Hippocampus as a generative circuit for predictive coding of future sequences

Yusi Chen*
Computational Neurobiology Laboratory
Salk Institute for Biological Sciences
La Jolla, CA
cyusi@ucsd.edu

Huanqiu Zhang
Computational Neurobiology Laboratory
Salk Institute for Biological Sciences
La Jolla, CA
huz015@ucsd.edu

Terrence J. Sejnowski
Computational Neurobiology Laboratory
Salk Institute for Biological Sciences
La Jolla, CA
terry@salk.edu

Abstract

The interaction between hippocampus and cortex is key to memory formation and representation learning. Based on anatomical wiring and transmission delays, we proposed a self-supervised recurrent neural network (PredRAE) with a predictive reconstruction loss to account for the cognitive functions of hippocampus. This framework extends predictive coding in the time axis and incorporates predictive features in Bayes filters for temporal prediction. In simulations, we were able to reproduce characteristic place cell features such as one-shot plasticity, localized spatial representation and replay, which marks the trace of memory formation. The simulated place cells also exhibited precise spike timing, evidenced by phase precession. Trained on MNIST sequences, PredRAE learned the underlying temporal dependencies and a **spontaneous** representation of digit labels and rotation dynamics in the linear transformation of its hidden unit activities. Such learning is robust against 16 different image corruptions. Inspired by the brain circuit, the simple and concise framework has great potential to approximate human performance with its capacity to robustly disentangle representations and generalize temporal dynamics.

1 Introduction

Hippocampus formation is deeply involved in important cognitive functions such as memory [1] and sensory representation [2]. Its anatomical and functional properties have been extensively studied [3] [4], fostering a wide range of theoretical speculations about its computation principles [5] [6] [7]. For example, Stachenfeld et al.[5] modeled the activities of place cells (representative hippocampus neurons) as the successor representation involved in reinforcement learning. Whittington et al. [6] proposed it as a statistical machine inferring structural properties of observations. Although their simulation results reproduced a handful of place cell statistics under various task settings, these speculations are hard to experimentally verify because they failed to map onto the anatomical structure, and thus couldn't make predictions directly related to neural response in sub-regions. In contrast, our framework, rooted in the anatomical structure and conductance delays in the circuit,

*Correspondence addressed to Y.C. and T.J.S

Preprint. Under review.

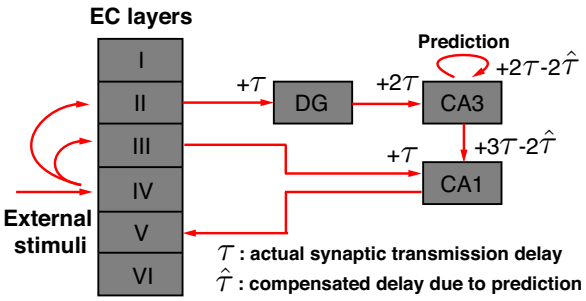

Figure 1: **Anatomical wiring and transmission delays.** Red arrows mark the signal transmission direction. External sensory stimuli start from different cortical layers in EC and reach CA1 through two pathways, forming a self-supervised structure. After passing one synapse (i.e. red arrow), the signal is delayed by τ. The accumulated transmission time difference ($+2\tau$) between two pathways was compensated by CA3 predicting into the future ($-2\hat{\tau}$).

models hippocampus as a generative machine. It predicts distinct statistical features of sub-regions in hippocampus.

Our approach was inspired by the theory of predictive mind [8][9]: the brain encodes a generative model of external observations and optimizes to decrease surprisal level (i.e. prediction errors). Such theory could be implemented through predictive coding, which has gained attention because of its popularity in modeling sensory responses such as visual and auditory responses [10][11]. Central to predictive coding is the presence of error encoding neurons, which enables Hebbian learning based model training and explains extra-classical neural responses [9]. On the other hand, our memory system, coordinated through hippocampus, also exhibits key components of predictive coding. First, hippocampus stores an internal representation, evidenced by replay [12] and preplay [13], where neurons are sequentially activated given no external stimuli. These sequential representations are usually abstractions of external stimuli such as spatial cues, auditory frequency, oder types and even time [14]. This implies another key function of hippocampus: extraction of latent variables from sensory inputs or disentanglement of representations. Secondly, there's evidence pointing hippocampus and its sub-region CA1 as a region detecting unexpected signals [15][16][17], arguing for the existence of neurons encoding for prediction errors. Several studies have also emphasized the distinct statistical properties in CA3 and CA1 encoding [18][19][20], which don't agree with the frameworks that model the function of all hippocampus neurons uniformly. Given the above evidence, we modeled hippocampus based on the anatomical wiring between sub-regions and formulated its computational principles under the framework of predictive coding with an additional temporal component.

In the sessions below, we formally stated our hypothesis and and implemented it as a recurrent auto-encoder with a novel predictive loss function (PredRAE). Then a set of results was displayed to answer two key questions: 1) Does the framework match neurological recordings of place cells under different task settings? 2) How does it implement cognitive functions of hippocampus described above: memory and representation learning? Lastly, we hope the proposed framework could provide inspirations for robust and interpretable artificial intelligence.

2 Hypothesis and PredRAE

Major connections linking cortex and hippocampus were summarized in Fig. 1. Entorhinal cortex (EC) is the major cortical region interacting with hippocampus, which contains of sub-region Dentate Gyrus (DG), CA1 and CA3. Among them, recurrently connected CA3 is ideal for storing internal representations like the Hopfield associative memory system. Interestingly, there are two pathways going from EC to CA1: the indirect one via DG and CA3 and the direct one. Moreover, the two pathways are delayed to different extents because most transmission delay happens at synapses (the end of red arrow in Fig. 1). Signals transmitting through the indirect pathway are delayed by 3τ given crossing one synapse takes time τ while those going through the direct pathway are only delayed by τ. The function of this seemingly redundant and asynchronized transmission setup remains to be illuminated. We hypothesize that, based on the delayed incoming signal, CA3 makes **future** predictions, which are then compared at CA1 with the less delayed actual signal from the direct pathway. Such comparison is similar to the update step in the Bayes filter procedure [21] where predictions based on previous information were compared against current observations to update the probabilistic model. Prediction errors are computed at CA1 and used to refine the internal model

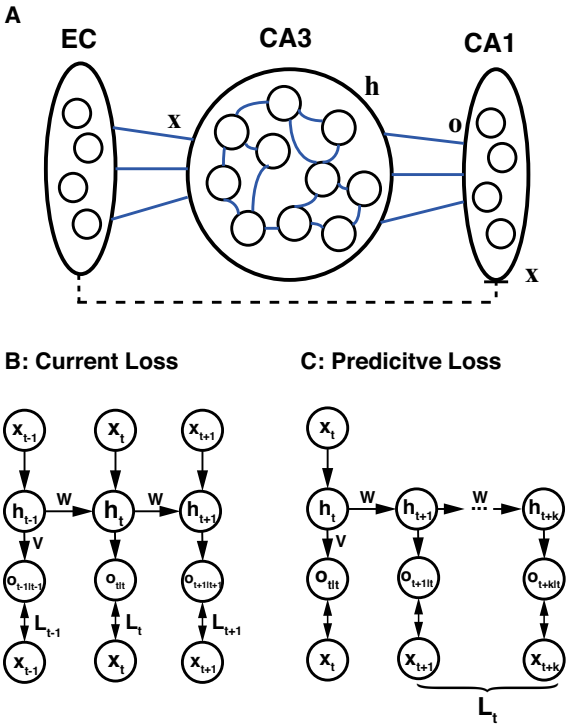

Figure 2: **PredRAE and predictive loss function.**(A) PredRAE architecture and its correspondence to hippocampus sub-regions. Dark blue weights are trainable (W, U, V and b in Eq. 1). (B) The computation graph of a conventional recurrent neural network whose loss function is MSE between prediction and actual signal at current time; (C) In contrast, a predictive loss function is calculated as the summation of MSEs based on k future predictions. $o_{t+k|t}$ means the prediction of time step $t+k$ given information at time step t. For simulations shown below, $k = 1$ unless otherwise specified.

stored at CA3. In such way, interactions between cortex and hippocampus form a self-supervised machine, which learns a generative model of external stimuli through a predictive loss.

To verify the above hypothesis, we implemented a predictive recurrent auto-encoder (PredRAE) with its output (o) trying to match its input (x) (Fig. 2A, Eq. 1). The input and recurrent layer models EC and CA3, respectively. CA1 response is computed as the difference between output prediction and actual signal (ReLU($x - o$)). Predictive loss function is implemented as mean square errors (MSE) between future predictions ($o_{t+1|t}$) and future teacher signals (x_{t+1}). The network was implemented in PyTorch (v1.11.0) and training was performed through stochastic gradient descent. Gradients were calculated as back-propagation through time (BPTT). The recurrent network has 200 units while the number input/output units depends on tasks. We stopped training the network when the process reached the maximum number (50,000) of epochs or the loss function didn't change more than 0.001% in 10 consecutive iterations. Isomura et al. [22] used a similar loss function to optimize for an auto-regressive and Principal Component Analysis (PCA) process.

$$
\begin{aligned}
h_t &= \tanh(Wh_{t-1} + Ux_t + b) \\
h_{t+1} &= \tanh(Wh_t + b) \\
o_{t+1|t} &= \mathrm{softmax}(Vh_{t+1}) \\
L &= \sum_t L_t = ||o_{t+1|t} - x_{t+1}||^2
\end{aligned}
\tag{1}
$$

PredRAE is not the first attempt to implement temporal predictive coding. PredNet [23][24], a hierarchy of CNNs and LSTMs, learnt to predict video scenes and captured a wide variety of neural responses in the visual processing stream. These two architectures differ in two aspects: 1) PredNet has built-in representation modules to extract latent variables while the disentanglement process in PredRAE occurred spontaneously; 2) PredNet uses a current loss function (Eq.4 in [23]) to capture temporal relationships.

3 Results

3.1 PredRAE matches key place cell features

To answer the first question, we simulated an agent running along a linear track with constant velocity (i.e. time and location are equivalent). Neurons in CA3 receive location specific Bell-shaped input according to their place fields (Fig. 3A, left). A new environment was modeled as a random shuffle of place fields (Fig. 3A, right). The agent remembered the first environment after training (i.e. PredRAE successfully reconstructed Env0). It was then exposed to a novel environment (i.e. the previously trained network continued to optimize over Env1). This led to an active response of CA1 error-encoding neurons, whose response gradually became weaker with more training epochs since the internal model was continuously updated to suppress errors. The simulated model response (Fig. 3B, upper) resembled the actual neural recordings (Fig. 3B, lower) [25] over a month. The rapid neural response upon exposure to a novel environment, pointed by the red arrow in Fig. 3B, is a well established place cell property called one-shot plasticity [26], whose mechanism is still debatable. Scientists have been looking for plasticity principles that modify weight abruptly [27], which might be dangerous for maintaining system stability if loosely regulated. Here we offered an alternative explanation to the one-shot plasticity phenomenon.

After being trained to remember two environments in the same batch, PredRAE formed attractor dynamics in its recurrent network, which explained two neuroscientific processes involved in memory formation: replay and prediction (Fig. 3C). Replay refers to the re-activation of place cells in the same order as they would fire during active locomotion. It usually happens when the animal is asleep or immobile, in other words, the agent receives no external sensory inputs. So we supplied low magnitude random noise to drive the trained network, it randomly reproduced one of the original input sequences (Fig. 3C, upper). Place cells also show predictive activities before making turning decisions: cells that code for possible future locations are activated [28]. This observation fits well under the pattern completion scenario in associative memory models. This time we supplied partial input to the network, it completed the remaining sequences (3C, lower). These two experiments are strong evidence of attractor dynamics in the system. In the language of state-space analysis, attractor dynamic means the system state would converge to a given (sequential) pattern, determined by network topology, regardless of inputs or perturbations [29]. To dig into network topology after training, we reordered the trained recurrent weight based on the activation order of the input sequence (3D, upper). The reordered matrix is similar to a toeplitz matrix, which enables a convolution operation after matrix multiplication. The approximate kernel (3D, lower) here is like a Mexican hat: positive values in the middle and negative values padding to both sides. This kernel has been showed to generate moving-bump attractor dynamics in recurrent neural networks [30]. Since we only trained 10% of the weights, the network could form multiple attractors, which only sparsely interfere with each other.

The trained weight is also precise enough to produce phase precession: the timing of place cell firing with respect to the oscillatory population activity (i.e. firing phase) gradually occurs earlier along traversal in the place field [31]. To generate biologically relevant action potentials, we transferred the trained recurrent weight to a network of leaky-integrate-fire (LIF) neurons following the procedure described in [32] and recorded the emitted spikes (Fig. 3E). Oscillatory activity was artificially enforced by injecting 8Hz inhibitory currents. Spike phases were calculated and plotted against their relative location to place field (PF) centers. We observed a trend of phase precession (Fig. 3F) similar to the one reported in experiments (Fig. 1 in [31]). As far as we know, PredRAE is the first **training-based** framework to take into account of spike timing accuracy in place cells.

In a more realistic simulation, PredRAE-simulated place cells exhibited a spatially localized representation. This is one of the most commonly recognized properties featured by place cells. We sampled the recorded trajectory of a rat running in a square arena (Fig. 4A) and simulated upstream neural response as the input to the model. Specifically, input was constructed as 200 grid cells [33] with various-sized grid fields periodically responding to spatial locations as the agent running along the trajectory. A firing cost was added to the original loss function ($L_t = ||o_{t+1|t} - x_{t+1}||^2 + \lambda||h_t||^2$). After training (Fig. 4B), we plotted place fields, at which locations neurons become active, of the recurrent units and observed localized representation (Fig. 4C).

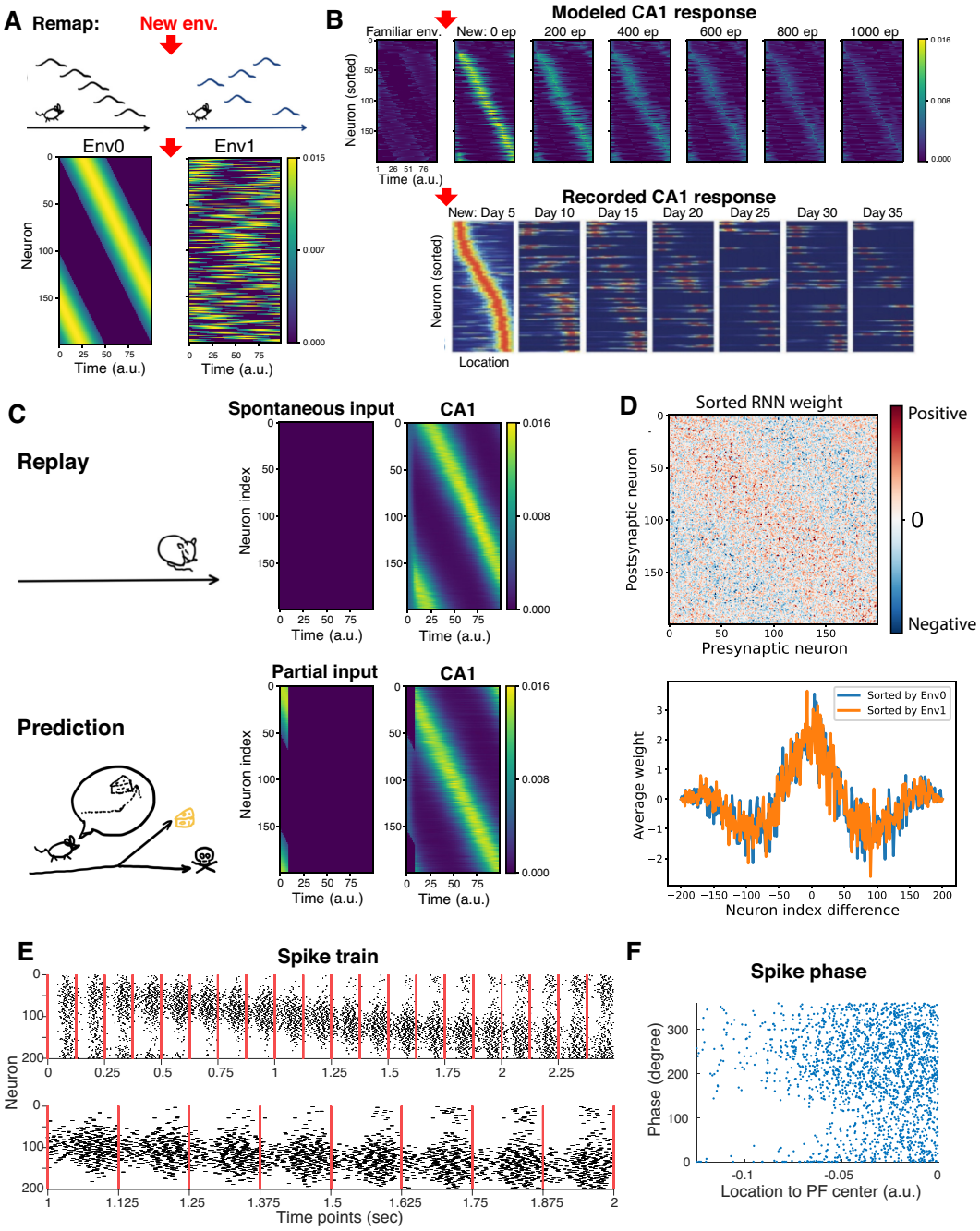

Figure 3: **PredRAE matches key place cell features, including one-shot plasticity, replay, prediction and phase precession.** (A) Input matrices (x) simulating different environments. With the equivalence of time and location, each row represents a Bell-shaped location specific input current. Env: environment. (B) Long term CA1 dynamics from the model (upper) and experimental recordings [25] (lower). Red arrow marks the initial exposure to a novel environment. The plotted neural response was sorted by their activation order. ep: epochs; (C) Replay: Given low magnitude random input simulating spontaneous activities, PredRAE output its previously remembered pattern as CA1 neuron response. Prediction: given input of the first 10 time steps, PredRAE performed pattern completion. (D) Upper: The trained recurrent weight matrix sorted by the activation order of Env0. Note similar entries on the diagonals; Lower: Average value of the matrix diagonals offset by the index shown on the horizontal axis. This would be the convolution kernel given a toeplitz matrix. The approximate kernels here look like Mexican hats. (E) Spike train recorded from a network of LIF neurons. Red vertical lines mark the trough of 8Hz population activity (i.e. phase=180 degree) (F) Spike phases (relative to 8Hz population activity) of all CA3 neurons plotted against their relative location in a place field (PF).

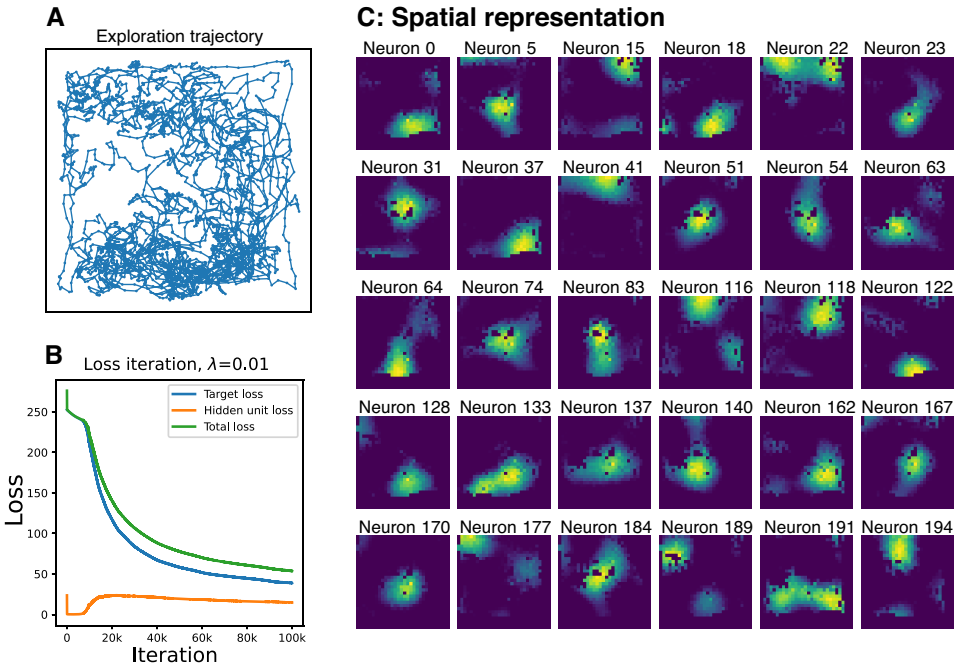

Figure 4: **Spatial representation in hidden units.** (A) Recorded exploration trajectory of a rat running in a 1.8m by 1.8m square arena sampled at 4Hz for 15 minutes from CRCNS hc-3, session ec014.277 (`http://crcns.org/data-sets/hc/hc-3/`)[34]. (B) Training curve of the regularized PredRAE. (C) Place fields of the recurrent units (CA3 neurons) after training.

3.2 PredRAE predicts temporal dynamics with an interpretable representation

Now we turned to the second question: how does PredRAE extract abstract sequential structure from complicated sensory inputs. In most cases, the extracted structure/representation is useful for downstream processing such as classification or prediction [35]. Popular latent space models such as Variational Auto-Encoder [36], Restricted Boltzmann Machines [37] and their many variants have achieved good performance in learning explicit representations. Here, we presented PredRAE as a natural and simple framework to achieve representation learning without sophisticated statistical priors or explicitly defined representation modules.

Consider visual cues as an example, we trained PredRAE to remember repetitive sequences of increasing MNIST digits (Fig. 5A). To minimize network structure, input was constructed as the top 68 principle components of the MNIST digits, which explain **87%** variance. After training, PredRAE output generic MNIST digits (Fig. 5B) since the different realizations of digits are hard to predict from previous time steps. When the input was stopped after digit 3, PredRAE continued to complete the sequences (Fig. 5C)[2]. Such continued prediction was not observed in a benchmark model with the same structure and a current loss function (RAE) (Fig. S1B). Interestingly, PredRAE not only learned a generative model to predict one time step ahead but also showed a sequential representation of input in the linear transformation space of hidden unit activities (Fig. 5D). Such interpretable representation was achieved without explicitly defined representation units as in the PredNet case. Each of the ten source components, revealed by independent component analysis (ICA) over hidden unit activities, coded one MNIST class (Fig. 5D upper, in contrast to Fig. S1C). This sequential activation of IC units along the time axis (Fig. 5D middle) indicates attractor dynamics in the transformed space. To quantify the accuracy of clustering, we performed k-means (number of clusters = 10) over the ICs. The metric of clustering accuracy shows the extent of representation separation/disentanglement. In both training and test cases, PredRAE could correctly label every digit while RAE had an error rate around 20% (Fig. 5D, bottom). Besides disentanglement, as most recurrent networks, PredRAE learns long-term temporal dependency. It was trained to remember a non-Markovian sequence of 0-1-2-3-4-5-4-3-2-1 with a prediction horizon equal to 1 ($k = 1$). When

[2]Readers could draw an analog to the predictive capacity of place cells in Fig. 3C

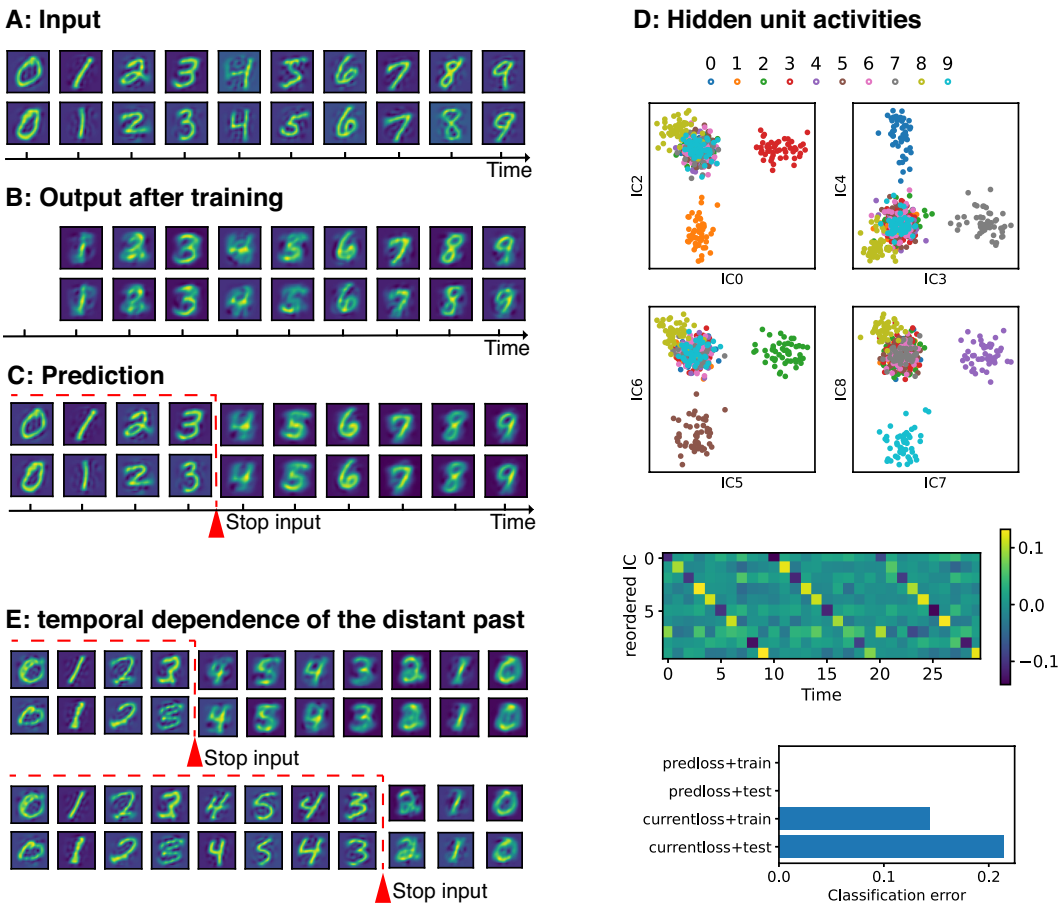

Figure 5: **PredRAE learns long-term temporal dependency and an interpretable representation.** (A) Input organized as repetitive sequences of MNIST images from 0 to 9. Each row is one batch consisting of 100 time steps. Two out of the five batches used in the input were shown. (B) PredRAE could reconstruct the given input with generic digits. (C) PredRAE continued to predict the sequence even when the input was stopped after digit 3. Input digits were plotted before the red vertical line while predictions were plotted afterwards. (D) Upper: Independent components (IC) of hidden unit activity. Most ICs represent one class (color-coded); Middle: Reordered IC activities over time. Note the sequential activation in 10 time steps (one cycle); Bottom: classification error using k-means clustering of the IC activities. Test case was calculated by supplying a different realization of MNIST sequences. (E) PredRAE was trained to remember repetitive MNIST sequences of 0-1-2-3-4-5-4-3-2-1. Top (Bottom): when the input was stopped at the first (second) 3, PredRAE continued to predict the increasing (decreasing) trend.

the input was stopped at digit 3, the network successfully used information from more than one step before to determine whether the sequence is increasing or decreasing (Fig. 5E). Trained to predict one step ahead, PredRAE learned more than just a mapping from time t to $t+1$.

Besides labels, PredRAE also learned the rotation operation and generalized it to test images (Fig. 6). This time, PredRAE was trained to reconstruct a sequence of MNIST digits rotating 30 degree counterclockwise at each time step (Fig. 6A). As before, it was able to rotate the given image after the input was stopped (Fig. 6B). Moreover, it generalized the rotation operation to test images although the reconstruction performance was not good because PredRAE was never trained to reconstruct these digits (Fig. 6C). The rotation dynamics could be recovered in the first three principal components (PC) of the hidden unit activity (Fig.6D). This cannot be observed in the control model with current loss function (Fig.S2). Although the rotation representation is in the PC space, it is still just a simple linear transformation, just as ICA, of the hidden unit activity. Our brain could easily extract it through a layer of feed-forward neural network.

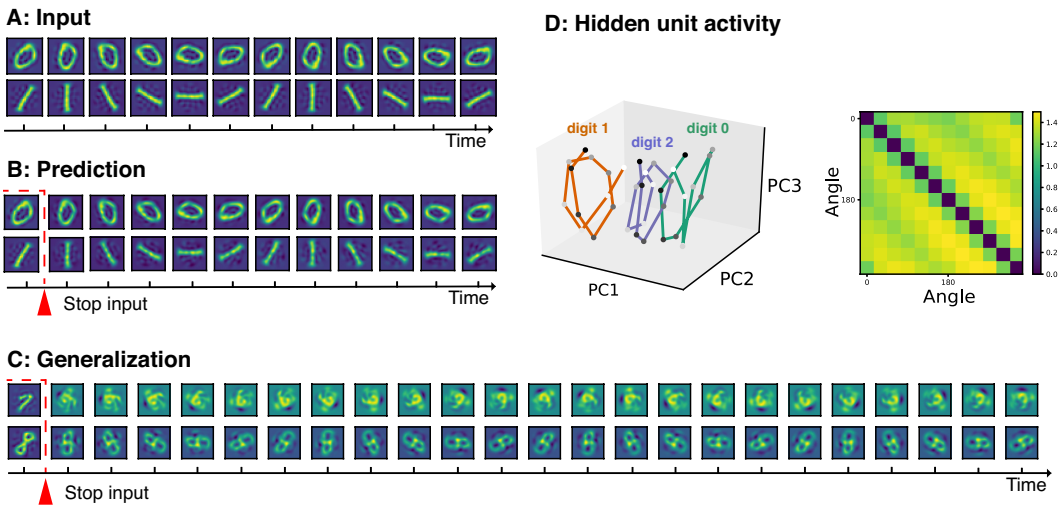

Figure 6: **PredRAE generalizes rotational dynamics.** (A) Input as a sequence of MNIST digits rotating 30 degree counterclockwise. (B) PredRAE learnt the rotation operation and kept rotating the digits after the input stops at the red vertical line. (C) Such operation worked well on novel digits (7 and 8 in this case). The entire class of 7 and 8 has never been shown to the network before. The reconstruction was not ideal because PredRAE has never learnt to reconstruct 7 or 8. It could also rotate the given image more than one cycle (two cycles shown here). (D) Left: Cycled dynamics of hidden unit activity shown in PC space. Right: the Euclidean distance matrix of hidden unit activities between different time steps averaged over all digits. The ± 1 off-diagonal stripe means cycled dynamics from current step to one step ahead. The ± 6 off-diagonal stripe appeared because some digits are symmetric (like 0, 1 and 8), therefore rotating 180 degree is the same as 360 degree.

3.3 PredRAE disentanglement is robust to multiple corruption types

Another advantage of the predictive loss function is its robustness against noise and corruptions because these are unpredictable. This may offer another layer of solutions to robust AI in addition to adversarial training [38] or data augmentation since most loss functions could be modified to incorporate a predictive component. Here we adopted 16 corruptions types reported in MNIST-C [39] to challenge PredRAE on its robustness. We first trained multiple PredRAEs over homogeneously corrupted MNIST sequences and measured its classification error as a metric of disentanglement. For most corruption types, the disentanglement performance is as good as clean MNIST sequences except for the "translate" and "fog" challenge (Fig. 7A). We believe the weakness to "translate" could be alleviated through data augmentation. Then we adopted a more challenging task: each digit in the input sequence was randomly corrupted by one of the 16 types (Fig. 7B). This is a hard task even for state-of-art Convolutional Neural Networks (CNN) [39]. After training, the reconstruction was not very satisfying but the disentanglement performance remained robust compared to that in the control network and in the original input space. Training with current loss didn't improve the disentanglement of representation in the hidden unit activity space.

4 Discussion

We proposed a temporal predictive coding framework to account for key cognitive functions of hippocampus: memory and representation learning. It was inspired by anatomical facts and successfully replicated a variety of experimental observations. In addition, the predictive feature of the architecture enables a robust representation learning. Hopefully, this neuroscience inspired algorithm could inspire research in human-like artificial intelligence. Below we briefly address some concerns and future directions of the current work.

The notion that CA1 neurons encode prediction error may be counter-intuitive to hippocampus scientists who routinely decode location information from CA1 place cell activities. Actually, the decaying neural activity in Fig.3B was recorded at a time scale of an entire month while most

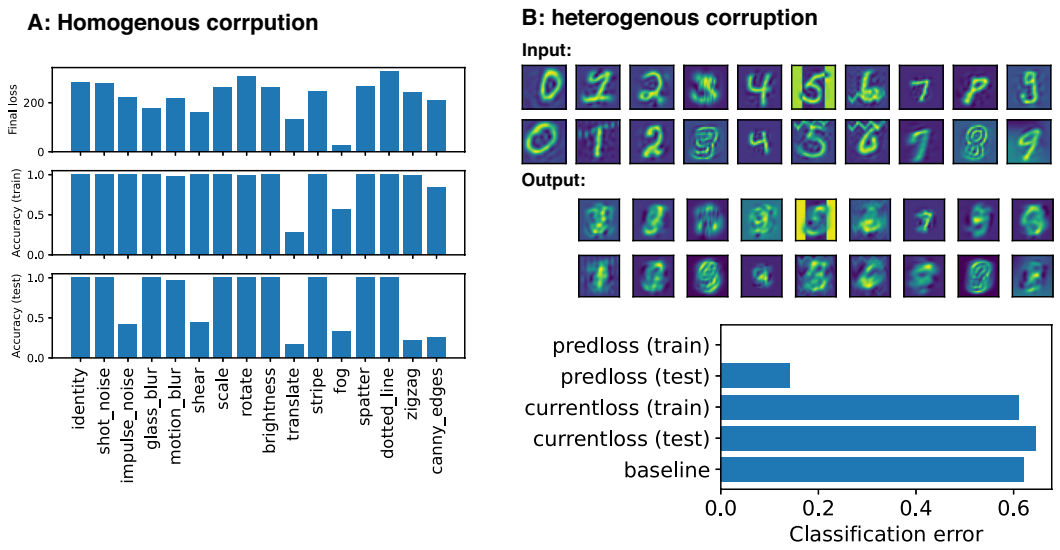

Figure 7: **PredRAE robustly learns temporal dynamics over MNIST-C images.** (A) PredRAE performance learning homogeneously corrupted MNIST sequences. Upper: the final loss after training; Middle: classification accuracy over training digits; Lower: classification accuracy over test digits. (B) Upper: A sequence of MNIST digits randomly corrupted by one type. Network output after training. Lower: k-means clustering error over the top 10 ICs of hidden unit activities. Baseline: clustering performance over the raw input sequences.

conventional recordings only last days, where the course of learning / error suppression hasn't been completed. A substantial amount of location information may still present in residual prediction errors. Besides the difference in time scale, the sensory input comes from multiple modalities, thus it is highly probabilistic and differs from trial to trial. For example, auditory or olfactory input could vary with time even the rat is still exploring the same environment. It's of low probability that we could actually observe null CA1 activities.

For predictive coding in sensory systems, error encoding neurons enable a biologically plausible learning rule, whose computation is local in both time and space. Although in this work, we adopted back-propagation through time (BPTT), it can be potentially replaced by local learning rules in reference to the derivation process in Rao et al. [10]. It might be difficult to accumulate gradients along the time axis. Nonetheless, there are increasing efforts to develop online training algorithms for recurrent neural networks through the approximation of target gradients [40][41]. In addition to training a memory model, errors computed in hippocampus could be further propagated to mid-brain dopaminergic neurons to enable training the value model during decision making [42].

PredRAE has minimal model structure to facilitate mechanistic investigations and interpretations. Future work could focus on scaling up the network or adding pre-processing modules to deal with real life problems such as semantic segmentation of video clips. Inspired by the brain circuit, PredRAE has great potential to approximate human performance with its capacity to robustly disentangle representations.

References

[1] Larry R Squire. The legacy of patient hm for neuroscience. *Neuron*, 61(1):6–9, 2009.

[2] John O'keefe and Lynn Nadel. *The hippocampus as a cognitive map*. Oxford university press, 1978.

[3] Jayeeta Basu and Steven A Siegelbaum. The corticohippocampal circuit, synaptic plasticity, and memory. *Cold Spring Harbor perspectives in biology*, 7(11):a021733, 2015.

[4] Thomas Klausberger. Gabaergic interneurons targeting dendrites of pyramidal cells in the ca1 area of the hippocampus. *European Journal of Neuroscience*, 30(6):947–957, 2009.

[5] Kimberly L Stachenfeld, Matthew M Botvinick, and Samuel J Gershman. The hippocampus as a predictive map. *Nature neuroscience*, 20(11):1643–1653, 2017.

[6] James CR Whittington, Timothy H Muller, Shirley Mark, Guifen Chen, Caswell Barry, Neil Burgess, and Timothy EJ Behrens. The tolman-eichenbaum machine: Unifying space and relational memory through generalization in the hippocampal formation. *Cell*, 183(5):1249–1263, 2020.

[7] Daniel C McNamee, Kimberly L Stachenfeld, Matthew M Botvinick, and Samuel J Gershman. Flexible modulation of sequence generation in the entorhinal–hippocampal system. *Nature Neuroscience*, 24(6):851–862, 2021.

[8] Jakob Hohwy. *The predictive mind*. OUP Oxford, 2013.

[9] Karl Friston. A theory of cortical responses. *Philosophical transactions of the Royal Society B: Biological sciences*, 360(1456):815–836, 2005.

[10] Rajesh PN Rao and Dana H Ballard. Predictive coding in the visual cortex: a functional interpretation of some extra-classical receptive-field effects. *Nature neuroscience*, 2(1):79–87, 1999.

[11] Gloria G Parras, Javier Nieto-Diego, Guillermo V Carbajal, Catalina Valdés-Baizabal, Carles Escera, and Manuel S Malmierca. Neurons along the auditory pathway exhibit a hierarchical organization of prediction error. *Nature communications*, 8(1):1–17, 2017.

[12] Hannah R Joo and Loren M Frank. The hippocampal sharp wave–ripple in memory retrieval for immediate use and consolidation. *Nature Reviews Neuroscience*, 19(12):744–757, 2018.

[13] George Dragoi and Susumu Tonegawa. Preplay of future place cell sequences by hippocampal cellular assemblies. *Nature*, 469(7330):397–401, 2011.

[14] György Buzsáki and David Tingley. Space and time: The hippocampus as a sequence generator. *Trends in cognitive sciences*, 22(10):853–869, 2018.

[15] Dharshan Kumaran and Eleanor A Maguire. An unexpected sequence of events: mismatch detection in the human hippocampus. *PLoS biology*, 4(12), 2006.

[16] Robert T Knight. Contribution of human hippocampal region to novelty detection. *Nature*, 383(6597):256–259, 1996.

[17] Katherine Duncan, Nicholas Ketz, Souheil J Inati, and Lila Davachi. Evidence for area ca1 as a match/mismatch detector: A high-resolution fmri study of the human hippocampus. *Hippocampus*, 22(3):389–398, 2012.

[18] Anna C Schapiro, Nicholas B Turk-Browne, Matthew M Botvinick, and Kenneth A Norman. 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, 2017.

[19] Inah Lee, Geeta Rao, and James J Knierim. A double dissociation between hippocampal subfields: differential time course of ca3 and ca1 place cells for processing changed environments. *Neuron*, 42(5):803–815, 2004.

[20] Stefan Leutgeb, Jill K Leutgeb, Alessandro Treves, May-Britt Moser, and Edvard I Moser. Distinct ensemble codes in hippocampal areas ca3 and ca1. *Science*, 305(5688):1295–1298, 2004.

[21] Wikipedia contributors. Recursive bayesian estimation — Wikipedia, the free encyclopedia, 2021. [Online; accessed 8-May-2022].

[22] Takuya Isomura and Taro Toyoizumi. Dimensionality reduction to maximize prediction generalization capability. *Nature Machine Intelligence*, 3(5):434–446, 2021.

[23] William Lotter, Gabriel Kreiman, and David Cox. Deep predictive coding networks for video prediction and unsupervised learning. *arXiv preprint arXiv:1605.08104*, 2016.

[24] William Lotter, Gabriel Kreiman, and David Cox. A neural network trained for prediction mimics diverse features of biological neurons and perception. *Nature Machine Intelligence*, 2(4):210–219, 2020.

[25] Yaniv Ziv, Laurie D Burns, Eric D Cocker, Elizabeth O Hamel, Kunal K Ghosh, Lacey J Kitch, Abbas El Gamal, and Mark J Schnitzer. Long-term dynamics of ca1 hippocampal place codes. *Nature neuroscience*, 16(3):264, 2013.

[26] James B Priestley, John C Bowler, Sebi V Rolotti, Stefano Fusi, and Attila Losonczy. Signatures of rapid plasticity in hippocampal ca1 representations during novel experiences. *Neuron*, 2022.

[27] Katie C Bittner, Aaron D Milstein, Christine Grienberger, Sandro Romani, and Jeffrey C Magee. Behavioral time scale synaptic plasticity underlies ca1 place fields. *Science*, 357(6355):1033–1036, 2017.

[28] Kenneth Kay, Jason E Chung, Marielena Sosa, Jonathan S Schor, Mattias P Karlsson, Margaret C Larkin, Daniel F Liu, and Loren M Frank. Constant sub-second cycling between representations of possible futures in the hippocampus. *Cell*, 2020.

[29] C. Eliasmith. Attractor network. *Scholarpedia*, 2(10):1380, 2007. revision #91016.

[30] Kechen Zhang. Representation of spatial orientation by the intrinsic dynamics of the head-direction cell ensemble: a theory. *Journal of Neuroscience*, 16(6):2112–2126, 1996.

[31] Misha V Tsodyks, William E Skaggs, Terrence J Sejnowski, and Bruce L McNaughton. Population dynamics and theta rhythm phase precession of hippocampal place cell firing: a spiking neuron model. *Hippocampus*, 6(3):271–280, 1996.

[32] Robert Kim, Yinghao Li, and Terrence J Sejnowski. Simple framework for constructing functional spiking recurrent neural networks. *Proceedings of the national academy of sciences*, 116(45):22811–22820, 2019.

[33] Edvard I Moser, Emilio Kropff, and May-Britt Moser. Place cells, grid cells, and the brain's spatial representation system. *Annu. Rev. Neurosci.*, 31:69–89, 2008.

[34] Kenji Mizuseki, Anton Sirota, Eva Pastalkova, and György Buzsáki. Theta oscillations provide temporal windows for local circuit computation in the entorhinal-hippocampal loop. *Neuron*, 64(2):267–280, 2009.

[35] Yoshua Bengio, Aaron Courville, and Pascal Vincent. Representation learning: A review and new perspectives. *IEEE transactions on pattern analysis and machine intelligence*, 35(8):1798–1828, 2013.

[36] Diederik P Kingma and Max Welling. An introduction to variational autoencoders. *arXiv preprint arXiv:1906.02691*, 2019.

[37] Ruslan Salakhutdinov, Andriy Mnih, and Geoffrey Hinton. Restricted boltzmann machines for collaborative filtering. In *Proceedings of the 24th international conference on Machine learning*, pages 791–798, 2007.

[38] Aleksander Madry, Aleksandar Makelov, Ludwig Schmidt, Dimitris Tsipras, and Adrian Vladu. Towards deep learning models resistant to adversarial attacks. *arXiv preprint arXiv:1706.06083*, 2017.

[39] Norman Mu and Justin Gilmer. Mnist-c: A robustness benchmark for computer vision. *arXiv preprint arXiv:1906.02337*, 2019.

[40] James M Murray. Local online learning in recurrent networks with random feedback. *ELife*, 8:e43299, 2019.

[41] Guillaume Bellec, Franz Scherr, Anand Subramoney, Elias Hajek, Darjan Salaj, Robert Legenstein, and Wolfgang Maass. A solution to the learning dilemma for recurrent networks of spiking neurons. *Nature communications*, 11(1):1–15, 2020.

[42] John E Lisman and Anthony A Grace. The hippocampal-vta loop: controlling the entry of information into long-term memory. *Neuron*, 46(5):703–713, 2005.