Disrupting dorsal hippocampus impairs category learning in rats

Matthew B. Broschard[1,2], Jangjin Kim[3], Bradley C. Love[4], Hunter E. Halverson[2], & John H. Freeman[2]

[1]The Picower Institute of Learning and Memory, Massachusetts Institute of Technology, Cambridge, MA, USA
[2]Department of Psychological and Brain Sciences, University of Iowa, Iowa City, IA, USA
[3]Department of Psychology, Kyungpool National University, Daegu, Korea
[4]Department of Experimental Psychology and The Alan Turing Institute, University College London, London, UK

Correspondence
John H. Freeman, Department of Psychological and Brain Sciences, University of Iowa, Iowa City, IA 52242. Email: john-freeman@uiowa.edu

Funding information
National Institutes of Health Grant P01-HD080679 to J.H.F. and B.C.L. and Wellcome Trust Investigator Award WT106931MA to B.C.L.

Abstract

Categorization requires a balance of mechanisms that can generalize across common features and discriminate against specific details. A growing literature suggests that the hippocampus may accomplish these mechanisms by using fundamental mechanisms like pattern separation, pattern completion, and memory integration. Here, we assessed the role of the rodent dorsal hippocampus (HPC) in category learning by combining inhibitory DREADDs (Designer Receptors Exclusively Activated by Designer Drugs) and simulations using a neural network model. Using touchscreens, we trained rats to categorize distributions of visual stimuli containing black and white gratings that varied along two continuous dimensions. Inactivating the dorsal HPC impaired category learning and generalization, suggesting that the rodent HPC plays an important role during categorization. Hippocampal inactivation had no effect on a control discrimination task that used identical trial procedures as the categorization tasks, suggesting that the impairments were specific to categorization. Model simulations were conducted with variants of a neural network to assess the impact of selective deficits on category learning. The hippocampal inactivation groups were best explained by a model that injected random noise into the computation that compared the similarity between category stimuli and existing memory representations. This model is akin to a deficit in mechanisms of pattern completion, which retrieves similar memory representations using partial information.

Introduction

Categorization involves grouping objects together according to perceptual or relational similarity. This requires mechanisms that can simultaneously *generalize* across within-category differences (e.g., different dog breeds vary in head shape, body size, and fur) and *discriminate* against between-category similarities (e.g., dogs and cats have similar body structure). Balancing generalization and discrimination can be accomplished by the hippocampus, which has been shown to 1) link experiences together according to overlapping features and 2) amplify differences between relatively similar memory traces (McNaughton & Morris, 1987; O'Reilly & McClelland, 1994; Hunsaker, 2013).

Early theories of categorization minimized the importance of the hippocampus in category learning (Ashby et al., 1998). This was largely because patients with amnesia did not show reliable learning impairments across multiple categorization tasks (Knowlton & Squire, 1993; Knowlton, Mangels, & Squire, 1996; Filoteo, Maddox, & Davis, 2001; Haslam, 1997; but see Zaki, 2004). However, more recent evidence from neuroimaging (Zeithamova, Dominick, & Preston, 2012; Kumaran, Summerfield, Hassabis, & Maguire, 2009; Mack, Love, & Preston, 2016), neurophysiology (Hampson, Pons, Stanford, & Deadwyler, 2004; Kraskov, Quiroga, Reddy, Fried, & Koch, 2007; Kreiman, Koch, & Fried, 2000), and rodent inactivation studies (Kim, Castro, Wasserman, & Freeman, 2018) have challenged this idea and argue that the hippocampus is central to categorization. Now, it is predicted that the hippocampus builds and maintains flexible category representations (Mack et al., 2018; Bowman & Zeithamova, 2018). This function mirrors the role of the hippocampus in maintaining structured memory representations, called 'schemas' (Tse et al., 2007; Baraduc, Duhamel, & Wirth, 2019; Guo, Chen, & Yang, 2023).

This new view has led to the development of theoretical frameworks that describe how well-documented mechanisms of the hippocampus could be leveraged during category learning. For example, EpCon (Episodes-to-Concepts), describes how pattern separation (i.e., separating similar memory traces to avoid interference; Marr, 1969; Leutgeb, Leutgeb, Moser, & Moser, 2007; Bakker, Kirwan, Miller, & Stark, 2008; Yassa & Stark, 2011; Kirwan et al., 2012), pattern completion (i.e., using partial information to retrieve memory traces; Horner, et al., 2015; Gold & Kesner, 2005; Guzman, Schlögl, Frotscher, & Jonas, 2016), and memory integration (i.e., integrating new memory traces into existing representations; Dusek & Eichenbaum, 1997; Eichenbaum, 2001; Backus, Schoffelen, Szebenyi, Hanslmayr, & Doeller, 2016; Schlichting & Preston, 2015; Pajkert et al., 2017) could all be relevant for learning new categories (Mack, Love, & Preston, 2018). EpCon posits that the hippocampus 1) retrieves memory representations that are similar to the stimulus being categorized (i.e., pattern completion), 2) integrates new stimuli into existing representations (i.e., memory integration), and 3) forms new representations after encountering surprising stimuli (i.e., pattern separation). Frameworks like EpCon are intuitive in that they build on decades of research. Nevertheless, few experiments have tested these predictions directly.

One approach to test the EpCon framework is to utilize a computational model of categorization that encompasses fundamental mechanisms of the hippocampus. One such model is SUSTAIN (Fig. 1; Supervised and Unsupervised STratified Adaptive Incremental Network; Love, Medin, & Gureckis, 2004; Love & Gureckis, 2007). SUSTAIN assumes that similar training experiences tend to cluster together in memory (Fig. 1A). Categories are represented by single or multiple ‘clusters’, where each cluster reflects a learned group of similar training experiences (Fig. 1B). Categorizing a new stimulus involves retrieving cluster representations

that are perceptually similar to that stimulus (i.e., pattern completion; Fig. 1C). After receiving feedback, the cluster representations are updated by 1) integrating the new stimulus into existing clusters (i.e., memory integration; Fig. 1D) and/or 1) forming a new cluster (i.e., pattern separation; Fig. 1E). We posit that SUSTAIN is a desirable model to bridge the fundamental mechanisms of the hippocampus with principles of category learning.

Indeed, there is growing evidence that activity in the hippocampus is functionally similar to the clustering mechanism of SUSTAIN. Multiple studies have demonstrated that the hippocampus creates 'cognitive maps' (Tolman 1948; Behrens et al., 2018) of non-spatial, multidimensional feature spaces (Eichenbaum & Cohen, 2014; Theves, Fernandez & Doeller, 2019; Solomon, Lega, Sperling, & Kahana, 2019; Constantinesceu, O'Reilly, & Behrens, 2016; Morton, Sherill, & Preston, 2017). These representations emphasize category-relevant stimulus information and reflect task goals (Theves, Fernandez & Doeller, 2020; Mack et al., 2016). Furthermore, Mok & Love, 2019 showed that a clustering model could simulate neural activity of place cells and grid cells as a rat navigated an environment. This suggests that similar mechanisms may be recruited to mediate both spatial navigation and concept learning. Expanding the investigation of the hippocampus to non-spatial paradigms like categorization may provide key insight regarding generalized hippocampal mechanisms that go beyond spatial navigation.

In the current experiment, we used inhibitory DREADDs (Designer Receptors Exclusively Activated by Designer Drugs; Roth, 2016) to examine the role of the dorsal hippocampus (HPC) in category learning. Using a touchscreen apparatus, rats were trained to categorize distributions of controlled visual stimuli derived from classic human paradigms that have been used for decades (Ashby et al., 1998). The category stimuli contained black and white

gratings that varied along two continuous dimensions (i.e., spatial frequency and orientation; Fig. 2A; Broschard, Kim, Love, Wasserman, & Freeman, 2019; Ashby et al., 1998). For some rats, categorizing the stimuli encouraged a shift of attention to a single stimulus dimension (i.e., 1D tasks; spatial frequency *or* orientation). For other rats, categorizing the stimuli required attention to both stimulus dimensions (i.e., 2D tasks; spatial frequency *and* orientation). Inactivation of the HPC impaired category learning and generalization for both the 1D tasks and 2D tasks. We then fit SUSTAIN to the learning data to test the role of the HPC in storing and retrieving category representations.

Material & Methods

Subjects

Thirty-eight Long Evans rats (twenty female; n = ~9 per group) were used for the following experiment. After arriving in the animal colony, rats were given *ad libitum* access to food and water and put on a 12-hour light/dark cycle. Food was restricted after a week of acclimating to the new environment. Weights were recorded daily so that the rats did not go below 85% of their free feeding weight. All procedures were approved by the Institutional Animal Care and Use Committee at the University of Iowa.

Touchscreen Apparatus

All experimental sessions were conducted in custom-built chambers outfitted with a touchscreen (36 × 41 × 36 cm). A computer monitor (Model 1550V, NEC, Melville, NY) was mounted on the right wall of each chamber and presented visual stimuli to the rats. A touchscreen (15-in, Elo Touch Systems, Fremont, CA) overlaid the computer monitor and allowed the rats to interact

with the screen. A food tray (6.5 × 13 × 4.5 cm) was positioned on the left wall of each chamber and delivered food pellets to the rats via a rotary pellet dispenser (Med Associates Inc., Georgia, VT, model ENV-203IR) that was controlled by an electrical board (Model RS-232, National Control Devices, Osceola, MO). A house light above the food tray was always on during experimental sessions. White noise was used in the experimental room to minimize distractions. All experimental sessions and procedures were controlled by custom-written MATLAB scripts (MathWorks, Natick, MA). Finally, a camera (model ELP-USB100W05MT-RL36) was mounted to the ceiling of each chamber to observe the rats' behavior.

Pre-Training Procedures

After acclimating to the animal colony, each rat was handled daily for one week to reduce the stress of interacting with experimenters. Then, each rat was placed on a laboratory cart and was encouraged to forage for 45-mg pellets scattered on the cart's surface. This procedure has been shown to accelerate habituation to the lab environment (Kim et al., 2018) and primes the rats to search for food pellets within the touchscreen chambers. This procedure was repeated daily until the rats consumed at least twenty pellets within fifteen minutes. Finally, each rat underwent a daily shaping procedure within the touchscreen chambers to learn to interact with the touchscreen (for details, see Broschard, Kim, Love, & Freeman, 2020). This procedure included four separate phases; each phase was incrementally similar to the trial sequence used during category training and testing sessions. All shaping procedures took about 14 days.

Surgery

After all pre-training procedures, each rat underwent stereotaxic surgery. Under isoflourane (1% - 4%) anesthesia, either AAV5-CaMKIIa-hM4D(Gi)-mCherry or AAV5-CaMKIIa-EGFP (Roth, 2016) was infused bilaterally into the HPC (1 µL per hemisphere; AP: -3.8; ML: ±2.5; DV: -3.2) using a Hamilton syringe (1 µL; 26 gauge). Viral constructs contained a CaMKIIα promoter that targeted excitatory neurons within the HPC. The inhibitory DREADD construct contained DNA for a GPCR (hM4Di; G Protein-Coupled Receptor) that hyperpolarizes neurons when activated by the synthetic ligand, clozapine-N-oxide (CNO). The control virus did not contain DNA for hM4Di. Viral constructs also contained a fluorescent tag (i.e., inhibitory virus: mCherry; control virus: GFP) so that viral expression and location could be observed after data collection was completed. Meloxicam (1 mg/ml) was administered during and 24 hours after surgery as an analgesic. Rats were placed on a heat pad immediately after surgery to prevent hypothermia. Rats were given one week to recover. Category training sessions began no sooner than three weeks after surgery to ensure adequate transduction of the viral construct.

Category Tasks

Across multiple training and testing sessions, rats categorized abstract visual stimuli into two categories (i.e., category 'A' and category 'B'). Briefly, on each trial, a unique stimulus was presented to the rat, and the rat decided the category membership of the stimulus by choosing one of two report keys. Food pellets were delivered after correct responses to reinforce the rats' behavior.

The visual stimuli (239 x 239 pixels; Fig. 2A) contained black and white gratings that, across stimuli, varied along two continuous dimensions: spatial frequency (0.2532 cycles per visual degree to 1.2232 cpd) and orientation (0 radians to 1.75 radians). The ranges of these

dimensions are within the perceptual limits of Long Evans rats using touchscreens (Crijns & Op de Beeck, 2019) and were determined to have roughly equal salience (Broschard et al., 2019). A two-dimensional stimulus space was created by performing linear transformations of these dimensions so that both dimensions had a common scale (i.e., 0 to 100; Broschard et al., 2019).

Category tasks were created by placing bivariate normal distributions on this stimulus space (Fig. 2A; Category A: $\mu_X = 30$, $\sigma_X = 2.5$, $\mu_Y = 50$, $\sigma_Y = 20$; Category B: $\mu_X = 70$, $\sigma_X = 2.5$, $\mu_Y = 50$, $\sigma_Y = 20$; Broschard et al., 2019). Each distribution constituted a category, and each point within a distribution represented a unique category stimulus. Three additional tasks were created by rotating these distributions in 45-degree increments (Fig. 2A). This rotation does not affect any physical property of the distributions (e.g., standard deviation, mean between-category distance, etc.; Ashby, Smith, & Rosedahl, 2020); however, it does affect how the distributions are oriented relative to the axes of the stimulus space. The 1D tasks had distributions that were perpendicular to a stimulus axis. For these tasks, only one dimension (i.e., the perpendicular dimension) was category-relevant, and the other dimension (i.e., the parallel dimension) could be ignored. 1D tasks are typically learned by shifting attention towards the category-relevant dimension (Broschard et al., 2019). Conversely, the 2D tasks had distributions that were not aligned with either stimulus axis. For these tasks, both dimensions were category-relevant. 2D tasks are typically learned by combining information from both stimulus dimensions[1].

Category Training

Rats were randomly assigned to learn one of the four category tasks (Broschard et al., 2019; Broschard et al., 2020) and were given ten training sessions to learn their respective task. Each

1.Humans typically learn the 1D tasks faster than the 2D tasks. This learning advantage has been attributed to humans' propensity for testing unidimensional strategies (Ashby et al., 1998), which is governed by the lateral prefrontal cortex (Wallis & Miller, 2001). Rats, on the other hand, typically learn the 1D tasks and the 2D tasks at the same rate, suggesting that rats have a smaller capacity for rule-based learning. Nevertheless, rats seem to learn the 1D tasks by orienting attention to the relevant stimulus dimension (Broschard et al., 2019). This form of selective attention is mediated by the rodent prelimbic prefrontal cortex (Broschard et al., 2021).

session included eighty training trials. On each trial, a star stimulus was presented at the center of the screen (Fig. 2B; Star Phase). After one touch of the star, a category stimulus was randomly selected from one of the training distributions and was presented at the center of the screen (Cue Phase). After three observing touches of this stimulus, copies of the stimulus were presented on the left and right sides of the screen, acting as report keys (Choice Phase). The rat touched either report key, depending on the category membership of the stimulus during the Cue Phase. The categories were mapped spatially such that members of category 'A' required a touch to the left report key and members of category 'B' required a touch to the right report key. If the rat chose the correct side, a white box appeared; one touch of this box delivered a food pellet (Reward Phase). If the rat chose the incorrect side, a correction trial was initiated. Here, the trial repeated from the Cue Phase after a five to ten second timeout. Correction trials continued without food reinforcement until the correct side was selected or after three consecutive correction trials. Inter-trial intervals ranged from five to ten seconds. IP injections of CNO (1.0 mg/ml) were administered thirty minutes before each training session to activate the GPCRs. All sessions were completed within two hours to ensure that the CNO was effective throughout the session. The CNO was dissolved in DMSO and was suspected in sterile saline. Remaining CNO was placed in an -4 degree C freezer and was used for up to seven days.

Category Generalization

After category training, rats were given five testing sessions to examine category generalization (Broschard et al., 2019; Fig. 6A). The testing stimuli had a grid configuration that spanned the entire stimulus space. Each testing session sampled from each point in the grid once (i.e., 84 trials). A third of the testing stimuli overlapped with the training distributions (i.e., Trained;

within two standard deviations), a third of the testing stimuli were closer to the category boundary relative to the training distributions (i.e., Proximal), and a third of the testing stimuli were farther from the category boundary relative to the training distributions (i.e., Distal). Generally, accuracy improves for stimuli farther from the boundary (Broschard et al., 2019). The trial sequence was identical to training sessions except that correction trials were not administered after incorrect responses. Therefore, all choices during the testing sessions were reinforced. IP injections of CNO (1.0 mg/ml) were administered thirty minutes before each session to activate the GPCRs. All sessions were completed within two hours to ensure that the CNO was effective throughout the session.

Simple Discrimination

Finally, rats underwent training sessions to learn a control discrimination task. Instead of categories of stimuli, only two images were presented during training sessions (i.e., a light box and a dark box; both images contained a common pattern of dots to add perceptual complexity Fig. 7A; Kim et al., 2018). The light stimulus was mapped to the left report key, and the dark stimulus was mapped to the right report key. All other training procedures were identical to the categorization sessions; therefore, this task acted a control to ensure that group differences were not caused by deficits in functions unrelated to categorization (e.g., movement, motivation, perception, spatial learning, etc.). Each training session contained 72 training trials. Sessions continued until the rat reached a learning criterion (i.e., at least 75% accuracy for both images on two consecutive sessions). IP injections of CNO (1.0 mg/ml) were administered thirty minutes before each session. All sessions were completed within two hours.

Statistical Analysis

Multiple dependent measures quantified the rats' performance during training and testing sessions. Session accuracy was defined as the proportion of correct responses during the Choice phase. Reaction time was calculated during the Cue phase and Choice phase to quantify the amount of time to 1) observe the stimulus and 2) make a category decision. Reaction times from incorrect trials were excluded from all analyses. Additionally, reaction times that exceeded two standard deviations of the mean were excluded from all analyses, a criterion that is commonly used to eliminate outliers (O'Donoghue et al., 2020). These outliers rarely occurred.

These dependent measures were analyzed using linear mixed effects modeling (R, version 3.4.2). Models used for training sessions included fixed effects for experimental group, training session, and a quadratic function across training sessions, as well as random effects for slope, intercept, and the quadratic function. Models for testing sessions included fixed effects for experimental group, trial type (Distal, Trained, and Proximal), and a quadratic function across trial types, as well as random effects for slope, intercept, and the quadratic function. Quadratic functions were used because they best fit the data, and higher order terms did not significantly improve these fits. Sex was added as a covariate for all models to check whether there were any significant differences between male and female rats. To find the simplest model that fit the data, we used a model simplification strategy (Crawley, 2007). We started with the full model and then systematically removed random effects one at a time. This continued until the estimates were significantly different from the larger model before it. Finally, a covariate for sex was added to each model to examine differences in performance between male and female rats.

Histology

After all behavioral testing was complete, rats were perfused to verify viral expression and placement. Rats were given a lethal dose of euthanasia solution (sodium pentobarbital) and then perfused with ~150 mL PBS and ~150 mL of 4% paraformaldehyde. Brains were covered in foil and stored at 4°C. Then, a sliding microtome made coronal sections (50 µm) of the target region. Slides were cover slipped and stored in a dark, cold environment. Sections were observed under a fluorescent microscope to ensure that viral expression was contained within the HPC. The boundary of the HPC was defined according to Paxinos & Watson, 1998. Rats with viral expression largely outside of the HPC were excluded from all analyses.

SUSTAIN Modeling

The network SUSTAIN has been useful in multiple contexts for mapping neural activity to specific cognitive processes (Love et al., 2004; Mack et al., 2016; Mack et al., 2020; Broschard et al., 2021; Fig. 8A). The current analysis used SUSTAIN to assess potential functions of the HPC during category learning (Fig. 1). This was accomplished by designing multiple model variants. Each model variant simulated the effect of the inhibitory DREADDs by disrupting a single computation of the network. A model comparison approach was used, such that the function of the HPC was inferred by determining which model variant produced the best fit of the learning data. This approach provided a top-down framework by which we could test the impact of selective learning deficits. The first three models tested whether the HPC is critical for maintaining category representations (i.e., Model 1: retrieving representations; Model 2: updating representations; Model 3: recruiting new representations; Mack et al., 2018; Love & Gureckis, 2007). Model 4 tested whether the HPC is critical for selective attention, presumably through interactions with the prefrontal cortex (Mack et al., 2020; Broschard et al., 2021).

Finally, Model 5 was a control model and assumed that the HPC was not critical for category learning.

SUSTAIN represents categories through single or multiple 'clusters'; each cluster reflects a learned group of similar training experiences (Love et al., 2004). On each training trial, the current stimulus is compared to existing clusters, and each cluster is activated according to its similarity to the stimulus. The cluster with the highest activation, the 'winning' cluster, sends its activation to an output layer, which makes a probabilistic decision regarding the category membership of the stimulus. *Model 1* (Pattern Completion) assumed that the HPC is critical for retrieving the winning cluster by comparing the similarity between the current stimulus and each cluster. In this model, hippocampal inactivation was simulated by adding a normal distribution of noise to the activation of each cluster, thereby increasing the probability that the model retrieved a cluster that was dissimilar to the current stimulus. The mean of this distribution was zero, and the standard deviation of this distribution was a positive free parameter.

After making a category decision, SUSTAIN receives feedback on its decision and updates the cluster representations accordingly. This is accomplished by moving the position of the winning cluster towards the position of the current stimulus. *Model 2* assumed that the HPC is critical for updating cluster representations (Memory Integration). For this model, hippocampal inactivation was simulated by moving the position of the winning cluster in a random direction (instead of towards the current stimulus). SUSTAIN can also update the representations by recruiting a new cluster. SUSTAIN contains a single cluster at the beginning of training and recruits new clusters after encountering 'surprising' stimuli (e.g., discovering that a bat is a mammal and not a bird). A cluster is recruited when the cluster activations exceed a threshold value, indicating that the model was especially confident in an incorrect decision.

Typically, 1D tasks are learned by recruiting a single cluster per category, and 2D tasks are learned by recruiting multiple clusters per category (~4-5; Broschard et al., 2020). *Model 3* (Pattern Separation) assumed that the HPC is critical for recruiting new clusters. In this model, hippocampal inactivation was simulated by increasing the threshold value, thereby limiting cluster recruitment.

Finally, SUSTAIN contains an attention mechanism that modulates the current stimulus before it is compared to the cluster representations. This mechanism allows stimulus information from category-relevant dimensions to contribute more to the cluster activations (and therefore the category decision). *Model 4* (Selective Attention) assumed that the HPC is critical for this mechanism, presumably through interactions with the prefrontal cortex. Hippocampal inactivation was simulated by shuffling the proportion of attention towards each stimulus dimension before each trial, thereby increasing the probability that attention was directed towards category-irrelevant dimensions. These models were compared to *Model 5* (Control), which assumed that the HPC was not necessary for category learning.

Using the MATLAB function *fmincon*, SUSTAIN was first fit to the average learning curves of the control groups by optimizing SUSTAIN's free parameters. This provided a baseline model that learned the category tasks at the same rate as a typical rat. The experimental models were derived from the baseline model; each experimental model was fit to the average learning curves of the inactivation groups. The quality-of-fit was determined for each experimental model by calculating the Bayesian Information Criterion (BIC; Neath & Cavanaugh, 2011). The experimental model that best fit the inactivation groups (i.e., the lowest BIC value) was used to infer the function of the HPC during category learning.

Perceptual Recency Effects

With the current experimental design, each rat completed a large number of training trials, which allowed us to track category learning on a trial-by-trial basis. This sensitivity was leveraged to observe how category performance was influenced by the identity of the most recent training exemplar (i.e., perceptual recency effects; Jones, Love, & Maddox, 2006). Recency effects often interact with the perceptual similarity between exemplars. For example, performance is facilitated if the exemplar is perceptually similar to the most recent exemplar (Jones et al., 2006). Therefore, we binned the accuracy of training trials according to the perceived similarity between the current exemplar (*n*) and the most recent exemplar (*n-1*; Nosofsky, 1986). Perceptual similarity between exemplars *i* and *j* was calculated as:

$$s_{ij} = e^{-d_{ij}},$$

where *d* is the psychological distance between exemplars *i* and *j*. Psychological distance was defined as,

$$d_{ij} = \sum_{m=1}^{M} w_m * \left| x_i - x_j \right|$$

where w_m was SUSTAIN's estimated attention weight for dimension *m* on trial *n,* and *x* was the physical value of the exemplar along dimension *m.* Trial effects were isolated by subtracting the binned accuracies by the average of 1,000 permutations where trial order was shuffled. Therefore, positive recency scores indicate increased accuracy due to trial order, negative scores indicate decreased accuracy due to trial order, and 0 indicates no effect of trial order.

CNO Control Experiment

Thirty-two rats (16 females; n = ~8 per group) were used for a control experiment to ensure that IP injections of CNO do not affect categorization by interacting with non-target receptor types.

For this experiment, rats were given ten training sessions and five testing sessions on either a 1D task or a 2D task. All procedures were the same as before except the rats did not undergo stereotaxic surgery. IP injections of either CNO (1.0 mg/ml) or PBS were administered 30 minutes before each session and each session did not exceed two hours. Accuracy and reaction time were measured to examine any effect of CNO on categorization.

DREADDs Verification

A control experiment was conducted such that *in vivo* single units were recorded in the HPC to verify that the inhibitory DREADD effectively suppressed neural activity. For this experiment, AAV was infused into the HPC of a male rat during stereotaxic surgery. Critically, the inhibitory DREADD (AAV5-CaMKIIa-hM4D(Gi)-mCherry; 1 µL) was infused into one hippocampal hemisphere (AP: -3.8; ML: -2.5; DV: -3.2) and the control DREADD (AAV5-CaMKIIa-EGFP; 1 µL) was infused into the other hemisphere (AP: -3.8; ML: +2.5; DV: -3.2). Meloxicam (1 mg/ml) was administered during and after the surgery as an analgesic. After a week of recovery, the rat underwent a second surgery to implant a custom-built microdrive supporting movable tetrodes (8 recording tetrodes, 2 reference tetrodes; final impedance of each wire was adjusted to 150-300 kΩ using a gold solution) that targeted both hippocampal hemispheres. Two exit tips were positioned over the HPC, and each tetrode was lowered 1.0 mm into the brain. Meloxicam was administered to increase recovery.

After recovery, the tetrodes were slowly lowered in 0.25 mm increments. The recording tetrodes were lowered to their target site (DV: -3.2 mm) and small adjustments were made until neural recordings were stable on the majority of the recording tetrodes. The reference tetrodes (one per hemisphere) were lowered until no single units were detectable (i.e., ~1.0 mm above the

HPC). Data were amplified and digitized using data acquisition software (Neuralynx). Single unit activity was sampled at 32 kHz. Spikes from single units were isolated off-line through cluster cutting software (MClust 4.4). Multiple parameters, including peak, width, height, and energy associated with the waveforms as well as the interspike interval histograms, were used to isolate single units.

Once single units were stable, an IP injection of CNO was administered to examine its effect on neural activity of each hemisphere. Spiking activity was recorded thirty minutes before the injection and three hours post injection. This procedure was repeated for several days. Once complete, the position of each individual tetrode was marked by electrolytic lesions (10 µA current for 10 s). The rat was then perfused, and hippocampal sections (50 µm) were observed under a fluorescent microscope to observe the spread of the AVV as well as the position of the tetrodes.

Results

Hippocampal inactivation impairs category learning

All rats had adequate viral expression within the HPC, as described by Paxinos & Watson, 1998 (see Fig. 3 for representative examples of the AAV position and spread). Therefore, data from all rats were included in the following analyses. AAV did not extend into the ventral hippocampus and was contained within AP: -2.8 and -4.9. For the majority of rats, AAV extended into each hippocampal subfield (i.e., CA1, CA3, and the DG). For a small subset of rats (three males and two females), AAV did not extend into the CA3. Accuracy and reaction time were not different for these rats.

We first examined session accuracy during category training. Accuracy significantly increased across the ten training sessions (Fig. 4; $t(34.45) = 7.95$, $p < .001$), suggesting that the rats reliably learned the 1D and 2D tasks. There were no significant differences between sexes (males vs. females: $t(35.20) = -1.94$, $p = .061$), as well as between task types (1D tasks vs. 2D tasks: $t(45.42) = -0.24$, $p = .981$), suggesting that all groups learned the tasks at the same rate and to equal levels. For both task types, rats with hippocampal inactivations had impaired accuracy compared to the controls (1D tasks: $t(68.12) = -3.30$, $p = .002$; 2D tasks: $t(96.74) = -2.35$, $p = .021$), suggesting that the rat HPC is critical for category learning.

The effect of hippocampal inactivation on reaction time

We next examined whether the hippocampal inactivations affected reaction time during each trial event (i.e., the Cue phase and the Choice phase). Across the training sessions, Choice RT decreased significantly (Fig. 5B; $t(34.55) = -2.92$, $p = .006$), but Cue RT did not change (Fig. 5A; $t(37.18) = 0.47$, $p = .644$). This suggests that the amount of time required to make each category decision decreased across training sessions, but the average time to observe each stimulus was consistent across training. Reaction time did not differ between the task types (1D tasks vs. 2D tasks; Cue RT: $t(52.83) = 0.37$, $p = .712$; Choice RT: $t(49.70) = .89$, $p = .376$). Interestingly, reaction time was significantly faster for the males compared to the females (Cue RT: $t(36.27) = 3.00$, $p = .005$; Choice RT: $t(34.91) = -2.86$, $p = .008$).

Compared to the control groups, rats with hippocampal inactivations had longer Choice RT (1D tasks: $t(96.35) = 2.32$, $p = .022$; 2D tasks: $t(109.10) = 2.07$, $p = .041$). This difference in Choice RT was present throughout training for rats learning the 1D tasks but emerged during later training sessions for rats learning the 2D tasks. Cue RT was significantly longer for the

inactivated rats learning the 1D tasks ($t(60.17) = 3.20$, $p = .002$), but not rats learning the 2D tasks ($t(52.83) = 0.37$, $p = .712$). Taken together, these results suggest that without the HPC, rats required more time to examine each stimulus and make category decisions. These differences were pervasive in rats learning the 1D tasks and emerged later in training in rats learning the 2D tasks.

Hippocampal inactivation impairs category generalization

The rats were then given five testing sessions to examine category generalization (Figs. 6A). Stimuli were configured into a grid that spanned the entire stimulus space. We first examined how accuracy changed across the space by generating heatmaps of the rats' accuracy (Fig. 6B). Each grid was rotated so that all category tasks had the same orientation (i.e., the x-axis was perpendicular to the category boundary and the y-axis was parallel to the category boundary). Accuracy was largely affected by distances along the relevant axis, such that accuracy increased for stimuli farther from category boundary and decreased for stimuli closer to the category boundary. Accuracy was unaffected by distances along the other, irrelevant axis. For rats with hippocampal inactivations, accuracy was impaired across the entire stimulus space.

We quantified these patterns by separating the stimuli into three trial types: 1) 'Trained' stimuli overlapped with the training distributions, 2) 'Distal' stimuli were farther from the category boundary, and 3) 'Proximal' stimuli were closer to the category boundary. As expected, accuracy was related to the distances between the testing stimuli and the category boundary. Compared to the Trained stimuli, accuracy was impaired for the Proximal stimuli ($t(74.97) = -13.80$, $p < .001$), and accuracy improved for the Distal stimuli ($t(74.97) = 5.32$, $p < .001$). Accuracy was not significantly different between task types ($t(110.00) = -0.48$, $p = .636$) or

between sexes ($t(21.84) = -0.91$, $p = .375$), replicating the training results. Compared to the controls, inactivating the hippocampus impaired category generalization for both task types (1D tasks: $t(110.00) = -2.17$, $p = .032$; 2D tasks: $t(110.00) = -3.42$, $p < .001$). There were no significant interactions across trial types (all $p > .05$), suggesting that performance was equally impaired across the stimulus space.

Finally, we examined reaction time during the testing sessions. Choice RT was significantly slower for Proximal stimuli compared to Trained stimuli (Proximal vs. Trained: $t(74.99) = 3.06$, $p = .003$; Distal vs. Trained: $t(74.99) = -0.06$, $p = .954$), suggesting that the rats perceived the Proximal stimuli as more difficult. Conversely, Cue RT did not differ across trial types (Proximal vs. Trained stimuli: $t(74.94) = 1.51$, $p = .135$; Distal vs. Trained stimuli: $t(74.94) = 1.26$, $p = .212$), suggesting that the rats required an equal amount of time to view each stimulus. There were no significant differences in reaction time between task types (Cue RT: $t(35.82) = 1.69$, $p = .100$; Choice RT: $t(36.86) = -0.35$, $p = .732$) or between sexes (Cue RT: $t(29.30) = 1.55$, $p = .101$; Choice RT: $t(30.05) = -0.12$, $p = .907$). The hippocampal inactivations had no effect on Cue RT (1D tasks: $t(107.75) = 1.21$, $p = .201$; 2D tasks: $t(31.85) = -0.14$, $p = .892$) or Choice RT (1D tasks: $t(98.99) = 1.55$, $p = .124$; 2D tasks: $t(33.93) = 1.29$, $p = .205$), suggesting that the HPC's contribution to decision-making is specific to early training sessions.

Hippocampal inactivation does not affect learning a control discrimination task

After category generalization, all rats were trained on a control discrimination task and learned to differentiate between a white stimulus and a black stimulus (both stimuli contained a common pattern of dots to add perceptual complexity; Fig 7A). Training sessions continued until each rat reached a learning criterion (i.e., at least 75% accuracy for both stimuli on two consecutive

sessions). Using a 2x2 ANOVA, there were no significant differences in the number of sessions to reach this criterion across experimental groups (Fig. 7B; $F(3,34) = .59$, $p = .626$). This suggests that the observed impairments during the category sessions were not related to deficits in irrelevant factors such as perception, motivation, movement, and stimulus-spatial response mapping.

Hippocampal impairments are best simulated by a deficit in pattern completion mechanisms

SUSTAIN was used to further examine the role of the HPC in category learning. This was accomplished by designing and fitting multiple experimental models to the learning data. Each model assumed that inactivating the HPC produced a unique deficit during learning. We inferred the role of the HPC according to the model that best fit the data (Fig. 8A; a complete description of each model can be found in Materials & Methods). Models 1-3 assumed that the HPC was critical for maintaining category representations (i.e., Pattern Completion, Memory Integration, and Pattern Separation, respectively). Model 4 (Selective Attention) assumed that the HPC was critical for selective attention. Model 5 (Control Model) served as a control and assumed that inactivating the HPC had no effect on category learning.

Figure 8B shows the BIC values of each model variant. First, all models produced a better fit of the data than the control model. Second, the models that targeted SUSTAIN's cluster layer (i.e., Models 1-3) produced a better fit of the data than the Selective Attention model, which failed to predict a learning impairment in the 2D tasks. This supports the general prediction that the HPC is important for maintaining abstract category representations (Mack et al., 2016; Love & Gureckis, 2007). Model 1 (Pattern Completion) produced the best fit of the learning data (Figs. 8B&C). This model assumed that the HPC was critical for retrieving

appropriate cluster representations. In SUSTAIN, cluster representations are activated according to their similarity to the current stimulus. Clusters that are strongly activated are retrieved and used to categorize the current stimulus. The Pattern Completion model simulated the hippocampal inactivations by adding a normal distribution of noise to these similarity judgements. This noise increased the probability that category decisions were based on cluster representations that were dissimilar to the current stimulus.

To assess how this learning deficit affected the underlying cluster representations, we examined the cluster layer of the winning model (Fig. 8D; Pattern Completion model). For the control groups, SUSTAIN recruited 1-2 clusters per category to learn the 1D tasks and 5-6 clusters per category to learn the 2D tasks (Broschard et al., 2020). This suggests that the category representations for the 1D tasks tended to be more prototype-based (Rosch & Mervis, 1975), whereas the representations for the 2D tasks tended to be more exemplar-based (Nosofsky, 1986). For the inactivation groups, SUSTAIN recruited about twice the number of clusters for both task types.

Hippocampal inactivation impairs perceptual recency effects

Broschard et al., 2021 demonstrated that rats' decisions were influenced by recent training experiences. Specifically, accuracy was facilitated if the current stimulus was perceptually similar to the most recent stimulus, and accuracy was impaired if the current stimulus was perceptually dissimilar to the most recent stimulus. Broschard et al., 2021 found that these recency effects were mediated by the rodent prelimbic cortex. Here, we tested the prediction that these recency effects are also mediated by the HPC.

As expected, we found strong recency effects for the control rats (Fig. 9). Specifically, performance was facilitated (i.e., a positive recency score) if the current stimulus was perceptually similar to the previous stimulus (i.e., similarity above the median), and performance was impaired (i.e., a negative recency score) if the current stimulus was perceptually dissimilar to the previous stimulus (i.e., similarity below the median). Importantly, these recency effects were reduced in rats with hippocampal inactivation (Fig 9; low similarity: $F(3,40) = 17.49$, $p < .001$; high similarity: $F(3,40) = 17.22$, $p < .001$). This suggests that the inactivation groups were less likely to make their decision according to the identity of the previous stimulus. This aligns with the SUSTAIN simulation results and suggests that the HPC is critical for comparing the current stimulus to previous training experiences.

IP injections of CNO do not affect categorization

The current experiment used CNO as a ligand to activate the designer receptors. Although unlikely, it is possible that the categorization impairments were caused by CNO binding to non-target receptors instead of the designer receptors. To rule out this possibility, we conducted a control experiment where rats were given IP injections of either CNO or PBS before learning the category tasks. All procedures were identical except that no AAV was infused into the HPC.

We first tested whether CNO influenced category learning (Fig. 10A). As before, sessions accuracy significantly improved across learning ($t(30.08) = 7.62$, $p < .001$), Choice RT significantly decreased across learning ($t(29.96) = -2.05$, $p = .049$), and Cue RT did not change across sessions ($t(30.25) = -1.71$, $p = .098$). There were no significant differences in category learning between rats injected with CNO vs. PBS. This was true for rats learning the 1D tasks (Accuracy: $t(27.14) = -0.58$, $p = .567$; Cue RT: $t(29.31) = -1.00$, $p = .324$; Choice RT: $t(26.92) =$

-0.69, p = .499) as well as rats learning the 2D tasks (Accuracy: t(26.89) = 0.61, p = .549; Cue RT: t(29.58) = 0.81, p = .425; Choice RT: t(26.97) = -1.32, p = .229). These results suggest that the CNO injections did not affect category learning by binding to non-target receptor types.

We next examined the effect of injecting CNO on category generalization (Fig. 10B). Stimuli were segregated into three trial types: Trained, Distal, and Proximal. As before, accuracy was significantly impaired for the Proximal stimuli (Trained vs. Proximal: t(60.00) = -11.74, p < .001), and significantly improved for the Distal stimuli (Trained vs. Distal: t(60.00) = 5.39, p < .001). Cue RT and Choice RT were not significantly different across trial types (all p > .05). There were no significant differences between rats administered injections of CNO compared to rats injected with PBS (Fig. 10B). This was true for rats that had learned the 1D tasks (Accuracy: t(33.53) = 0.05, p = .961; Cue RT: t(41.10) = 0.27, p = .787; Choice RT: t(29.49) = 0.75, p = .460) as well as the 2D tasks (Accuracy: t(33.53) = -0.26, p = .793; Cue RT: t(41.10) = 1.42, p = .163; Choice RT: t(29.49) = -0.54, p = .594). This suggests that the IP injections of CNO themselves did not affect category generalization.

DREADDs AAV inhibits hippocampal activity

As a separate control, we recorded *in vivo* single unit activity in a single rat to confirm that the inhibitory DREADD effectively suppressed neural activity in the HPC (Fig. 11). This experiment used a within-subject design, such that the inhibitory DREADD construct was infused into one hippocampal hemisphere, and the control virus was infused into the other hippocampal hemisphere (see Fig. 11A for the virus spread). Spiking activity was recorded before and after the administration of an IP injection of CNO to examine its effect on neural activity. Figure 11B shows the firing rate of two representative cells transduced by the inhibitory DREADD. Firing

rate decreased from baseline thirty minutes after the CNO injection (Wilcoxon Rank Sum Test; $p < .001$) and remained suppressed two hours after the injection ($ps < .001$). Firing rate was significantly suppressed for six out of fourteen neurons recorded ($ps < .01$). Three of the recording tetrodes were within the DG and the fourth tetrode was within CA1; firing rate decreased in neurons from both subregions. Conversely, the firing rate of all neurons containing the control virus (in total, sixteen recorded neurons) remained at baseline throughout the recording period and was not affected by the CNO injection (see Fig. 11C for two representative examples; all $ps > .05$). This confirms that the inhibitory DREADD suppressed neural activity in the HPC.

Discussion

Early theories of category learning posit that the hippocampus serves a relatively minor role in categorization (e.g., Ashby et al., 1998). The current experiment adds to a growing literature that challenges this view (e.g., Mack et al., 2018; Mack et al., 2016; Kim et al., 2018). We found that selective inactivation of the HPC using inhibitory DREADDs impaired category learning and category generalization in rats learning both 1D tasks and 2D tasks. Rats with hippocampal inactivation had lower accuracy (Fig. 4) and longer reaction time (Fig. 5). Inactivating the HPC did not affect performance on a control discrimination task (Fig. 7), suggesting that the impairments were not likely caused by unrelated processes (e.g., perception, motivation, or motion).

EpCon posits that the hippocampus is central to category learning by building and maintaining category representations (Mack et al., 2018). We tested this prediction using SUSTAIN by designing and testing multiple models that simulated the effect of the inhibitory

DREADDs (Fig. 8). Effectively, these model variants injected noise into specific model components. This allowed us to directly test the impact of specific functions while preserving the rest of the network. This manipulation was thought to be functionally similar to the effect of the inhibitory DREADDs.

The models that assumed the HPC functioned as part of SUSTAIN's clustering mechanism (i.e., Models 1-3) produced better fits of the data than the other models (i.e., Models 4&5; Selective Attention and the Control Model; Fig. 8). This suggests that the function of the HPC may be related to SUSTAIN's clustering mechanism; however, other categorization models would need to be systematically fit to the data to rule out alternative functions of the HPC (e.g., Gluck & Myers, 1993; Kumaran & McClelland, 2012). From the models tested, Model 1 (Pattern Completion) produced the best fit of the data. This model simulated the inhibitory DREADDs by adding noise to the calculation that compared the current stimulus to existing category representations. In SUSTAIN, representations are retrieved according to their similarity to the stimulus being categorized. Without the HPC, category decisions were based on representations that were dissimilar to the current stimulus. This model variant recruited about twice the number of clusters as the controls, suggesting that the cluster representations were not used appropriately or efficiently.

This deficit may be related to the observed impairments in the perceptual recency effects (Fig. 9). For controls, behavior was influenced by the similarity between the current stimulus and the previous training trial. Accuracy was larger if this similarity was high and smaller if this similarity was low. For the inactivation groups, these effects were reduced, suggesting that the inactivation groups had difficulty comparing the similarity between the current stimulus and the previous stimulus. This impairment is similar to the deficit described by the best-fitting

SUSTAIN model variant. In both cases, the inactivation groups had an impairment in comparing stimuli to previous training experiences.

Together, we hypothesize that HPC inactivation caused impairments in the mechanisms of pattern completion, which use auto-association to retrieve similar memory traces (Horner, et al., 2015; Gold & Kesner, 2005; Guzman, Schlögl, Frotscher, & Jonas, 2016). A recent experiment using patients with amnesia supports this interpretation (Cutler, Duff, & Polyn, 2019). Participants were asked to generate relevant features of common concepts (e.g., 'berry'). Compared to the healthy comparisons, recall in amnesic patients was restricted to features close to the target concept in semantic space. Pattern completion mechanisms within the hippocampus may be critical for extrapolating features that are associated to each category (Solomon & Schapiro, 2020).

An important next step in this research is to directly examine how categories are represented in the hippocampus. Conventionally, the hippocampus has been associated with representing memories of single events, akin to exemplar theory (Nosofsky, 1986; Gluck & Myers, 1993). Growing evidence suggests that the hippocampus may use recurrent connections to also support more prototype-based representations (Rosch & Mervis, 1975; Kumaran & McClelland, 2012; Bowman & Zeithamova, 2018). Most likely, both exemplar and prototype representations are available to the brain (Bowman, Iwashita, & Zeithamova, 2020) and can be used differentially according to the task demands.

Clustering models like SUSTAIN are somewhat of a hybrid between these representational schemes. After every trial, SUSTAIN decides whether a new cluster is recruited (i.e., exemplar-based representations) or the stimulus is integrated into an existing cluster (i.e., prototype-based representations). At its extremes, SUSTAIN behaves like an exemplar model if

new clusters are always recruited and a prototype model if new stimuli are always integrated. An alternative possibility is that both exemplar and prototype representations are maintained within the hippocampus in parallel. For example, the C-HORSE model posits that a trisynaptic pathway involving the dentate gyrus creates exemplar representations, whereas a monosynaptic pathway to CA1 creates prototype representations (Sucevic & Schapiro, 2023). The presence of multiple representation types can be tested using projection-specific inactivation techniques.

We can better understand the nature of representations in the hippocampus by recording neural activity from the hippocampus during category learning. We predict that single neurons would show increased firing rate for stimuli sampled from specific portions of the stimulus space (i.e., cluster-like selectivity; Aronov, Nevers, & Tank, 2017). For animals learning the 1D tasks, this selectivity would likely generalize over an entire category (i.e., prototype-based). For animals learning the 2D tasks, the selectivity of each neuron would only cover a small portion of a category (i.e., exemplar-based). By recording from multiple hippocampal subfields, we could test whether multiple representational schemes are supported by the hippocampal simultaneously. Specifically, we would expect to see more exemplar-based representations in the dentate gyrus, and more prototype-based representations in the CA1.

Another line of research can examine how the HPC interacts with other brain regions during category learning. Multiple studies have implicated the ventromedial prefrontal cortex (vmPFC) in category learning (Zeithamova, Dominick, & Preston, 2012; Mack et al., 2020; Kumaran, Summerfield, Hassabis, & Maguire, 2009; Bowman & Zeithamova, 2018). A general prediction is that the prefrontal cortex biases representations in the hippocampus according to current goals (Mack et al., 2020; Love & Gureckis, 2007). This is supported by Broschard et al., 2021, which concluded that the rodent prelimbic prefrontal cortex (PL) maintains attention to

relevant stimulus information and decides when to create new representations. Together, we suspect that rodent category learning involves a close interaction between the HPC and the PL. Future experiments can examine this interaction by neural recordings and inactivation approaches. For example, we predict that inactivating the PL would result in decreased selectivity in the HPC.

In the current experiment, the inactivation groups had longer reaction times compared to the controls (Fig. 5). Generally, the HPC inactivation produced larger RT impairments for rats learning the 1D tasks compared to rats learning the 2D tasks. For example, differences in Choice RT were present throughout training for the rats learning the 1D tasks but emerged towards the end of training for the rats learning the 2D tasks. Additionally, rats learning the 1D tasks, but not the 2D tasks, had significantly longer Cue RT than controls. Learning the 1D tasks requires more top-down signaling from regions like the prelimbic prefrontal cortex (Broschard et al., 2021). We speculate that greater RT differences in the 1D tasks may have resulted from weakened connectivity with these upstream regions. However, this interpretation should be taken cautiously, considering there were no accuracy differences between the task types.

Finally, it is important to note that rats without the HPC were still able to learn both task types, although at a slower rate compared to controls. There are many possible explanations for this. First, the current task design may not have targeted another potentially key function of the HPC: pattern separation (Leutgeb, Leutgeb, Moser, & Moser, 2007; Bakker, Kirwan, Miller, & Stark, 2008; Yassa & Stark, 2011). We expect to see larger impairments if the category tasks put more strain on pattern separation mechanisms (e.g., adding additional irrelevant stimulus dimensions or moving the means of the training distributions closer together). Second, inhibitory DREADDs only infects about half of the target cells (Roth, 2016). Therefore, it is possible that

some functions of the HPC may have been at least partially intact. Third, other neural systems (e.g., the dorsal striatum; Ashby et al., 1998) may have compensated in the absence of the HPC. Fourth, the ventral HPC may also be important for category learning (Moser & Moser, 1998; Fanselow & Dong, 2010).

To conclude, the current experiment supports the hypothesis that the HPC serves a critical role in category learning. Inactivation of the HPC through inhibitory DREADDs impaired learning for both 1D and 2D categorization tasks. Simulation results from SUSTAIN suggest that the HPC may use pattern completion mechanisms to retrieve relevant category representations. These representations are used to make category decisions. Future experiments will investigate these representations at a more mechanistic level and examine how they interact with other brain regions. We are optimistic that the current paradigm offers exciting innovations that will allow for a thorough understanding of mechanisms that underlie complex behavior.

References

Aronov, D., Nevers, R., & Tank, D. W. (2017). Mapping of a non-spatial dimension by the hippocampal–entorhinal circuit. *Nature*, *543*(7647), 719–722. https://doi.org/10.1038/nature21692

Ashby, F. G., Alfonso-Reese, L. A., Turken, A. U., & Waldron, E. M. (1998). A neuropsychological theory of multiple systems in category learning. *Psychological Review*, *105*(3), 442–481. https://doi.org/10.1037//0033-295X.105.3.442

Ashby, F. G., Smith, J. D., & Rosedahl, L. A. (2020). Dissociations between rule-based and information-integration categorization are not caused by differences in task difficulty. *Memory & Cognition*, *48*(4), 541–552. https://doi.org/10.3758/s13421-019-00988-4

Backus, A. R., Schoffelen, J., Szebényi, S., Hanslmayr, S., & Doeller, C. F. (2016). Hippocampal-Prefrontal Theta Oscillations Support Memory Integration. *Current Biology*, *26*(4), 450-457. doi:10.1016/j.cub.2015.12.048

Bakker, A., Kirwan, C. B., Miller, M., & Stark, C. E. (2008). Pattern Separation in the Human Hippocampal CA3 and Dentate Gyrus. *Science*, *319*(5870), 1640-1642. doi:10.1126/science.1152882

Baraduc, P., Duhamel, J. R., & Wirth, S. (2019). Schema cells in the macaque hippocampus. *Science*, *363*(6427), 635–639. https://doi.org/10.1126/science.aav5404

Behrens, T. J., Muller, T. H., Whittington, J. R., Mark, S., Baram, A. B., Stachenfeld, K. L., & Kurth-Nelson, Z. (2018). What Is a Cognitive Map? Organizing Knowledge for Flexible Behavior. *Neuron*, *100*(2), 490–509. https://doi.org/10.1016/j.neuron.2018.10.002

Bowman, C. R., Iwashita, T., & Zeithamova, D. (2020). Tracking prototype and exemplar representations in the brain across learning. *eLife, 9,* e59360. https://doi.org/10.7554/eLife.59360

Bowman, C.R., Zeithamova, D. (2018). Abstract memory representations in the ventromedial prefrontal cortex and hippocampus support concept generalization. *Journal of Neuroscience, 38*(10), 2605-2614.

Broschard, M. B., Kim, J., Love, B. C., & Freeman, J. H. (2020). Category learning in rodents using touchscreen-based tasks. *Genes, Brain and Behavior*. doi:10.1111/gbb.12665

Broschard, M. B., Kim, J., Love, B. C., Wasserman, E. A., & Freeman, J. H. (2019). Selective attention in rat visual category learning. *Learning & Memory, 26*(3), 84-92. doi:10.1101/lm.048942.118

Broschard, M. B., Kim, J., Love, B. C., Wasserman, E. A., & Freeman, J. H. (2021). Prelimbic cortex maintains attention to category-relevant information and flexibly updates category representations. *Neurobiology of Learning and Memory, 185*, 107524. doi:10.1016/j.nlm.2021.107524

Constantinescu, A. O., OReilly, J. X., & Behrens, T. E. (2016). Organizing conceptual knowledge in humans with a gridlike code. *Science*, *352*(6292), 1464–1468. https://doi.org/10.1126/science.aaf0941

Crawley MJ. 2007. The R book. John Wiley & Sons Ltd., Chichester.

Crijns, E., & Op de Beeck, H. (2019). The visual acuity of Rats in Touchscreen Setups. *Vision*, *4*(1). doi: 10.3390/vision4010004.

Cutler, R. A., Duff, M. C., & Polyn, S. M. (2019). Searching for Semantic Knowledge: A Vector Space Semantic Analysis of the Feature Generation Task. *Frontiers in Human Neuroscience, 13*. https://doi.org/10.3389/fnhum.2019.00341

Dusek J. A., & Eichenbaum H. (1997). The hippocampus and memory for orderly stimulus relations. *Proc Natl Acad Sci, 94*, 7109 – 14.

Eichenbaum, H. 2001. The hippocampus and declarative memory: cognitive mechanisms and neural codes. *Behavioural Brain Research, 127*(1), 199-207. https://doi.org/10.1016/S0166-4328(01)00365-5.

Eichenbaum, H., & Cohen, N. J. (2014). Can we reconcile the declarative memory and spatial navigation views on hippocampal function?. *Neuron, 83*(4), 764–770. https://doi.org/10.1016/j.neuron.2014.07.032

Fanselow, M. S., & Dong, H. W. (2010). Are the dorsal and ventral hippocampus functionally distinct structures?. *Neuron, 65*(1), 7–19. https://doi.org/10.1016/j.neuron.2009.11.031

Filoteo, J. V., Maddox, W. T., & Davis, J. D. (2001). Quantitative modeling of category learning in amnesic patients. *Journal of the International Neuropsychological Society, 7*(1), 1-19. doi:10.1017/s1355617701711010

Gluck, M. A., & Myers, C. E. (1993). Hippocampal mediation of stimulus representation: a computational theory. *Hippocampus*, *3*(4), 491–516. https://doi.org/10.1002/hipo.450030410

Gold, A. E., & Kesner, R. P. (2005). The role of the CA3 subregion of the dorsal hippocampus in spatial pattern completion in the rat. *Hippocampus*, *15*(6), 808-814.

Guzman, S. J., Schlögl, A., Frotscher, M., & Jonas, P. (2016). Synaptic mechanisms of pattern completion in the hippocampal CA3 network. *Science*, *353*(6304), 1117-1123.

Guo, D., Chen, G. & Yang, J. (2023). Effects of schema on the relationship between post-encoding brain connectivity and subsequent durable memory. *Scientific Reports, 13*, 8736. https://doi.org/10.1038/s41598-023-34822-4

Hampson, R. E., Pons, T. P., Stanford, T. R., & Deadwyler, S. A. (2004). Categorization in the monkey hippocampus: A possible mechanism for encoding information into memory. *Proceedings of the National Academy of Sciences, 101*(9), 3184-3189. doi:10.1073/pnas.0400162101

Haslam, C. (1997). Preserved category learning in amnesia. *Neurocase, 3*(5). doi:10.1093/neucas/3.5.337-a

Horner, A., Bisby, J., Bush, D. *et al.* (2015). Evidence for holistic episodic recollection via hippocampal pattern completion. *Nature Communications, 6*, 7462. https://doi.org/10.1038/ncomms8462

Hunsaker, M. R., & Kesner, R. P. (2013). The operation of pattern separation and pattern completion processes associated with different attributes or domains of memory. *Neuroscience and biobehavioral reviews*, *37*(1), 36–58. https://doi.org/10.1016/j.neubiorev.2012.09.014

Jones, M., Love, B.C., & Maddox, W.T. (2006). Recency effects as a window to generalization: separating decisional and perceptual sequential effects in category learning. *Journal of Experimental Psychology. Learning, Memory, and Cognition, 32*(2), 316-32.

Kim, J., Castro, L., Wasserman, E. A., & Freeman, J. H. (2018). Dorsal hippocampus is necessary for visual categorization in rats. *Hippocampus, 28*(6), 392-405. doi:10.1002/hipo.22839

Kirwan, C. B., Hartshorn, A., Stark, S. M., Goodrich-Hunsaker, N. J., Hopkins, R. O., & Stark, C. E. (2012). Pattern separation deficits following damage to the hippocampus. *Neuropsychologia, 50*(10), 2408-2414. doi:10.1016/j.neuropsychologia.2012.06.011

Knowlton, B. J., Mangels, J. A., & Squire, L. R. (1996). A Neostriatal Habit Learning System in Humans. *Science, 273*(5280), 1399-1402. doi:10.1126/science.273.5280.1399

Knowlton, B., & Squire, L. (1993). The learning of categories: Parallel brain systems for item memory and category knowledge. *Science, 262*(5140), 1747-1749. doi:10.1126/science.8259522

Kraskov, A., Quiroga, R. Q., Reddy, L., Fried, I., & Koch, C. (2007). Local Field Potentials and Spikes in the Human Medial Temporal Lobe are Selective to Image Category. *Journal of Cognitive Neuroscience, 19*(3), 479-492. doi:10.1162/jocn.2007.19.3.479

Kreiman, G., Koch, C., & Fried, I. (2000). Category-specific visual responses of single neurons in the human medial temporal lobe. *Nature Neuroscience, 3*(9), 946-953. doi:10.1038/78868

Kumaran, D., & McClelland, J. L. (2012). Generalization through the recurrent interaction of episodic memories: a model of the hippocampal system. *Psychological review, 119*(3), 573–616. https://doi.org/10.1037/a0028681

Kumaran, D., Summerfield, J. J., Hassabis, D., & Maguire, E. A. (2009). Tracking the Emergence of Conceptual Knowledge during Human Decision Making. *Neuron, 63*(6), 889–901. doi: 10.1016/j.neuron.2009.07.030

Leutgeb, J. K., Leutgeb, S., Moser, M., & Moser, E. I. (2007). Pattern Separation in the Dentate Gyrus and CA3 of the Hippocampus. *Science, 315*(5814), 961-966. doi:10.1126/science.1135801

Love, B. C., & Gureckis, T. M. (2007). Models in search of a brain. *Cognitive, Affective, & Behavioral Neuroscience, 7*(2), 90-108. doi:10.3758/cabn.7.2.90

Love, B. C., Medin, D. L., & Gureckis, T. M. (2004). Sustain: A network model of category learning. *Psychological Review, 111*(2), 309-332. doi:10.1037/0033-295x.111.2.309

Mack, M. L., Love, B. C., & Preston, A. R. (2016). Dynamic updating of hippocampal object representations reflects new conceptual knowledge. *The Proceedings of the National Academy of Sciences, 113*(46). doi:10.1101/071118

Mack, M. L., Love, B. C., & Preston, A. R. (2018). Building concepts one episode at a time: The hippocampus and concept formation. *Neuroscience Letters, 680*, 31-38. doi:10.1016/j.neulet.2017.07.061

Mack, M. L., Preston, A. R., & Love, B. C. (2020). Ventromedial prefrontal cortex compression during concept learning. *Nature Communications, 11*(1), 1-11. doi: 10.1038/s41467-019-13930-8

Marr, D. (1969), A theory of cerebellar cortex. *The Journal of Physiology, 202.* doi:10.1113/jphysiol.1969.sp008820.

McNaughton, B. L., & Morris, R. G. (1987). Hippocampal synaptic enhancement and information storage within a distributed memory system. *Trends in Neurosciences, 10*(10), 408-415. https://doi.org/10.1016/0166-2236(87)90011-7.

Mok, R. M., & Love, B. C. (2019). A non-spatial account of place and grid cells based on clustering models of concept learning. *Nature Communications, 10*(1). https://doi.org/10.1038/s41467-019-13760-8

824 Morton, N. W., Sherrill, K. R., & Preston, A. R. (2017). Memory integration constructs maps of
825 space, time, and concepts. *Current Opinion in Behavioral Sciences, 17*, 161–168. doi:
826 10.1016/j.cobeha.2017.08.007
827 Moser, M. B., & Moser, E. I. (1998). Functional differentiation in the hippocampus.
828 *Hippocampus, 8*. 608-619. https://doi.org/10.1002/(SICI)1098-1063(1998)8:6<608::AID-
829 HIPO3>3.0.CO;2-7
830 Neath, A. A., & Cavanaugh, J. E. (2011). The Bayesian information Criterion: Background,
831 derivation, and applications. *Wiley Interdisciplinary Reviews: Computational Statistics,*
832 *4*(2), 199-203. doi:10.1002/wics.199
833 Nosofsky, R. M. (1986). Attention, similarity, and the identification–categorization relationship.
834 *Journal of Experimental Psychology: General, 115*(1), 39–57.
835 https://doi.org/10.1037/0096-3445.115.1.39
836 O'Donoghue, E. M., Broschard, M. B., Wasserman, E. A. (2020). Pigeons exhibit flexibility but
837 not rule formation in dimensional learning, stimulus generalization, and task switching.
838 *Journal of Experimental Psychology Animal Learning & Cognition, 46*(2), 107-23. doi:
839 10.1037/xan0000234.
840 O'Reilly, R. C., & McClelland, J. L. (1994). Hippocampal conjunctive encoding, storage, and
841 recall: avoiding a trade-off. *Hippocampus*, *4*(6), 661–682.
842 https://doi.org/10.1002/hipo.450040605
843 Pajkert, A., Finke, C., Shing, Y. L., Hoffmann, M., Sommer, W., Heekeren, H. R., & Ploner, C.
844 J. (2017). Memory integration in humans with hippocampal lesions. *Hippocampus,*
845 *27*(12), 1230-1238. doi:10.1002/hipo.22766

Paxinos, G. and Watson, C. (1998) The Rat Brain in Stereotaxic Coordinates. Academic Press, San Diego.

Rosch, E., & Mervis, C. B. (1975). Family resemblances: Studies in the internal structure of categories. *Cognitive Psychology, 7*(4). 573-605. https://doi.org/10.1016/0010-0285(75)90024-9.

Roth, B. L. (2016). Dreadds for neuroscientists. *Neuron, 89*(4), 683-694. doi:10.1016/j.neuron.2016.01.040

Schapiro, A. C., Turk-Browne, N. B., Botvinick, M. M., & Norman, K. A. (2017). Complementary learning systems within the hippocampus: a neural network modelling approach to reconciling episodic memory with statistical learning. *Philosophical Transactions of the Royal Society B: Biological Sciences, 372*(1711), 20160049. https://doi.org/10.1098/rstb.2016.0049

Schlichting, M. L., & Preston, A. R. (2015). Memory integration: Neural mechanisms and implications for behavior. *Current Opinion in Behavioral Sciences, 1*, 1-8. doi:10.1016/j.cobeha.2014.07.005

Solomon, E. A., Lega, B. C., Sperling, M. R., & Kahana, M. J. (2019). Hippocampal theta codes for distances in semantic and temporal spaces. *Proceedings of the National Academy of Sciences*, *116*(48), 24343–24352. https://doi.org/10.1073/pnas.1906729116

Solomon, S. H., & Schapiro, A. C. (2020). Semantic Search as Pattern Completion across a Concept. *Trends in Cognitive Sciences, 24*(2), 95–98. https://doi.org/10.1016/j.tics.2019.12.003

Sučević, J., & Schapiro, A. C. (2023). A neural network model of hippocampal contributions to category learning. *eLife, 12*, e77185. https://doi.org/10.7554/eLife.77185

Theves, S., Fernandez, G., & Doeller, C. F. (2019). The Hippocampus Encodes Distances in Multidimensional Feature Space. *Current Biology, 29*(7). https://doi.org/10.1016/j.cub.2019.02.035

Theves, S., Fernández, G., & Doeller, C. F. (2020). The Hippocampus Maps Concept Space, Not Feature Space. *The Journal of Neuroscience, 40*(38), 7318–7325. https://doi.org/10.1523/jneurosci.0494-20.2020

Tolman E. C. (1948). Cognitive maps in rats and men. *Psychological review, 55*(4), 189–208. https://doi.org/10.1037/h0061626

Tse, D., Langston, R. F., Kakeyama, M., Bethus, I., Spooner, P. A., Wood, E. R., Witter, M. P., & Morris, R. G. (2007). Schemas and memory consolidation. *Science, 316*(5821), 76–82. https://doi.org/10.1126/science.1135935

Wallis, J. D., Anderson, K. C., & Miller, E. K. (2001). Single neurons in prefrontal cortex encode abstract rules. *Nature, 411*(6840), 953–956. https://doi.org/10.1038/35082081

Yassa, M. A., & Stark, C. E. (2011). Pattern separation in the hippocampus. *Trends in neurosciences, 34*(10), 515–525. https://doi.org/10.1016/j.tins.2011.06.006

Zaki, S. R. (2004). Is categorization performance really intact in amnesia? A meta-analysis. *Psychonomic Bulletin & Review 11*(6), 1048-54.

Zeithamova, D., & Bowman, C. R. (2020). Generalization and the hippocampus: More than one story?. *Neurobiology of learning and memory, 175*, 107317. https://doi.org/10.1016/j.nlm.2020.107317

Zeithamova, D., Dominick, A. L., & Preston, A. R. (2012). Hippocampal and ventral medial prefrontal activation during retrieval-mediated learning supports novel inference. *Neuron, 75*(1), 168–179. https://doi.org/10.1016/j.neuron.2012.05.010

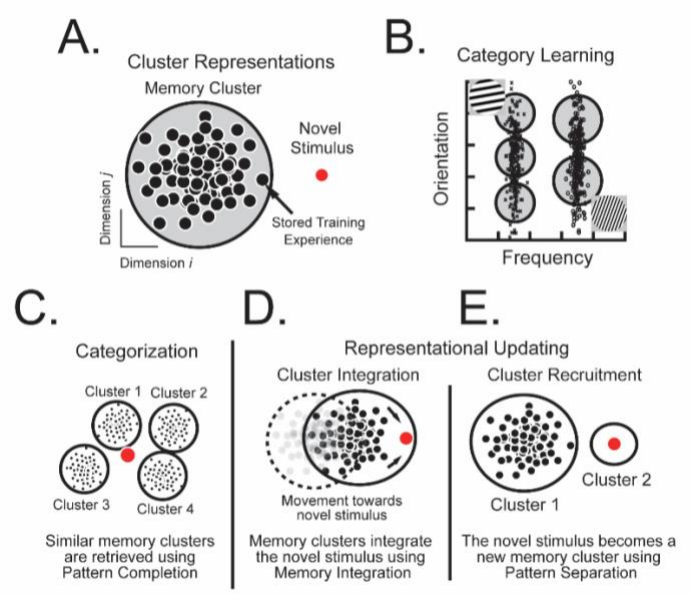

Figure 1. Clustering models such as SUSTAIN encompass fundamental mechanisms of the hippocampus. **A.** SUSTAIN assumes that perceptually similar training experiences tend to cluster together in memory. **B.** Categories are represented by one or multiple clusters, where each cluster surrounds a portion of stimuli within the stimulus space. **C.** Categorizing a new stimulus involves retrieving cluster representations that are similar to that stimulus. This is similar to pattern completion mechanisms that use auto-association to retrieve similar memory representations. **D-E.** After each trial, SUSTAIN updates the cluster representations through two mechanisms. **D.** First, the current stimulus can be integrated into an existing cluster, such that the center of that cluster moves towards the stimulus. This is similar to memory integration, where new memory traces are integrated into existing representations. **E.** Second, the current stimulus can become the center of a new cluster. This is similar to pattern separation, which maximizes the distance between training experiences to keep them separate.

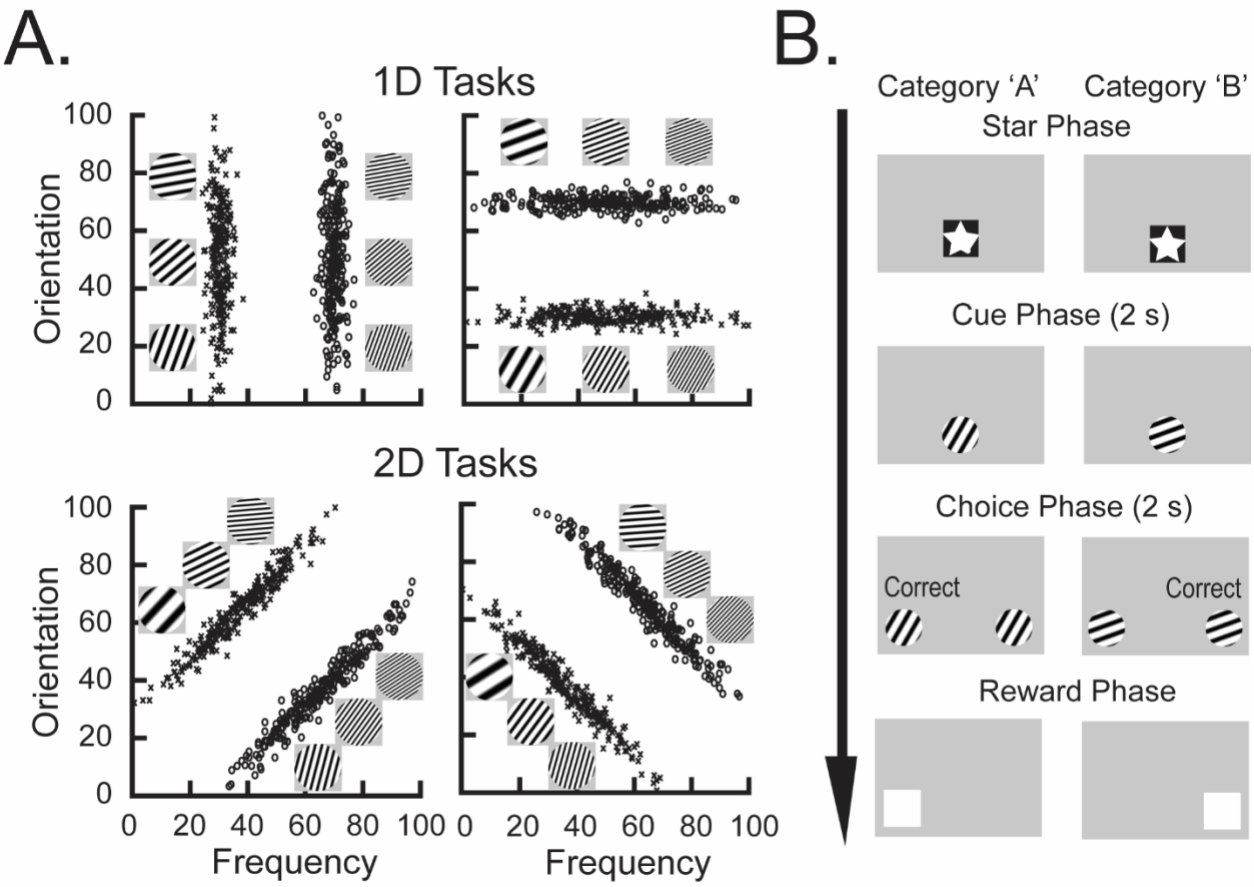

Figure 2. Category tasks and trial procedure. Rats were trained to categorize visual stimuli containing gratings that changed in spatial frequency and orientation. Categories were created using this two-dimensional space. **A.** 1D task had distributions that were perpendicular to one of the stimulus dimensions; learning a 1D task encouraged a shift in attention to the perpendicular dimension. 2D tasks had distributions that were not perpendicular to a stimulus dimension; learning a 2D task required attention to both stimulus dimensions. **B.** Trial procedure used for all training and testing sessions. First, rats touched a star to start the trial (Star phase). Next, an exemplar was presented at the center of the screen (Cue phase). After three touches of the exemplars, copies of the exemplar appeared on the left and right sides of the screen acting as response keys (Choice phase). Rats touched either response key depending on the category membership of the exemplar. After a correct response, a white box appeared. One touch of the white box delivered a food reward (Reward phase).

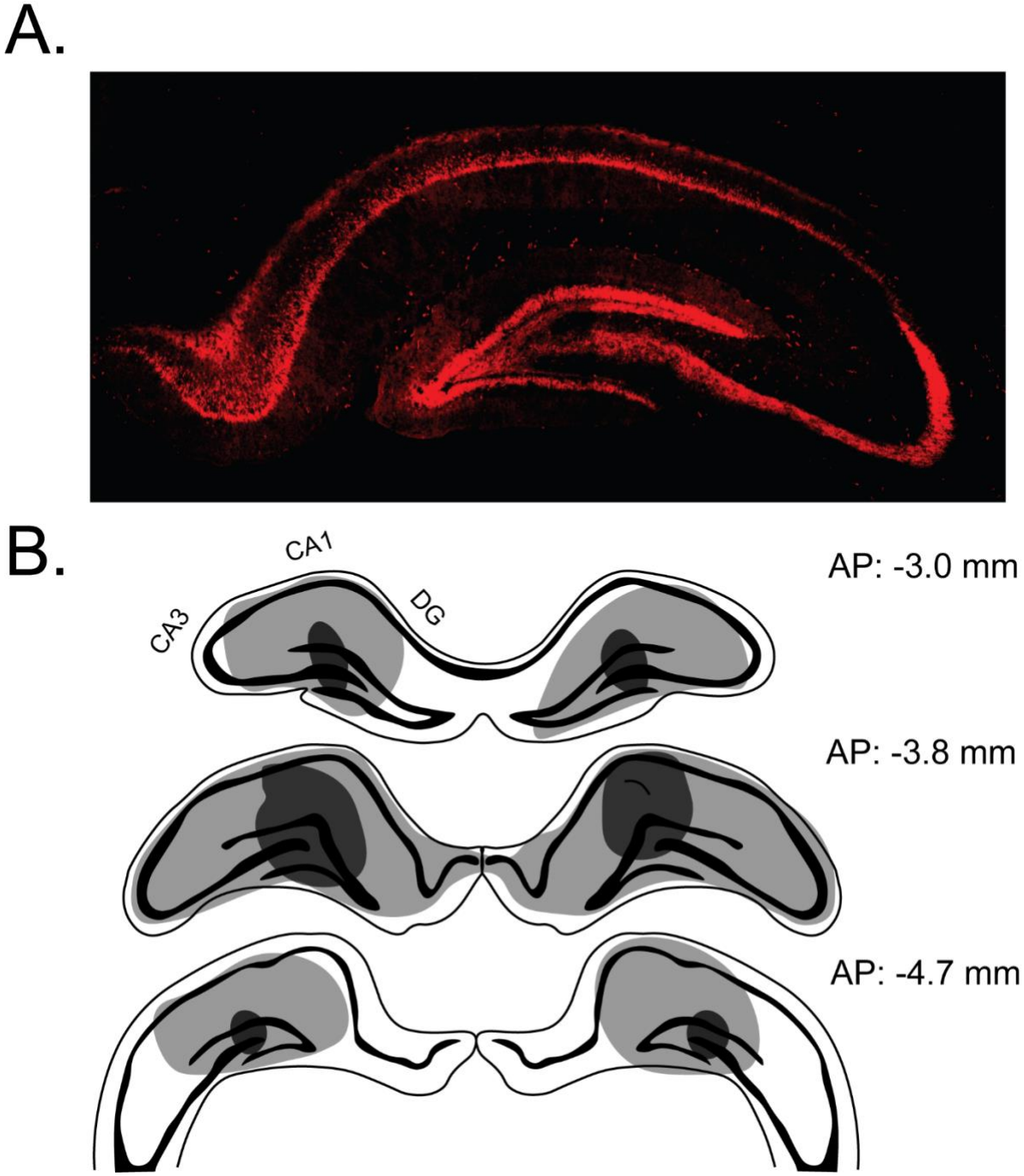

Figure 3: AAV expression in the HPC. Virus expression and location was observed for each rat to ensure adequate transduction. The boundary of the HPC was defined according to Paxinos & Watson, 1998. AAV did not extend into the ventral hippocampus and was contained within AP: -2.8 and – 4.9. **A.** Representative AAV expression in the HPC. **B.** Minimum (dark gray) and maximum (light gray) AAV expression within the HPC. For a small subset of rats, AAV did not extend into the CA3.

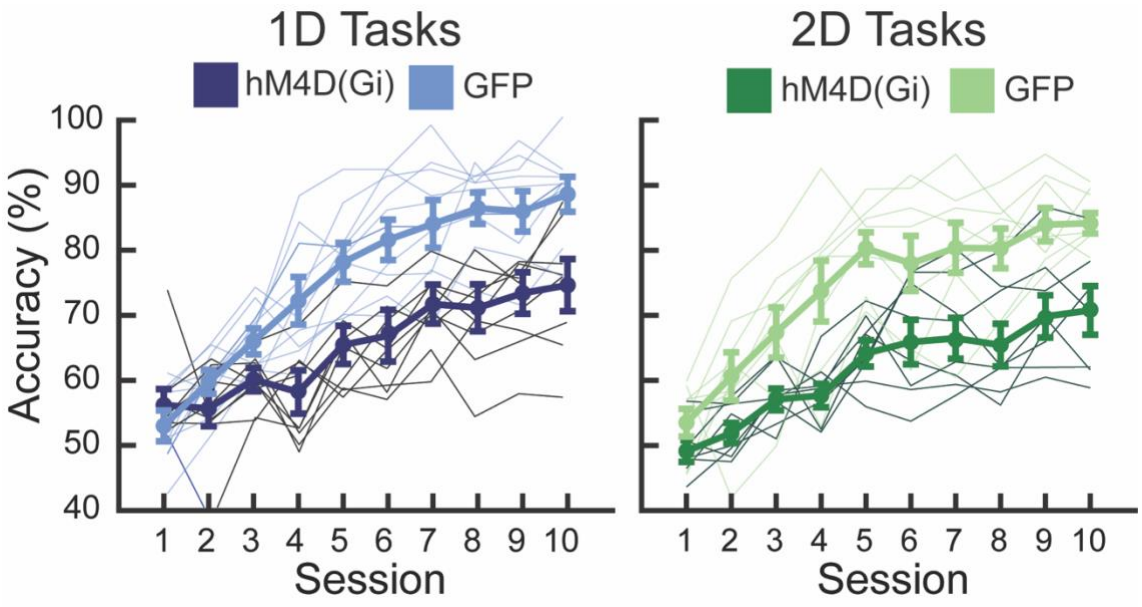

Figure 4. Category training. Rats were given ten training sessions to learn either a 1D task or a 2D task (n = 9 per group). Hippocampal inactivation impaired accuracy across training sessions for rats learning both the 1D tasks. All error bars indicate the *SEM*. hM4D(Gi) indicates the inhibitory DREADD, and GFP indicates the control virus. Background datapoints indicate individual learning curves.

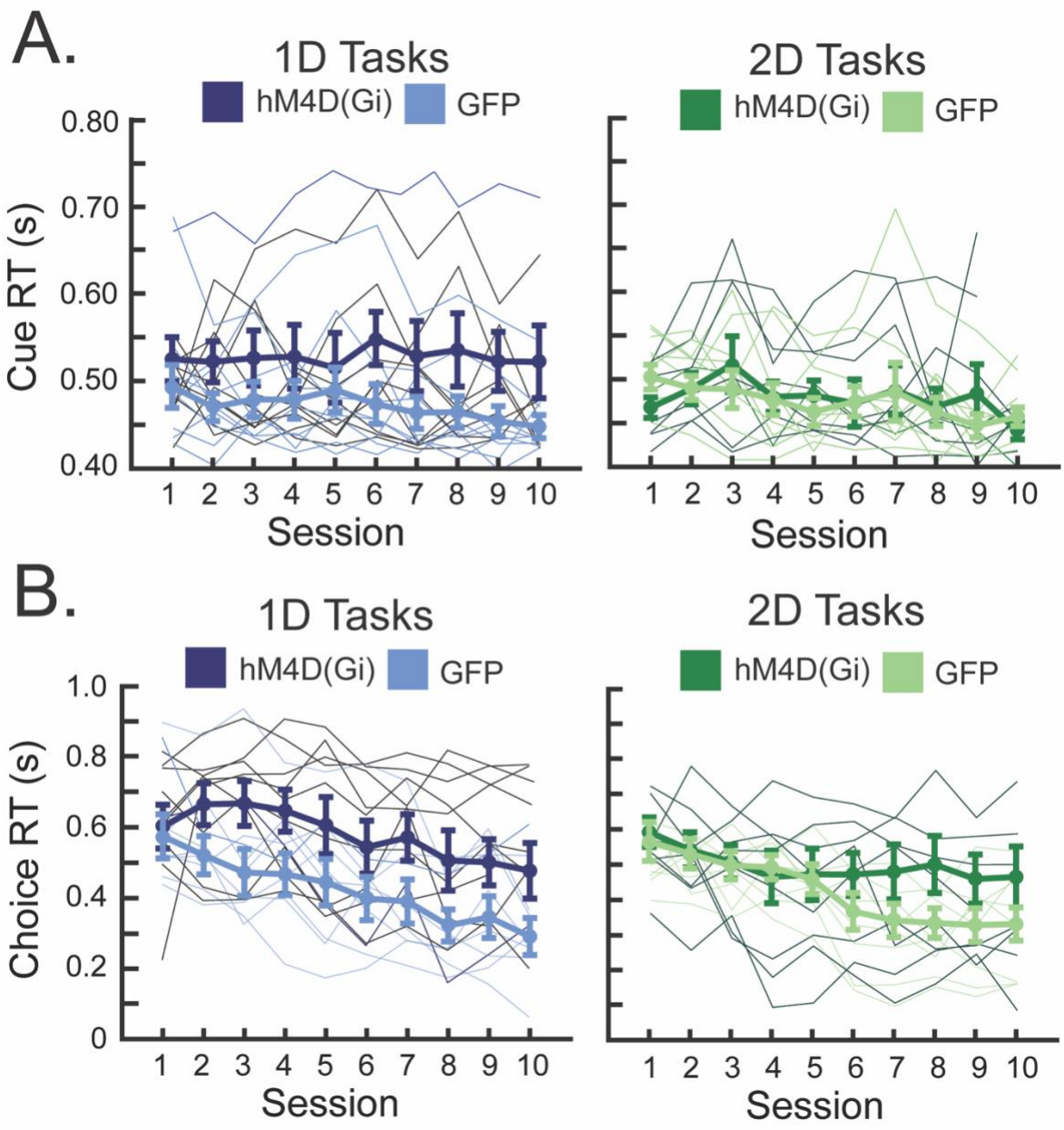

Figure 5. Reaction time during the Cue phase and Choice phase of training trials (Cue RT and Choice RT, respectively). **A.** Rats with hippocampal inactivation had increased Cue RT compared to controls when learning the 1D tasks, but not the 2D tasks. **B.** Rats with hippocampal inactivation had increased Choice RT compared to controls when learning both the 1D tasks and the 2D tasks. These results suggest that the HPC was critical for making category decisions. All error bars indicate the *SEM*. hM4D(Gi) indicates the inhibitory DREADD, and GFP indicates the control virus. Background datapoints indicate individual learning curves.

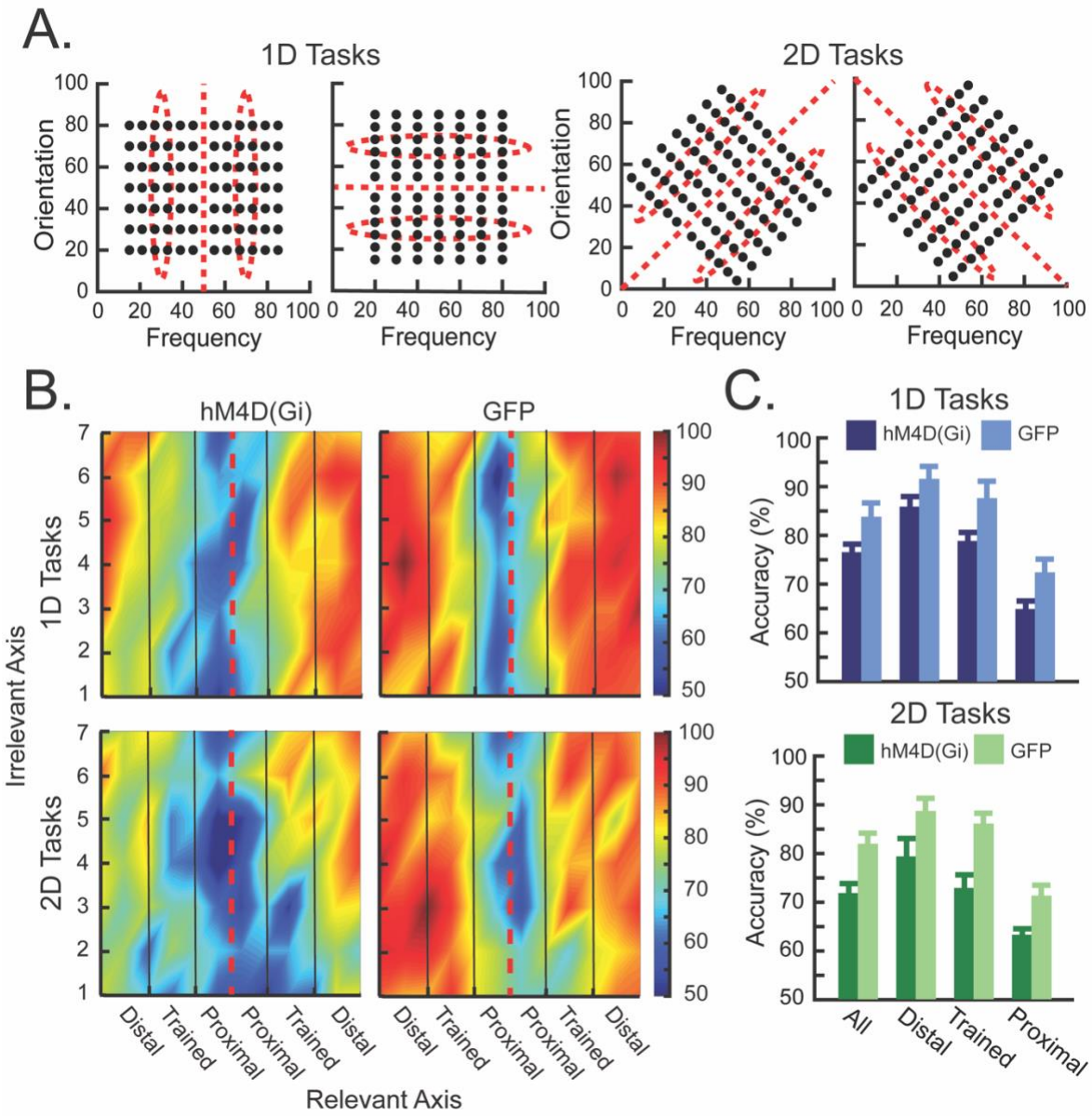

Figure 6. Category generalization. **A.** After training, rats were given five testing sessions to examine category generalization. Stimuli were separated into three trial types: stimuli that overlapped with the training distributions (i.e., Trained), novel stimuli that were farther from the category boundary relative to the training distributions (i.e., Distal), and novel stimuli that were close to the category boundary (i.e., Proximal). **B.** Heatmaps of the rats' performance were generated by averaging the accuracy for each testing stimulus within the grid. Each task was rotated in stimulus space so that the relevant axis was parallel to the x-axis, and the irrelevant axis was parallel to the y-axis. **C.** Average accuracy for each trial type. Compared to Training stimuli, accuracy was improved for Distal stimuli, and accuracy was impaired for Proximal stimuli. Hippocampal inactivation impaired accuracy for rats that had learned the 1D tasks and 2D tasks. There were no significant interactions between trial types. All error bars indicate the *SEM*. hM4D(Gi) indicates the inhibitory DREADD, and GFP indicates the control virus.

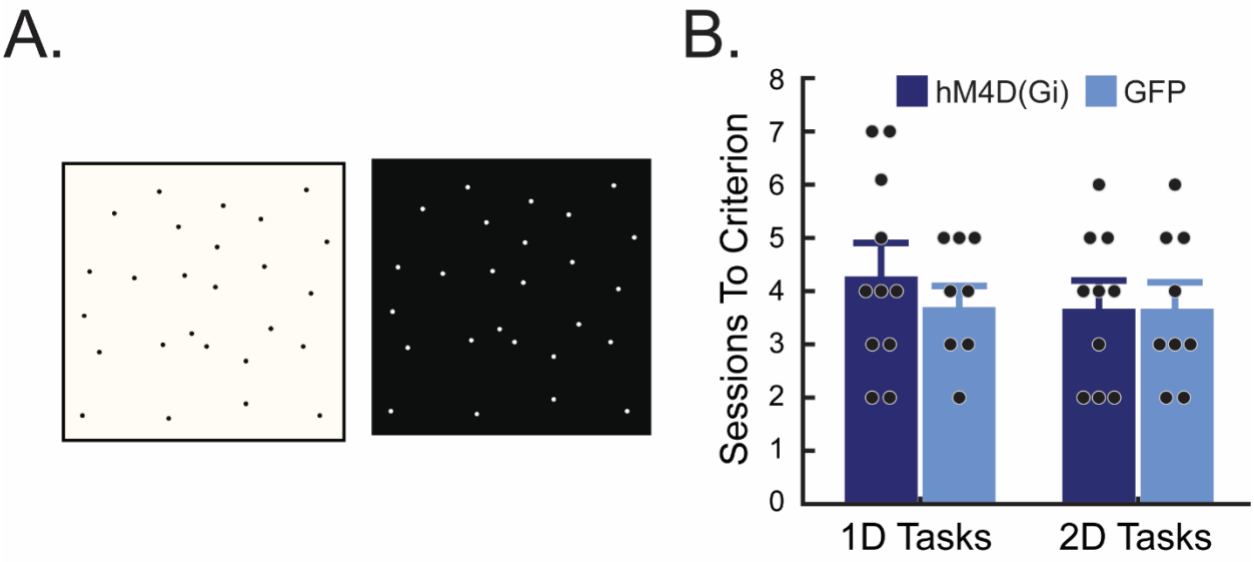

Figure 7. Simple discrimination control task. **A.** Rats were trained to discriminate a white box from a black box as a control experiment. Rats were given training sessions until a learning criterion was reached (i.e., at least 75% accuracy for both stimuli across two consecutive sessions). **B.** Hippocampal inactivation did not affect the number of sessions required to reach criterion. These results suggest that impairments were specific to categorization and were not caused by unrelated factors (e.g., perceptual, motivational, or motor deficits). All error bars indicate the *SEM*. hM4D(Gi) indicates the inhibitory DREADD, and GFP indicates the control virus. Scatterplots indicate individual subjects.

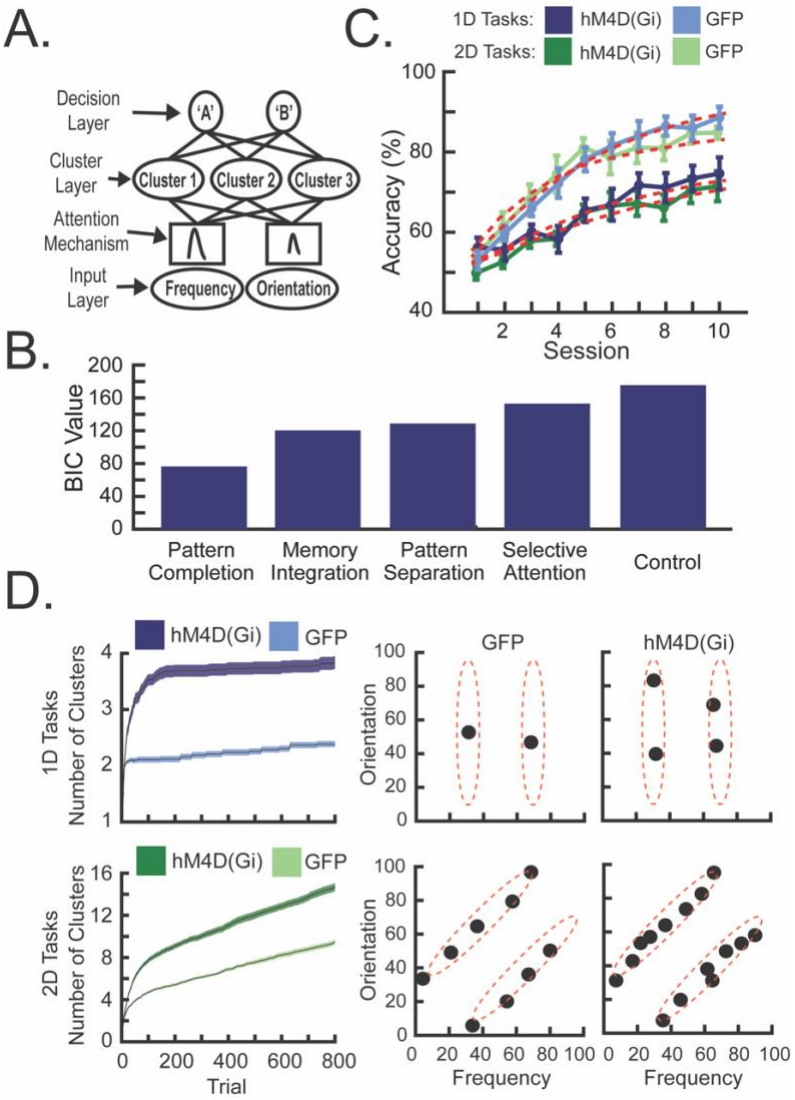

Figure 8. SUSTAIN modeling. **A.** A diagram of the computational model SUSTAIN. Briefly, SUSTAIN contains three layers, an input layer that loads the stimulus information, a cluster layer that stores category representations, and a decision layer that makes category decisions. **B.** Multiple models were designed to test the function of the HPC in category learning. Each model was fit to the averaged group data. The goodness-of-fit was determined for each model by calculating BIC, where low BIC values indicate a better fit of the data. All models that assumed the HPC served a function related to SUSTAIN's clustering mechanism (i.e., Pattern Completion, Memory Integration, and Pattern Separation) produced better fits of the data than the other models (i.e., Selective Attention and Control). The Pattern Completion model produced the best fit of the learning data. **C.** SUSTAIN's predictions for the best fitting model (Pattern Completion). All error bars indicate the *SEM*. hM4D(Gi) indicates the inhibitory DREADD, and GFP indicates the control virus. **D.** Left: The average number of unique clusters across training generated from the Pattern Completion model. The 1D tasks were typically solved by recruiting 1-2 clusters per category, whereas the 2D tasks were typically solved by recruiting 4-6 clusters per category. For both task types, rats with hippocampal inactivations recruited more clusters than the controls. Right: Example arrangement of cluster positions for rats learning the 1D tasks (top) and 2D tasks (bottom). The red ellipses indicate the position of the training distributions.

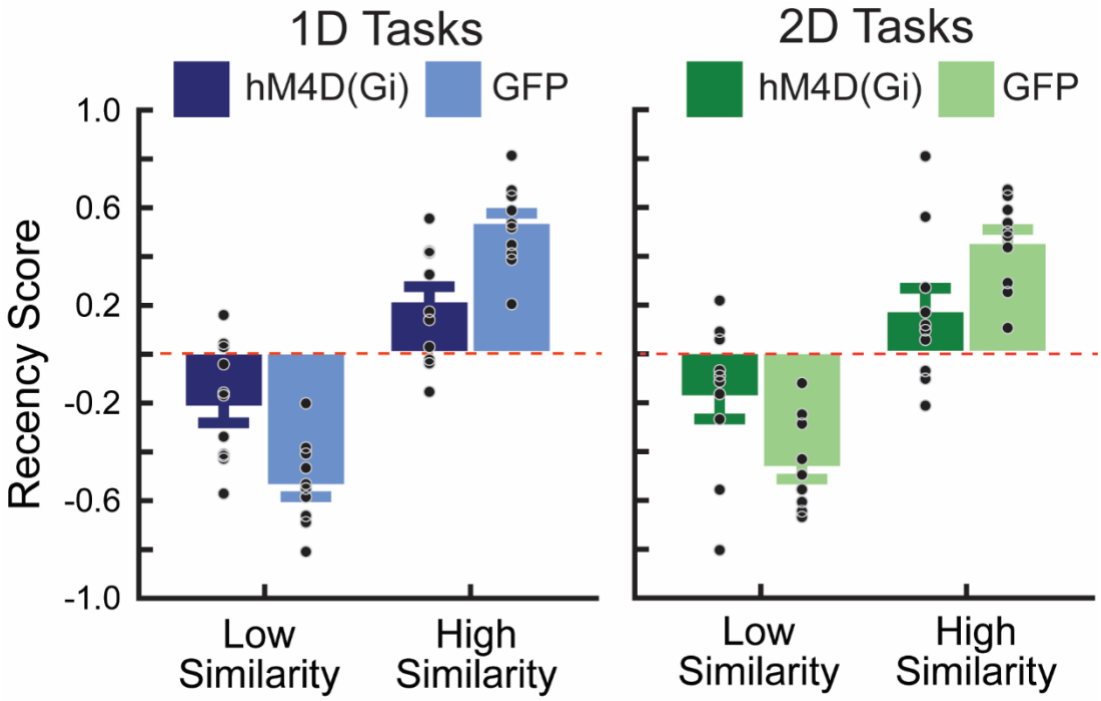

Figure 9. HPC is critical for updating category representations. Trial accuracy was analyzed on a trial-by-trial basis and was segregated according to the similarity between the current stimulus and the stimulus of the most recent trial. Positive scores indicate facilitated accuracy due to trial order, negative scores indicate impaired accuracy due to trial order, and 0 means no effect of trial order. For controls learning both the 1D tasks and the 2D tasks, accuracy was facilitated when the current trial was perceptually similar to the most recent trial, and accuracy was impaired when the current trial was perceptually dissimilar to the most recent trial. The effect of trial order was impaired for rats with hippocampal inactivation learning both the 1D tasks and the 2D tasks. These results suggest that the HPC is important for updating category representations and biasing decisions according to the most previous trial. All error bars indicate the *SEM*. hM4D(Gi) indicates the inhibitory DREADD, and GFP indicates the control virus. Scatterplots indicate individual subjects.

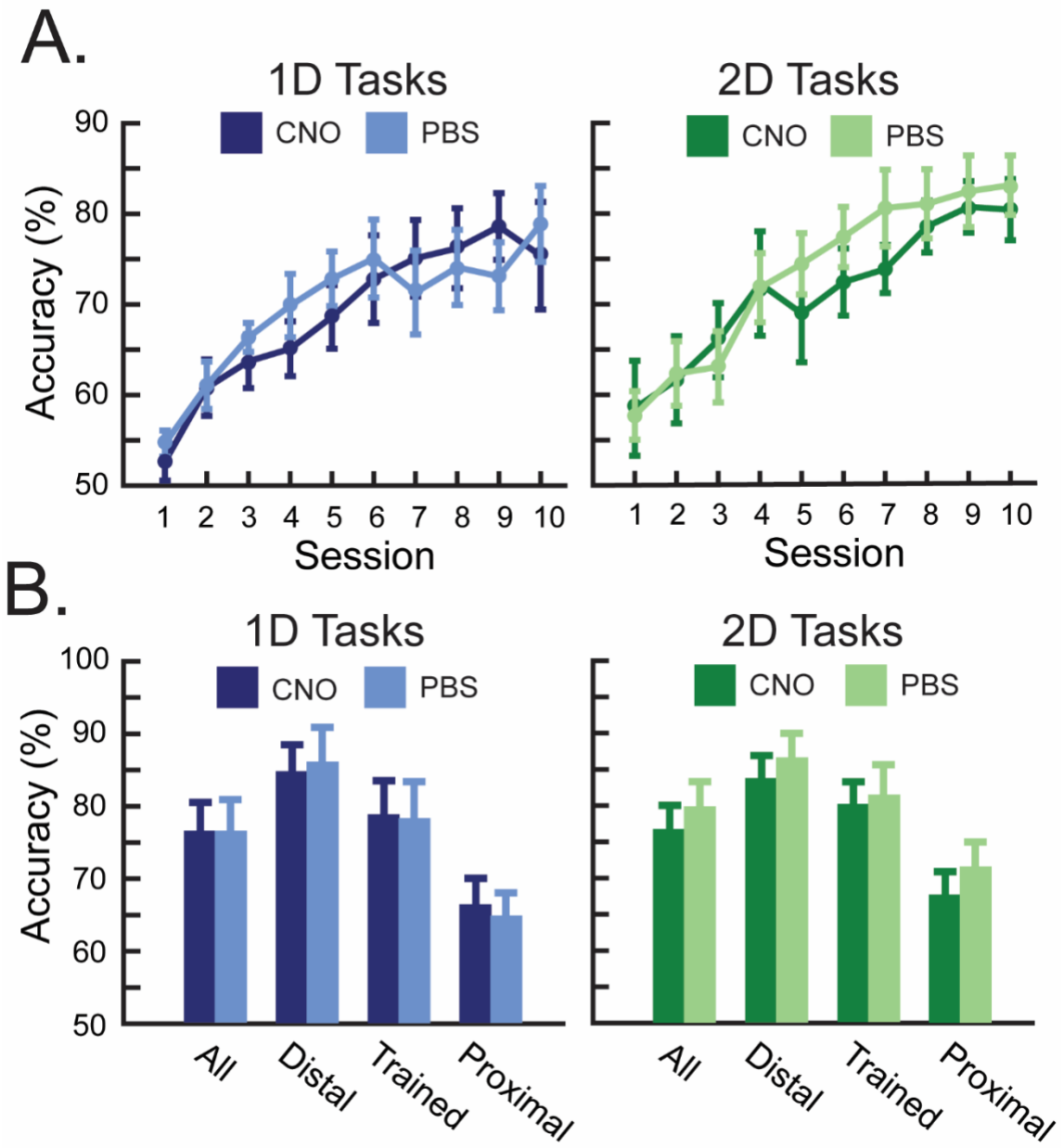

Figure 10. CNO control experiment. To ensure that IP injections of CNO did not affect categorization, rats were given training and testing sessions on either a 1D task or a 2D task. IP injections of either CNO or PBS were administered before each session. **A.** CNO injections did not affect category learning for rats learning the 1D tasks or the 2D tasks. **B.** CNO injections also did not affect category generalization for rats that learned the 1D tasks or the 2D tasks. Together, these results support that IP injections of CNO did not affect categorization. All error bars indicate the *SEM*.

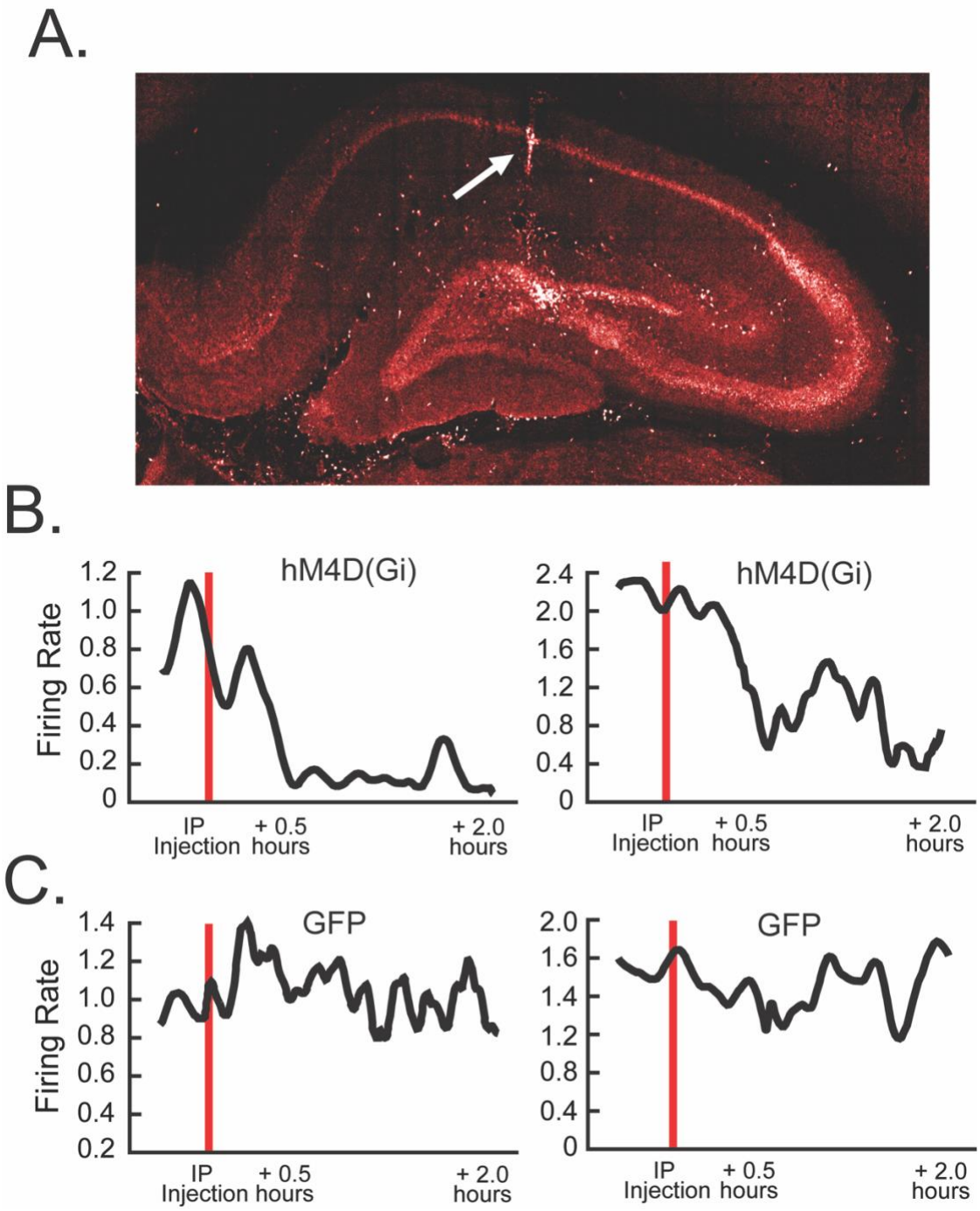

Figure 11. DREADDs verification. **A.** In a single rat, the inhibitory DREADD virus was infused into one hippocampal hemisphere and the control virus was infused into the other hippocampal hemisphere. A custom built microdive containing multiple recording tetrodes was implanted to record *in vivo* single units within the HPC. The white arrow indicates the placement of a recording tetrode. **B-C.** The firing rate of representative cells transduced by the inhibitory virus (**B**) or control virus (**C**) after the administration of an IP injection of CNO. **B.** The firing rate of neurons transduced with the inhibitory DREADD decreased below baseline 30 minutes after injection and remained suppressed 2 hours after injection. **C.** The firing of neurons transduced with the control AAV remained at baseline throughout the recording period.