Neuron

Predictive sequence learning in the hippocampal formation

Highlights

- Analysis of neural recordings confirms that CA3 neurons predict the next input
- Self-supervised learning of sequences uses prediction error computed by CA1 neurons
- Simulations explain the distinctly different place field dynamics in CA1 and CA3
- A biologically plausible learning algorithm can train the predictive recurrent network

Authors

Yusi Chen, Huanqiu Zhang, Mia Cameron, Terrence Sejnowski

Correspondence

chenyusi151201@gmail.com (Y.C.), terry@salk.edu (T.S.)

In brief

Chen et al. simulated hippocampal circuits that learned to predict sequences of sensory inputs and validated the model with analysis of neural recordings. CA1 neurons in the model compute prediction error, using local self-supervised learning, consistent with the differential fading of CA1 and CA3 place cells.

Chen et al., 2024, Neuron *112*, 1–14
August 7, 2024
https://doi.org/10.1016/j.neuron.2024.05.024

CellPress

Neuron

CellPress

Article

Predictive sequence learning in the hippocampal formation

Yusi Chen,[1,2,4,*] Huanqiu Zhang,[1,3] Mia Cameron,[1,2] and Terrence Sejnowski[1,2,5,*]
[1]Computational Neurobiology Laboratory, Salk Institute for Biological Sciences, La Jolla, CA 92037, USA
[2]Department of Neurobiology, University of California, San Diego, La Jolla, CA 92093, USA
[3]Neurosciences Graduate Program, University of California, San Diego, La Jolla, CA 92093, USA
[4]Computational Neuroscience Center, University of Washington, Seattle, WA 98195, USA
[5]Lead contact
*Correspondence: chenyusi151201@gmail.com (Y.C.), terry@salk.edu (T.S.)
https://doi.org/10.1016/j.neuron.2024.05.024

SUMMARY

The hippocampus receives sequences of sensory inputs from the cortex during exploration and encodes the sequences with millisecond precision. We developed a predictive autoencoder model of the hippocampus including the trisynaptic and monosynaptic circuits from the entorhinal cortex (EC). CA3 was trained as a self-supervised recurrent neural network to predict its next input. We confirmed that CA3 is predicting ahead by analyzing the spike coupling between simultaneously recorded neurons in the dentate gyrus, CA3, and CA1 of the mouse hippocampus. In the model, CA1 neurons signal prediction errors by comparing CA3 predictions to the next direct EC input. The model exhibits the rapid appearance and slow fading of CA1 place cells and displays replay and phase precession from CA3. The model could be learned in a biologically plausible way with error-encoding neurons. Similarities between the hippocampal and thalamocortical circuits suggest that such computation motif could also underlie self-supervised sequence learning in the cortex.

INTRODUCTION

The representation of sensory information in cortical structures is encoded in the spatiotemporal patterns of spikes in populations of neurons. During locomotion, the spike timing of neurons in area CA1 of the hippocampus precesses relative to the local phase of the theta wave.[1–4] Spike timing is also precisely regulated at the millisecond level to engage spike-timing-dependent plasticity (STDP).[5–7] This regulation must take into account time delays for both the conduction of spikes between neurons and transmission delays at synapses. We focus here on the functional implications of this precision for how temporal sequences of spikes are shaped by neural circuits. We show how the temporal precision of spike timing coupled with anatomical wiring could support the learning and replay of temporal sequences in the hippocampal formation.

Cognitive maps are created in the hippocampus, with place cells in rodents responding not only to locomotion signals[8] but also to other sensory stimuli, such as reward,[9] auditory tones,[10] odors,[11] and time.[12–15] These stimuli are high dimensional and highly redundant, yet only a few hippocampal neurons are reliably and repetitively activated in a short time interval, forming a relatively low-dimensional dynamical trajectory in activity space.[16] The hippocampus therefore learns how to encode high-dimensional sensory and motor signals at the apex of cortical hierarchies into low-dimensional, latent, non-redundant, sequential representations that ultimately support abstract representational learning. After learning sequences of events, the hippocampus then replays them during sleep and immobility when external inputs to the cerebral cortex are suppressed.[17,18]

Existing computational frameworks[19–22] have successfully modeled cognitive functions of the hippocampus and reproduced the statistics of place cell under various task conditions. However, these models do not provide implementation of these cognitive functions based on neural mechanisms or account for the distinct encoding and firing properties of neurons in CA3, CA1, and the dentate gyrus (DG).[23–26] For example, CA1 neurons are more responsive to unexpected signals than neurons in other hippocampal areas,[27–29] and their activity decays over a timescale of weeks in familiar environments, faster than neurons in other subregions (Figure S1). In contrast, recurrent circuits in CA3 store an internal representation of sequences that are regenerating during replay[17,18] and preplay.[30] Neural place fields emerge faster in CA1 but are generally more stable in CA3 upon remapping.[24,26] Schapiro et al.[31] proposed a complementary learning system for CA1 and CA3 that reconciled statistical learning with episodic memory. We exploit these functional differences for a temporal predictive learning theory of sequences in the hippocampus.

Predictive coding efficiently encodes visual features in lower cortical layers, enabling higher layers to represent more abstract features.[32,33] This study builds upon these findings by extending predictive coding into the temporal domain to model interactions among hippocampal subregions. We confirmed the temporal

Please cite this article in press as: Chen et al., Predictive sequence learning in the hippocampal formation, Neuron (2024), https://doi.org/10.1016/j.neuron.2024.05.024

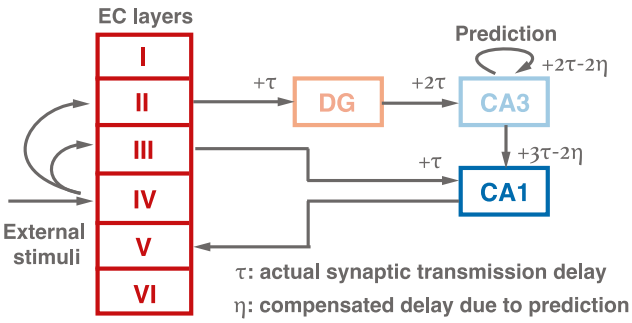

Figure 1. The circuit within the hippocampal formation

Anatomical wiring and interregional delays. External sensory stimuli start from different cortical layers in EC and reach CA1 through two pathways, forming a self-supervised structure. Assume that spikes are delayed by τ after one synaptic transmission, they will be delayed by 3τ and τ at CA1 through the indirect and direct pathways, respectively. We hypothesize that CA3 predicts the future (-2η) to compensate for the accumulated transmission time difference ($+2\tau$).

prediction hypothesis by analyzing neural recordings. We were unable to replicate experimental findings with a recurrent network model of CA3 that simply learned sequences. However, when we trained a network to make temporal predictions, we were able to successfully replicate the observed statistics of neural activity, the qualitatively distinct dynamics in CA1 and CA3 place cells, and representational learning to generate replay. We further demonstrated a biologically plausible predictive learning rule.

RESULTS

Temporal prediction hypothesis

Figure 1 summarizes the major connectivity in the hippocampal formation. The entorhinal cortex (EC) is the major cortical input to the hippocampus and is the major recipient of its output. Among hippocampal subregions, recurrently connected CA3 is ideal for storing internal states in the form of attractor dynamics.[34] Area CA1 receives inputs from two pathways projecting from the EC to CA1: an indirect pathway via DG and CA3 and a direct pathway from the EC. Moreover, the two pathways are delayed to different extents because there are more synaptic delays in the indirect path through CA3.[35] Assuming a synaptic transmission delay $\tau > 0$, signals transmitted through the indirect pathway to CA1 are delayed by 3τ, while those going through the direct pathway are only delayed by τ. An *in vitro* electrophysiology study[36] measured a 2.5-ms delay from EC to CA1 through the direct pathway and a 9- to 17-ms delay through the trisynaptic indirect pathway. The delay from EC to DG was 1.7 ms.

The function of this seemingly redundant and asynchronous transmission from EC to CA1 suggests that CA3 may be making predictions about future inputs, which can then be compared at CA1 with the less delayed teacher signal from the direct pathway. This comparison is similar to a Bayesian filter[37] where future predictions based on currently available information are compared with future observations to update the model. Therefore, we hypothesize that prediction errors are computed at CA1 and can be used to refine the internal model stored in CA3. In this way, interactions between the cortex and the hippocampus form a self-supervised loop, which enables the circuit motif to learn and remember the latent variables of a predictive autoencoder represented in CA3 as sequences.

Neural evidence for transmission delay and predicting ahead

To verify the above hypothesis, we analyzed simultaneously recorded neural activities from these subregions for evidence of transmission delay and predicting ahead. Assuming that neural signal propagation strictly follows the anatomical organization of the hippocampal formation in Figure 1, signals encoded by a region should be correlated with the upstream signal shifted by an interregional time delay. Ideally, if a location-sensitive neuron in EC has a bell-shaped response curve $f(x)$, where x represents any arbitrary physical variable such as location, its direct downstream DG neuron should exhibit a response $f(x-\tau)$, where τ refers to the default interregional delay (Figure 2A). Similarly, the response curves of their downstream neurons in CA3 and CA1 should be $f(x-2\tau)$ and $f(x-3\tau)$, respectively (dashed lines in Figure 2A).

Alternatively, if according to our hypothesis, CA3 is predicting future signals to match the signal arrived from the direct pathway, CA3 and CA1 would have response curves of $f(x)$ and $f(x-\tau)$, respectively (solid lines in Figure 2A), given similar interregional delays. Although the recordings are unlikely to be from directly connected neurons, evaluating the similarity measures between distributions of temporally shifted neural activities should reveal interregional spike coupling properties (Figures 2B–2D, upper).

According to our hypothesis, CA3 spike trains should couple tightly with leftward-shifted DG spike train (Figure 2C, upper), indicating that CA3 firing leads DG. This suggests that CA3 is predicting ahead since it is anatomically downstream of the DG. For both the prediction and non-prediction scenarios, CA1 activity should always follow CA3 by one synaptic delay (Figure 2B, upper). Moreover, despite the challenges associated with measuring CA1-DG coupling due to their lack of direct connection, similarity measures for CA1 and DG spike trains for the prediction scenario are expected to reach a maximum at approximately zero delay, since both these areas are one synapse away from the EC. A peak at zero signifies information synchrony between these two regions and would highlight the predominance of signals delayed by τ in CA1 (see Figure 2D, upper).

We used the visual encoding neuropixel dataset from the Allen Brain Observatory.[38] This dataset contains simultaneous recordings of neural spikes sampled at 30 kHz in DG, CA3, and CA1 from mice performing passive visual perception of natural movies, i.e., sequences of natural images (STAR Methods). The high temporal precision enabled us to investigate spiking timing accuracy on a millisecond timescale.

Following the methods in Siegle et al.,[38] we calculated the jitter-corrected cross-correlogram (CCG) of spike trains between pairs of subregions over all stimulus conditions and plotted the distribution of optimal shifts where CCG peaks (Figures 2B–2D) (STAR Methods). To access higher-order statistical relationships, we also calculated the mutual information (MI) between the shifted spike trains, since we are interested in the amount

Please cite this article in press as: Chen et al., Predictive sequence learning in the hippocampal formation, Neuron (2024), https://doi.org/10.1016/j.neuron.2024.05.024

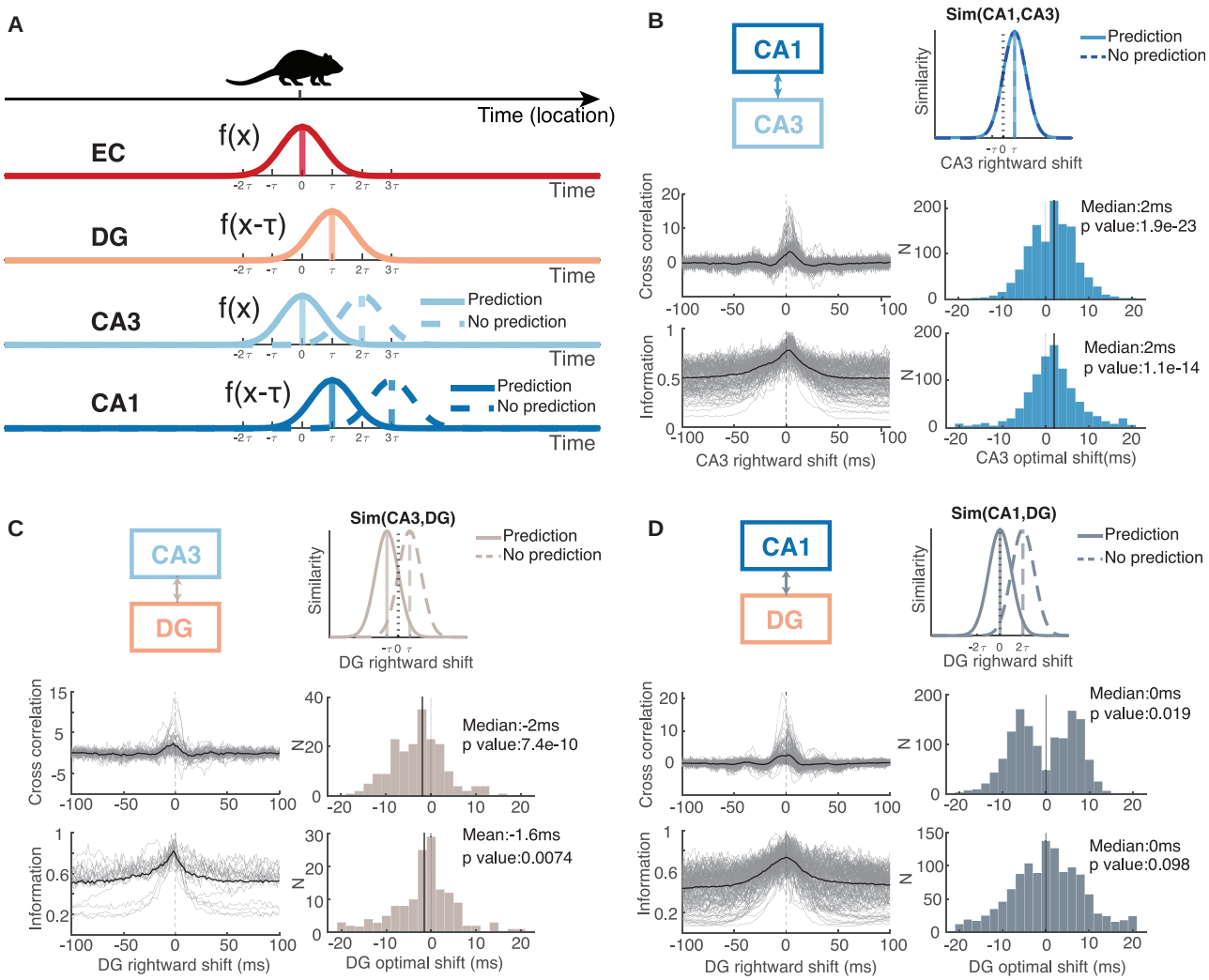

Figure 2. Neural evidence of transmission delay and predicting ahead

(A) Schematics of delayed neural response and hypothesized predicting effect. Assume a rat running with constant velocity (time = location), one representative location-sensitive neuron in EC exhibits bell-shaped response curve peaked at $t = 0$. Given there's no prediction, its direct downstream DG and CA3 neuron will peak at $t = \tau$ and $t = 2\tau$, respectively. Meanwhile, CA1 would receive mixed signals, delayed by τ and 3τ, from dual pathways. If there is prediction ahead, CA3 would instead peak at $t = 0$, and CA1 would only respond to signals peaked at $t = \tau$.

(B) Spike coupling from CA3 to CA1. Top: schematics of spike train similarity with respect to CA3 neural activity shifts. Positive shift means shifting CA3 spike train toward the right and then computing its similarity with the unshifted CA1 spike train. Middle (bottom left): traces of corrected cross correlogram (mutual information) from an example session. Each gray trace represents the prediction from a population of CA3 neurons to one CA1 neuron. The solid black trace is the average across all CA1 neurons in the session. Middle (bottom right): histogram of optimal shift, where similarity measure peaks, pooled across 12 recording sessions. (*p* value: t test of population mean equals to zero).

(C) Spike coupling from DG to CA3.

(D) Spike coupling from DG to CA1. DG is synchronized with CA1, while CA3 leads DG by 2 ms. This confirms the hypothesis that CA3 is predicting ahead. See also Figure S2.

of delay in the information transmitted by the spike trains. When we used shifted CA3 spike trains to predict unshifted CA1 spike train in Figure 2B, they coupled most strongly when CA1 was shifted to the right. Thus, unsurprisingly, CA3 was ahead of CA1 activity by 2 ms, which matched the previously reported synaptic delay,[35,36] validating our approaches to calculate synaptic coupling at the precision of milliseconds.

In Figure 2C, we compared the unshifted CA3 spike train with shifted DG spike trains. We found that both similarity measures peaked when DG shifted significantly toward the left by a median of 2 ms. This directly supports the CA3 predictive-ahead hypothesis as explained above. In Figure 2D, we compared the unshifted CA1 spike train with shifted DG spike trains. From the MI analysis, the distribution of optimal shifts is approximately a normal distribution with median value of zero. This means that neurons in CA1, despite some randomness, are synchronized with those in DG assessed by MI. They were both delayed by one synapse with respect to EC. The identification of a leftward

Please cite this article in press as: Chen et al., Predictive sequence learning in the hippocampal formation, Neuron (2024), https://doi.org/10.1016/j.neuron.2024.05.024

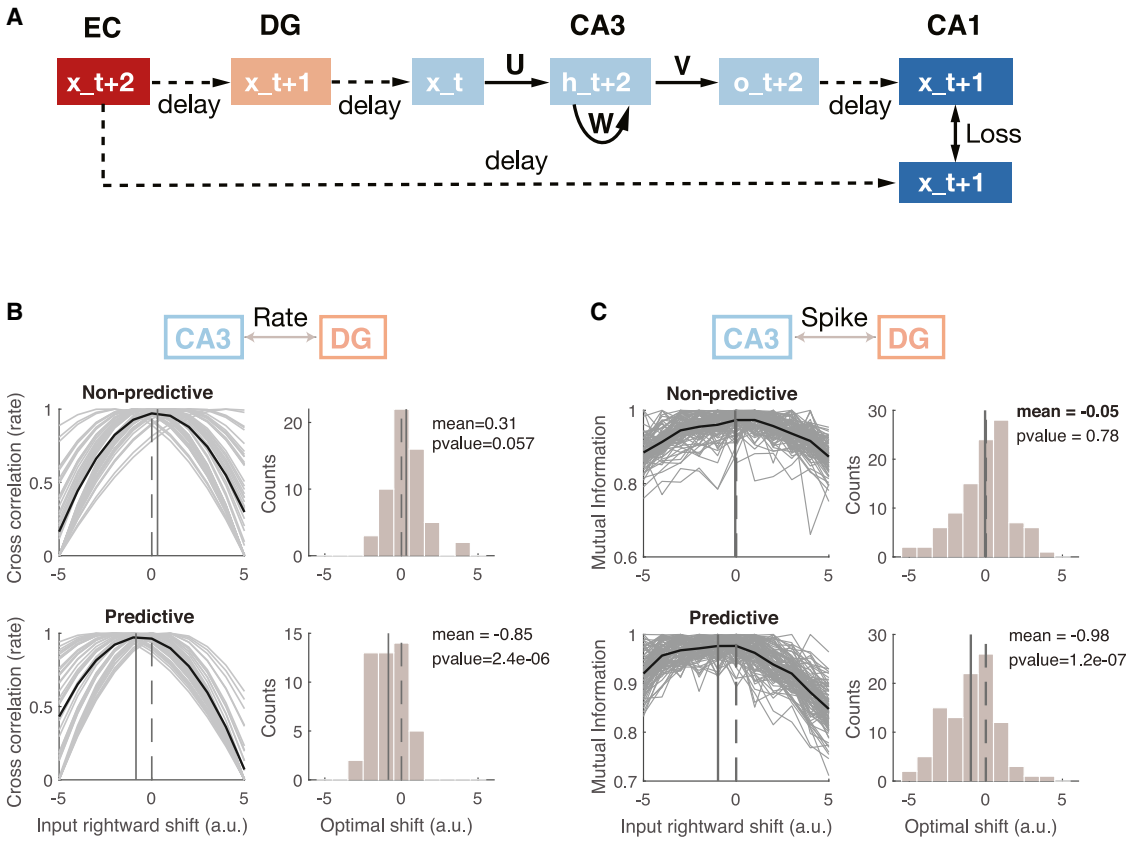

Figure 3. A predictive RNN explains observed statistics

(A) Summary of the temporal relationship observed in Figure 2. The time stamps are labeled from the perspective of CA3 input. Dashed and solid lines indicate the delay operation and weighted computation, separately. We believe prediction happened through the recurrent weight W, forming a conventional recurrent neural network as described by the equations. EC/DG, CA3, and CA3 function as the input and recurrent and output layer, separately. To enforce the recurrent units to predict ahead, we adopted a predictive loss function.

(B) Rate coupling from DG to CA3, equivalently, cross-correlation analysis between x_t and h_t for non-predictive networks (first row) and predictive networks (second row). In a predictive network, recurrent units correspond better with x_{t+1}.

(C) Mutual information analysis of spikes from DG to CA3. Spikes are generated through a Poisson process with the rate given by the trained networks and Δt = 0.02 s.

peak (CA1 preceding DG) in the cross-correlation analysis and a synchronized peak in the MI analysis strongly supports the predictive-ahead hypothesis.

The distribution of optimal shifts between CA1 and DG from cross-correlogram analysis was bimodal (Figure 2D). We acknowledge that comparing the correlated activity in CA1 and DG may be problematic, considering their lack of a direct connection and the inherent difficulty in recovering coupling at the millisecond level. However, the leftward peak could only appear with the predictive component in the circuit, otherwise CA1 response from the indirect pathway is always going to lag behind DG response. The presence of a rightward peak in the cross-correlation analysis could be a consequence of fast oscillations in the 100–200 Hz range in local field potential (LFP) recordings in CA1. Fast oscillations were not observed in DG on the same probe (see Figure S2). An oscillation could induce the bimodal peaks observed in Figure 2D. The MI analysis, which is more robust to firing rate fluctuations, was not bimodal, partially supporting this hypothesis.

Explaining observed spike coupling with a predictive recurrent neural network

We developed a predictive recurrent autoencoder model of the hippocampus to compare the time delays in the model with those observed between CA3 and DG. In Figure 3A, we illustrated the temporal relationships identified in the preceding section. Notably, the recurrent units in CA3 encode information at $t+2$ due to the predict-ahead training. Dashed lines represent delay operations, while solid lines signify network computations governed by the equations in Equation 1. The input signal x originates from the EC and DG, with DG serving solely as a delay operator in our model. The recurrent signal h models CA3 activities, and the CA1 response is computed as a concatenation of prediction errors and predictions (o). The dynamics of CA3 and CA1 activities will be explored in subsequent sessions.

To train the model for predicting its next-step input x, we adjusted the recurrent weight W using backpropagation through time (BPTT)[39–41] over a predictive loss function (Equation 1). (Since we don't have a physical time scale in this model, we

Please cite this article in press as: Chen et al., Predictive sequence learning in the hippocampal formation, Neuron (2024), https://doi.org/10.1016/j.neuron.2024.05.024

CellPress

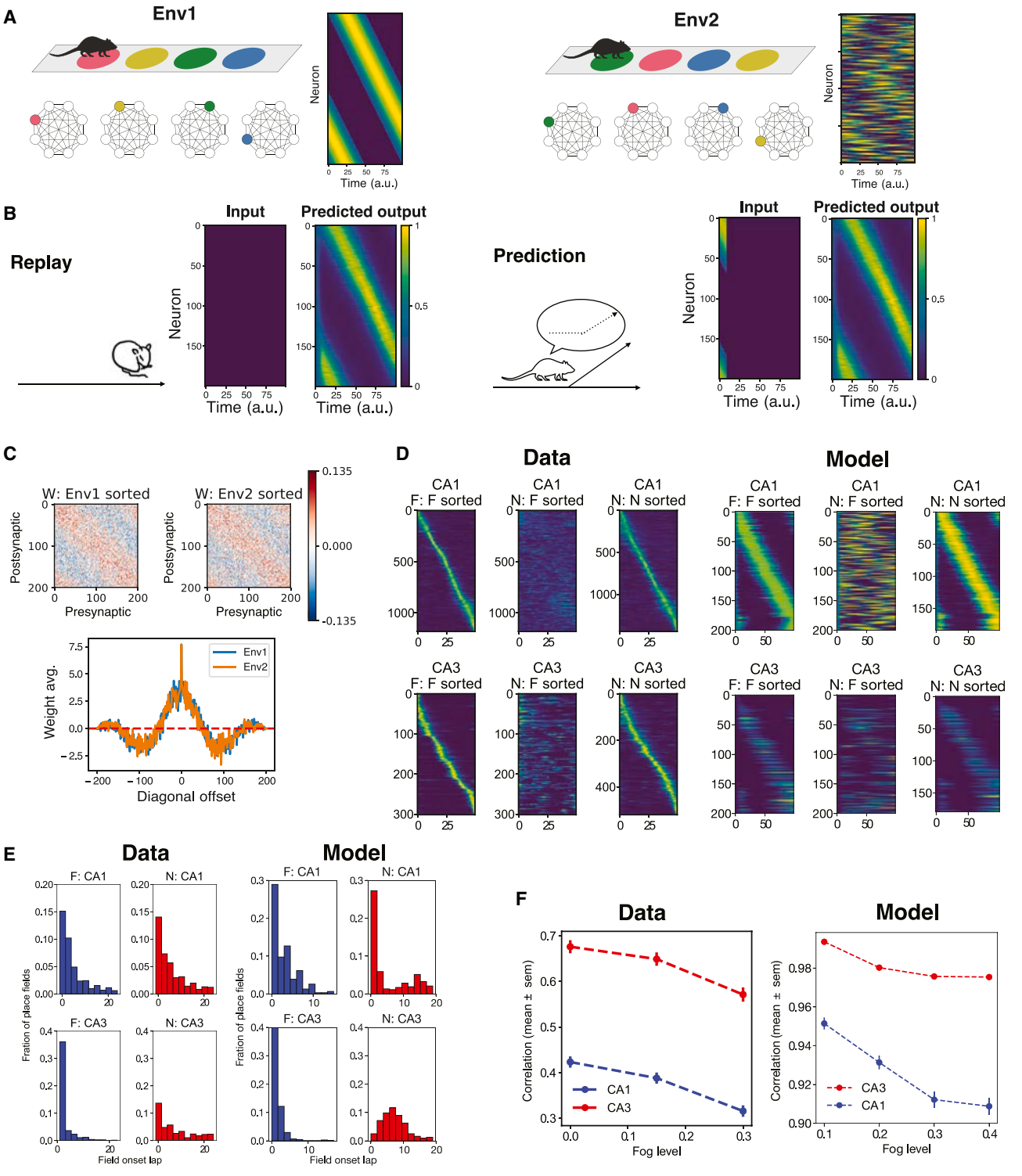

Figure 4. CA1 error-encoding neurons facilitate the learning of internal model and explain distinct CA1 and CA3 place field dynamics during remapping

(A) Input matrices (x_t) used during remapping. With the equivalence of time and location, each row represents a bell-shaped location-specific input current. Env, environment. A second environment was modeled as the complete shuffling of the first familiar environment.

(B) Replay: given low-magnitude random input simulating spontaneous activities, a predictive recurrent autoencoder outputs its previously remembered pattern. Prediction: given input of the first 10 time steps, the network performed pattern completion.

(legend continued on next page)

Please cite this article in press as: Chen et al., Predictive sequence learning in the hippocampal formation, Neuron (2024), https://doi.org/10.1016/j.neuron.2024.05.024

assumed one unit of time for the number of temporal delays from EC to CA3.) For comparison, we also trained the network using a non-predictive loss function. In the later section, we show that the error signal leads to comparable learning performance with biologically plausible learning rules.

$$\begin{aligned} &\text{Network Dynamics} && : h_t = \sigma(Wh_t + Ux_t); o_t = \sigma(Vh_t) \\ &\text{Predictive loss} && : L = \sum_t \|o_t - x_{t+1}\|^2 \\ &\text{Non-predictive loss} && : L = \sum_t \|o_t - x_t\|^2 \end{aligned} \quad \text{(Equation 1)}$$

where inputs x_t project to the hidden units h_t with weights U. The hidden units are connected with recurrent weights W and project to the outputs o_t through weights V. We tasked both networks with learning a bell-shaped pulse spanning 100 time steps, centered at $t = 50$, and subsequently calculated the cross-correlation of x and h, as illustrated in Figure 3B. In the network trained using a predictive loss function, the recurrent units exhibited a notable leftward shift, compared with the input (bottom panels). This shift was not observed in the control network trained using the non-predictive loss function. We then used the firing rates derived from the artificial networks to generate Poisson spikes and performed cross-MI analysis on the simulated spike trains. The same results were found in the spiking network model as those obtained in the rate-based model (Figure 3C).

Facilitating learning of internal models through sequence prediction

Our first test of the predictive recurrent network model was to explain place cell dynamics. We first simulated a rat running along a circular track with constant velocity (where time and location are equivalent). All neurons in the recurrent layer (CA3) received location-specific bell-shaped input activity (x_t) representing their respective place fields on the track (Figure 4A, left). A new environment was modeled as a random shuffle of place fields (Figure 4A, right). The network was trained on Env1 and Env2, using the predictive loss function.

The trained models successfully reproduced the input sequences and exhibited replay and prediction (Figure 4B). Replay refers to the re-activation of place cells in the same order as they would during active exploring. Typically, this occurs when the animal is in a state of sleep or immobility, meaning that the simulated agent is not receiving any external sensory inputs. When low-magnitude random noise was used to drive the trained network, it randomly reproduced one of the learned sequences (Figures 4B, left, and S3).

Place cells in the model also showed predictive activities that have been reported for cells that are activated before making turns and code for possible future locations.[42,43] This could be a consequence of pattern completion by the recurrent network model. To demonstrate this, following a partial input sequence to the network, it completed the remaining sequences (Figure 4B, right).

This is strong evidence for line attractor dynamics[44] in the network. We reordered the learned recurrent weight matrix based on the activation order of the hidden units in either Env1 or Env2 (Figure 4C, upper). The reordered matrix has an approximately symmetric Toeplitz form resembling a one-dimensional chain of neurons, each connected to its nearest neighbors with positive values and more distant neurons with negative values (Figure 4C, lower). The mathematical significance of Toeplitz connectivity is elaborated in the discussion.

The trained network also exhibited phase precession, which occurs in recordings where the timing of place cell firing with respect to the phase of the oscillatory population activity becomes progressively earlier when traversing a place field.[1] To generate biologically relevant action potentials, we transferred the weights from a trained recurrent weight to a network of leaky-integrate-fire (LIF) neurons, following the procedure described in Kim et al.,[45] and recorded the emitted spikes (Figure S4A). Oscillatory activity was artificially enforced by injecting 8 Hz inhibitory currents, mimicking oscillatory inputs onto inhibitory neurons originating in the septal nucleus. Spike phases were calculated and plotted against their relative location to place field centers. Analysis of the LIF neurons during simulated running on the track exhibited precession of the spike timing (Figure S4B) similar to phase precession recorded from neurons *in vivo* (see Figure 1 in Tsodyks et al.[46]).

CA1 error-encoding neurons explain distinct CA1 and CA3 place field dynamics

Although differences in the encoding properties of place cells in CA3 and CA1 are well known,[47] they have been overlooked in most hippocampal models. In our model, CA3 stores an internal model of the world, while CA1 not only inherits CA3 output but also simultaneously encodes prediction error. Supporting evidence for error-encoding neurons involves earlier experimental observations that CA1 neurons respond more than neurons in other regions to unexpected signals[27–29] and that CA1 place fields decay slowly in familiar environments (Figure S1). At the same time, acute silencing of CA3 drastically reduces CA1 response,[48] suggesting that CA3 is the predominant driver of CA1 place cells under normal conditions. (We want to note the debate regarding the primary pathway driving CA1, as noted in

(C) Upper: the trained recurrent weight matrix sorted by the activation order of Env1 or Env2. Note similar entries on the diagonals; lower: average value of the matrix diagonals offset by the index shown on the horizontal axis. The approximate kernel here looks like a Mexican hat, pointing to the existence of line attractor dynamics.

(D) During remapping, different ensembles of neurons are activated in both CA1 and CA3 populations. Place fields of CA1 (first row) and CA3 (second row) neurons in familiar (F) or novel (N) environment, sorted by activation order. (Data reproduced from Dong et al.,[23] Figure 2A.)

(E) Histogram of CA1 (first row) and CA3 (second row) place field onset laps in F (blue) or N (red) environments. Both regions showed instant onset place fields in F, but only CA1 neurons responded instantaneously in N. (Data reproduced from Dong et al.,[23] Figure 2D.)

(F) Correlation of CA1 and CA3 place fields between F environment and its noisy/foggy version. See the table in the STAR Methods, network training, for details about the network and training. (Data reproduced from Shin et al.,[24] Figure 5.)

See also Figures S3 and S4.

the studies by Brun et al.[49,50] and Nakashiba et al.[51] This discrepancy may stem from the methodological differences in lesion studies, specifically whether the lesions were induced acutely or over a prolonged period.) In a later section, we present a biologically plausible predictive learning algorithm for CA3 based on local prediction error.

To model remapping from a familiar environment (F) to a novel environment (N), we instructed a network that had already memorized Env1 to optimize toward Env2. Neural responses in CA3 and CA1 were recorded as described above and sorted based on experimental observations (Figure 4D). For both CA1 and CA3 neurons, distinct ensembles of place cells were activated in the two environments, consistent with previous experimental data.[23] Importantly, in Figure 4E (right), upon the switch to the novel environment, CA1 place cells emerged more rapidly, as error information initially reflected the structure of the novel environment. Over time, CA1 place cells transitioned to representing CA3 output. In contrast, CA3 place cells emerged more slowly, suggesting that the internal representation might require multiple traversals to learn. The rapid neural response upon exposure to a novel environment aligns with the well-established place cell property called one-shot learning,[52] the mechanism of which is still debated. We propose here that this may be attributed to the coding of error signals rather than to an abrupt increase in weights. Abruptly modifying synaptic weights to a large population of neurons[53] could pose risks to system stability if not tightly regulated. In this study, we present an alternative explanation for apparent one-shot learning: neural activity immediately increases in a new environment because CA1 reflects one-shot error, not one-shot learning.

Next, we modeled place cell remapping back to a familiar environment by optimizing toward a noisy version of Env1 (Figure 4E, left). Consistent with experimental data,[23] both regions exhibited instant place fields, as the internal representation stored in the network matched the familiar environment. We also examined the relationship between the correlation of neural activities and the noise level in the Env1 environment (i.e., fog level in the data panel); noise decorrelated the representation in both CA3 and CA1, but CA3 was more robust to the presence of noise (Figure 4F).

Learning future sequences promotes interpretable latent representation

We next simulated a rat in a random foraging task consisting of straight trajectories and random turns in a square arena (Figure 5A). Input was handcrafted as a nonlinear mixture of body direction, world direction, path-integrated distance, and distance to the closest wall, following the approach described in Benna and Fusi[54] (STAR Methods). The selection of these input dimensions was based on the existence of head direction cells, path-integration signals, and border cells in EC. Any one of these inputs alone is not able to determine the agent's current location, as evidenced by the low MI per second between input unit activity and location (Figure 5C). To enforce a sparse representation in the hidden layer, we added a regularization penalty of unit activity to the loss function (STAR Methods).

After training using the predictive loss function, hidden units showed spatially localized representations similar to place fields (Figure 5B). In comparison to networks trained with the non-predictive loss function, hidden units in networks trained with the predictive loss function displayed significantly higher MI (Figure 5C). Both networks contributed to the extraction of locations as hidden units have much higher MI, compared with the original input signal. This indicates predicting ahead is beneficial for extracting location information from upstream inputs. The same results were found when the regularization strength was varied (Figure S6), suggesting that a temporal predictive loss function consistently aids in forming a localized representation.

To investigate the potential of the predictive loss function for achieving representational learning—specifically, forming sequential representations from high-dimensional sensory inputs—we organized image sequences using an increasing order of MNIST (Modified National Institute of Standards and Technology Database) handwritten digits (Figure S5). We first extracted the first 68 components of the images through principal-component analysis (PCA), which account for 86% variance, as the input to keep the minimal network structure. Image reconstruction was based on the network output and inverse transformation of PCA. After training, the network continued to complete the sequences when the input was stopped after digit 3 (Figure 6C). Predictive completion was not observed in control models trained with the non-predictive loss function (Figure S5).

Interestingly, the trained network not only constructed a generative model to predict sequences but also automatically clustered the digits according to their labels. In Figure 6D, we plotted the independent components (ICs) of the hidden unit activity for each digit and colored the digits according to their labels. For the top 10 ICs (sorted by the contribution of demixing matrix), each component represents one group of digits, approximately.

This interpretable representation was achieved without explicitly defined labels. The sequential activation of ICs along the time axis (Figure 6D, bottom) mimics the sequential activation of place cells in the linear transformed space. Imagine a rat running on tiles of MNIST patterns: Figure 6 shows how the interaction between cortex and hippocampus transforms the complicated sensory input into sequential activation of hippocampus neurons as reported in numerous experimental studies.

Error neurons facilitate a biologically plausible learning algorithm

We devised a predictive recirculation learning algorithm for a predictive autoencoder, consisting of a set of three local learning rules for the input weights (U), recurrent weights (W), and output weights (V) (Equation 2). These learning rules approximate the gradient of a predictive mean square error loss under certain assumptions (see STAR Methods for derivations).

The precise gradient of the output weight (ΔV in Equation 2) can be directly assessed as Hebbian learning between error-encoding neurons in CA1 (δx) and the recurrent neurons in CA3 (h).

The exact gradients of the input and recurrent weights pose challenges to achieving locality in both time and space. This temporal dependence was mitigated by truncating the temporal gradient beyond the current time step. Additionally, to preserve spatial locality, we avoided backpropagating errors by using recirculation. This strategy, inspired by the original recirculation algorithm proposed by Hinton and McClelland[55] for a three-layer

Please cite this article in press as: Chen et al., Predictive sequence learning in the hippocampal formation, Neuron (2024), https://doi.org/10.1016/j.neuron.2024.05.024

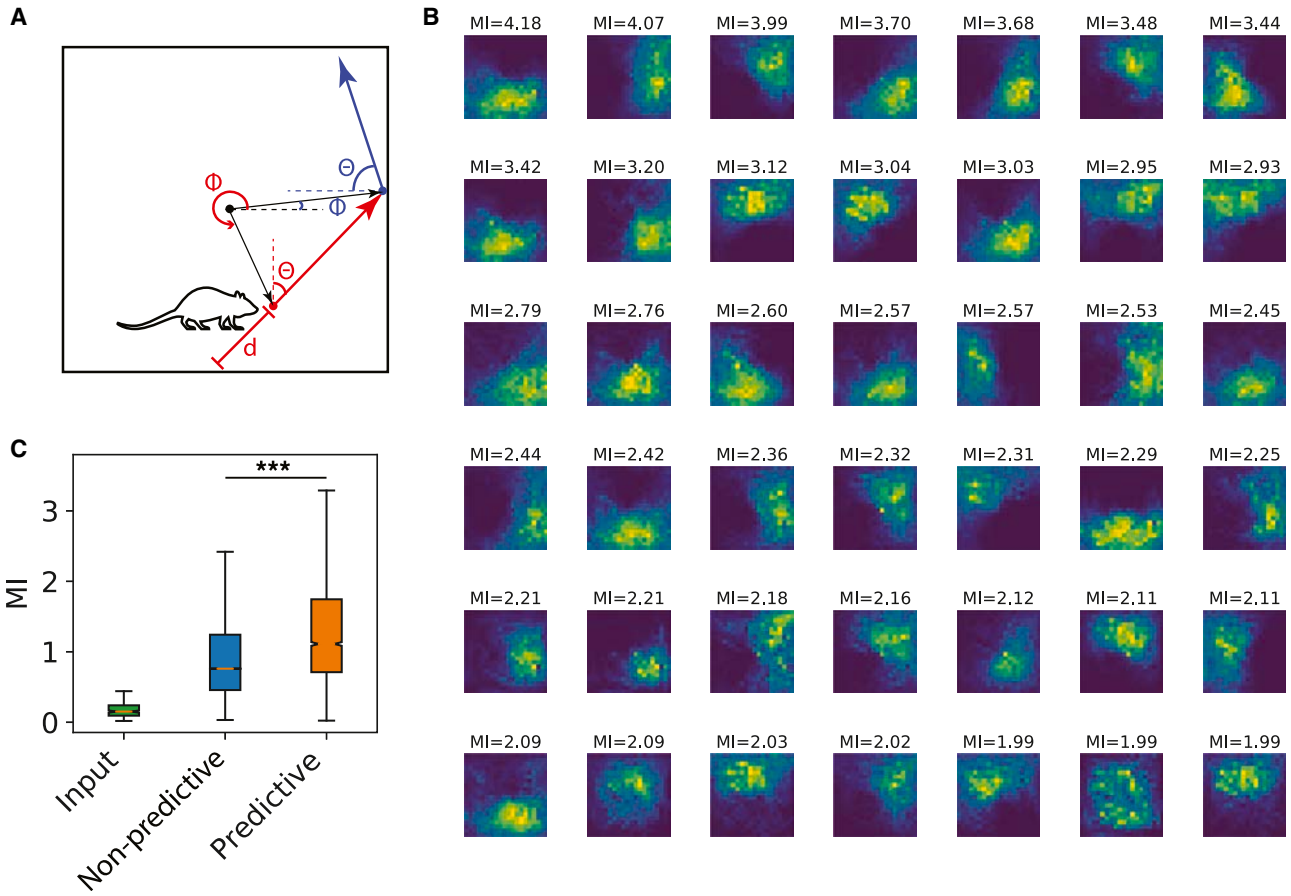

Figure 5. Predictive networks led to more localized representation in a random foraging task

(A) Schematic of foraging task simulation. The agent was running straight in an open arena until it hit a wall and make a random turn. The red and blue straight arrows indicate two straight trajectories. Network input defined as a random nonlinear mixture of body direction with respect to the norm of the last hit wall (θ), world direction with respect to east (φ), and path-integrated distance (d).

(B) Place fields of the recurrent units (CA3 neurons) after training. Extent of localization was quantified by mutual information (MI) rate per second between firing rate and location.

(C) MI of the input units and recurrent units in 10 networks trained by either the current loss function as a control or the predictive loss function. Recurrent units trained by the predictive loss function showed significantly higher MI (t test, $p < 0.001$).

See also Figure S6.

feedforward autoencoder, facilitates local learning for both input and output weights by feeding back reconstructed inputs to the encoder (Equation 4). As the authors of the recirculation algorithm noted, the input weights converge approximately to the transpose of the output weight ($U = V^T$). We confirmed that during learning, the matrix entries in U and V^T in our predictive autoencoder also converged: predictive recirculation learning effectively drove the weights from random initialization to approximate transposition (Figure S7). In the hippocampus, we propose that this recirculation process could be implemented through the feedback projections from CA1 to EC (Figure 7A).

In Figure 7, we trained a network using predictive recirculation algorithm to reproduce and recall a sequence of MNIST handwritten digits.

Predictive recirculation learning algorithm

$$\begin{aligned}
&\text{Network Dynamics :} && h_t = \tanh(Wh_{t-1} + Ux_{t-1}) \\
& && \hat{x}_t = Vh_t \\
&\text{Predictive loss function :} && L = \sum_t L_t = \sum_t \|\hat{x}_t - x_t\|^2 \\
&\text{Learning rules :} && \delta x_t = x_t - \hat{x}_t, \delta h_t = U\delta x \\
& && \Delta W \propto -\ \text{diag}\left(1 - h_t^2\right) \delta h_t\, h_{t-1}^T \\
& && \Delta V \propto -\ \delta x_t\, h_t^T \\
& && \Delta U \propto -\ \text{diag}\left(1 - h_t^2\right) \delta h_t\, x_{t-1}^T
\end{aligned} \qquad \text{(Equation 2)}$$

where $\text{diag}(y)$ is a diagonal matrix with its diagonal equal to the vector argument y. A derivation of these learning algorithms is given in the STAR Methods.

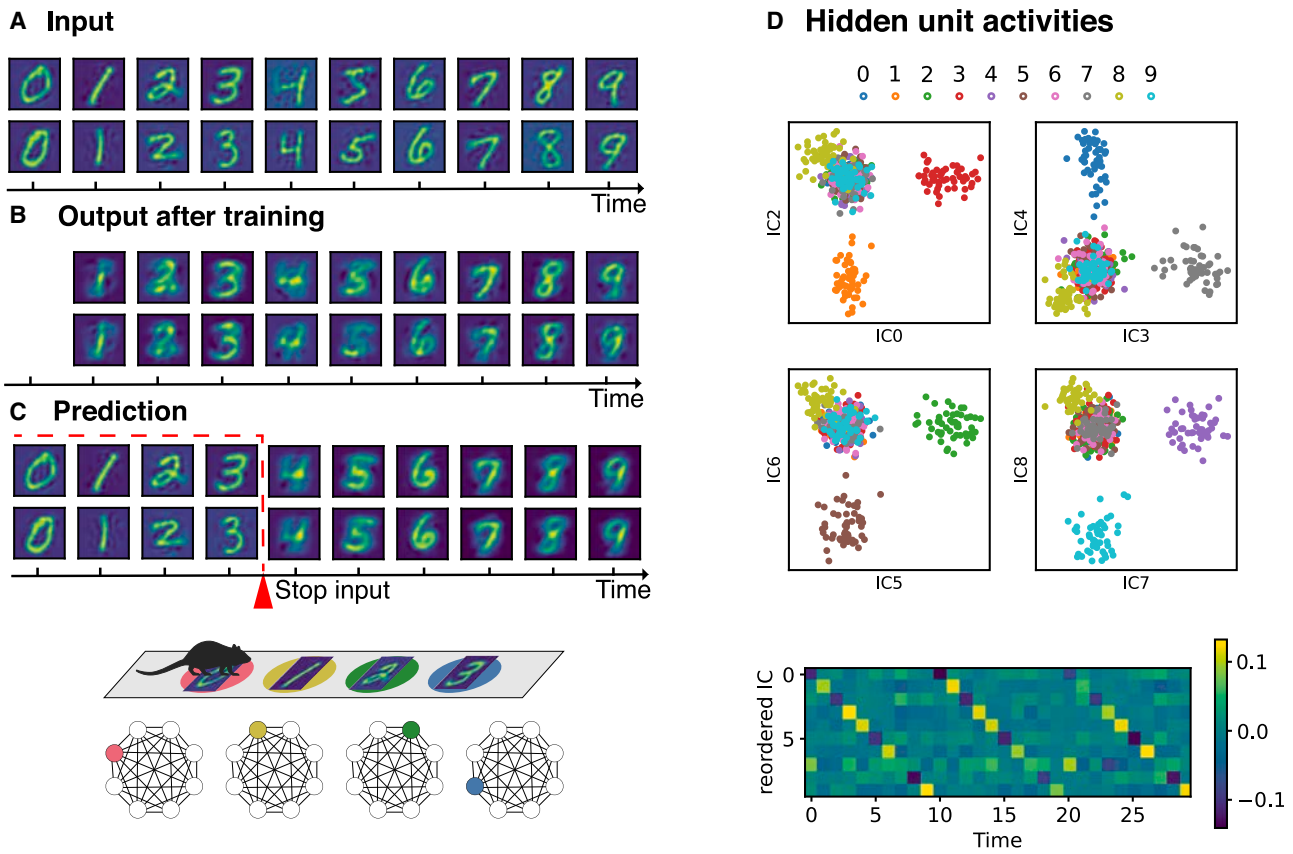

Figure 6. Predictive networks compress visual cues into sequential activation of hippocampal neurons

(A) Top: input organized as repetitive sequences of MNIST images from 0 to 9. Two sequences show individual differences.

(B) Trained network initiated with the first input could reconstruct the subsequent generic digits.

(C) When the input was stopped after digit 3, the network continued to predict the rest of the digits. Input digits were plotted before the red dashed line, while predictions were plotted afterward.

(D) Top: independent components (ICs) of hidden unit activity. Most ICs represent one class (color-coded); bottom: reordered IC activities over time. Note the sequential activation in 10 time steps (one cycle). ICA does not extract a unique sign, so diagonal entries can have both large positive and negative values. See also Figure S5.

DISCUSSION

The temporal predictive coding framework proposed here to account for sequence memory and representation learning was inspired by the anatomy of the hippocampus, validated by neural recordings, and successfully replicated a variety of experimental observations.

This theoretical framework makes several experimental predictions. First, the ablation of the direct pathway is expected to suppress the formation of new place fields upon remapping. This is partially supported by findings from Grienberger and Magee,[56] where the optogenetic inhibition of EC3 input activity led to a significant reduction in experience-dependent shaping of CA1 representations. Second, our analysis suggests that the significance of CA3 predictions may grow slowly during the early stages of sequential learning, requiring multiple epochs of training to achieve accurate prediction. Third, our model makes detailed predictions that could be tested with simultaneous long-term recordings from CA1, CA3, and EC recordings before and after learning a sequential task. Analysis of time course of coupling between different regions would reveal the amount of prediction and how it changes over time.

We found that training on sequences yielded a recurrent network with a sparse Toeplitz form that can store multiple sequences. Toeplitz matrices have a diagonal structure that performs a matrix temporal convolution. Toeplitz matrices also support traveling waves, which speed up learning of sequential tasks by two orders of magnitude and over much longer timescales.[57] The Toeplitz convolutional kernel underlies moving-bump line attractor dynamics in recurrent neural networks, including the dynamics of network models for compass cells in rodents,[58] neural integrators, and other neural systems.[59,60] A connectomic analysis of the rodent area CA3 could potentially confirm the predicted Toeplitz connectivity, providing further validation for our proposed model.

Predictive loss functions are routinely used in state-space models, such as model-based control and Bayesian filtering.

Please cite this article in press as: Chen et al., Predictive sequence learning in the hippocampal formation, Neuron (2024), https://doi.org/10.1016/j.neuron.2024.05.024

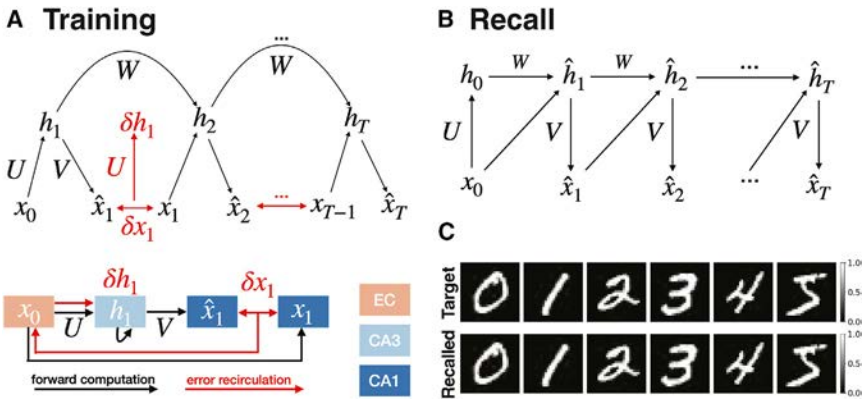

Figure 7. Predictive recurrent network that can learn and playback sequences of images using biologically plausible local learning rules

(A) During training, inputs x_t project to the hidden units h_t in the recurrent network. Three local learning algorithms update the weights U from the inputs to the hidden units, the weights V that feedback to the inputs, and the weights W between the hidden units in the recurrent network (STAR Methods). All three weight updates can be computed by prediction error δx and its recirculated error $\delta h = U\delta x$ through the feedback pathway from CA1 to EC.

(B) During recall, an initial input x_0 generates a sequence of outputs from the hidden layer.

(C) Example of a network trained by local learning algorithms on an MNIST sequence of handwritten digits (target). The output following the 0 input without further input replays the sequence in the trained order (recalled). See the table in the STAR Methods, network training, for details about the network and training hyperparameters.

See also Figure S7.

By predicting the future observations, generative models are continually updated to make more accurate predictions.[61] This approach has recently been incorporated in several model-based artificial intelligence systems.[62,63] The performance of these systems is more robust and has superior generalizability when there is an internal model of the system. Our predictive model-based approach to the hippocampus has similar advantages and is supported by evidence from neural recordings.

Recanatesi et al.[62] also explored predictive network models with both state and action as inputs to predict the state on the next time step. They demonstrated that representation compression, including localized representation, could be achieved through the addition of action signals. However, the evidence for action signal encoding in the EC, the main input to the hippocampus, is minimal. Without the action input, we also observed compression of redundant representation in Figures 5 and 6 as long as the input signal is redundant, indicating that this is a robust computational advantage inherent in having a predictive loss function. Our predictive model without action aligns closely to a one-step successor representation.[64] Building on the extensive exploration of hippocampal neuron firing properties and their theoretical underpinnings,[19–21,54,65,66] our model contributes a more mechanistic perspective, grounded in the analysis of neural recordings from hippocampal subregions. The simplicity and minimal realization of our model could also serve as a critical building block for most statistical inference models proposed for hippocampus.

Representing prediction errors in some CA1 neurons could not only facilitate the learning of the internal model stored in the recurrent CA3 network, but it could also regulate the release of dopamine in novelty-dependent firing of cells in ventral tegmental area (VTA) through subiculum, accumbens, and ventral pallidum.[67] This could explain why novelty detection is an essential function of the hippocampus. Temporal prediction error has already been established for learning sequences of actions in the basal ganglia to obtain future rewards.[68] If temporal predictive coding principles for learning sequences are also found in the cortex, as suggested in Figure 8, then predicting the next input in every cortical area may be an important design principle for human cortical function.

Self-supervised models, such as variational autoencoders,[72] and unsupervised Boltzmann machines[73,74] and their many variants have avoided labor-intensive supervised input labeling. The recent success of self-supervised transformers like GPT were trained by predicting the next word in a sentence.[75,76] These sophisticated models implicitly learn semantic representations. In the same way, our predictive network achieves representation learning of sequences without needing sophisticated statistical priors or explicitly defined representation modules. Its minimal model structure facilitates detailed investigation and interpretation. Future work could focus on scaling up the network or adding preprocessing modules to handle more realistic problems such as the semantic segmentation of video clips.

There is an analogy between the superficial and deep layers of the six-layered neocortex with areas CA3 and CA1 of the hippocampus, respectively. This is illustrated in Figure 8. Upon comparison with Figure 1, parallels emerge between the indirect pathway through the DG in the hippocampus, projecting to CA3, and the thalamic inputs to layer 4 of the neocortex, projecting to layers 2/3. Similarly, the direct pathway from the EC to area CA1 in the hippocampus corresponds in the cortex to direct inputs from the thalamus to layer 5. As in the DG, layer 4 neurons are small and numerous, creating input representation that separates similar patterns; neurons in layers 2/3 form a highly recurrent network, similar to that in CA3; neurons in layer 5 are output neurons, like CA1 neurons. This similarity has been noted by others (D. Feldman, personal communication). We go further and suggest that all cortical areas, as well as the hippocampus, may be predictive autoencoders.

In this cortical model, the recurrent network in layers 2/3 is trained as a predictive autoencoder to remember sequences of inputs arising from the thalamus, which, like the EC in the

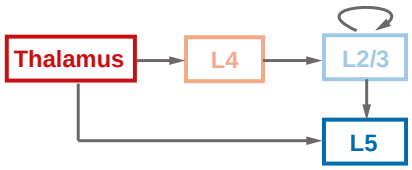

Figure 8. Universality of the circuit for computing temporal prediction error in the cortex

Signals from thalamus reach cortex layer 5 through two different pathways: the direct pathway[69] and the indirect pathway via layer 4 and recurrent layers 2/3.[70] This cortical circuit resembles the pathways in the hippocampus (Figure 1): the small stellate cells in layer 4 may have the same preprocessing function as DG granule cells; recurrent layers 2/3 learns sequences by making temporal predictions; pyramidal neurons in layer 5 compute the temporal prediction error between these direct and indirect pathways and propagated globally to subcortical structures and through layer 6 to the reticular nucleus of the thalamus. A hierarchy of sequences learned by temporal prediction may be computed in the neocortex,[71] since the canonical circuit is similar throughout.

hippocampal model, serves as the input and output of the autoencoder. The latent sequences are further compressed in the downstream areas, becoming more abstract while ascending the hierarchy. The EC in Figure 8 is at the top of converging hierarchies of cortical areas, each recapitulating the same architecture, forming stacks of predictive autoencoders.

This generalization of our hippocampal model could be tested by recording simultaneously from the thalamus, layer 4, layers 2/3, and layer 5 and analyzing the spikes the same way we analyzed recordings from the corresponding areas of the hippocampus, DG, CA3, and CA1.

If temporal predictive coding principles for learning sequences are found in the hippocampus and the cortex, as suggested in Figure 8, then predicting the next input in every cortical area may be an important design principle for human cortical function. Predictive learning using conserved circuits could underlie the robustness and flexibility of human intelligence. Transformers in large language models achieve remarkable performance with predictive self-supervised learning. Inspired by brains, our model has potential for further improvements in robustly disentangling representations in artificial intelligence and approaching human levels of performance.

Although we used BPTT as a way to construct networks that learn sequences, we showed that it could potentially be replaced by local learning rules by combining multiple biologically plausible learning algorithms that we call predictive recirculation. Our local learning rule computes a temporally truncated version of the gradient computed by BPTT. It might nonetheless be difficult to accumulate gradients for long sequences. It is possible that the pathway from the EC to CA3 might facilitate the learning of longer sequences. The addition of grid cells from EC might also make it easier to learn long sequences.[77] CA3 may also behave like a reservoir,[78] generating a wide range of time varying signals, and the prediction error signal could be used to select inputs and weight them to reduce the prediction error. We will pursue these possibilities in a subsequent study.

STAR★METHODS

Detailed methods are provided in the online version of this paper and include the following:

- KEY RESOURCES TABLE
- RESOURCE AVAILABILITY
  - Lead contact
  - Materials availability
  - Data and code availability
- METHOD DETAILS
  - Neural evidence of transmission delay and predicting ahead
  - Analysis of CA1 activity with respect to environment familiarity
  - Network training
  - Localization
  - Learning MNIST sequences
  - Predictive recirculation: A biologically plausible learning algorithm
- QUANTIFICATION AND STATISTICAL ANALYSIS

SUPPLEMENTAL INFORMATION

Supplemental information can be found online at https://doi.org/10.1016/j.neuron.2024.05.024.

ACKNOWLEDGMENTS

This work has been supported by DARPA W911NF1820, ONR N00014-23-1-2069, and the Swartz Foundation. We thank Homero Esmeraldo and Vikrant Jaltare for early exploration of the project. We thank Jorge Aldana for technical support. We thank Dr. David Kleinfeld and Dr. Johnatan (Yonatan) Aljadeff for scientific input.

AUTHOR CONTRIBUTIONS

Y.C., H.Z., and T.S. conceptualized the study and wrote and revised the manuscript; Y.C. simulated the computational models, and Y.C. and H.Z. analyzed the experimental data; M.C. designed and simulated the models using local learning rules.

DECLARATION OF INTERESTS

The authors declare no competing interests.

DECLARATION OF GENERATIVE AI AND AI-ASSISTED TECHNOLOGIES IN THE WRITING PROCESS

During the preparation of this work the author(s) used ChatGPT to scan for grammar errors and typos. After using this tool/service, the author(s) reviewed and edited the content as needed and take(s) full responsibility for the content of the publication.

Received: April 23, 2023
Revised: January 21, 2024
Accepted: May 22, 2024
Published: June 24, 2024

REFERENCES

1. O'Keefe, J., and Recce, M.L. (1993). Phase relationship between hippocampal place units and the EEG theta rhythm. Hippocampus *3*, 317–330. https://doi.org/10.1002/hipo.450030307.
2. Skaggs, W.E., McNaughton, B.L., Wilson, M.A., and Barnes, C.A. (1996). Theta phase precession in hippocampal neuronal populations and the compression of temporal sequences. Hippocampus *6*, 149–172. https://doi.org/10.1002/(SICI)1098-1063(1996)6:2<149::AID-HIPO6>3.0.CO;2-K.
3. O'Keefe, J., and Burgess, N. (2005). Dual phase and rate coding in hippocampal place cells: theoretical significance and relationship to

entorhinal grid cells. Hippocampus *15*, 853–866. https://doi.org/10.1002/hipo.20115.

4. Buzsáki, G. (2006). Rhythms of the Brain (Oxford University Press).
5. Markram, H., Lübke, J., Frotscher, M., and Sakmann, B. (1997). Regulation of synaptic efficacy by coincidence of postsynaptic APs and EPSPs. Science *275*, 213–215. https://doi.org/10.1126/science.275.5297.213.
6. Bi, G.Q., and Poo, M.M. (1998). Synaptic modifications in cultured hippocampal neurons: dependence on spike timing, synaptic strength, and postsynaptic cell type. J. Neurosci. *18*, 10464–10472. https://doi.org/10.1523/JNEUROSCI.18-24-10464.1998.
7. Abbott, L.F., and Nelson, S.B. (2000). Synaptic plasticity: taming the beast. Nat. Neurosci. *3*, 1178–1183. https://doi.org/10.1038/81453.
8. O'Keefe, J., and Nadel, L. (1978). The Hippocampus as a Cognitive Map (Oxford University Press).
9. Gauthier, J.L., and Tank, D.W. (2018). A dedicated population for reward coding in the hippocampus. Neuron *99*, 179–193.e7. https://doi.org/10.1016/j.neuron.2018.06.008.
10. Aronov, D., Nevers, R., and Tank, D.W. (2017). Mapping of a non-spatial dimension by the hippocampal–entorhinal circuit. Nature *543*, 719–722. https://doi.org/10.1038/nature21692.
11. Eichenbaum, H., Kuperstein, M., Fagan, A., and Nagode, J. (1987). Cue-sampling and goal-approach correlates of hippocampal unit activity in rats performing an odor-discrimination task. J. Neurosci. *7*, 716–732. https://doi.org/10.1523/JNEUROSCI.07-03-00716.1987.
12. Buzsáki, G., and Tingley, D. (2018). Space and time: The hippocampus as a sequence generator. Trends Cogn. Sci. *22*, 853–869. https://doi.org/10.1016/j.tics.2018.07.006.
13. Eichenbaum, H. (2014). Time cells in the hippocampus: a new dimension for mapping memories. Nat. Rev. Neurosci. *15*, 732–744. https://doi.org/10.1038/nrn3827.
14. Manns, J.R., Howard, M.W., and Eichenbaum, H. (2007). Gradual changes in hippocampal activity support remembering the order of events. Neuron *56*, 530–540. https://doi.org/10.1016/j.neuron.2007.08.017.
15. Pastalkova, E., Itskov, V., Amarasingham, A., and Buzsáki, G. (2008). Internally generated cell assembly sequences in the rat hippocampus. Science *321*, 1322–1327. https://doi.org/10.1126/science.1159775.
16. Nieh, E.H., Schottdorf, M., Freeman, N.W., Low, R.J., Lewallen, S., Koay, S.A., Pinto, L., Gauthier, J.L., Brody, C.D., and Tank, D.W. (2021). Geometry of abstract learned knowledge in the hippocampus. Nature *595*, 80–84. https://doi.org/10.1038/s41586-021-03652-7.
17. Wilson, M.A., and McNaughton, B.L. (1994). Reactivation of hippocampal ensemble memories during sleep. Science *265*, 676–679. https://doi.org/10.1126/science.8036517.
18. Skaggs, W., Mcnaughton, B., and Gothard, K. (1992). An information-theoretic approach to deciphering the hippocampal code. In Advances in Neural Information Processing Systems, S. Hanson, J. Cowan, and C. Giles, eds. (Morgan-Kaufmann).
19. Stachenfeld, K.L., Botvinick, M.M., and Gershman, S.J. (2017). The hippocampus as a predictive map. Nat. Neurosci. *20*, 1643–1653. https://doi.org/10.1038/nn.4650.
20. Whittington, J.C., Muller, T.H., Mark, S., Chen, G., Barry, C., Burgess, N., and Behrens, T.E. (2020). The Tolman-Eichenbaum machine: Unifying space and relational memory through generalization in the hippocampal formation. Cell *183*, 1249–1263.e23. https://doi.org/10.1016/j.cell.2020.10.024.
21. McNamee, D.C., Stachenfeld, K.L., Botvinick, M.M., and Gershman, S.J. (2021). Flexible modulation of sequence generation in the entorhinal–hippocampal system. Nat. Neurosci. *24*, 851–862. https://doi.org/10.1038/s41593-021-00831-7.
22. George, D., Rikhye, R.V., Gothoskar, N., Guntupalli, J.S., Dedieu, A., and Lázaro-Gredilla, M. (2021). Clone-structured graph representations enable flexible learning and vicarious evaluation of cognitive maps. Nat. Commun. *12*, 2392. https://doi.org/10.1038/s41467-021-22559-5.
23. Dong, C., Madar, A.D., and Sheffield, M.E. (2021). Distinct place cell dynamics in CA1 and CA3 encode experience in new environments. Nat. Commun. *12*, 2977. https://doi.org/10.1038/s41467-021-23260-3.
24. Shin, J., Lee, H.-W., Jin, S.-W., and Lee, I. (2022). Subtle visual change in a virtual environment induces heterogeneous remapping systematically in CA1, but not CA3. Cell Rep. *41*, 111823. https://doi.org/10.1016/j.celrep.2022.111823.
25. Lee, I., Yoganarasimha, D., Rao, G., and Knierim, J.J. (2004). Comparison of population coherence of place cells in hippocampal subfields CA1 and CA3. Nature *430*, 456–459. https://doi.org/10.1038/nature02739.
26. Lee, I., Rao, G., and Knierim, J.J. (2004). A double dissociation between hippocampal subfields: Differential time course of CA3 and CA1 place cells for processing changed environments. Neuron *42*, 803–815. https://doi.org/10.1016/j.neuron.2004.05.010.
27. Kumaran, D., and Maguire, E.A. (2006). An unexpected sequence of events: Mismatch detection in the human hippocampus. PLOS Biol. *4*, e424. https://doi.org/10.1371/journal.pbio.0040424.
28. Knight, R.T. (1996). Contribution of human hippocampal region to novelty detection. Nature *383*, 256–259. https://doi.org/10.1038/383256a0.
29. Duncan, K., Ketz, N., Inati, S.J., and Davachi, L. (2012). Evidence for area CA1 as a match/mismatch detector: A high-resolution fMRI study of the human hippocampus. Hippocampus *22*, 389–398. https://doi.org/10.1002/hipo.20933.
30. Dragoi, G., and Tonegawa, S. (2011). Preplay of future place cell sequences by hippocampal cellular assemblies. Nature *469*, 397–401. https://doi.org/10.1038/nature09633.
31. Schapiro, A.C., Turk-Browne, N.B., Botvinick, M.M., and Norman, K.A. (2017). Complementary learning systems within the hippocampus: A neural network modelling approach to reconciling episodic memory with statistical learning. Philos. Trans. R. Soc. Lond. B Biol. Sci. *372*, 20160049. https://doi.org/10.1098/rstb.2016.0049.
32. Lotter, W., Kreiman, G., and Cox, D. (2020). A neural network trained for prediction mimics diverse features of biological neurons and perception. Nat. Mach. Intell. *2*, 210–219. https://doi.org/10.1038/s42256-020-0170-9.
33. Rao, R.P., and Ballard, D.H. (1999). Predictive coding in the visual cortex: A functional interpretation of some extra-classical receptive-field effects. Nat. Neurosci. *2*, 79–87. https://doi.org/10.1038/4580.
34. Hopfield, J.J. (1982). Neural networks and physical systems with emergent collective computational abilities. Proc. Natl. Acad. Sci. USA *79*, 2554–2558. https://doi.org/10.1073/pnas.79.8.2554.
35. Sabatini, B.L., and Regehr, W.G. (1999). Timing of synaptic transmission. Annu. Rev. Physiol. *61*, 521–542. https://doi.org/10.1146/annurev.physiol.61.1.521.
36. Leung, L.S., Roth, L., and Canning, K.J. (1995). Entorhinal inputs to hippocampal CA1 and dentate gyrus in the rat: A current-source-density study. J. Neurophysiol. *73*, 2392–2403. https://doi.org/10.1152/jn.1995.73.6.2392.
37. Kalman, R.E. (1960). A new approach to linear filtering and prediction problems. Journal of Basic Engineering *82*, 35–45. https://doi.org/10.1115/1.3662552.
38. Siegle, J.H., Jia, X., Durand, S., Gale, S., Bennett, C., Graddis, N., Heller, G., Ramirez, T.K., Choi, H., Luviano, J.A., et al. (2021). Survey of spiking in the mouse visual system reveals functional hierarchy. Nature *592*, 86–92. https://doi.org/10.1038/s41586-020-03171-x.
39. Bryson, A.E. (1961). A gradient method for optimizing multi-stage allocation processes (Proceedings of the Harvard University Symposium on Digital Computers and Their Applications).
40. Werbos, P.J. (1990). Backpropagation through time: What it does and how to do it. Proceedings of the IEEE *78*, 1550–1560. https://doi.org/10.1109/5.58337.

41. Rumelhart, D.E., Hinton, G.E., and Williams, R.J. (1986). Learning representations by back-propagating errors. Nature *323*, 533–536. https://doi.org/10.1038/323533a0.

42. Johnson, A., and Redish, A.D. (2007). Neural ensembles in CA3 transiently encode paths forward of the animal at a decision point. J. Neurosci. *27*, 12176–12189. https://doi.org/10.1523/JNEUROSCI.3761-07.2007.

43. Pfeiffer, B.E., and Foster, D.J. (2013). Hippocampal place-cell sequences depict future paths to remembered goals. Nature *497*, 74–79. https://doi.org/10.1038/nature12112.

44. Eliasmith, C. (2007). Attractor network. Scholarpedia *2*, 1380. https://doi.org/10.4249/scholarpedia.1380.

45. Kim, R., Li, Y., and Sejnowski, T.J. (2019). Simple framework for constructing functional spiking recurrent neural networks. Proc. Natl. Acad. Sci. USA *116*, 22811–22820. https://doi.org/10.1073/pnas.1905926116.

46. Tsodyks, M.V., Skaggs, W.E., Sejnowski, T.J., and McNaughton, B.L. (1996). Population dynamics and theta rhythm phase precession of hippocampal place cell firing: A spiking neuron model. Hippocampus *6*, 271–280. https://doi.org/10.1002/(SICI)1098-1063(1996)6:3<271::AID-HIPO5>3.0.CO;2-Q.

47. Leutgeb, S., Leutgeb, J.K., Treves, A., Moser, M.-B., and Moser, E.I. (2004). Distinct ensemble codes in hippocampal areas CA3 and CA1. Science *305*, 1295–1298. https://doi.org/10.1126/science.1100265.

48. Davoudi, H., and Foster, D.J. (2019). Acute silencing of hippocampal CA3 reveals a dominant role in place field responses. Nat. Neurosci. *22*, 337–342. https://doi.org/10.1038/s41593-018-0321-z.

49. Brun, V.H., Leutgeb, S., Wu, H.-Q., Schwarcz, R., Witter, M.P., Moser, E.I., and Moser, M.-B. (2008). Impaired spatial representation in CA1 after lesion of direct input from entorhinal cortex. Neuron *57*, 290–302. https://doi.org/10.1016/j.neuron.2007.11.034.

50. Brun, V.H., Otnass, M.K., Molden, S., Steffenach, H.-A., Witter, M.P., Moser, M.-B., and Moser, E.I. (2002). Place cells and place recognition maintained by direct entorhinal-hippocampal circuitry. Science *296*, 2243–2246. https://doi.org/10.1126/science.1071089.

51. Nakashiba, T., Young, J.Z., McHugh, T.J., Buhl, D.L., and Tonegawa, S. (2008). Transgenic inhibition of synaptic transmission reveals role of CA3 output in hippocampal learning. Science *319*, 1260–1264. https://doi.org/10.1126/science.1151120.

52. Priestley, J.B., Bowler, J.C., Rolotti, S.V., Fusi, S., and Losonczy, A. (2022). Signatures of rapid plasticity in hippocampal CA1 representations during novel experiences. Neuron *110*, 1978–1992.e6. https://doi.org/10.1016/j.neuron.2022.03.026.

53. Bittner, K.C., Milstein, A.D., Grienberger, C., Romani, S., and Magee, J.C. (2017). Behavioral time scale synaptic plasticity underlies CA1 place fields. Science *357*, 1033–1036. https://doi.org/10.1126/science.aan3846.

54. Benna, M.K., and Fusi, S. (2021). Place cells may simply be memory cells: Memory compression leads to spatial tuning and history dependence. Proc. Natl. Acad. Sci. USA *118*, e2018422118. https://doi.org/10.1073/pnas.2018422118.

55. Hinton, G.E., and McClelland, J. (1987). Learning representations by recirculation. In Neural Information Processing Systems, D. Anderson, ed. (American Institute of Physics).

56. Grienberger, C., and Magee, J.C. (2022). Entorhinal cortex directs learning-related changes in CA1 representations. Nature *611*, 554–562. https://doi.org/10.1038/s41586-022-05378-6.

57. Keller, T.A., Muller, L., Sejnowski, T., and Welling, M. (2023). Traveling waves encode the recent past and enhance sequence learning. Preprint at arXiv. https://doi.org/10.48550/arXiv.2309.08045.

58. Zhang, K. (1996). Representation of spatial orientation by the intrinsic dynamics of the head-direction cell ensemble: A theory. J. Neurosci. *16*, 2112–2126. https://doi.org/10.1523/JNEUROSCI.16-06-02112.1996.

59. Gardner, R.J., Hermansen, E., Pachitariu, M., Burak, Y., Baas, N.A., Dunn, B.A., Moser, M.-B., and Moser, E.I. (2022). Toroidal topology of population activity in grid cells. Nature *602*, 123–128. https://doi.org/10.1038/s41586-021-04268-7.

60. Khona, M., and Fiete, I.R. (2022). Attractor and integrator networks in the brain. Nat. Rev. Neurosci. *23*, 744–766. https://doi.org/10.1038/s41583-022-00642-0.

61. Isomura, T., and Toyoizumi, T. (2021). Dimensionality reduction to maximize prediction generalization capability. Nat. Mach. Intell. *3*, 434–446. https://doi.org/10.1038/s42256-021-00306-1.

62. Recanatesi, S., Farrell, M., Lajoie, G., Deneve, S., Rigotti, M., and Shea-Brown, E. (2021). Predictive learning as a network mechanism for extracting low-dimensional latent space representations. Nat. Commun. *12*, 1417. https://doi.org/10.1038/s41467-021-21696-1.

63. Lotter, W., Kreiman, G., and Cox, D. (2016). Deep predictive coding networks for video prediction and unsupervised learning. Preprint at arXiv. https://doi.org/10.48550/arXiv.1605.08104.

64. Fang, C., Aronov, D., Abbott, L.F., and Mackevicius, E.L. (2023). Neural learning rules for generating flexible predictions and computing the successor representation. eLife *12*, e80680. https://doi.org/10.7554/eLife.80680.

65. Basu, J., and Siegelbaum, S.A. (2015). The corticohippocampal circuit, synaptic plasticity, and memory. Cold Spring Harb. Perspect. Biol. *7*, a021733. https://doi.org/10.1101/cshperspect.a021733.

66. Klausberger, T. (2009). GABAergic interneurons targeting dendrites of pyramidal cells in the CA1 area of the hippocampus. Eur. J. Neurosci. *30*, 947–957. https://doi.org/10.1111/j.1460-9568.2009.06913.x.

67. Lisman, J.E., and Grace, A.A. (2005). The hippocampal-VTA loop: Controlling the entry of information into long-term memory. Neuron *46*, 703–713. https://doi.org/10.1016/j.neuron.2005.05.002.

68. Watabe-Uchida, M., Eshel, N., and Uchida, N. (2017). Neural circuitry of reward prediction error. Annu. Rev. Neurosci. *40*, 373–394. https://doi.org/10.1146/annurev-neuro-072116-031109.

69. Gökçe, O., Bonhoeffer, T., and Scheuss, V. (2016). Clusters of synaptic inputs on dendrites of layer 5 pyramidal cells in mouse visual cortex. eLife *5*, e09222. https://doi.org/10.7554/eLife.09222.

70. Markov, N.T., Ercsey-Ravasz, M., Van Essen, D.C., Knoblauch, K., Toroczkai, Z., and Kennedy, H. (2013). Cortical high-density counterstream architectures. Science *342*, 1238406. https://doi.org/10.1126/science.1238406.

71. Haeusler, S., and Maass, W. (2007). A statistical analysis of information-processing properties of lamina-specific cortical microcircuit models. Cereb. Cortex *17*, 149–162. https://doi.org/10.1093/cercor/bhj132.

72. Kingma, D.P., and Welling, M. (2019). An introduction to variational autoencoders. Foundations and Trends in Machine Learning *12*, 307–392. https://doi.org/10.1561/2200000056.

73. Ackley, D.H., Hinton, G.E., and Sejnowski, T.J. (1985). A learning algorithm for Boltzmann machines. Cogn. Sci. *9*, 147–169.

74. Hinton, G.E., and Sejnowski, T.J. (1986). Learning and relearning in Boltzmann machines. In Parallel Distributed Processing: Explorations in the Microstructure of Cognition, *1*, M.I. Jordan and T.J. Sejnowski, eds. *(MIT Press), p. 2.*

75. Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A.N., Kaiser, L., and Polosukhin, I. (2017). Attention is all you need. In Advances in Neural Information Processing Systems, *30*, I. Guyon, U.V. Luxburg, S. Bengio, H. Wallach, R. Fergus, S. Vishwanathan, and R. Garnett, eds. *(Curran Associates, Inc).*

76. Sejnowski, T.J. (2023). Large language models and the reverse Turing test. Neural Comput. *35*, 309–342. https://doi.org/10.1162/neco_a_01563.

77. Chandra, S., Sharma, S., Chaudhuri, R., and Fiete, I. (2023). High-capacity flexible hippocampal associative and episodic memory enabled by

prestructured "spatial" representations. Preprint at bioRxiv. https://doi.org/10.1101/2023.11.28.568960.

78. Jaeger, H., and Haas, H. (2004). Harnessing nonlinearity: Predicting chaotic systems and saving energy in wireless communication. Science *304*, 78–80. https://doi.org/10.1126/science.1091277.

79. Mizuseki, K., Sirota, A., Pastalkova, E., Diba, K., and Buzsáki, G. (2013). Multiple single unit recordings from different rat hippocampal and entorhinal regions while the animals were performing multiple behavioral tasks. CRCNS.org. https://doi.org/10.6080/K09G5JRZ.

80. Mizuseki, K., Diba, K., Pastalkova, E., Teeters, J., Sirota, A., and Buzsáki, G. (2014). Neurosharing: Large-scale data sets (spike, LFP) recorded from the hippocampal-entorhinal system in behaving rats. F1000Research *3*, 98.

STAR★METHODS

KEY RESOURCES TABLE

REAGENT or RESOURCE	SOURCE	IDENTIFIER
Deposited data		
CRCNS, hc-6	https://doi.org/10.6080/K0NK3BZJ	https://crcns.org/data-sets/hc/hc-6/about-hc-5 (Retrieved April 2024)
Allen Visual Coding Dataset	Allen Institute	https://portal.brain-map.org/circuits-behavior/visual-coding-neuropixels (Retrieved April 2024)
Software and algorithms		
Python version 3.6.7	Python	https://www.python.org
MATLAB	MATLAB	https://matlab.mathworks.com
PyTorch 1.4.0	PyTorch	https://pytorch.org
Analysis scripts v1.0.0	Contributed by Y.C.	https://github.com/yschen13/HCPrediction; DOI: https://doi.org/10.5281/zenodo.10989139

RESOURCE AVAILABILITY

Lead contact

Further information and requests for data and code should be directed to and will be fulfilled by the lead contact Terrence Sejnowski (terry@salk.edu).

Materials availability

No material was generated in this study.

Data and code availability

- This paper analyzes existing, publicly available data. These accession numbers for the datasets are listed in the key resources table. The Allen NeuroPixel and CRCNS (hc-6) datasets used in the analysis is publicly available through their website. All data reported in this paper will be shared by the lead contact upon request.
- All neural data analysis scripts and simulation scripts are available in https://github.com/yschen13/HCPrediction with the release version https://doi.org/10.5281/zenodo.10989139.
- Any additional information required to reanalyze the data reported in this paper is available from the lead contact upon request.

METHOD DETAILS

Neural evidence of transmission delay and predicting ahead

We used the publicly accessible visual encoding NeuroPixel dataset[38] from the Allen Brain Observatory. Neuropixels were used to simultaneously record the spiking activity of thousands of neurons in mice passively perceiving standard visual stimuli such as drifting gratings, natural scenes, natural movies, etc. We pre-selected recording sessions that involves recordings from DG, CA3 and CA1 in "Functional Connectivity" in WT mice. Very few neurons were recorded from EC. Number of units being used was summarized in the table below.

Session	CA1	CA3	DG
766640955	170	17	62
771160300	282	48	32
767871931	104	20	32
768515987	102	27	25
771990200	70	6	31
778240327	251	24	42

(*Continued on next page*)

Continued

Session	CA1	CA3	DG
778998620	133	45	16
779839471	142	25	49
781842082	114	19	28
793224716	159	31	29
821695405	56	17	40
847657808	189	6	70

For mutual information calculation, only recordings from the natural movie viewing sessions (30 s × 80 repeats) were used while for cross correlation, recordings from all stimuli sessions were concatenated to increase signal-to-noise ratio.

The processed spike train was binned at 2 ms. To compute cross-correlogram (CCG) between N neurons in region A and M neurons in region B, we first calculated jitter-corrected correlograms[38] between $N \times M$ neuron pairs using a jittering window of 20 ms. Jitter-correction was performed by randomly shuffling the spike train within the chosen time window, calculating the jitter-CCG repeatedly for 100 times, and then subtracting its average from the original CCG. Corrected-CCG was further normalized by the geometric firing rate of the neuron pair. In this way, slower time scale correlations, such as the strong theta oscillation in hippocampus or nonstationary trend could be removed and then we could focus on fast time scale neural coupling. To increase signal-to-noise ratio for prediction of one neuron in region A, we used the CCG average of all M neurons in region B. Time shifting was performed in region B neurons. For a spike train denoted by $f(t)$, a positive τ shift would lead to a rightward shifted spike train of $f(t - \tau)$. The optimal time shift is defined as the time shift that maximizes the M-to-1 averaged cross correlation. We focused on time shifts from -20 ms to 20 ms as any shifted coupling above this range would be scrambled by jittering.

We compute the mutual information between the spike train of one neuron from region A and the shifted spike trains of 10 neurons from region B. The one-dimensional spike train from region A was treated as a random variable A. The latter high-dimensional spike train is treated as a random vector B which has 2^{10} states being sampled at different time steps. The mutual information is then calculated as $I(A;B) = H(B) - H(B|A)$. For each neuron in A, we compute the information between that neuron and 10 neurons in B and repeat 100 times for different randomly sampled subsets of 10 recorded neurons from B. To show the results for one neuron in A, for each subset of 10 neurons from B, information over time shift is normalized by its maximal value. Then the average of the 100 normalized information curves is taken to reveal the effect of time shift on the mutual information. The optimal time shift is defined as the time shift that maximizes mutual information.

Analysis of CA1 activity with respect to environment familiarity

Datasets were obtained from http://crcns.org/data-sets/hc/hc-3, contributed by the Buzsáki laboratory at New York University.[79,80] See http://crcns.org/files/data/hc3/crcns-hc3-processing-flowchart.pdf for more details about experiments, recording and data pre-processing. For rats exploring a 180 × 180-cm box, all sessions that have more than 50 simultaneously recorded CA1 neurons were included for analysis. We excluded neurons that are marked as inhibitory or not identified. For each session, we compute spike rate of the neurons during the last two-thirds of the sessions for stability of the responses.

Network training

The network was implemented in PyTorch (v1.11.0) and training was performed through stochastic gradient descent of samples split into mini batches with a fixed learning rate, as shown in the below table. Gradients were calculated with backpropagation through time (BPTT). We used 'sigmoid' and 'tanh' nonlinearities for the activation of output and recurrent units, unless otherwise mentioned. We stopped training when the process reached the maximum number of epochs or the loss function reached less than 1% of its initial value and did not change more than 0.001% in 10 consecutive iterations. Network structure and related hyperparameters are summarized in the table below, where S = sample/batch size; N = number of input units; T = sequence length; H = number of hidden units; L = loss function; η = learning rate.

Task	S	N	T	H	L	η	Total epochs
Figure 4	1 or 2	200	100	200	MSE	0.01	50,000
Figure 5	50	200	100	500	MSE + Reg(h_t)	0.01	50,000
Figure 6	5	68	100	200	MSE	0.01	50,000
Figure 7	1	68	6	100	MSE	0.0001	100,000

Please cite this article in press as: Chen et al., Predictive sequence learning in the hippocampal formation, Neuron (2024), https://doi.org/10.1016/j.neuron.2024.05.024

Localization

The open arena was simulated as a 2 m × 2 m environment. Exploratory trajectory was generated as straight lines of 0.1 m step size until hitting a border. Then a random turnaround angle will be generated to continue exploration. Altogether 5,000 time steps split into 50 samples were used to train the network. Following Benna and Fusi[54] (Supplementary Equation 1), the location information (path integrated distance, distance to the closest border, world direction and head direction) was randomly and nonlinearly expanded into higher dimensions ($N = 200$) as input and target signal. We switched to 'ReLU' nonlinearity for hidden unit activation as we would like to avoid negative responses in terms of place field calculation. To enable a sparse representation, a penalty of hidden unit firing was added to the loss function (Equation 3)

$$L = \sum_t L_t = \sum_t \left(\| o_{t+1|t} - x_{t+1} \|^2 + \lambda \| h_t \|^2 \right) \quad \text{(Equation 3)}$$

Mutual information of a hidden unit place field was calculated following Skaggs et al.[18] as the mutual information between firing rate and the arena location discretized into 25 × 25 grids. Specifically, It was calculated as $\mathrm{MI} = \sum_i \lambda_i \log(\lambda_i) p_i$ in bits/second where i represents location grid, λ_i is the neuron's firing rate at location grid i and p_i is the occupancy probability in grid i.

Learning MNIST sequences

Input was constructed as the top 68 principal components (PC) of the entire MNIST dataset, which explain 87% variance. Input was organized as sequences consisting of 100 time steps, which repeats from digit 0 to digit 9 for 10 times. Five randomly sampled batches of digit images were used for training to predict the next time step PC vector. Independent component analysis (ICA) was performed to reduce the dimension of hidden unit activation from the number of hidden units to the number of chosen ICs (i.e. 10). We manually ordered the ICs by the contribution (column L2 norm) of the converged demixing matrix. For the local learning rule, the input was a single sequence consisting of 6 time steps, where the PC's were normalized to be between 0 and 1.

Predictive recirculation: A biologically plausible learning algorithm

Recirculation

The recirculation learning algorithm[74] for a three-layer feedforward autoencoder approximates gradient descent without the need to backpropagate (BP) errors under certain conditions:

$$\begin{aligned} &\text{Network dynamics :} && h = \sigma(Ux) \\ &&& \hat{x} = \lambda x + (1 - \lambda) V h \\ &&& \tilde{h} = \lambda h + (1 - \lambda) \sigma(U\hat{x}) \\ &\text{Update rules :} && \Delta V \propto -(x - \hat{x})\, h^T \\ &&& \Delta U \propto -(h - \tilde{h})\, x^T \\ &\text{To approximate BP :} && U = V^T \end{aligned} \quad \text{(Equation 4)}$$

Using this set of learning rules, the symmetry between the input and output weights (up to scaling) is almost guaranteed. A new predictive recirculation learning is derived here based on Equation 4, and assuming that $U = V^T$.

- **Output weights**. With the dynamics defined in Equation 2, the exact gradient of the output weight V can be obtained using only local information, assisted by error-encoding neurons:

$$\Delta V \propto -\frac{\partial L(t)}{\partial V} = -[x(t) - \hat{x}(t)]\, h(t)^T \quad \text{(Equation 5)}$$

- **Input weights**. The exact gradient of input weight (U) is given by:

$$\begin{aligned} \frac{\partial L(t)}{\partial U_{mn}} &= \sum_i \frac{\partial L(t)}{\partial h_i(t)} \frac{\partial h_i(t)}{\partial U_{mn}} \\ &= \left[\sum_i \frac{\partial L(t)}{\partial h_i(t)} \left(1 - h_i^2(t)\right) \delta_{im} \right] x_n(t-1) \end{aligned} \quad \text{(Equation 6)}$$

where in our simulations we chose $\sigma = \tanh$, and $\sigma' = 1 - \tanh^2$.

$$\frac{\partial L(t)}{\partial h(t)} = W^T \mathrm{diag}\left(1 - h^2(t+1)\right) \frac{\partial L(t)}{\partial h(t+1)} + V^T \frac{\partial L(t)}{\partial \hat{x}} = U\left[x(t) - \hat{x}(t)\right], \quad \text{(Equation 7)}$$

where the first term containing the temporal dependency of L with respect to W was truncated and $V^T = U$ from the original recirculation algorithm. Combining the above two equations:

$$\Delta U_{mn} \propto -\left[1 - h_m^2(t)\right] \sum_k U_{mk} \left[x_k(t) - \hat{x}_k(t)\right] x_n(t-1) \quad \text{(Equation 8)}$$

which in the vector notation is:

$$\Delta U \propto -\mathrm{diag}\left[1 - h^2(t)\right] U\left[x(t) - \hat{x}(t)\right] x^T(t-1) \quad \text{(Equation 9)}$$

This learning algorithm only involves locally available information: x, $\hat{x}$, h and U. Assuming the input to hidden weights are linear, the term $[h - \tilde{h}]$ in Equation 4 can be replaced with $U[x(t) - \hat{x}(t)]$ in Equation 9, thus making our learning rule for input and output weights approximate the recirculation learning rule described in Equation 4. As a result, the input and output weights trained from our learning algorithm is approximately the transpose of each other up to scaling (Figure S7).

- **Recurrent weight**. The exact gradient of the recurrent weight (W) is:

$$\frac{\partial L(t)}{\partial W_{mn}} = \sum_i \frac{\partial L(t)}{\partial h_i(t)} \frac{\partial h_i(t)}{\partial W_{mn}} = \sum_i \frac{\partial L(t)}{\partial h_i(t)} \left[1 - h_i(t)^2\right] \delta_{im} h_n(t-1) \quad \text{(Equation 10)}$$

Following the same derivation used to approximate $\partial L/\partial h$ in Equation 7:

$$\Delta W \propto -\mathrm{diag}\left[1 - h^2(t)\right] U\left[x(t) - \hat{x}(t)\right] h^T(t-1) \quad \text{(Equation 11)}$$

This Hebbian rule between the postsynaptic prediction error and the previous presynaptic input from a recurrent unit is a biologically plausible mechanism for updating the recurrent weights.

The coefficient $\sigma' = \mathrm{diag}[1 - h^2]$ modulates the learning rate and is only significant around threshold, acting like a gate that restricts weight change to the currently active neurons. In real neurons, this could correspond to backpropagating action potentials that gate synaptic plasticity in dendrites.

QUANTIFICATION AND STATISTICAL ANALYSIS

Statistical significance was defined by alpha pre-set to 0.01. For all panels in Figures 2 and 3, we used one-sample t test with the null hypothesis that null data comes from a normal distribution with zero mean and unknown but fixed variance. For Figure 2, we used cross-region pairwise statistics, thus the number of samples used for statistical comparison could be calculated from the table in neural evidence of transmission delay and predicting ahead. For example, for Figure 2B (CA3 vs. CA1), the number of samples can be calculated as the cell number in the second column times that in the third column and add up all rows. The exact N number is 2057, 741 and 2228 for Figures 2B–2D. For Figures 3B and 3C, the exact N number of 200. For Figures 5C and S6, we used two sample t test with the null hypothesis that the difference between points sampled from two populations are normally distributed with zero mean and fixed variance. $N = 200 \times 10$ where N stands for the pooling of recurrent units from 10 randomly initialized networks. All the statistical tests are described in the figure legends and each test was selected based on data distributions using histograms. Detailed statistical procedures are described in each subsection of method details.