A Computational Model for Latent Learning based on Hippocampal Replay

Pablo Scleidorovich, Martin Llofriu, Jean-Marc Fellous, and Alfredo Weitzenfeld, *Senior Member, IEEE*

***Abstract*—We show how hippocampal replay could explain latent learning, a phenomenon observed in animals where unrewarded pre-exposure to an environment, i.e. habituation, improves task learning rates once rewarded trials begin. We first describe a computational model for spatial navigation inspired by rat studies. The model exploits *offline* replay of trajectories previously learned by applying reinforcement learning. Then, to assess our hypothesis, the model is evaluated in a "multiple T-maze" environment where rats need to learn a path from the start of the maze to the goal. Simulation results support our hypothesis that pre-exposed or habituated rats learn the task significantly faster than non-pre-exposed rats. Results also show that this effect increases with the number of pre-exposed trials.**

I. Introduction

More than a century ago, Small [1] recognized that unrewarded pre-exposure of rodents to the training environment significantly facilitated learning. His experimental protocols included letting rats explore a maze overnight before they began to learn a task.

Nearly three decades later, Blodgett [2] studied this phenomenon quantitatively, showing how the unrewarded pre-exposure of a rat to the environment can increase the learning rate once rewarded trials begin, a phenomenon that came to be recognized as a form of *latent learning*. The task consisted in travelling a "multiple T-maze", in order to get food. Rats were divided into three groups. The Control group started to learn the task without any pre-exposure to the maze. The other two groups had three and seven days of pre-exposure respectively, where no food was given (non-rewarded trials). Blodgett concluded that pre-exposure improved the learning rate of rats in the rewarded trials by observing a fast decrease in error counts.

Since no reward was provided during habituation, Blodgett's results posed an interesting challenge to the stimulus-response theory of learning. Tolman [3] proposed that the difference in learning rates was due to rats with pre-exposure building a 'cognitive map' of the environment that allowed them to learn the task faster. A number of experiments and mazes were designed to test Tolman's different hypotheses (see Olton [4] for a review of these original mazes and experiments).

In our present work, we hypothesize that the latent learning observed in Blodgett's experiment could be explained in terms of enhancing hippocampal replay [5], [6]. Hippocampal replay is a phenomenon in which place cells involved in a task reactivate during periods of inactivity in the same or reversed order in which they were observed during the task [6], [7]. Replay during awake immobility at key decision points has been shown to be in part correlated with future paths taken by rats, and thus with decision making [8], [9], [10]. On the other hand, replay during sleep has been linked to memory consolidation [11], [12]. Here we hypothesize that by pre-exposing a rat to an unrewarded environment (habituation), intrahippocampal connections involved in replay are formed or strengthened. In turn, these connections later facilitate task learning by improving the quality of replay sequences during rewarded trials. Consequently, replay could be a mechanism capable of explaining the latent learning observed by Blodgett.

To assess our hypothesis, we extend the computational model of replay presented by Johnson and Redish [13] and we use it to compare the learning rates of rats with and without habituation. During both rewarded and unrewarded trials, the model updates the connection strengths between hippocampal place cells in a Hebbian fashion. This creates a topological map of the environment stored in a connectivity matrix, which is then used to generate replay events. These events, in turn, are used to perform batch reinforcement learning [14].

The simulations show that faster learning rates can be obtained through pre-exposure and replay events. As a result, this work provides a plausible mechanism for explaining the latent learning phenomenon observed by Blodgett [2].

In the remainder of this document, section II presents the task, section III the model, section IV the simulation experiments and results and section V the conclusions and discussion.

II. Task

The task consists on having two groups of rats learn a path from the starting position to the goal (location of the food) in a multiple T-maze configuration consisting of four T's, as shown in Fig. 1. At each trial, rats start from the bottom T, and food is located at the top T's left arm. The "No Habituation" group of rats receives food from the very first trial, while the "Habituation" group receives food only after trial *N*. Following the original experiment by Blodgett [2], in both rewarded and unrewarded trials, rats are removed from the maze when they reach the food location. A rat has reached the food location when it's distance to the feeder is smaller or equal than 8cm (distance covered by the robot after each action or step). Alternatively, a trial ends if the rat performs a total of 2,000 steps without reaching the goal. This timeout allows simulated rats to traverse a distance of 80m corresponding to

*NSF IIS Robust Intelligence research collaboration grant #1703340

Pablo Scleidorovich is at the Computer Science and Eng Dept, University of South Florida, Tampa, FL (pablos@mail.usf.edu).

Martin Llofriu is at Fing, UdelaR, and was at the Computer Science and Eng Dept, University of South Florida, Tampa, FL (mllofriualon@mail.usf.edu).

Jean-Marc Fellous is at the Psychology Dept, University of Arizona Tucson, AZ (fellous@email.arizona.edu).

Alfredo Weitzenfeld is at the Computer Science and Eng Dept, University of South Florida, Tampa, FL (aweitzenfeld@usf.edu).

978-1-7281-6926-2/20/$31.00 ©2020 IEEE

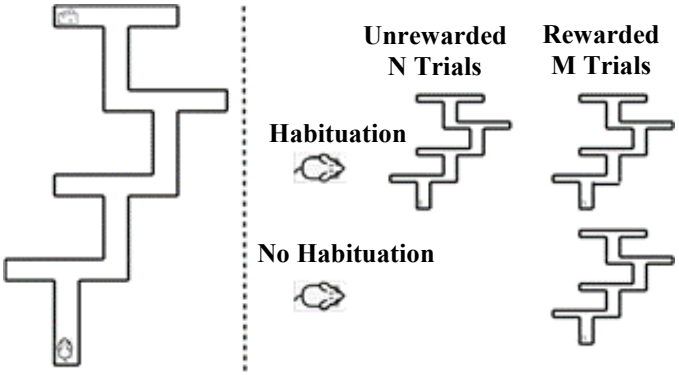

Fig. 1 Multiple T-Maze environment and task. Rats start from the bottom of the first "T" and need to reach the food at the left end of the last "T". Two groups are compared: "Habituation" and "No Habituation". Rats in group "Habituation" perform N unrewarded trials (no food is present), followed by M rewarded trials. Rats in group "No Habituation" only perform the M rewarded trials.

approximately 4 times the entire maze. Note that the horizontal portion of the maze measures 3.6m, the vertical 2.6m and the corridors have a width of 0.2m.

III. Model

The model consists of three components: 1) an actor critic component [14], associating preferences to all (state, action) pairs; 2) an action selection component that selects the next action to be performed after every step; and, 3) a replay component for offline learning.

A. Actor Critic

In our work, we model the spatial navigation process of a rat as an actor critic algorithm [14]. The algorithm executes over a continuous 2D state space, and a discrete action space with 8 allocentric actions (move north, northeast, east, etc.). The state space is represented using Gaussian radial basis functions [14] which model our place cells, and are used with linear function approximation to compute the state and action value functions [14] as illustrated in Fig. 2.

Place cells are modelled using Gaussians whose values are set to 0 outside of a given radius. Each Gaussian represents the average firing rate of a place cell whose preferred firing location corresponds to the Gaussian's center. In total, 2000 place cells are used per rat. Their centers are chosen randomly from a uniform distribution over the area covered by the multiple T-maze.

In order to calculate the state and action value functions for an arbitrary state (position), first the activation of each place cell is calculated according to eq 1.

$$A_i(\vec{x}) = \begin{cases} 0 & if\ \left|\left|\vec{x} - \vec{x}_i\right|\right| > 2.3\sigma \\ N\left(\left|\left|\vec{x} - \vec{x}_i\right|\right|, 0, \sigma\right) & \text{otherwise} \end{cases} \quad (1)$$

where $A_i(x)$ is the activation of place cell i for state $\vec{x}$, $\vec{x}_i$ is the center of place field i, and N is a gaussian with mean 0 and standard deviation σ.

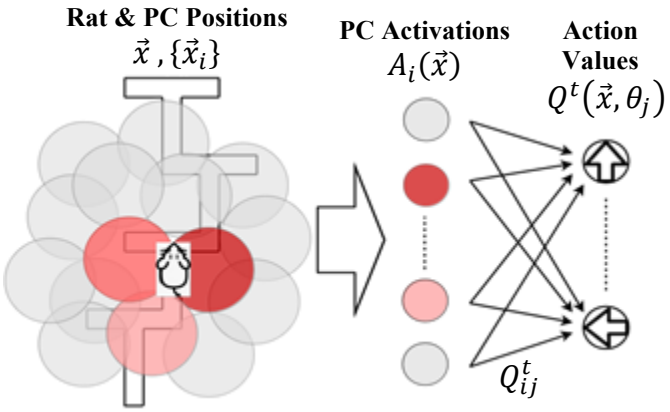

Fig. 2 Place cell activation in the rat hippocampus and action values for the current state. The variables shown in the image correspond to the variables used in equations 1-6.

Then, the state values can be calculated using eq 2.

$$V^t(\vec{x}) = \frac{\sum_i A_i(\vec{x}) V_i^t}{\sum_i A_i(\vec{x})} \quad (2)$$

where $V^t(x)$ represents the state value function at time t for state x, and V_i^t represents the associated state value for place cell i at time t. In a similar way, the action value function is given by eq 3.

$$Q^t(\vec{x}, \theta_j) = \frac{\sum_i A_i(\vec{x}) Q_{ij}^t}{\sum_i A_i(\vec{x})} \quad (3)$$

where $Q^t(\vec{x}, \theta_j)$ represents the action value at time t for state $\vec{x}$, with action θ_j, while Q_{ij}^t represents the action value associated with cell i, at time t, for action θ_j.

After calculating the action values for the current state, the values are sent to an action selection module that chooses the next action to be performed, denoted by a_t. Note that we use the symbols θ_j and a_t to differentiate the possible actions from the actual actions performed.

After performing action a_t, and observing reward r_t, the error in the estimated value $\delta(t)$ is calculated and used to update the values V_i^t and Q_{ij}^t, according to eq 4, 5, and 6 [14].

$$\delta(t) = r_t + \gamma V^t(\vec{x}_{t+1}) - V^t(\vec{x}_t) \quad (4)$$

$$V_i^{t+1} = V_i^t + \lambda\, \delta(t)\, A_i\,(\vec{x}_t) \quad (5)$$

$$Q_{ij}^{t+1} = \begin{cases} Q_{ij}^t + \lambda\, \delta(t)\, A_i\,(\vec{x}_t) & \text{if } a_t = \theta_j \\ Q_{ij}^t & \text{otherwise} \end{cases} \quad (6)$$

where, γ is the discounting factor, and λ is the learning rate.

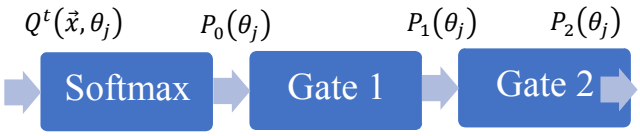

Fig. 3 Action selection process consisting of a softmax process followed by two gate transformations ("Gate 1" and "Gate 2").

B. Action Selection

Actions are chosen randomly from a probability distribution obtained by first applying the *softmax* function to the action values, and then applying 2 transformations or "gates"that convert one distribution into another. The action selection process is shown in Fig. 3.

The *softmax* function is shown in eq 7.

$$P_0(\theta_j) = \frac{e^{Q(\vec{x},\theta_j)}}{\sum_j e^{Q(\vec{x},\theta_j)}} \tag{7}$$

Each gate redistributes the probabilities by assigning weights to the actions and normalizing the results as described by eq 8.

$$P_k(\theta_j) = \frac{w_j^k P_{k-1}(\theta_j)}{\sum_j w_j^k P_{k-1}(\theta_j)} \tag{8}$$

where $P_k(\theta_j)$ is the probability for action θ_j after applying gate number k (or the output of the *softmax* if $k = 0$), and w_j^k is the weight given to action θ_j by gate k.

The first gate uses the concept of affordances [15], weighting each probability by a function proportional to the distance to the wall in its corresponding direction (see Fig. 4). This assigns a probability of 0 to actions that cannot be performed, and gives preference to directions pointing along the corridors of the maze.

The weights of the first gate are given by eq 9.

$$w_j^1 = \begin{cases} \min\{d_j, D\} & d_j > d \\ 0 & otherwise \end{cases} \tag{9}$$

where d_j represents the distance to the wall along direction j (direction associated to action θ_j), d represents the minimum distance to consider an action as possible, and D is a constant which limits the maximum weight given to an action.

The second gate biases the movement of the rat using the concept of "coherent actions". We define two actions to be coherent if their respective angles differ at most by 90 degrees. The gate biases the actions so that if the last two actions were coherent, then it gives preference to actions that are coherent with both of them. If the previous two actions were not coherent then the gate gives preference to actions coherent with the last action. The idea of this gate is to model a rat whose exploratory behavior is not a completely random walk, so that it looks closer to navigation paths observed in rats. Without this gate, one would observe trajectories such as "one step left, one step right, ..." and so on, which are not realistic.

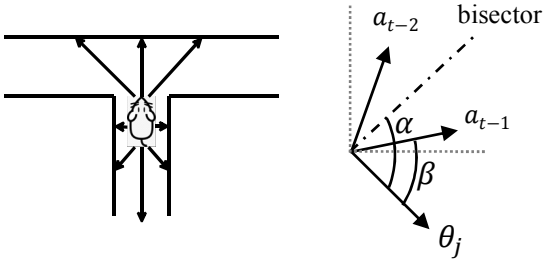

Fig. 4 Left: Affordances of a rat near an intersection. There are 8 possible directions. Right: Definition of the $\boldsymbol{\alpha}$ and $\boldsymbol{\beta}$ angles used on gate 2. $\boldsymbol{\alpha}$ is the angle between action $\boldsymbol{\theta_j}$ and the angle bisector of $\boldsymbol{a_{t-1}}$ and $\boldsymbol{a_{t-2}}$. $\boldsymbol{\beta}$ is the angle between action $\boldsymbol{\theta_j}$ and $\boldsymbol{a_{t-1}}$.

The weights for the second gate are computed using equations 10 and 11. Note that by assigning a minimum value to each weight, equation 11 prevents this gate from assigning 0 probability to any action.

$$b_j = \begin{cases} (1-\frac{\alpha}{\pi})^k & \text{if } a_{t-2}, a_{t-1} \text{ and } \theta_j \text{ are all coherent} \\ (1-\frac{\beta}{\pi})^k & \text{if } \theta_j \text{ and } a_{t-1} \text{ are coherent, but not } a_{t-2} \text{ and } a_{t-1} \\ 0 & \text{otherwise} \end{cases} \tag{10}$$

$$w_j^2 = \epsilon + (1-\epsilon)\frac{b_j}{\sum_i b_i} \tag{11}$$

where k is a constant, α is the angle between θ_j and the bisector angle between a_{t-2} and a_{t-1}, and β is the angle between θ_j and a_{t-1}. Fig. 4 illustrates the definition of α and β.

Once the output of the *softmax* has passed through both gates, the resulting probability distribution is used to sample the next action to be performed (action a_t).

C. Replay

In our work, replay is used as a means to provide offline training sequences to a reinforcement learning algorithm. The replay model has been extended from the one originally presented by Johnson and Redish [13].

To do so, during each trial, the model keeps track of a square matrix representing connection strengths between pairs of place cells. The matrix is meant to encode information regarding the paths traversed by the rat, thus we call it "the path matrix".

After each rewarded trial, while resting, the path matrix is used to generate sequences of place cell reactivations that, by converting them to sequences of (state, action) pairs, can be used to train the reinforcement learning algorithm. Fig. 5 illustrates the concept.

At the start of the simulation, the weights of the path matrix are initialized to 0. Then, as the rat moves through the maze, the weights are updated at every time step according to eq 12.

$$w_{ij}^{t+1} = w_{ij}^t + tan^{-1}\left(\frac{PC_i(x_{t+1}) + PC_i(x_t)}{2} \cdot \left(PC_j(x_{t+1}) - PC_j(x_t)\right)\right) \tag{12}$$

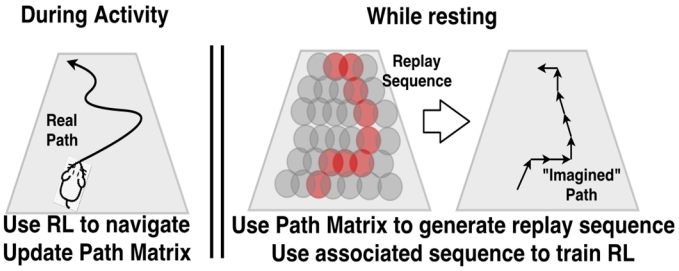

Fig. 5 Replay Model. The rat generates a path during activity. The path is used to train RL and update the path matrix. After the activity, while resting, the path matrix is used to generate a replay sequence of place cells. The sequence is then converted to a sequence of positions and acions that are used to train RL

where w_{ij} is the connection strength from place cell i to place cell j.

Eq 11 is a small variant of Johnson and Redish's formula [13], that is in turn a discretization of the formula used by Blum and Abbott [16]. The difference between eq 11 and the original model is that we use the average value of the presynaptic neuron between times t and $t + 1$, rather than the value at time $t + 1$. In both models, the more often the rat moves from place cell i to place cell j, the higher the connection strength w_{ij} will be. Thus, frequently traversed paths will be better represented by the matrix.

After each rewarded trial, when the rat is "resting", the path matrix is used to generate 200 replay sequences. Each sequence is generated by first choosing a random place cell i_0 to activate (chosen from a uniform distribution), and then using the path matrix to recursively propagate the activity until a termination criterion. Each time, the activity is propagated from the currently active place cell to its neighbor with highest connectivity as specified by eq 13.

$$i_{k+1} = argmax_j\{w_{i_k j}\} \tag{13}$$

where i_k and i_{k+1} represent the indexes of the currently active place cell and its successor, and $w_{i_k j}$ is the path matrix weight between cells i_k and j.

The activity propagation is repeated until the replay sequence reaches a cell representing the place where the rat received food, the sequence forms a loop, or until no connections surpass the propagation threshold ($\max_j w_{ij} > T$). Fig. 6 exemplifies the process.

After each replay sequence $\{i_k\}$ is generated, the sequence is converted to another sequence of state action pairs $\{(\vec{x}_k, a_k)\}$. Here, $\vec{x}_k$ represents the center of place cell i_k, and a_k represents the action whose angle best matches the direction of the displacement vector $(\vec{x}_{k+1} - \vec{x}_k)$. Finally, this newly created sequence is used to train the reinforcement learning algorithm offline by updating the state and action values according to the equations presented in the "Actor Critic" section.

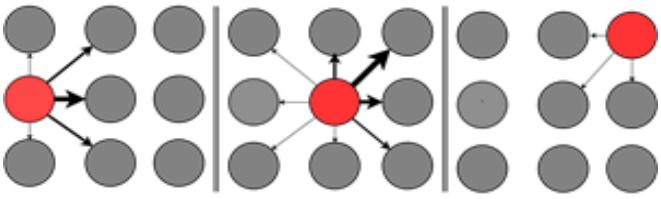

Fig. 6 PC Activity Propagation. The red circles represents the active place cells. Arrow thickness represents the connection strength. On each step the activation propagates to the neighboring cell with the strengest connection. The propagation stops when: a) no connections surpass a given threshold, b) when a loop is formed, or c) the cell representing the place where the rat received food is reached

IV. Experiments & Results

In total 4 experiments were performed sharing a similar setup. In all cases, the performance of several rat groups was compared by measuring the number of steps taken to complete the task in the multiple T-maze. Each group consisted of one hundred simulated rats. The details of each experiment are given in the following subsections. Also, table 1 provides a summary of the values used for each model parameter, and a statistical analysis of the results is provided at the end of this section.

Param	Value
σ	0.08
r_t	1 if found food 0 otherwise
γ	0.99
λ	0.6
D	2
k	1.5
ϵ	0.001
T	1

Table 1 Model parameters used in the simulations.

A. Habituation

To test whether replay could account for latent learning, the first experiment compared two groups of rats: "Habituation" vs "No Habituation". The "Habituation" group was allowed to explore the maze 10 times before rewarded trials begun, while the "no habituation" group received food from the very beginning. During both rewarded and unrewarded trials, rats were removed from the maze once they reached the feeder or once they moved 2000 steps. In total, simulated rats in both groups performed 20 rewarded trials each.

Fig. 7 shows the results for the first experiment and contrasts the performance between the two groups ("Hab" and "NoHab"). The figure shows the median number of steps taken by each group along with the upper and lower quartiles as a function of the trial number. In both groups, trial 1 corresponds to the first rewarded trial. As it can be observed, the median number of steps decreases faster for the habituation group than

for the no habituation group. This suggests that habituation rats tend to learn faster, thus exhibiting latent learning, consistent with the results obtained by Blodgett [2].

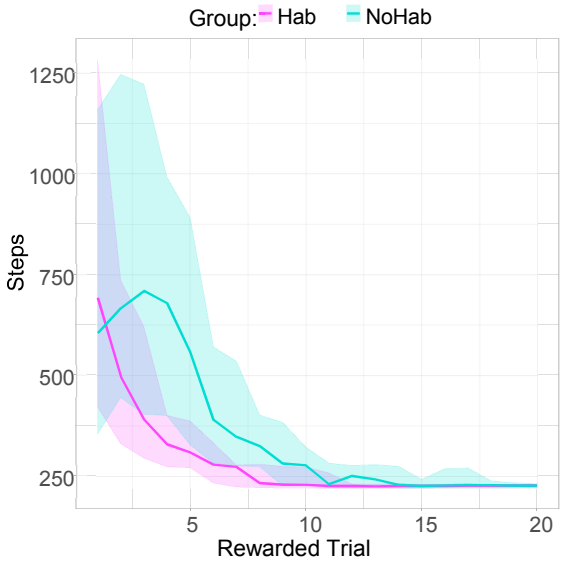

Fig. 7 The effect of habituation trials on learning. The plot shows the median completion times in number of actions for the group with habituation trials (Hab) and without (NoHab). The shaded areas indicate the upper and lower quartiles.

B. *Pre-exposure*

The second experiment performed consisted in repeating the previous experiment, but varying the amount of pre-exposure, corresponding to the pre-exposure days in the original rat experiments, to evaluate their effect on performance. In this experiment, 5 groups were compared ("Hab00", "Hab01", "Hab03", "Hab05" and "Hab10"). Each group received 0, 1, 3, 5 and 10 pre-exposure days, respectively.

Fig. 8 shows the median number of steps taken by the groups with 0, 1, 3, 5 and 10 habituation days ("Hab00", "Hab01", "Hab03", "Hab05", and "Hab10"). The figure shows that as the number of habituation trials increases, the rate at which the rats learn the task also increases.

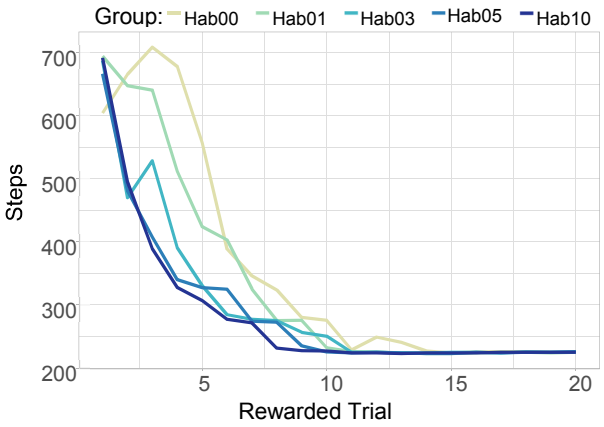

Fig. 8 The effect of the amount of habituation trials. The plot shows the median number of steps taken by the groups with 0, 1, 3, 5 and 10 habituation trials.

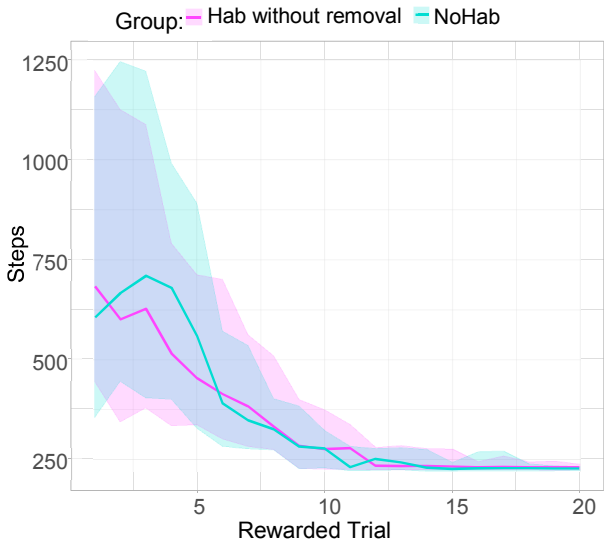

Fig. 9 The effect of not removing the animal at the rewarding location during habituation trials. The plot shows completion times of the rewarded trials for the habituation group without removal at the end of the maze (Hab without removing), and for the group without habituation (NoHab).

C. *Removal*

Blodgett's [2] original experiment removed rats at the end of the maze during unrewarded trials. In real rats, we would expect to see latent learning regardless of this condition being true. Thus, we repeated the experiment to evaluate if the model would exhibit latent learning even if the rats where not removed at the end of the maze during habituation trials. For this experiment, rats were removed only after walking 2,000 steps during habituation trials.

Fig. 9 shows the results of the experiment "Rat removal vs no removal". In the figure we only compare the groups using 10 habituation trials and the group with no habituation. The figure shows the median number of steps taken by each group as a function of the trials, along with their upper and lower quartiles (shaded regions). As opposed to Fig. 7, Fig. 9 shows no significant difference between the two groups (see next subsection for a statistical analysis). Thus, the latent learning capabilities of the model seem to fade if habituation trials do not end at the final location. This result was also true for all other groups using fewer habituation days except for one (although it seems likely to be an outlier).

To better assess the difference between the removal and non-removal groups, Fig. 10 compares the path matrices of both groups after the 10 habituation trials. As observed in the figure, particularly in the expanded views, there are three main differences between the matrices. First, the connections of the no-removal group are generally stronger than the removal group. This is likely due to the fact that no-removal rats spent more time exploring the maze than removal rats. Secondly, the connection strengths in the no-removal group look uniform

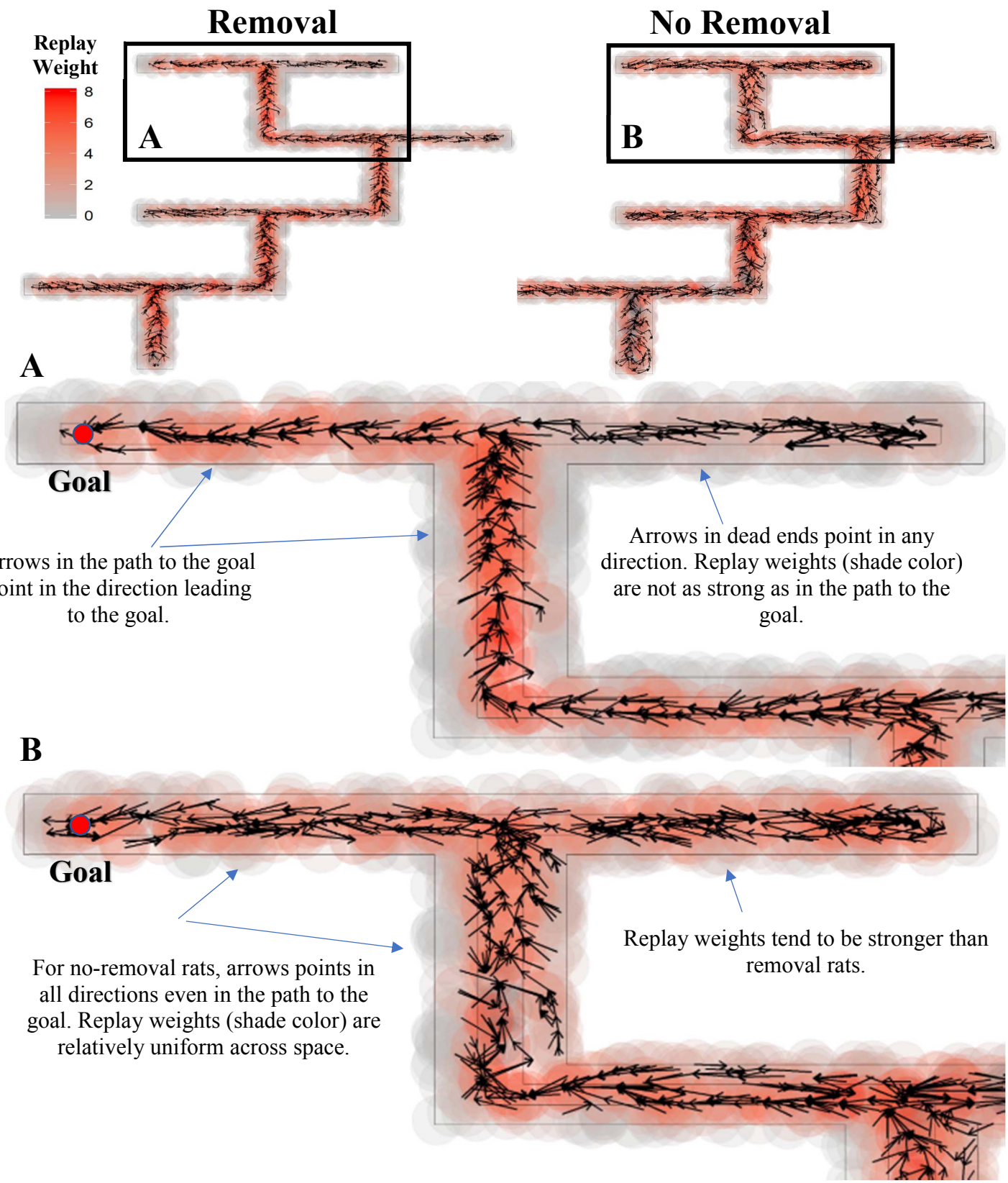

Fig. 10 Path matrix for the removal and no-removal groups with 10 habitation days. For all cells that have a connection above the propagation threshold, an arrow is drawn from its center to its neighbor with highest conectivity (replay weight) according to equation 13. Arrows indicate the direction in which the replay will propagate. Shaded circles indicate the value of the connection, the redder the color, the stronger the connection. The first row shows the matrix in the full maze. The second and third rows show an expanded view of boxes A and B respectively which correspond with the last T of the maze for each group.

across space, whereas in the removal group, cells corresponding to the optimal path have stronger connections than cells in the incorrect arms (dead ends in the maze) of the multiple T-maze. Finally, and most importantly, the directions of replay propagation for rats in the removal group have a distinct orientation pointing towards the goal in the optimal path. On the other hand, arrows for no-removal rats show no clear orientation. This is especially true along the last T of the multiple T-maze (expanded zone B), where arrows pointing in all directions can be observed. These differences are most likely due to where the rat is removed. When the rat is removed at the goal, connections oriented in the direction leading to the goal are likely to be strengthened more than in other direction.

D. Group Statistical Analysis

A Dunn test [17] was performed to test completion times during the 3rd rewarded trial in order to assess the learning rates after the rewards have started. A Benjamini-Hochberg adjustment [18] was made to account for "type I errors" [19].

Fig. 11 shows the results of the test. All but one of the habituation groups learn significantly faster than the group without habituation (i.e. have lower completion times during the 3rd rewarded trial). The only habituation group that didn't, was the group with a single habituation trial. Regarding removal, latent learning fades out when the rats are not removed at the end of the maze during non-rewarding trials ("NoRem" groups) for all but one number of habituation trials. Consistently, the Habituation groups with larger number of habituation trials perform significantly better than groups without removal.

V. Conclusions and Discussion

The computational model presented in this paper demonstrates how non-rewarded trials can speed up learning rate of rewarded trials. Consequently, we show another way in which replay may enhance spatial cognition. The model integrates replay events and reinforcement between place cells that fire concurrently, thus, forming a topological map of the environment [20], [21].

The original experiments presented by Blodgett [2] imposed restrictions on rat navigation. Once a rat had navigated an intersection in the correct direction, the intersection was closed. As a result, non-rewarded trials always ended in the "correct" location at the end of the maze. This restriction was imposed in order to decrease the amount of time taken by the experiment. However, the author argued that this restriction was not teaching the rats the path to the goal as they showed no significant decrease in errors throughout non-rewarded trials. At this time, the model reproduces the observed behaviors by requiring rats to be removed at the goal but the effect fades when the restriction is removed. It would be interesting to perform additional experiments to see if rats would display latent learning even under these circumstances.

Assessing the differences between removal and no removal of rats at the goal, we observed that the path matrix from the model does not precisely learn a map but instead it learns a specific path on each trial (thus the name we chose for the matrix). The path it learns on each trial is "a summary" of

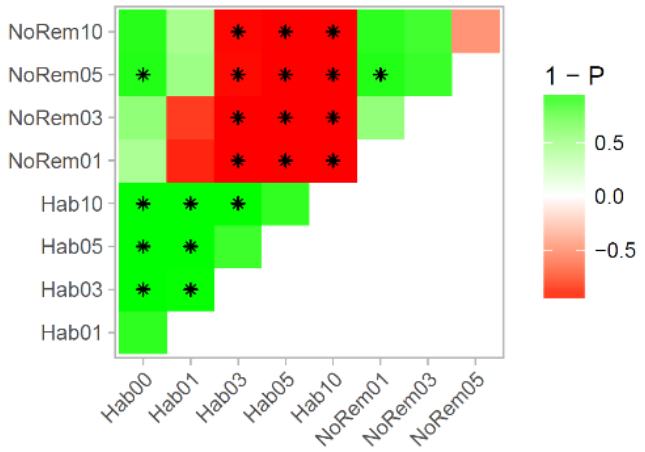

Fig. 11 The result of the Dunn test for the completion times of a rewarded trial. The color codes for the sign of the difference between groups, and the intensity codes for the statistical soundness of the conclusion. Green means the group in the row took less step than the group in the column. An asterisk was added for corrected p-values lower than 0.05.

the path traversed. For example, if the rat moves left and then goes back and decides it wanted to go right, the matrix will only remember the move to the right. When removing the rat at the end of the maze, the matrix will learn the path from start to end, while if the rat is left to roam free, the matrix will learn the path from the starting position to the location where it was removed. In order for replay to display latent learning under more general circumstances, it would be desirable to have a model that stores both the topological relationship between places (the map), as well as the traveled paths. In such case, pre-exposure should help build the map, but not the path, as there is no obvious rewarding path in the absence of food. We hypothesize that a model that benefits from a more consolidated topological map to store the paths would show the same results with respect to improved replay events and faster learning.

Additional latent learning experiments have shown how animals are able to learn the spatial distribution of different rewarding stimuli (e.g. food and water) during habituation, and then use the knowledge during tasks to navigate directly to a specific stimuli when in need, e.g. to food when hungry and to water when thirsty [22]. Our limited model is not able to reproduce these results as it lacks the concept of contextual valuation of actions. However, we argue that a learned topological map and pre-play events could assist downstream structures in the model-based decision of where to go. That is, a similar explanation as the one presented in this work could be used along with pre-play (which has been linked to action selection [9],[10]) to explain other latent learning phenomenon.

In [23], an alternative explanation is provided for the latent learning discussed in this paper. Their work suggests that place cells may not encode the current rat's location, but rather predictions of future states. To support their view, a computational model is provided and used to explain a wide variety of phenomena observed in rats, including latent learning. In their model, place cells provide a population code which encodes the successor representation in reinforcement learning [24] using a square matrix referred to as the successor

representation (SR) matrix. The SR matrix coefficients are proportional to the expected number of visits to the states represented by the place cells, starting from the same set, and using the current movement policy. Much like in our work pre-exposure allows to build the path matrix, in this framework pre-exposure allows for the building of the SR matrix. This results in reduced learning time once rewarded trials begin as exemplified by their experiments. The main differences between both explanations is that our path matrix encodes intra-hippocampal weights used for replay, while the SR matrix encodes place fields used to compute the value function.

In our current replay model, replay sequences starting from the same "location" in the same episode generate identical sequences preventing the model from "exploring" alternative routes. This is the result of propagating a place cell's activity to its neighbor with highest connectivity. Thus, in future research, it may be of interest to evaluate how different propagation methods affect the end-results.

As part of future work, we plan to verify whether rats present latent learning in multiple T-mazes without the backward-movement restrictions. Furthermore, we would like to assess two predictions that can be derived from the results of the model. First, we predict that disrupting replay events will strongly decrease the speed-ups observed in latent learning. Secondly, we expect that rats habituated to an environment will present either longer or more replay sequences as compared with non-habituated rats. Finally, depending on the results we would also like to extend the model to include a topological map as well as pre-play events.

ACKNOWLEDGMENT

This work is funded by NSF IIS Robust Intelligence research collaboration grant #1703340 at USF and U. Arizona entitled "RI: Medium: Collaborative Research: Experimental and Robotics Investigations of Multi-Scale Spatial Memory Consolidation in Complex Environments".

REFERENCES

[1] W. S. Small, "Experimental study of the mental processes of the rat. II," *Am. J. Psychol.*, pp. 206–239, 1901.

[2] H. C. Blodgett, "The effect of the introduction of reward upon the maze performance of rats.," *Univ. Calif. Publ. Psychol.*, 1929.

[3] E. C. Tolman, "Cognitive maps in rats and men.," *Psychol. Rev.*, vol. 55, no. 4, p. 189, 1948.

[4] D. S. Olton, "Mazes, maps, and memory.," *Am. Psychol.*, vol. 34, no. 7, p. 583, 1979.

[5] M. A. Wilson and B. L. McNaughton, "Reactivation of hippocampal ensemble memories during sleep," *Science (80-.).*, vol. 265, no. 5172, pp. 676–679, 1994.

[6] W. E. Skaggs, B. L. McNaughton, M. A. Wilson, and C. A. Barnes, "Theta phase precession in hippocampal neuronal populations and the compression of temporal sequences," *Hippocampus*, vol. 6, no. 2, pp. 149–172, 1996.

[7] D. J. Foster and M. A. Wilson, "Reverse replay of behavioural sequences in hippocampal place cells during the awake state," *Nature*, vol. 440, no. 7084, p. 680, 2006.

[8] M. F. Carr, S. P. Jadhav, and L. M. Frank, "Hippocampal replay in the awake state: a potential substrate for memory consolidation and retrieval," *Nat. Neurosci.*, vol. 14, no. 2, p. 147, 2011.

[9] A. C. Singer, M. F. Carr, M. P. Karlsson, and L. M. Frank, "Hippocampal SWR activity predicts correct decisions during the initial learning of an alternation task," *Neuron*, vol. 77, no. 6, pp. 1163–1173, 2013.

[10] A. Johnson and A. D. Redish, "Neural ensembles in CA3 transiently encode paths forward of the animal at a decision point," *J. Neurosci.*, vol. 27, no. 45, pp. 12176–12189, 2007.

[11] G. R. Sutherland and B. McNaughton, "Memory trace reactivation in hippocampal and neocortical neuronal ensembles," *Curr. Opin. Neurobiol.*, vol. 10, no. 2, pp. 180–186, 2000.

[12] A. P. Vorster and J. Born, "Sleep and memory in mammals, birds and invertebrates," *Neurosci. Biobehav. Rev.*, vol. 50, pp. 103–119, 2015.

[13] A. Johnson and A. D. Redish, "Hippocampal replay contributes to within session learning in a temporal difference reinforcement learning model," *Neural Networks*, vol. 18, no. 9, pp. 1163–1171, 2005.

[14] R. S. Sutton and A. G. Barto, *Reinforcement Learning: An Introduction*. MIT press, 2018.

[15] A. Guazzelli, M. Bota, F. J. Corbacho, and M. A. Arbib, "Affordances, motivations, and the world graph theory," *Adapt. Behav.*, 1998.

[16] K. I. Blum and L. F. Abbott, "A model of spatial map formation in the hippocampus of the rat," *Neural Comput.*, vol. 8, no. 1, pp. 85–93, 1996.

[17] O. J. Dunn, "Multiple comparisons using rank sums," *Technometrics*, vol. 6, no. 3, pp. 241–252, 1964.

[18] Y. Benjamini, "Discovering the false discovery rate," *J. R. Stat. Soc. Ser. B (statistical Methodol.*, vol. 72, no. 4, pp. 405–416, 2010.

[19] R. A. Fisher, "The design of experiments., 8th edn.(Oliver and Boyd: Edinburgh)," 1966.

[20] R. U. Muller, M. Stead, and J. Pach, "The hippocampus as a cognitive graph.," *J. Gen. Physiol.*, vol. 107, no. 6, pp. 663–694, 1996.

[21] A. D. Redish and D. S. Touretztky, "The role of the hippocampus in the Morris water maze," in *Computational Neuroscience*, Springer, 1998, pp. 101–106.

[22] K. W. Spence and R. Lippitt, "An experimental test of the sign-gestalt theory of trial and error learning.," *J. Exp. Psychol.*, vol. 36, no. 6, p. 491, 1946.

[23] K. L. Stachenfeld, M. M. Botvinick, and S. J. Gershman, "The hippocampus as a predictive map - supplemental material," *Nat. Neurosci.*, vol. 20, no. 11, pp. 1643–1653, 2017.

[24] P. Dayan, "Improving Generalization for Temporal Difference Learning: The Successor Representation," *Neural Comput.*, vol. 5, no. 4, pp. 613–624, 1993.