version from Zhou et al. (3), which divides CA1 into proximal and distal components. Input/output layers in Simulation 1 consisted of seven feature layers, each corresponding to a satellite feature's high level visual and verbal representation in Entorhinal Cortex (EC). Simulation 2 had a separate input and output layer, corresponding to superficial and deep EC layers (3, 4).

Training for Simulation 1. Each satellite consisted of seven features (five visual parts, a class name, and a code name) of which some were shared across exemplars from a category and others were unique to each exemplar. Satellites were chosen randomly in each training trial and one feature was held out. Features were presented as one-hot inputs on the corresponding input/output feature layer and were held out at a ratio of unique:shared = 99:1 (to better match unique and shared learning speed, unique features need to be queried the vast majority of the time). Each training trial consisted of a minus and plus phase. In the minus phase the satellite was presented to the model with a feature held out and the model generated a prediction for the identity of the feature. In the plus phase, the full satellite was presented to the model. The model was trained to a criterion of 66% feature completion performance for shared and for unique features.

Training for Simulation 2. Each training trial minus phase involved clamping an Env 1 or Env 2 item on the input layer and requiring the model to reproduce that item on the output layer. Each environment had 10 items and each item consisted of seven units. Env 1 and Env 2 item units were chosen such that there was a 2/7 unit overlap between corresponding items from Env 1 and Env 2 (between Item 1 from Env 1 and Item 1 from Env 2, Item 2 from Env 1 and Item 2 from Env 2, and so on). There was no overlap between items from the same environment. The same set of items were used across all initializations of the model. In the plus phase, the item was clamped onto both the input and output layer.

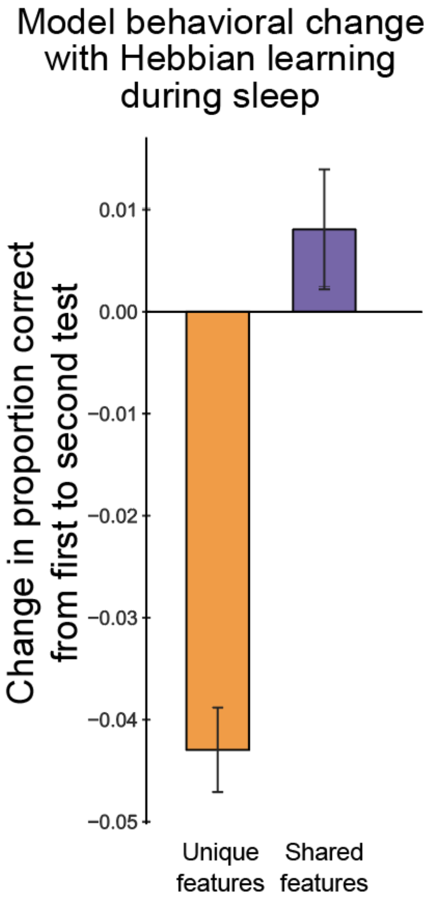

Fig S1. Simulation 1 was run with a simple Hebbian learning rule (weights were updated based on plus phase coactivity exclusively) during sleep instead of Contrastive Hebbian Learning. The learning rate for the Hebbian simulations was lowered such that the average weight change magnitude per sleep trial was approximately matched between the two conditions. We found that the Hebbian learning rule results in substantial forgetting of unique information and less robust increases in shared feature memory.

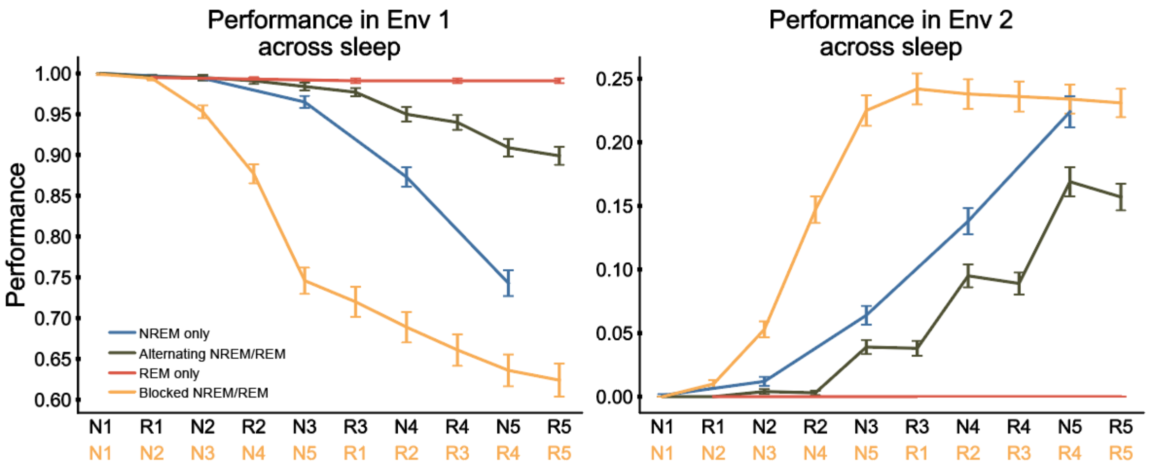

Fig S2. We ran a blocked NREM/REM schedule (5 epochs of NREM followed by 5 epochs of REM) to test whether REM can protect Env 1 memories without sleep stage alternations. This condition resulted in catastrophic forgetting of Env 1, indicating that REM cannot repair Env 1 memories after extensive damage and that the alternating schedule is necessary for Env 1 protection.

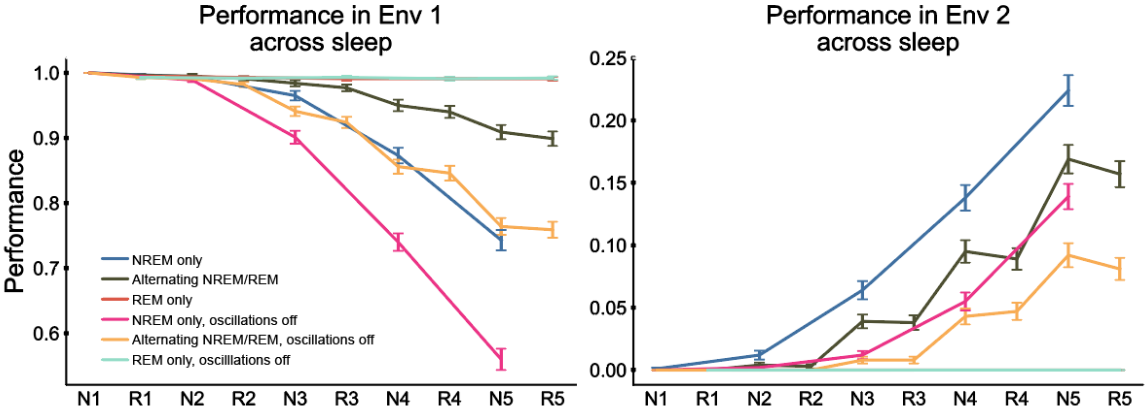

Fig S3. We ran each of the three Simulation 2 conditions (NREM only, Alternating NREM/REM, and REM only) without oscillating inhibition. The REM only, oscillations off condition is plotted on top of the REM only condition. Both the NREM only, oscillations off and Alternating NREM/REM, oscillations off conditions are substantially worse at improving Env 2 performance relative to their counterparts with oscillations, indicating that oscillations play an important role in NREM learning. They also lead to substantially more damage to Env 1 performance relative to their counterparts with oscillations. The REM only and REM only, oscillations off conditions are nearly identical and cause neither Env 1 damage nor Env 2 improvement.

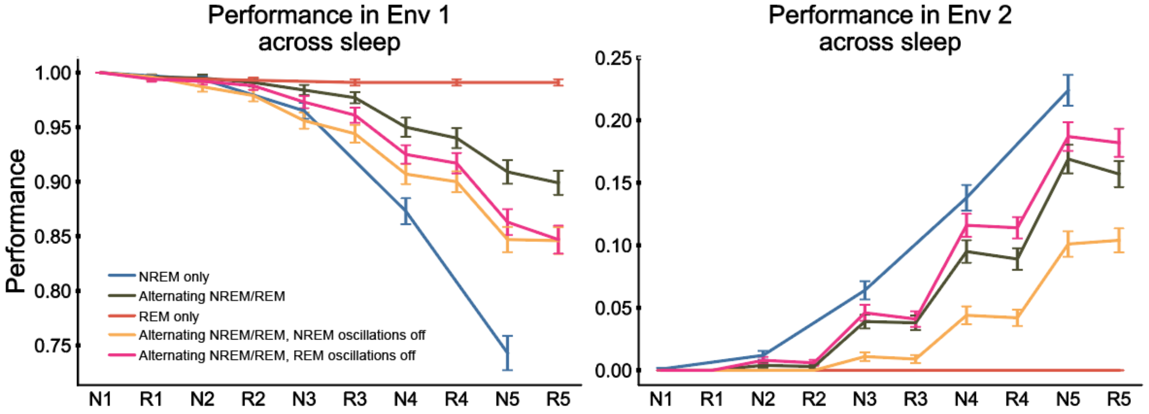

Fig S4. The Alternating NREM/REM condition was run with oscillating inhibition selectively turned off in either REM or NREM epochs to further clarify the stage-specific contributions of oscillating inhibition to performance under alternating stages. Both conditions cause more damage to Env 1 performance relative to the alternating condition with oscillations. When NREM oscillations are turned off, there is substantially less Env 2 performance improvement. These simulations reveal that oscillations in both sleep stages facilitate better protection of Env 1, while oscillations in NREM facilitate better Env 2 learning (consistent with Fig. S3 results).

Table S1. Layer Parameters (Simulation 1)

Layer	**Parameter**	**Value**
pCA1	# Units	50
	Inhibition	FFFB (Gi = 2.2)
dCA1	# Units	50
	Inhibition	FFFB (Gi = 2.2)
DG	# Units	225
	Inhibition	FFFB Gi = 2.8
CA3	# Units	100
	Inhibition	FFFB Gi = 5
Feature Layers	# Units	6/3/90 (Visual/Classname/Codename)
	Inhibition	FFFB Gi = 2.1/1.6/2.6 (Visual/Classname/Codename)
Neocortex	# Units	400
	Inhibition	FFFB Gi = 2.4

Table S2. Projection Parameters (Simulation 1)

Projection	**Parameter**	**Value**
Default	Weight Init Distribution	0.25 - 0.75
Default	Connectivity	Full
Default	Learning rate	0.05
Feature Layers -> DG	Learning rate	0.1
Feature Layers -> DG	Connectivity Sparsity	0.6
Feature Layers -> CA3	Learning rate	0.1
Feature Layers -> CA3	Connectivity Sparsity	0.1
DG -> CA3	Learning rate	0
DG -> CA3	Connectivity Sparsity	0.1
DG -> CA3	Weight Init Distribution	0.89 - 0.91
CA3 -> CA3	Learning rate	0.1
CA3 -> CA3	Weight Init Distribution	0 - 0.6
Feature Layers <-> Neocortex	Learning rate	0.0001

Table S3. Sleep Parameters (Simulation 1)

Projection	**Parameter**	**Value**
Feature Layers <-> Neocortex	Learning rate	0.01
Feature Layers -> DG	Learning rate	0
Feature Layers -> CA3	Learning rate	0
CA3 -> pCA1	Learning rate	0
pCA1 -> Feature Layers	Learning rate	0
Feature Layers <-> dCA1	Learning rate	0
Synaptic Depression	Inc	0.00035
	Dec	0.00025

Table S4. Layer Parameters (Simulation 2)

Layer	**Parameter**	**Value**
pCA1	# Units	50
	Inhibition	FFFB (Gi = 2.2)
dCA1	# Units	50
	Inhibition	FFFB (Gi = 2.2)
DG	# Units	225
	Inhibition	FFFB Gi = 2.8
CA3	# Units	100
	Inhibition	FFFB Gi = 5
Input/Output	# Units	120
	Inhibition	FFFB Gi = 3.2/2.8 (Input/Output)
Neocortex	# Units	400
	Inhibition	FFFB Gi = 2.4

Table S5. Projection Parameters (Simulation 2)

Projection	**Parameter**	**Value**
Default	Weight Init Distribution	0.25 - 0.75
Default	Connectivity	Full
Default	Learning rate	0.05
Input -> DG	Learning rate	0.1
Input -> DG	Connectivity	0.6
Input -> CA3	Learning rate	0.1
DG -> CA3	Learning rate	0
DG -> CA3	Connectivity Sparsity	0.1
DG -> CA3	Weight Init Distribution	0.89 - 0.91
CA3 -> CA3	Learning rate	0.1
CA3 -> CA3	Weight Init Distribution	0 - 0.6
Input -> Neocortex	Learning rate	0.0001
Neocortex -> Output	Learning rate	0.0001

Table S6. Sleep Parameters (Simulation 2)

Projection	**Parameter**	**Value**
Input <-> Neocortex	Learning rate	0.01
Neocortex -> Output	Learning rate	0.01
pCA1 -> Neocortex	Learning rate	0
dCA1 -> Neocortex	Learning rate	0
Feature Layers -> DG	Learning rate	0
Feature Layers -> CA3	Learning rate	0
CA3 -> pCA1	Learning rate	0
pCA1 -> Ouput	Learning rate	0
Input <-> dCA1	Learning rate	0
Synaptic Depression	Inc	0.0006
	Dec	0.0003

SI References

1. B. Aisa, B. Mingus, R. O'Reilly, The Emergent neural modeling system. *Neural Netw.* **21**, 1146–1152 (2008).
2. R. C. O'Reilly, Y. Munakata, M. J. Frank, T. E. Hazy, Contributors, *Computational Cognitive Neuroscience*, 4th Ed. (2020).
3. Z. Zhou, D. Singh, M. C. Tandoc, A. C. Schapiro, Distributed representations for human inference. bioRxiv (2021). https:/doi.org/10.1101/2021.07.29.454337.
4. A. C. Schapiro, N. B. Turk-Browne, M. M. Botvinick, K. A. Norman, Complementary learning systems within the hippocampus: a neural network modelling approach to reconciling episodic memory with statistical learning. *Philos. Trans. R. Soc. B Biol. Sci.* **372**, 20160049 (2017).
5. J. Sučević, A. C. Schapiro, A neural network model of hippocampal contributions to category learning. bioRxiv (2022). https:/doi.org/10.1101/2022.01.12.476051.
6. N. Ketz, S. G. Morkonda, R. C. O'Reilly, Theta Coordinated Error-Driven Learning in the Hippocampus. *PLoS Comput. Biol.* **9**, e1003067 (2013).
7. K. A. Norman, R. C. O'Reilly, Modeling hippocampal and neocortical contributions to recognition memory: A complementary-learning-systems approach. *Psychol. Rev.* **110**, 611–646 (2003).
8. R. C. O'Reilly, K. A. Norman, J. L. McClelland, "A Hippocampal Model of Recognition Memory" in *Advances in Neural Information Processing Systems*, M. Jordan, M. Kearns, S. Solla, Eds. (MIT Press, Cambridge, MA, 1997), pp. 73-79.