

Unification of cognitive maps and relational memory via generalization in the hippocampal formation

Young Joon Kim



Wadham

This dissertation is submitted on July, 2022 for the degree of Master of Science

Declaration

This dissertation is the result of my own work and includes nothing which is the outcome of work done in collaboration except as declared in the Preface and specified in the text. It is not substantially the same as any that I have submitted, or am concurrently submitting, for a degree or diploma or other qualification at the University of Oxford or any other University or similar institution except as declared in the Preface and specified in the text. I further state that no substantial part of my dissertation has already been submitted, or is being concurrently submitted, for any such degree, diploma or other qualification at the University of Oxford or any other University or similar institution except as declared in the Preface and specified in the text. This dissertation does not exceed the prescribed limit of 30,000 words.

Young Joon Kim
July, 2022

Abstract

Unification of cognitive maps and relational memory via generalization in the hippocampal formation

Young Joon Kim

The hippocampal formation has been implicated in both learning and generalization. The large variety of neural representations, ranging from grid and border cells to place and object-vector cells, observed in the hippocampus, entorhinal cortex, and surrounding brain regions are believed to form the neural bases for intelligent learning. In conjunction with countless experimental studies, several computational models of the hippocampal formation have been proposed as explanations for how the neural representations can support learning. Most such models, however, exhibit one of two critical flaws: they either fail to generalize and transfer what they have learned from one environment to another or they can only do so after having observed many training environments and become useless in environments with different underlying structure. Needless to say, a unified explanation for how the hippocampal formation can enable generalizable learning remains elusive. Here, I present a more unified model that combines the two aforementioned model types and reciprocally addresses each other's weaknesses. I demonstrate that this combined model enables faster generalization while maintaining flexibility across different environment or task structures, rendering this model a potentially more complete description of hippocampal learning.

Acknowledgements

I would like to sincerely thank Prof. Timothy E.J. Behrens for all the guidance and mentorship he provided me throughout the duration of the project. From its inception to its conclusion, my research would not have been feasible without much of his willingness to help me think through challenges and to shed light on the most confusing results. I would also like to thank Dr. James C.R. Whittington for the great deal of technical insight and trouble-shooting assistance he offered this past year. In particular, without his help, I would not have been able to proceed with the modeling component of my project. Finally, I would like to thank my mother and father, who have been my most ardent supporters throughout my entire life. Their encouragement helped me push through the hurdles that this past year presented.

Contents

1 Introduction **11**

1.1 Computational models of the hippocampal formation 11

1.2 Hippocampus as map versus memory 12

1.3 Proposed joint model 13

2 Results **15**

2.1 Model and training description 15

2.2 Generalization across tasks with same task structure 18

2.3 Generalization across different task structures 19

2.4 Potential neural correlates of map and memory in CA1 and CA3, respectively . 20

3 Discussion **23**

3.1 Unaddressed technical challenges 23

3.2 Unverified experimental predictions 24

References **25**

Chapter 1

Introduction

Generalization, which refers to an agent's ability to extract transferrable knowledge from one particular setting and apply it to another, is a hallmark of natural intelligence. Investigating the mechanisms underlying generalization has been of interest to both neuroscientists uncovering the neural bases of learning and machine learning researchers designing more intelligent models. At the heart of this discourse lies the hippocampus and surrounding brain regions. In the past few decades, the hippocampal formation has been implicated in learning and generalization for both spatial and nonspatial tasks.[22, 5]

A large variety of neural representations, including place, grid, object-vector, and border cells, have been identified in experimentally recorded brain activity from both animals and humans. Neuroscientists have hypothesized that these, along with a potpourri of other hippocampal phenomena, such as remapping and drift, are the neural tools that enable intelligent learning. [22, 10, 3, 31] Naturally, formulating a realistic model that links the aforementioned neural phenomena to complex learning and behavior is of great interest to the neuroscience community.

1.1 Computational models of the hippocampal formation

Accordingly, computational neuroscientists have taken on the challenge of creating an algorithmic summary of hippocampal learning. While there exist several recent counter-examples (e.g. Tolman-Eichenbaum Machine, see below)[31], most computational models fail to account for the existence of such a large variety of hippocampal representations and to mimic naturalistic, intelligent learning behavior observed in animals and humans.[30] Indeed with the ever-increasing library of experimental studies of the hippocampus, formulating such a unifying model is no trivial feat.

In the following study, I focus on two complementary weaknesses of extant computational models that I hope to address in a joint model. The first shortcoming involves the model's ability to replicate the capacity for generalization via transfer learning that animals and humans possess. One class of hippocampal models, such as the successor representation (SR)[27], attribute the

existence of place and grid cells as evidence of the hippocampus' ability to distinguish and identify latent states of the current task or environment. These models learn a transition matrix that associates the identified latent states with each other so that the agent can use this "map" to navigate through the environment. One such model, the Clone-Structure Cognitive Graph (CSCG), also provides a learning algorithm that is capable of rapid latent state identification in any spatial or non-spatial world.[9]

Despite their flexibility (i.e. latent state disambiguation is useful in any environment) and speed (i.e. learning of the latent states and their relationships amongst each other can occur within a single environment), these models critically lack the ability to generalize. Both CSCG and SR store the information on identified latent states in a transition matrix that is specific to one environment and that cannot be used to any degree outside of this single context.

A second class of models makes up for the above models' inability to generalize by learning not just the relationship between specific sensory observations, which could differ between environments, but also the relationships between abstracted latent states, which is assumed to be constant between different environments. In short, each world is construed to be a collection of latent states in which each state is a composition of an abstract "location" in physical space or task space and a sensory observation. Generalization becomes possible once the rules between abstract spaces are learned and transferred while a memory component is used to temporarily store the sensory observations and their placements on the abstract locations for each environment. The Tolman-Eichenbaum Machine (TEM) indeed uses grid cells to learn the abstract transition rules between states and place cells to temporarily memorize the conjunction between sensory observations and abstract locations.[31]

Nevertheless, models of generalization exhibit one major flaw: they can navigate environments with sufficient accuracy only after having seen and learned from thousands of training environments. We know that animals and humans, however, are still capable of rudimentary navigation even when facing unfamiliar tasks or environment types. This slow learning speed is further impeded by sensory ambiguation, which is the presence of multiple copies of an identical sensory observation in different abstract locations. Models like TEM must learn directly from the sensory sequence rather than from the potentially more informative sequence of (unobserved) latent states, slowing down learning.

1.2 Hippocampus as map versus memory

As mentioned above, models of latent state identification do so by rapidly learning a transition matrix "map" at hand that permits navigation within a single, specific environment. Meanwhile, models of generalization focus on learning the transitions between abstract locations and relegate the "learning" of the environment-specific sensory details to their memory modules. The memory components on their own do not possess navigational capabilities, but must inherit them from the

entorhinal cells' transferrable transition map. This functional differentiation of map and memory models mirrors the decades-long debate in neuroscience on the role of the hippocampus that continues to this day.

On one hand, based on its strongly supported role in spatial navigation, the hippocampal formation is thought to primarily navigate through environments by "navigating" through the corresponding internal cognitive map.[22] Once learned, these internal maps afford flexible planning and behavior.[3, 28, 21] On the other hand, Eichenbaum and colleagues have proposed that the hippocampus acts purely as a relational memory that links either sensory objects together or binds sensory objects to locations while spatial and non-spatial navigational properties arise elsewhere, such as the entorhinal cortex.[25, 8]

While many perspectives have contributed to this debate, with opinions ranging the full gamut of the map-memory spectrum, I chose to focus less on the dichotomy of hypothesized hippocampal roles and more on how both map and memory could be used together to synergistically accelerate learning and generalization. Indeed the idea of combining different, complementary learning models is not a novel suggestion (see McClelland's Complementary Learning Systems[20] and O'Keefe's Local and Taxon systems[22]).

1.3 Proposed joint model

In the following study, I propose that a unified account of the hippocampal formation can be formulated through the lens of generalization. In the process, I show that the two aforementioned, incomplete descriptions of hippocampal learning can be combined to reciprocally address each other's weaknesses. Briefly, the model combines TEM and CSCG in such a way that the two components learn to improve each other's generalization capabilities in sensory prediction tasks. This cooperation allows for not only generalization across environments with the *same* task structure (i.e. TEM), but also enables generalization across environments with *different* task structures. Finally, I offer suggestive parallels between our model's map (CSCG) and memory (TEM) units and CA1 and CA3 neuronal populations by demonstrating that our model replicates several key experimental findings that differentiate CA1 and CA3 population behavior during learning.

Chapter 2

Results

2.1 Model and training description

As background, we begin with brief descriptions of TEM and CSCG, respectively. TEM assumes the factorized generative model $P_\theta\left(x_{1:T}, p_{1:T}, g_{1:T}\right) = \prod_{t=1}^{T} P_\theta(x_t \mid p_t) P_\theta(p_t \mid g_t, M_{t-1}) \prod_{t=2}^{T} P_\theta(g_t \mid g_{t-1}, a_t)$ where $x_t, p_t,$ and g_t represent the sensory, place, and grid cells, respectively, while M_t and a_t represents the memories and action at time t (**2.1B**). Due to the intractability of the posterior $P(p_{1:T}, g_{1:T} \mid x_{1:t}, a_{1:T})$, TEM uses approximate inference via an inference neural network in a variational autoencoder (VAE) setup to learn the latent $p_{1:T}$ and $g_{1:T}$.

In contrast, CSCG is a Hidden Markov Model (HMM) representing the generative model $P_\theta\left(x_{1:T}, p_{1:T}\right) = \prod_{t=1}^{T} P_\theta(x_t \mid p_t) \prod_{t=2}^{T} P_\theta(p_t \mid p_{t-1}, a_t)$. Note that the transitions occur between p_{t-1} and p_t instead of between g_{t-1} and g_t as in TEM. CSCG makes an additional simplifying assumption $P(x_t = j \mid p_t = i) = 1$ only for all $i \in C(j)$ (and 0 otherwise) where $C(j)$ represents the subset of place units (i.e. *clones*) that emit to the sensory observation i (**2.1A**). Since CSCG is functionally identical to a regular HMM, it can be learned efficiently and exactly via the Expectation Maximization (EM) algorithm.

As briefly mentioned in the introduction, the two aforementioned models exhibit complementary abilities. TEM, on the one hand, abstracts the structure of the given task away from the environment-specific sensory observations, allowing it to generalize the learned task structure across environments with different sensory observations. Furthermore, using its place unit memory bank, it is able to remember *what* was *where* and path integrate (i.e. predict the sensory observation at the next time step if the model is returning to a previously visited location). However, in order to learn to perfectly path integrate, TEM must observe a large number of training environments. Moreover, once TEM has learned to handle one specific task structure, it cannot adequately navigate environments where the underlying task structure is even slightly different.

On the other hand, CSCG exhibits infinite flexibility in that it can learn any task structure with a fixed transition matrix upon first exposure without suffering from interference from previously

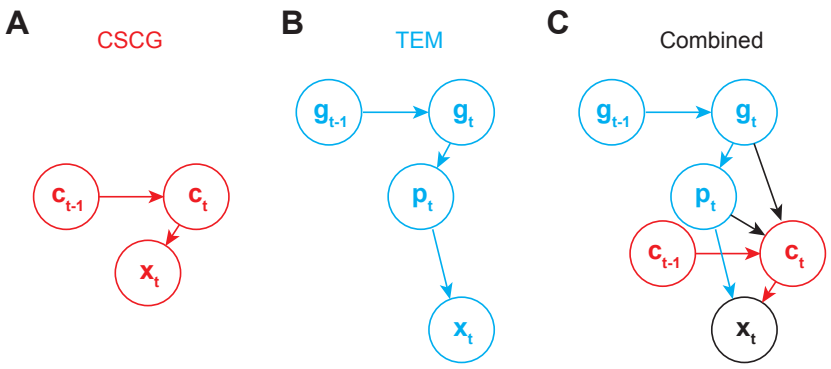


Figure 2.1: **Graphical models** Probablistic graphical models of CSCG (**A**), TEM (**B**), and the combined (**C**) models.

learned task structures. It is able to do so by completely forgetting all that it has learned from before and completely re-learning from scratch every time it enters a new environment. Moreover, it is able to create a perfectly usable map for navigation within a single exposure to the environment. Unfortunately, this flexibility comes at the cost of the lack of generalizability: it cannot transfer anything it has learned from past environments to new ones even if they may exhibit identical underlying task structures.

Therefore, the proposed model is a fusion of TEM and CSCG so that it may combine the benefits from each, respectively. We implement the following combined generative model:

$$
\begin{aligned}
P_\theta(x_{1:T}, p_{1:T}, g_{1:T}) = &\prod_{t=1}^{T} P_\theta(x_t \mid p_t) P(x_t \mid c_t) P_\theta(p_t \mid g_t, M_{t-1}) \\
&\prod_{t=2}^{T} P_\theta(g_t \mid g_{t-1}, a_t) P(c_t \mid c_{t-1}, a_t) P(c_t \mid g_t, a_t)
\end{aligned}
\tag{2.1}
$$

We separate the TEM memory units ($p_{1:T}$) from the CSCG map units ($c_{1:T}$) to enable efficient learning of the model's latent variables and a clear demonstration of how the two sub-systems synergistically assist each other throughout training (**2.1C**). On one hand, the map units help disambiguate aliased sensory observations to accelerate the learning of the TEM grid units as if the latter were trained to predict latent locations and not just sensory observations. On the other hand, as the grid units gradually improve, they can provide better (than random) initializations of the CSCG transition matrix so that maps of novel environments can be formed with fewer sensory observations and EM iterations. Furthermore, the EM training of the map units depends on the memory bank of sensory sequences stored in the memory units $p_{1:T}$.

Practically, implementing the combined model proved tricky due to the fact that CSCG learns quickly and exactly through the Expectation Maximization (EM) algorithm while TEM learns slowly and approximately through stochastic gradient descent. While gradient information can flow through $p_{1:T}$ and $g_{1:T}$ (as in TEM), $c_{1:T}$ and its EM iterations prevent joint, fully gradient-based training of the combined model. Consequently, we adopt the alternating scheme shown

below.

For each environment:

1. Generate $g_{1:T}$ with the current transition matrix: $g_t = \tanh(W_{g,g}^{(a_t)} g_{t-1})$
2. Predict CSCG clone activations: $\tilde{c}_t = \text{softmax}(W_{g,c} g_t)$
3. Initialize CSCG transition matrix, $\tilde{W}_{c,c}$, with $\tilde{c}_{1:T}$
4. (CSCG EM) Learn $c_{1:T}$ via EM algorithm and $\tilde{W}_{c,c}$
5. (TEM forward) For each time step, populate outer product memory $p_t = g_t h_t^T$ where $h_t = \text{concat}(x_t, c_t)$ and infer g_t
6. (TEM backward) Backpropagate sensory, $\mathcal{L}(x_t, \hat{x}_T)$, and clone, $\mathcal{L}(c_t, \hat{c}_T)$, prediction losses to learn $W_{g,g}$

Step 4 corresponds to CSCG's EM algorithm while steps 5 and 6 corresponds to TEM's forward and backward passes, respectively. Notably, this functional separation affords flexibility in what type of environments each sub-system is trained on. For instance, grid units that are trained on purely spatial tasks can be utilized to assist map-building in related tasks where the previously learned task structure can still serve as a useful starting point for the novel task.

While TEM uses a biologically plausible continuous attractor that is learned via Hebbian means as its memory module, I adopt the *attention* mechanism borrowed from the machine learning community's *transformer*.[29] This modification led to two major benefits. To begin with, a single self-attention layer can be parallelized and is significantly more computationally efficient than several steps of a continuous attractor. More importantly, it allows for our model to look up memories very far back in time (more than 1000 time steps) and path integrate . This was a core issue due to the fact that TEM (as it is implemented in [31]) is only trained to path integrate several dozen steps before the grid cells begin drifting (i.e. revisiting the same location in two points very far apart in time is not amenable to path integration). Using self-attention allowed for the combined model to path-integrate several thousand steps without drifting. I note that while the *attention* mechanism may seem less biologically plausible than a hebbian attractor, the former has been shown to be equivalent to a continuous attractor with an exponentially converging energy function.[23]

The combined model is trained on open grid world environments. For each environment, I sample a $5,000$ step random diffusion walk. At test time, I sample $5,000$ step random walks from either standard or modified grid worlds (see below). Due to GPU memory constraints, I use batch lengths of $1,000$ steps. Although this trains the model to path integrate up to $1,000$ steps, this requirement was enough to successfully path integrate all $5,000$ steps in each environment. A learning rate of 0.0005 was used along with the Adam optimizer with the following loss function.

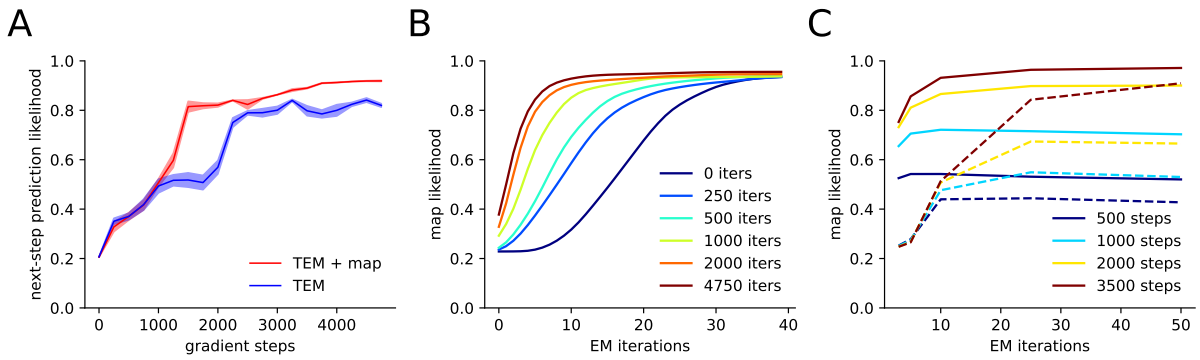


Figure 2.2: **Within-task generalization performance of model.** (**A**) Next-step sensory prediction performance in a novel test environment for both TEM and TEM augmented with latent state maps. (**B**) CSCG EM learning curves in test environments when initialized by TEM after different numbers of training gradient iterations. (**C**) CSCG EM learning curves for different numbers of sensory observations in a test environment either with fully trained TEM (dotted line) or TEM with latent maps (solid line).

$$\mathcal{L} = \mathcal{L}(g_{1:T}, \tilde{g}_{1:T}) + \mathcal{L}(x_{1:T}, \tilde{x}_{1:T}) + \mathcal{L}(c_{1:T}, \tilde{c}_{1:T}) \tag{2.2}$$

For CSCG's EM algorithm, we use 100 EM iterations.

2.2 Generalization across tasks with same task structure

While TEM on its own is capable of generalization across tasks with the same task structure, it only learns to do so well relatively slowly (after training on many different environments). TEM's slow learning rate is significantly exacerbated by the fact that it must distinguish on its own different latent states with the same sensory observations (i.e. sensory de-aliasing). Naturally, if we are to train TEM to predict locations (i.e. the latent states themselves) rather than sensory observations, I have eased the learning task for TEM since, rather than having to implicitly differentiate aliased latent states, we are now giving TEM explicit instructions to perform path integration.

Since I cannot inform TEM of its true location, the learned latent states from CSCG will take on the role of the de-aliased TEM inputs. When TEM is trained to not just predict the next sensory observation, but also the next CSCG latent state, it achieves $\geq 90\%$ predictive capability, indicative of accurate path integration, by 4000 gradient steps (800 training environments) (**2.2A, red**). Meanwhile, standard TEM (i.e. only predicting sensory observations) takes far longer to achieve the same next-step sensory prediction performance, only reaching $\sim 80\%$ at 5000 gradient steps (1000 training environments) (**2.2B, blue**).

On the flipside, TEM can be leveraged to accelerate map creation once it knows how to path integrate. When initializing the CSCG transition matrix with the clone activity predictions

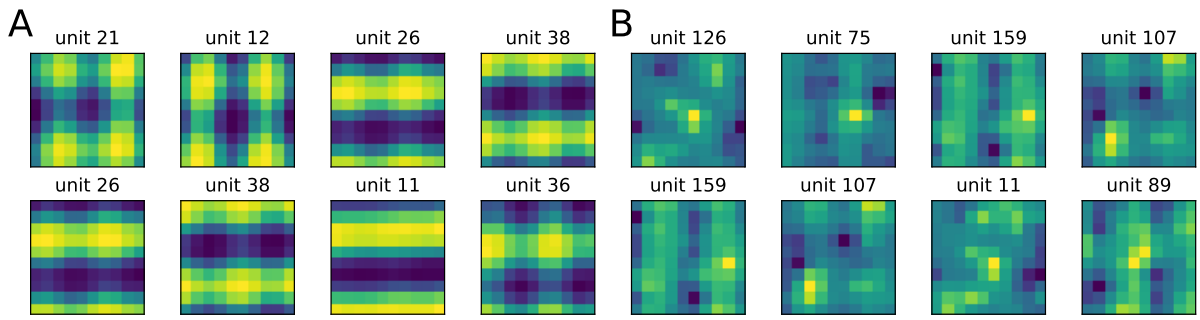


Figure 2.3: **Spatial representations of model.** Representations of the fully trained model produces both grid units (**A**) and place units (**B**) as expected of a model trained in open grid worlds.

outputted by TEM, we observe that map building is accelerated by TEM's inputs. Namely, both the initial quality of the map (as measured by the likelihood of the observed test sensory sequence) and the number of EM iterations required to fully learn the map both improve alongside TEM's training and the magnitude of the improvements are correlated with the amount of learning TEM underwent. (**2.2B**). Additionally, TEM's clone predictions also allow accurate CSCG maps to be formed with fewer sensory observations in a novel test environment. Specifically, we see that, armed with TEM-initialized transition matrices, CSCG can learn equally good maps with 1000 and 2000 steps versus 2000 and 3500 steps in an unfamiliar environment, respectively (**2.2C**).

I confirmed that the combined model is able to learn the standard representations known to appear in open spaces, namely grid and place cells (**2.3**). The above is a clear demonstration of reciprocal acceleration between map building and path integration across different environments with the same underlying task structure (open grid worlds in this case).

2.3 Generalization across different task structures

The above demonstration, however, paints an incomplete picture of generalization. After all, we are not only interested in transferring knowledge about the same underlying task across its different instantiations, but also in transferring knowledge across different, but partially related, tasks. Therefore, I train my model on open grid worlds, but test it on semi-spatial worlds, such as the T-maze, that contain both spatial and non-spatial latent states for accurate navigation. It is predicted, then, that the TEM component of the model should offer partially correct sensory predictions, since its learned task structure only partially matches that of the test tasks. Nevertheless, due to the spatial nature of the T-maze, the combined model should still provide better-than-random initializations to CSCG, which will then create a fully accurate map of the novel task at hand.

The combined model is able to predict sensory observations relatively well in the T-maze task despite the task's non-spatial component (**2.4A**). Nevertheless, we see the same patterns as

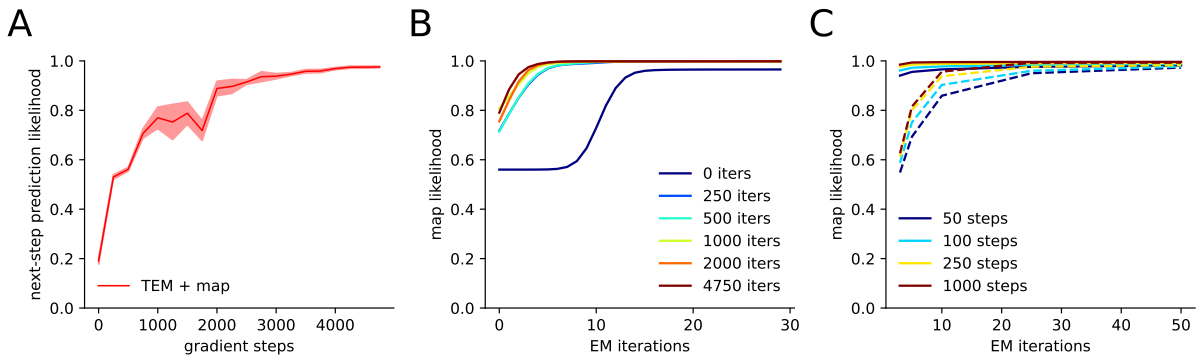


Figure 2.4: **Across-task generalization performance of model.** (**A**) Next-step sensory prediction performance in a novel test environment for TEM augmented with latent state maps. (**B**) CSCG EM learning curves in test environments when initialized by TEM after different numbers of training gradient iterations. (**C**) CSCG EM learning curves for different numbers of sensory observations in a test environment either with fully trained TEM (dotted line) or TEM with latent maps (solid line).

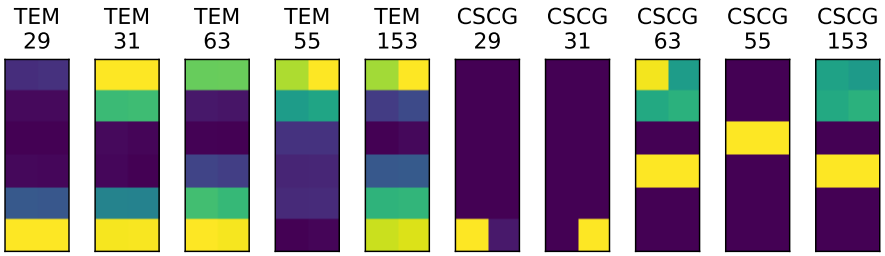


Figure 2.5: **TEM and CSCG place units for T-maze** Five representative place units from TEM and CSCG, respectively.

before when it comes to facilitating map building: training the path integration part of our model leads to improvements in initial map quality (**2.4B**), reductions in the number of EM iterations required to complete map creation (**2.4B,C**), and reductions in the number of test environment observations need to build accurate maps (**2.4C**). Interestingly, we observe both purely spatial place units and trajectory-dependent "splitter" units in both TEM and CSCG (**2.5**). This is likely due to the fact that I supply far more CSCG clones than latent locations, resulting in some of the clones picking up spurious, non-zero activations for both left and right trials at the same location on the trunk of the T-maze.

2.4 Potential neural correlates of map and memory in CA1 and CA3, respectively

The differences between the map and memory modules in their training and roles in generalization suggest that they may map to disparate neuronal populations of the hippocampal formation.

Namely, cells participating in map building would be much more flexible and rapid in incorporating novel latent state information as the subject is exploring their world while those involved in memory storage should exhibit greater stability as they must inherit much of their activity patterns from the slowly learning entorhinal cortex.

Indeed, such a distinction can be found in a several studies that delineate the representational differences in CA1 and CA3 of the hippocampus. Specifically, neuroscientists have discovered that CA1 cells specifically represent latent states (e.g. overlapping trajectories in T-maze or reward locations) more than CA3 cells [11, 7, 18, 16]. Furthermore, CA1 fields emerge rapidly and exhibit low stability while CA3 fields emerge more slowly but then maintain higher stability than their CA1 counterparts [6, 24, 2, 1, 18, 17, 19, 15]. Finally, CA1 cells alter their firing rates in novel environments while CA3 cells do not, suggesting that this novelty signaling indicates a form of rapid learning that is later passed on to the entorhinal cortex. [15, 13, 33, 14]

We show that assigning our model's map and memory units to be CA1 and CA3 cells reconstructs several of the aforementioned experimental findings, supporting the hypothesis that CA1 engages in map building while CA3 is more involved in memory storage. I would like to note that the following is not irrefutable evidence that this hypothesis is correct. Nevertheless, the parallels between the two types of studies serves as the basis for future lines of inquiry. Firstly, CA1 cells exhibit overall lower spatial correlations between two consecutive presentations of a novel environment compared to CA3 cells (**2.6A**). [6] Our model replicates this result with our map and memory units where each "presentation" of a novel environment is treated as one forward-backward sweep of the map building EM algorithm (**2.6D**). [19] Likewise, our model replicates CA1's relative lack of field stability compared to CA3 when repeatedly presented with a novel environment (i.e. with each iteration of the EM algorithm) (**2.6B,E**). Finally, the combined model's map units exhibit the novelty-related firing rate changes observed in CA1 cells during novel environment exploration (**2.6C**), in which cells with intermediate firing rates either, for the most part, suppress their firing rates or, occasionally, increase their firing rates (**2.6F**). [12]

While not all experimental differences between CA1 and CA3 can be neatly explained by our model, the fact that treating CA1 as a map builder and CA3 as memory storage reproduces several key findings suggest that our model may help provide an organizing principle for the division of labor between CA1 and CA3.

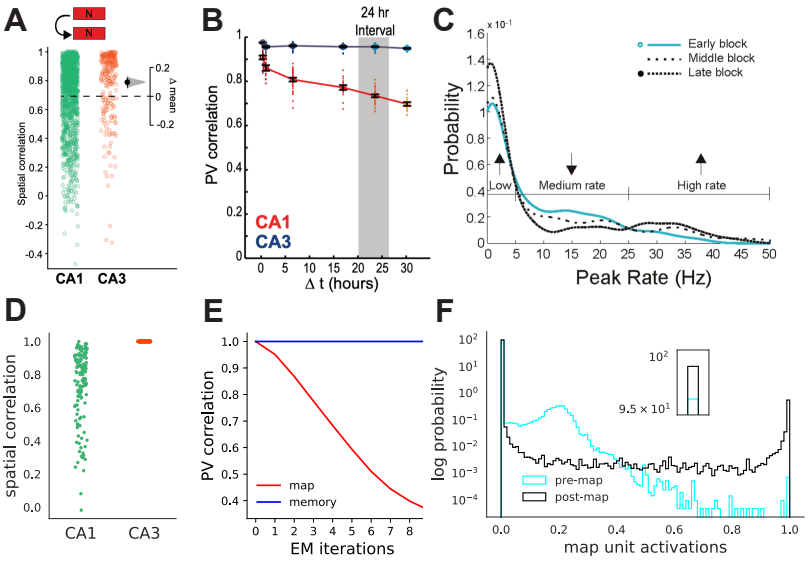


Figure 2.6: **Replication of CA1 vs. CA3 experimental results** (**A**) Spatial correlations of CA1 and CA3 cells' firing rates between two consecutive presentations of an unfamiliar environment (adopted from [6]). (**B**) Population vector correlations of CA1 and CA3 cells' firing rates throughout repeated exposures of a novel environment over time (adopted from [19]). (**C**) Novelty signaling of CA1 cells where the majority of CA1 cells suppress their activity during familiarization of an unfamiliar environment while a small number of CA1 cells increase their firing rates (adopted from [12]). (**D,E,F**) Recreations of **A,B,C** with the combined model's map and memory units.

Chapter 3

Discussion

Overall, my work represents another incremental step in a long line of computational modeling studies that attempt to provide a normative description for the large variety of hippocampal phenomena and their role in facilitating learning and generalization. Specifically, I fused two distinct, yet complementary, computational models of the hippocampal formation, TEM and CSCG, to produce and demonstrate that rapid, flexible map-building and slower, generalizable path integration accelerate generalization in cooperation.

I show that when this novel model is trained to predict upcoming sensory observations in open grid environments with sensory aliasing, it learns to do so with fewer gradient updates. In turn, the TEM component of the model aids in producing accurate CSCG maps utilizing fewer test environment sensory observations and EM iterations. The same trend is observed for when the model is trained on open grid worlds, but evaluated on the T-maze, which is a semi-spatial task. This demonstrates the model's ability to marry the generalization capabilities of TEM with the *ad hoc*, flexible learning of CSCG in generalizing across different task structures.

Finally, I offer the suggestion that CA1 and CA3 of the hippocampus could correspond to the map and memory components of my model by replicating experimental evidence highlighting the distinct characteristics of the two hippocampal sub-regions. My model's CSCG place units exhibit lower field stability across time than the TEM place units just as CA1 place cells do compared to CA3 place cells. Additionally, CSCG place units show similar novelty-dependent firing rate patterns as experimentally recorded CA1 place cells.

3.1 Unaddressed technical challenges

There remains several key technical issues that my model does not completely resolve. The most significant of these has to do with the map-building component. To learn the environment-specific transition matrices, CSCG, just as any HMM would, utilizes the widely adopted EM algorithm. As mentioned at the top of the Results chapter, the EM algorithm prevents my model to be learned jointly via gradient descent, which forced me to resort to training the TEM and CSCG

components separately in an alternating fashion. Furthermore, and more critically, since CSCG latent units directly encode a discrete probability distribution, it is difficult to interpret them as mimicking arbitrary continuous firing rate vectors. In other words, the CSCG latent units strictly reside on a simplex since the sum of their activations must be equal to 1. A novel map-building module that can utilize all real-valued vectors and can be trained jointly with the path integration component of the full model would be ideal.

A second future line of inquiry could incorporate policy- or decision-making into the model. Currently, I feed a stream of either random (in the open grid worlds) or strictly ideal (in the T-maze) action sequences, using which the model must then predict the upcoming sensory observations. A more realistic and complete model, however, would not be a merely passive agent, but also would learn to predict the future action sequences by leveraging its knowledge of upcoming sensory observations. This could done either by combining my model with a reinforcement learning agent or to engage in "planning by inference,"[4] in which the probabilistic graphical model incorporating actions and sensory observations could be used to infer the best course of action between two sensory objects.

3.2 Unverified experimental predictions

There also remains several unanswered questions when it comes to the neural predictions that my model makes. Most notably, while there are several studies suggesting representational differences between CA1 and CA3, they do not provide a clear picture about the different roles of the two hippocampal sub-regions.[2, 1, 11, 6] For instance, the studies, some of which are mentioned in the Results chapter, do not address the fact that representational differences may only be specific to either the left or right sides of the hippocampus.[26] Furthermore, the studies rely on different experimental tasks, ranging from T-maze [11] to circular arenas [15] and multi-armed environments.[2, 1] The disparate environmental conditions of the various studies make it difficult to draw clear conclusions about representational and functional differences between CA1 and CA3.

Finally, the model's predictions do not clearly match known connectivity differences that exist betwen CA1 and CA3. While CA1 exhibits little recurrent connectivity, CA3's strongly recurrent connections have pushed neuroscientists to designate its role as a memory attractor. [32] This well-established dichotomy does not neatly coincide with the prediction that CA1 corresponds to a map-building module since it could be assumed that the state-to-state transitions of the CA1 map may most easily be encoded via recurrent connections. Ultimately, to truly verify or refute the hypothesis that CA1 and CA3 occupy distinct computational roles, connectivity differences, hippocampal lateralization, and task differences must all be taken into account within a controlled and comprehensive experimental study.

References

[1] A. S. Bahar and M. L. Shapiro. Remembering to Learn: Independent Place and Journey Coding Mechanisms Contribute to Memory Transfer. *Journal of Neuroscience*, 32(6): 2191–2203, February 2012. ISSN 0270-6474, 1529-2401. doi: 10.1523/JNEUROSCI.3998-11.2012. URL https://www.jneurosci.org/lookup/doi/10.1523/JNEUROSCI.3998-11.2012.

[2] A. S. Bahar, P. R. Shirvalkar, and M. L. Shapiro. Memory-Guided Learning: CA1 and CA3 Neuronal Ensembles Differentially Encode the Commonalities and Differences between Situations. *Journal of Neuroscience*, 31(34):12270–12281, August 2011. ISSN 0270-6474, 1529-2401. doi: 10.1523/JNEUROSCI.1671-11.2011. URL https://www.jneurosci.org/lookup/doi/10.1523/JNEUROSCI.1671-11.2011.

[3] Timothy E.J. Behrens, Timothy H. Muller, James C.R. Whittington, Shirley Mark, Alon B. Baram, Kimberly L. Stachenfeld, and Zeb Kurth-Nelson. What Is a Cognitive Map? Organizing Knowledge for Flexible Behavior. *Neuron*, 100(2):490–509, October 2018. ISSN 08966273. doi: 10.1016/j.neuron.2018.10.002. URL https://linkinghub.elsevier.com/retrieve/pii/S0896627318308560.

[4] Matthew Botvinick and Marc Toussaint. Planning as inference. *Trends in Cognitive Sciences*, 16(10):485–488, October 2012. ISSN 13646613. doi: 10.1016/j.tics.2012.08.006. URL https://linkinghub.elsevier.com/retrieve/pii/S1364661312001957.

[5] A. O. Constantinescu, J. X. OReilly, and T. E. J. Behrens. Organizing conceptual knowledge in humans with a gridlike code. *Science*, 352(6292):1464–1468, June 2016. ISSN 0036-8075, 1095-9203. doi: 10.1126/science.aaf0941. URL https://www.sciencemag.org/lookup/doi/10.1126/science.aaf0941.

[6] Can Dong, Antoine D. Madar, and Mark E. J. Sheffield. Distinct place cell dynamics in CA1 and CA3 encode experience in new environments. *Nature Communications*, 12 (1):2977, December 2021. ISSN 2041-1723. doi: 10.1038/s41467-021-23260-3. URL http://www.nature.com/articles/s41467-021-23260-3.

[7] David Dupret, Joseph O'Neill, Barty Pleydell-Bouverie, and Jozsef Csicsvari. The reorganization and reactivation of hippocampal maps predict spatial memory performance. *Nature Neuroscience*, 13(8):995–1002, August 2010. ISSN 1097-6256, 1546-1726. doi: 10.1038/nn.2599. URL http://www.nature.com/articles/nn.2599.

[8] Howard Eichenbaum. The role of the hippocampus in navigation is memory. *Journal of Neurophysiology*, 117(4):1785–1796, April 2017. ISSN 0022-3077, 1522-1598. doi: 10.1152/jn.00005.2017. URL https://www.physiology.org/doi/10.1152/jn.00005.2017.

[9] Dileep George, Rajeev V. Rikhye, Nishad Gothoskar, J. Swaroop Guntupalli, Antoine Dedieu, and Miguel Lázaro-Gredilla. Clone-structured graph representations enable flexible learning and vicarious evaluation of cognitive maps. *Nature Communications*, 12(1): 2392, December 2021. ISSN 2041-1723. doi: 10.1038/s41467-021-22559-5. URL http://www.nature.com/articles/s41467-021-22559-5.

[10] Torkel Hafting, Marianne Fyhn, Sturla Molden, May-Britt Moser, and Edvard I. Moser. Microstructure of a spatial map in the entorhinal cortex. *Nature*, 436(7052):801–806, August 2005. ISSN 0028-0836, 1476-4687. doi: 10.1038/nature03721. URL http://www.nature.com/articles/nature03721.

[11] Hiroshi T. Ito, Sheng-Jia Zhang, Menno P. Witter, Edvard I. Moser, and May-Britt Moser. A prefrontal–thalamo–hippocampal circuit for goal-directed spatial navigation. *Nature*, 522(7554):50–55, June 2015. ISSN 0028-0836, 1476-4687. doi: 10.1038/nature14396. URL http://www.nature.com/articles/nature14396.

[12] M. P. Karlsson and L. M. Frank. Network Dynamics Underlying the Formation of Sparse, Informative Representations in the Hippocampus. *Journal of Neuroscience*, 28(52): 14271–14281, December 2008. ISSN 0270-6474, 1529-2401. doi: 10.1523/JNEUROSCI.4261-08.2008. URL https://www.jneurosci.org/lookup/doi/10.1523/JNEUROSCI.4261-08.2008.

[13] James J. Knierim, Inah Lee, and Eric L. Hargreaves. Hippocampal place cells: Parallel input streams, subregional processing, and implications for episodic memory. *Hippocampus*, 16 (9):755–764, September 2006. ISSN 1050-9631, 1098-1063. doi: 10.1002/hipo.20203. URL https://onlinelibrary.wiley.com/doi/10.1002/hipo.20203.

[14] I. Lee and J. J. Knierim. The relationship between the field-shifting phenomenon and representational coherence of place cells in CA1 and CA3 in a cue-altered environment. *Learning & Memory*, 14(11):807–815, November 2007. ISSN 1072-0502. doi: 10.1101/lm.706207. URL http://www.learnmem.org/cgi/doi/10.1101/lm.706207.

[15] Inah Lee, D. Yoganarasimha, Geeta Rao, and James J. Knierim. Comparison of population coherence of place cells in hippocampal subfields CA1 and CA3. *Nature*, 430(6998): 456–459, July 2004. ISSN 0028-0836, 1476-4687. doi: 10.1038/nature02739. URL http://www.nature.com/articles/nature02739.

[16] Sung-Hyun Lee, Namjung Huh, Jong Won Lee, Jeong-Wook Ghim, Inah Lee, and Min W. Jung. Neural Signals Related to Outcome Evaluation Are Stronger in CA1 than CA3. *Frontiers in Neural Circuits*, 11:40, June 2017. ISSN 1662-5110. doi: 10.3389/fncir.2017.00040. URL http://journal.frontiersin.org/article/10.3389/fncir.2017.00040/full.

[17] S. Leutgeb and J. K. Leutgeb. Pattern separation, pattern completion, and new neuronal codes within a continuous CA3 map. *Learning & Memory*, 14(11):745–757, November 2007. ISSN 1072-0502. doi: 10.1101/lm.703907. URL http://www.learnmem.org/cgi/doi/10.1101/lm.703907.

[18] Stefan Leutgeb, Jill K. Leutgeb, Alessandro Treves, May-Britt Moser, and Edvard I. Moser. Distinct Ensemble Codes in Hippocampal Areas CA3 and CA1. *Science*, 305(5688): 1295–1298, August 2004. ISSN 0036-8075, 1095-9203. doi: 10.1126/science.1100265. URL https://www.science.org/doi/10.1126/science.1100265.

[19] Emily A. Mankin, Fraser T. Sparks, Begum Slayyeh, Robert J. Sutherland, Stefan Leutgeb, and Jill K. Leutgeb. Neuronal code for extended time in the hippocampus. *Proceedings of the National Academy of Sciences*, 109(47):19462–19467, November 2012. ISSN 0027-8424, 1091-6490. doi: 10.1073/pnas.1214107109. URL https://pnas.org/doi/full/10.1073/pnas.1214107109.

[20] James L. McClelland, Bruce L. McNaughton, and Randall C. O'Reilly. Why there are complementary learning systems in the hippocampus and neocortex: Insights from the successes and failures of connectionist models of learning and memory. *Psychological Review*, 102(3):419–457, July 1995. ISSN 1939-1471, 0033-295X. doi: 10.1037/0033-295X.102.3.419. URL http://doi.apa.org/getdoi.cfm?doi=10.1037/0033-295X.102.3.419.

[21] Yael Niv. Learning task-state representations. *Nature Neuroscience*, 22(10):1544–1553, October 2019. ISSN 1097-6256, 1546-1726. doi: 10.1038/s41593-019-0470-8. URL http://www.nature.com/articles/s41593-019-0470-8.

[22] John O'Keefe and Lynn Nadel. *The hippocampus as a cognitive map*. Clarendon Press ; Oxford University Press, Oxford : New York, 1978. ISBN 978-0-19-857206-0.

[23] Hubert Ramsauer, Bernhard Schäfl, Johannes Lehner, Philipp Seidl, Michael Widrich, Thomas Adler, Lukas Gruber, Markus Holzleitner, Milena Pavlović, Geir Kjetil Sandve, et al. Hopfield networks is all you need. *arXiv preprint arXiv:2008.02217*, 2020.

[24] Eric D. Roth, Xintian Yu, Geeta Rao, and James J. Knierim. Functional Differences in the Backward Shifts of CA1 and CA3 Place Fields in Novel and Familiar Environments. *PLoS ONE*, 7(4):e36035, April 2012. ISSN 1932-6203. doi: 10.1371/journal.pone.0036035. URL https://dx.plos.org/10.1371/journal.pone.0036035.

[25] W. B. Scoville and B. Milner. LOSS OF RECENT MEMORY AFTER BILATERAL HIPPOCAMPAL LESIONS. *Journal of Neurology, Neurosurgery & Psychiatry*, 20(1): 11–21, February 1957. ISSN 0022-3050. doi: 10.1136/jnnp.20.1.11. URL https://jnnp.bmj.com/lookup/doi/10.1136/jnnp.20.1.11.

[26] Da Song, Deheng Wang, Qinghu Yang, Tianyi Yan, Zhe Wang, Yan Yan, Juan Zhao, Zhen Xie, Yuchen Liu, Zunji Ke, Talal Jamil Qazi, Yanhui Li, Yili Wu, Qing Shi, Yiran Lang, Heao Zhang, Tao Huang, Chunjian Wang, Zhenzhen Quan, and Hong Qing. The lateralization of left hippocampal CA3 during the retrieval of spatial working memory. *Nature Communications*, 11(1):2901, December 2020. ISSN 2041-1723. doi: 10.1038/s41467-020-16698-4. URL http://www.nature.com/articles/s41467-020-16698-4.

[27] Kimberly L Stachenfeld, Matthew M Botvinick, and Samuel J Gershman. The hippocampus as a predictive map. *Nature Neuroscience*, 20(11):1643–1653, November 2017. ISSN 1097-6256, 1546-1726. doi: 10.1038/nn.4650. URL http://www.nature.com/articles/nn.4650.

[28] Edward C. Tolman. Cognitive maps in rats and men. *Psychological Review*, 55(4): 189–208, 1948. ISSN 1939-1471, 0033-295X. doi: 10.1037/h0061626. URL http://doi.apa.org/getdoi.cfm?doi=10.1037/h0061626.

[29] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. Attention is all you need. In *Advances in Neural Information Processing Systems*, pages 5998–6008, 2017.

[30] James C. R. Whittington, David McCaffary, Jacob J. W. Bakermans, and Timothy E. J. Behrens. How to build a cognitive map: insights from models of the hippocampal formation. *arXiv:2202.01682 [cs, q-bio]*, February 2022. URL http://arxiv.org/abs/2202.01682. arXiv: 2202.01682.

[31] James C.R. Whittington, Timothy H. Muller, Shirley Mark, Guifen Chen, Caswell Barry, Neil Burgess, and Timothy E.J. Behrens. The Tolman-Eichenbaum Machine: Unifying

Space and Relational Memory through Generalization in the Hippocampal Formation. *Cell*, 183(5):1249–1263.e23, November 2020. ISSN 00928674. doi: 10.1016/j.cell.2020.10.024. URL https://linkinghub.elsevier.com/retrieve/pii/S009286742031388X.

[32] M. P. Witter. Intrinsic and extrinsic wiring of CA3: Indications for connectional heterogeneity. *Learning & Memory*, 14(11):705–713, November 2007. ISSN 1072-0502. doi: 10.1101/lm.725207. URL http://www.learnmem.org/cgi/doi/10.1101/lm.725207.

[33] D. Yoganarasimha. Head Direction Cell Representations Maintain Internal Coherence during Conflicting Proximal and Distal Cue Rotations: Comparison with Hippocampal Place Cells. *Journal of Neuroscience*, 26(2):622–631, January 2006. ISSN 0270-6474, 1529-2401. doi: 10.1523/JNEUROSCI.3885-05.2006. URL https://www.jneurosci.org/lookup/doi/10.1523/JNEUROSCI.3885-05.2006.