Supplementary material for:

Complementary learning systems within the hippocampus: A neural network modeling approach to reconciling episodic memory with statistical learning

Anna C. Schapiro, Nicholas B. Turk-Browne, Matthew M. Botvinick, & Kenneth A. Norman

Phil. Trans. R. Soc. B. doi: 10.1098/rstb.2016.0049

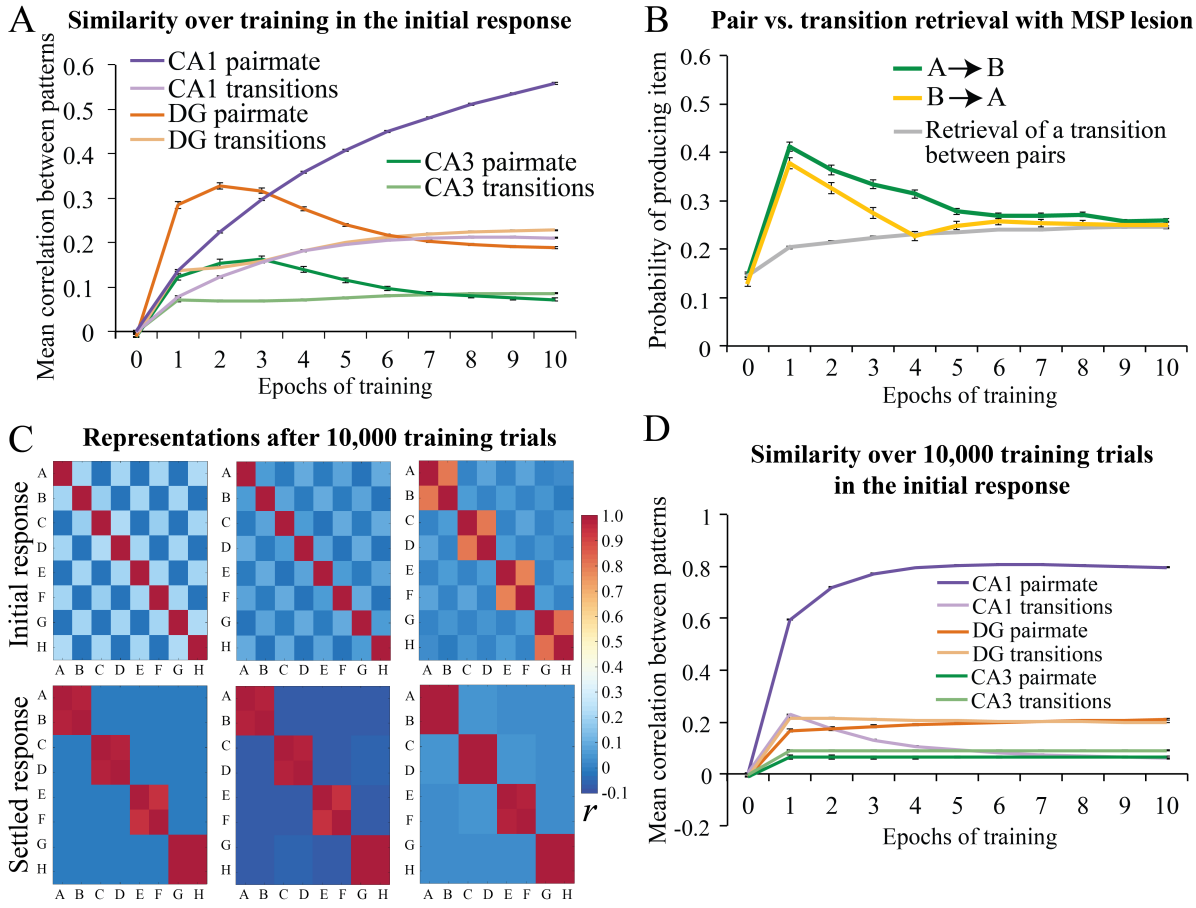
Slots in CA1

The Ketz et al. (2013) model and previous versions of this model (e.g., 1) grouped the units in EC_{in}, EC_{out}, and CA1 into distinct "slots", in which particular subsets of units in EC_{in}/EC_{out} connected only to particular subsets of units in CA1. This was intended to reflect the topographic organization of the MSP — ventral, dorsal, medial, and lateral projections remain relatively segregated (2). The topography is coarse, though, and it seems likely that the stimuli used in prior statistical learning experiments (e.g., 3) would have representations that fall within a small enough portion of EC to make the topography largely irrelevant. We therefore did not use slots in the current simulations.

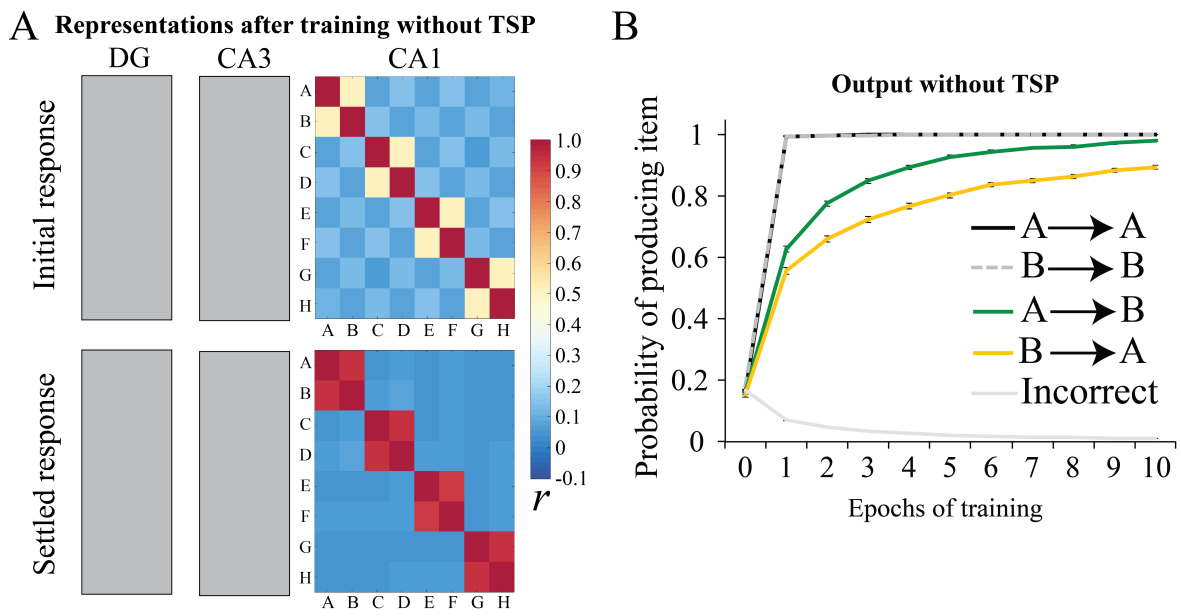
However, it remains possible that slots played an important functional role in prior episodic memory simulations using this model. To test this, we ran simulations using the classic AB-AC interference paradigm. In this paradigm, the model learns a set of item pairs AB followed by an interfering set AC. This blocked presentation causes catastrophic interference in cortical models but poses little challenge for the hippocampus model. We used the hip.proj network and inputs from (4). We modified the project to use a new seed for each run and to re-initialize the sparse projections and weights randomly for each run. We then ran 20 networks with the default slot setup, and 20 networks with slots removed. We implemented the slot removal by changing all projections in and out of CA1 that used GpOneToOne to FullPrjn, and setting CA1 inhib to operate over the entire layer with gi=2.0.

We ran 6 epochs of AB training, followed by 6 epochs of AC training. Surprisingly, the model without slots exhibited *less* interference: At the end of the 12 epochs, the slot model had mean AB performance of 0.66 and mean AC performance of 0.98, whereas the no-slot model achieved performance levels of 0.83 and 1.0, respectively. Both AB and AC performance were significantly better without slots (AB: $t[38]=4.21$, $p<0.001$; AC: $t[38]=2.08$, $p=0.04$). Thus, slots are not likely to be necessary for this kind of episodic memory model, at least when simulating domains with stimuli processed within small patches of EC.

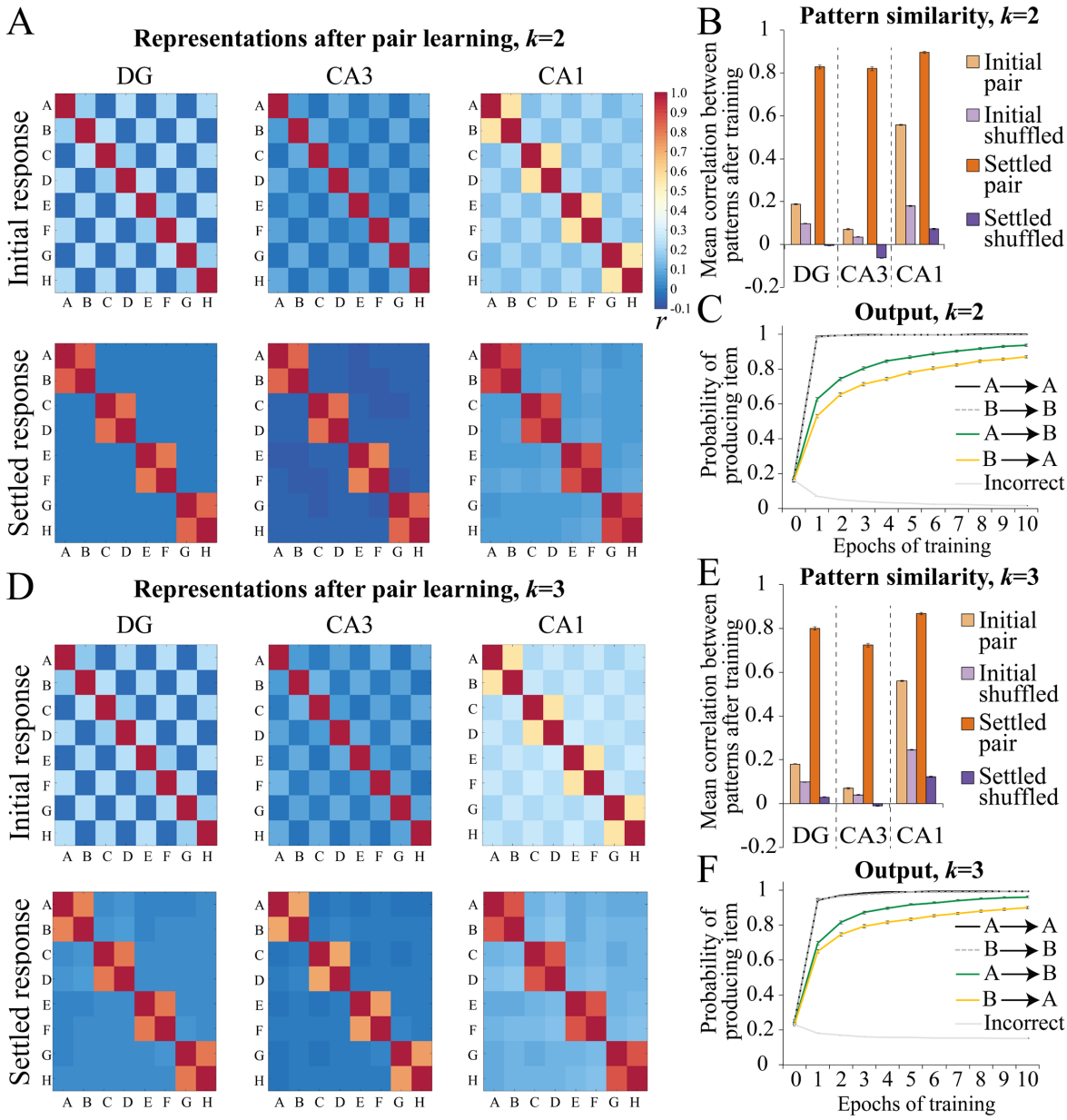
We did not test earlier versions of the model that used a pre-trained MSP (discussed in Methods), so it is possible that those models did benefit from the slot structure, especially given their use of Hebbian learning. However, the simulations above suggest that the latest model, which uses online error-driven MSP training and has a greater memory capacity than previous versions, may benefit from more diffuse connectivity. It is also worth noting that our MSP findings can generalize to slot-based architectures insofar as slots contain at least partial representations of more than one item.



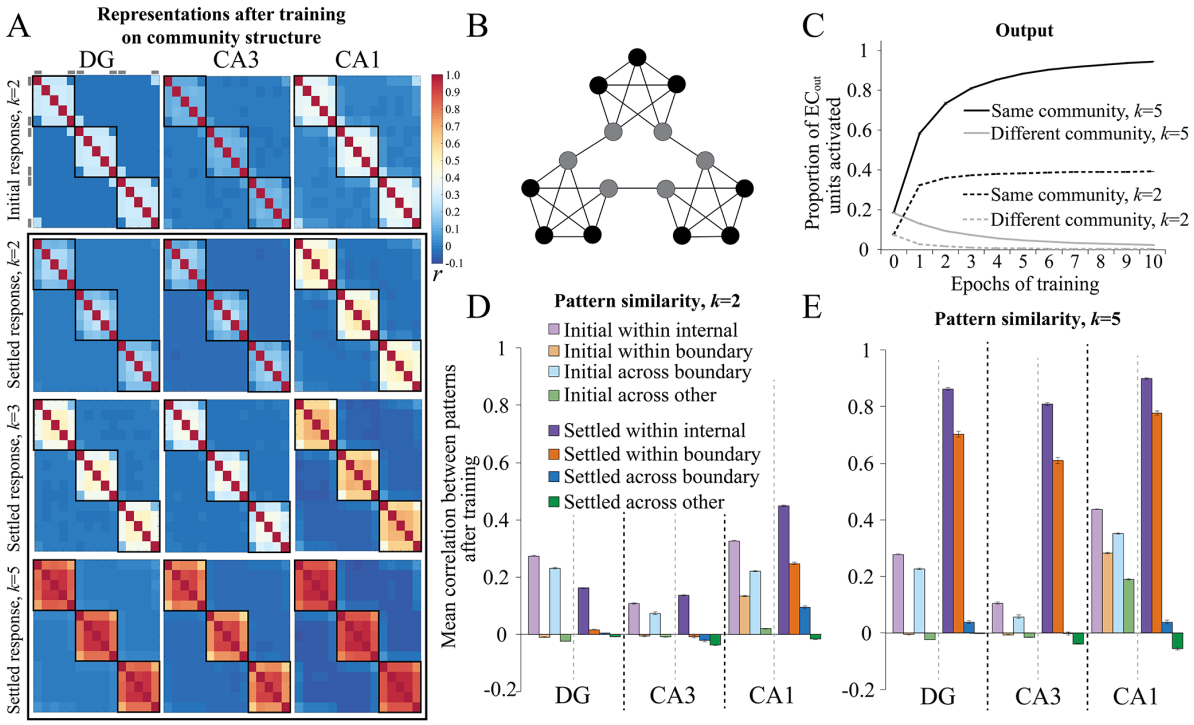
Supplementary Figure 1. Timecourse of pair structure learning. **(A)** Pattern similarity in the initial response in each hidden layer between two members of a pair (e.g., *AB*) and between two items that transitioned between pairs (e.g., *DA*), over the course of the 10 epochs of training. Pair similarity increases initially in DG and CA3 but weakens over time due to memorization of transition pairs, which fully catch up to and even slightly surpass the pairmate similarity. The non-monotonic change in the TSP initial response also occurred when pure Hebbian learning was used in the TSP. **(B)** Retrieval in EC_{out} of an item given its pairmate and retrieval of an item given an adjacent across-pair item (*D* → *A*), with the MSP lesioned. It is worth noting that although the MSP, in addition to the TSP, is important for episodic learning, MSP lesions were much less detrimental to learning the sequences that did not require statistical learning (i.e., with pairs presented separately). In that case, the highest mean probability of producing *B* in EC_{out} given *A* as input was 0.93 and the highest mean probability of producing *A* in EC_{out} given *B* as input was also 0.93. (Without a lesion, both probabilities reached 0.97.) **(C)** Pattern similarity after 10,000 training trials, with 1,000 trials per epoch over 10 epochs. (By default there are 800 total trials, with 80 trials per epoch over 10 epochs.) CA1 structure did not deteriorate over time, as was seen with CA3 and DG. Initial pair structure was very strong in CA1 after this extensive training, and settled structure was even stronger throughout the network. **(D)** Timecourse of pattern similarity in the initial response in each region over 10,000 trials.



Supplementary Figure 2. Undeveloped TSP. **(A)** Average representational similarity after training with a TSP lesion. DG and CA3 did not have any activity, so there was no structure in those regions. **(B)** Average probability of activating a particular item on the output given a particular item on the input, over training. Output and CA1 similarity structure were very similar to the intact model (Fig. 2D, 2F).



Supplementary Figure 3. Inhibition and pair learning. **(A, B, C)** For reference purposes, exact copies of Figure 2D-F, where k=2 in EC_{in} and EC_{out} at test. **(D)** Average representational similarity for the initial and settled response with k=3 at test (k=2 was used in all cases during training). The results are very similar, with slightly higher similarity for shuffled pairs in CA1. **(E)** Average representational similarity by pair type. **(F)** Average probability of activating a particular item on the output given a particular item on the input, over training. With k=3, incorrect responses are more frequent. This is the cause of the higher shuffled similarity in CA1, as the network is encouraged to activate a third unit (despite it always being incorrect).



Supplementary Figure 4. Inhibition and community structure. **(A)** Average representational similarity for the initial response using k=2 at test, and for the settled response using k=2, k=3, and k=5 at test (initial response is qualitatively similar across inhibition values). Top two rows correspond to Figure 5A, visualized on a different scale. Lowering inhibition at test strengthened community structure — each additional activated member of the community helped emphasize their common structure. **(B)** The graph used to generate the sequences, for reference (identical to Figure 3B). **(C)** Average proportion of output units activated that were from the same versus different community as the test item, over training. The k=2 variant cannot activate more than 2 out of 5 community members, so the best possible performance is 0.4, whereas the k=5 variant can activate all members and achieve 1.0 (k=3, not shown to reduce clutter, falls between k=2 and k=5). **(D, E)** Average representational similarity with k=2 and k=5 (k=3 is intermediate) for two neighboring nodes from the same community (within internal), the two boundary nodes from the same community (within boundary), two adjacent boundary nodes from different communities (across boundary), and all other pairs of items from different communities (across other). Higher-level structure is more prominent after settling with k=5: within boundary is almost as similar as within internal, and across boundary is almost as dissimilar as across other. Thus, across both the associative inference simulations (described in the main paper) and the community structure simulations, allowing activation to spread with lower inhibition (i.e., higher k) is useful in promoting transitive inference. In fact, spreading activation with low inhibition at test is, in theory, *sufficient* to uncover transitive associates in both the community structure and associative inference paradigms, which means that the TSP can potentially support successful behavior in these cases. This is difficult to probe in our model, however, because lesions to the MSP result in difficulties conveying information from the TSP to EC. Another crucial note here is that, while allowing activation to spread further at test (with learning turned off) is useful, allowing activation to spread further *during learning* can have deleterious consequences in the current version of the model. With the model's current learning rule, allowing the transitive associate to pop up in minus phases during training is detrimental to transitive behavior and representations because the plus phase (the target pattern) always contains only the direct associates — the two items actually presented together. In this situation, since the transitive associate is present during the minus phase but not the plus phase, the learning rule will act to weaken the transitive associate. Future modeling work could explore modifications to the learning rule such that strong retrieval of an associate that is not currently presented might lead to encoding of the transitive association, as has been proposed in integrative encoding accounts (5). Alternatively, the hippocampus may have periods of relatively low-

inhibition retrieval in which no learning takes place, akin to the idea that low levels of acetylcholine in the hippocampus can allow retrieval without new learning (6, 7). This modulation would occur at a lower frequency than the theta oscillation (8, 9) and might be initiated by explicit or strategic attempts to retrieve additional associates.

Area	# Units	kWTA type	Proportion activity (kWTA pct)	kWTA pt
EC_{in} and EC_{out}	8 / 15 / 9	kWTA Inhib	*k*=2 (varying pct)	0.5
DG	400	kWTA Avg Inhib	0.01	0.9
CA3	80	kWTA Avg Inhib	0.06	0.7
CA1	100	kWTA Avg Inhib	0.25	0.7

Supplementary Table 1. Parameters for layer sizes and inhibition, as implemented in the Emergent simulation environment (4). Units in EC refer to pair learning / community structure / associative inference simulations.

Projection	Weight range	Scale (abs / rel)	Connectivity	lrate
Input → EC_{in}	0.25 – 0.75	1 / 1	1 to 1	0
EC_{in} → DG	0.25 – 0.75	1 / 1	25%	0.2
EC_{in} → CA3	0.25 – 0.75	1 / 1	25%	0.2
DG → CA3 (*mossy fiber*)	0.89 – 0.91	1 / 8	5%	0
CA3 → CA3	0.25 – 0.75	1 / 1	100%	0.2
CA3 → CA1 (*Schaffer*)	0.25 – 0.75	1 / 1	100%	0.05
EC_{in} → CA1	0.25 – 0.75	3 / 1	100%	0.02
CA1 → EC_{out}	0.25 – 0.75	1 / 1	100%	0.02
EC_{out} → CA1	0.25 – 0.75	1 / 1	100%	0.02
EC_{out} → EC_{in}	0.49 – 0.51	2 / .5	1 to 1	0

Supplementary Table 2. Parameters for projections between layers. Weight range = range for uniform distribution over which weights were initialized. Scale = scaling of the projection relative to others (abs = absolute multiplier on weights in the projection; rel = relative weighting taking into account other projections to the layer). Connectivity = percent of units in sending layer projecting to a given unit in the receiving layer (rounded to nearest integer). For example, 25% connectivity from EC_{in} to DG in the associative inference paradigm (9 EC_{in} units) means that each DG unit receives input from 2 EC_{in} units. "1 to 1" means that each unit in one layer connects to one unit in the other layer. lrate = learning rate. Note that by default, the CA1 → EC_{out} pathway has scale abs = 4. We removed this because the kWTA dynamics in the model tend to cause full activation of the two stimuli in EC_{out}, even when the input and target are graded (which is the case here, where activity is decayed for the previous item in a sequence). Decreasing the strength of this pathway reduces the tendency to fully activate both stimulus output units, which helps resist the kWTA dynamics and allows the model to produce activity in the minus phases that is better matched to the graded activity in the plus phase. We also chose a small decay (previous stimulus clamped to 0.9) because a steeper decay would cause more tension with these kWTA dynamics. These modifications help the network handle graded values given its particular implementation of competitive inhibition; however, we do not view them as assumptions or predictions about actual hippocampal architecture or dynamics.

References

1. O'Reilly RC, Rudy JW. (2001). Conjunctive representations in learning and memory: principles of cortical and hippocampal function. *Psychol Rev*. 108(2):311-45.

2. Amaral DG, Witter MP. (1995). Hippocampal formation. In *The rat nervous system* (ed G Paxinos), pp. 443-493. San Diego: Acad. Press Inc.

3. Fiser J, Aslin RN. (2002). Statistical learning of higher-order temporal structure from visual shape sequences. *J Exp Psychol Learn Mem Cogn*. 28(3):458-67.

4. O'Reilly RC, Munakata Y, Frank MJ, Hazy TE, Contributors. (2014). *Computational Cognitive Neuroscience, 2nd Edition*. Chapter 8. https://grey.colorado.edu/CompCogNeuro/index.php/CCNBook/Main.

5. Shohamy D, Wagner AD. (2008). Integrating memories in the human brain: hippocampal-midbrain encoding of overlapping events. *Neuron.* 60(2):378-89.

6. Hasselmo ME, Bower JM. (1993). Acetylcholine and memory. *Trends Neurosci*. 16(6):218-22.

7. Kumaran D, McClelland JL. (2012). Generalization through the recurrent interaction of episodic memories: a model of the hippocampal system. *Psychol Rev*. 119(3):573-616.

8. Duncan K, Sadanand A, Davachi L. (2012). Memory's penumbra: episodic memory decisions induce lingering mnemonic biases. *Science.* 337(6093):485-7.

9. Hasselmo ME, Fehlau BP. (2001). Differences in time course of ACh and GABA modulation of excitatory synaptic potentials in slices of rat hippocampus. *J Neurophysiol.* 86(4):1792-802.