PHILOSOPHICAL TRANSACTIONS B

royalsocietypublishing.org/journal/rstb

Research

Article submitted to journal

Subject Areas:
computational neuroscience, systems neuroscience, neural dynamics, learning, memory

Keywords:
hippocampus, RNN, CA3, sequence generation, replay, prediction, symmetry breaking, traveling waves, inductive bias

Author for correspondence:
Margot Wagner
e-mail: mwagner@ucsd.edu

Balancing Stability and Flow in Hippocampal Networks via Inductive Bias and Learned Symmetry Breaking

Margot Wagner[1], Yusi Chen[2], Arjun Karuvally[3], Mia Cameron[4], and Terrence J. Sejnowski[1,3]

[1]Institute for Neural Computation, University of California San Diego, La Jolla, CA
[2]Institute of Neuroscience, Chinese Academy of Sciences, Shanghai, China
[3]Computational Neurobiology Laboratory, Salk Institute for Biological Studies, La Jolla, CA
[4]Computational Neuroscience Center, University of Washington, Seattle, WA

The hippocampus must balance stable memory representations with internally generated sequential dynamics underlying replay, prediction and traveling waves. How hippocampal circuitry achieves both remains unclear. Here, we show that recurrent neural networks trained on continuous prediction tasks converge to a mixed-symmetry dynamical regime, in which dominant symmetric recurrence stabilizes a continuous attractor while a weaker antisymmetric component induces directed flow along it. This structure accelerates learning and supports robust replay and prediction. We further show that such symmetry breaking can arise from biologically plausible spike-timing-dependent plasticity (STDP) rules, yielding a tilted Mexican-hat connectivity profile without explicit architectural constraints. Importantly, initializing networks with CA3-like structured connectivity biases learning toward this regime, improving optimization efficiency and performance. These results suggest that hippocampal computation reflects an interaction between biologically constrained circuit structure and task-driven learning, with partial symmetry breaking providing a low-dimensional control principle for balancing stability and flow in sequence generation.

© The Authors. Published by the Royal Society under the terms of the Creative Commons Attribution License http://creativecommons.org/licenses/by/4.0/, which permits unrestricted use, provided the original author and source are credited.

THE ROYAL SOCIETY PUBLISHING

1. Introduction

(a) Relationship between hippocampus memory and directed sequential activation

The hippocampus plays a central role in learning, storing, and retrieving sequential experience. From its earliest characterization as a spatial map [1] to contemporary views of the hippocampus as a generator of predictive internal models [2,3], a recurring challenge has been to reconcile memory stability with temporally ordered neural activity. In particular, how can neural circuits preserve memories while also supporting structured sequences that unfolds over time? The hippocampus and, especially its sub-region CA3, have long been proposed as recurrent associative memory systems [4–6], capable of rapid pattern completion through dense and plastic recurrence. While this framework successfully explains stable memory retrieval, accumulating experimental and theoretical evidence suggests that a purely static view of hippocampal function is incomplete [7–12]. Rather than simply reinstating stored representations, hippocampal activity frequently unfolds as internally generated sequences, implying intrinsic temporal organization beyond static recall [13–19].

Converging experimental evidence supports the view that hippocampal population dynamics are inherently sequential and directional. During both sleep and awake behavior, hippocampal populations exhibit temporally ordered reactivation of place cell sequences that are biased toward behaviorally relevant past and future trajectories [6,10,13]. These replay events reflect structured progression through internal representations rather than noise-driven fluctuations around a fixed point and are thought to support functions such as prediction and planning [10,11,14]. At shorter timescales, theta phase precession, traveling population waves, and directionally biased sequence expression further indicate that hippocampal circuits operate in a regime that supports both representational persistence and controlled temporal progression [7–9,15,18,20]. These observations motivate a closer examination of classical recurrent network models focused on memory stability that have historically been used for the hippocampus.

Classical theories of recurrent neural computation emphasize symmetry as a prerequisite for stable memory. Fully symmetric recurrent networks converge to fixed-point attractors, providing a principled mechanism for content-addressable associative memory and pattern completion [21–24]. This framework strongly influenced computational theories of hippocampal area CA3, which has long been modeled as a densely recurrent, symmetric autoassociative network optimized for rapid storage and retrieval of episodic patterns [4,24,25]. Extensions of this approach to continuous attractor networks, including ring and bump models, demonstrated how symmetric, distance-dependent recurrent excitation combined with broader inhibition (Mexican-hat) can stabilize low-dimensional manifolds representing continuous variables, yielding explanations for spatial coding, path integration, and head-direction representations [17,26–30]. Despite their success in explaining persistent activity and pattern completion, these models are fundamentally time-reversible and intrinsically static as activity relaxes toward stored states rather than progressing through them in a directed manner. As a consequence, purely symmetric attractor networks lack an intrinsic mechanism for sequence propagation and struggle to account for key hippocampal phenomena involving ordered temporal structure. Introducing asymmetric recurrent interactions enables temporal transitions between memory states [26,31,32], but unstructured asymmetry generally destabilizes network dynamics, leading to drift or high-dimensional chaos rather than reliable trajectories [33]. This highlights a fundamental balance: while symmetry supports memory stability, directionality is required for sequence generation, leaving open the question of how hippocampal circuits balance these competing demands.

Together, these observations point to the need for a unifying theoretical framework that can reconcile stable memory representations with controlled temporal progression, and explain how diverse hippocampal phenomena such as attractor-like persistence, sequential replay, prediction, and wave-like dynamics can emerge from a common circuit principle.

(b) The need to naturally break symmetry through neural mechanisms

A growing body of theory suggests that controlled symmetry breaking provides the missing link between stable memory and directed temporal dynamics. In continuous attractor models, symmetric (even) recurrent interactions stabilize a low-dimensional representational manifold, while a carefully structured antisymmetric (odd) component induces smooth motion along that manifold without distorting the underlying representational structure [34]. Even modest asymmetry can thus convert a static attractor into a system capable of controlled flow through state space. This framework has been used to explain head-direction integration, phase precession, and traveling population activity, in which weak antisymmetric or non-reciprocal interactions generate drift, rotation, or waves atop a stabilizing recurrent backbone [9,15,16,34–37]. Crucially, the structure of asymmetry matters: antisymmetric components aligned with spatial derivatives preserve representational structure and produce coherent translation, whereas unstructured asymmetry degrades stability or induces chaotic dynamics [32,33].

In hippocampal models, such directed dynamics have been attributed to several interacting sources of symmetry breaking, including anatomically biased recurrent connectivity, learning-induced asymmetries shaped by temporally asymmetric plasticity, and transient dynamical effects arising from state-dependent modulation or short-term synaptic processes [10,17,35,38,38,39,39–45]. Experimental observations are consistent this view, showing that hippocampal dynamics are organized as traveling waves rather than globally synchronized attractors, consistent with internally generated flow along representational axes [15,37]. These results highlight a central unresolved problem: hippocampal circuits must break symmetry to support replay and prediction, yet do so in a constrained manner tha preserves memory stability and dynamical robustness. How these mechanisms are balanced, combined, or selectively engaged remains an open question for understanding CA3 computation.

One biologically plausible route to controlled symmetry breaking is through synaptic plasticity. In particular, mixed symmetry in recurrent CA3 connectivity may emerge from the diversity of spike-timing-dependent plasticity (STDP) rules expressed

across hippocampal synapses. Early formulations of STDP emphasize a temporally asymmetric learning window in which pre-before-post activity potentiates synapses and post-before-pre activity depresses them, based on influential neocortical and hippocampal culture experiments that established STDP as a causality-sensitive Hebbian mechanism [46,47]. However, accumulating experimental and theoretical work indicates that plasticity rules are heterogeneous across synapses, cell types, and circuit motifs, implying that hippocampal recurrent circuits, and in particular the CA3, are governed by a mixture of symmetric and asymmetric effective couplings rather than a single canonical STDP kernel.

Consistent with this view, direct measurements at CA3-CA3 recurrent excitatory synapses reveal largely symmetric STDP rule, a regime well matched to CA3's autoassociative memory storage and recall. At the same time, theoretical and modeling studies demonstrates that temporally asymmetric plasticity can coexist within hippocampal-like circuits, giving rise to both stabilizing (more symmetric) and sequence-propagating (more asymmetric) components that jointly support pattern completion and temporal prediction [38]. Recent theory further formalizes this heterogeneity, showing that subpopulations with differing degrees of temporal asymmetry in their learning rules can regulate replay speed and directionality, with symmetric components stabilizing activity and asymmetric components driving transitions [48]. Finally, emerging CA3 microcircuit data indicate structural and subtype-specific asymmetries in recurrent excitation, providing an anatomical substrate that may interact with heterogeneous plasticity to support both stable attractor dynamics and controlled sequence generation [49–51].

(c) Inductive biases in hippocampal RNN models

The hippocampus is frequently modeled as a recurrent neural network (RNN) due to its dense recurrence, plasticity, and central role in sequence learning and prediction [2,38,39,52]. A central unresolved question, however, is whether hippocampal-like dynamics arise primarily through learning that sculpts initially unstructured connectivity, or whether circuit-level inductive biases imposed by development of evolution fundamentally constrain the space of attainable dynamics. Recent RNN studies have demonstrated that predictive and sequential structure can, in principle, emerge from training networks initialized with random recurrent weights [2,33,35,52,53]. Yet because random connectivity lacks biological specificity, it remains unclear which aspects of the learned structure are essential for computation, which are incidental, and which admit a meaningful interpretation in terms of hippocampal circuitry. In contrast, converging experimental and theoretical work suggests that hippocampal area CA3 is not randomly wired, but instead exhibits structured, distance-dependent recurrent connectivity that supports coordinated neural assemblies and continuous attractor dynamics [17,22,54,55]. Canonical center-surround (Mexican-hat-like) connectivity stabilizes representations through symmetric interactions, while modest symmetry breaking has long been predicted to enable directed flow [16,34]. These considerations motivate a systematic examination of inductive bias in hippocampal RNN models with the open question being whether biologically motivated recurrent structure shapes learning efficiency, dynamical stability, and the emergence of replay and prediction.

(d) Contribution

The central open question in hippocampal theory is how recurrent circuits can simultaneously stabilize internal representations while generating ordered temporal dynamics in a biologically plausible manner. We hypothesize that hippocampal CA3 operates in an intermediate regime where symmetric connectivity stabilizes representations while a weaker antisymmetric component drives sequential propagation. Here, we address this question by systematically examining how the balance between symmetric and antisymmetric recurrent connectivity governs learning, replay, and prediction in CA3-like recurrent networks.

First, using an Elman-style RNN trained on a continuous circular prediction task inspired by place-cell navigation, we show that networks trained from random connectivity reliably self-organize into a structured regime dominated by a symmetric Mexican-hat-like component with a smaller but functionally essential antisymmetric contribution, which we term a tilted Mexican hat. Despite not being imposed architecturally, this mixed-symmetry organization consistently emerges through learning and supports stable continuous attractors, while generating traveling-wave dynamics that underlie replay and prediction (Figure 1).

Second, we demonstrate that the balance between symmetric and antisymmetric connectivity constitutes a low-dimensional control parameter governing recurrent network dynamics. By explicitly manipulating the symmetry-antisymmetry ratio of structured recurrent weight matrices, we reveal sharp transitions between qualitatively distinct regimes such as static attractors under excessive symmetry, unstable drift or oscillation under excessive antisymmetry, and robust flow-stabilized attractors at intermediate values (Figure 2). Networks initialized near this intermediate regime learn faster, exhibit more reliable autonomous replay, and produce more accurate predictions than networks initialize with random, purely symmetric, or purely directed connectivity motifs.

Third, we show that biologically inspired CA3-like inductive biases improve both optimization and computation (Table 1 and 2). Tilted Mexican hat initialization accelerates early learning, reinstates a two-phase optimization regime absent under random connectivity, yields tighter alignment between replay and prediction dynamics, and uncovers a spatial derivative drift component. These results demonstrate that structured recurrent connectivity is not merely interpretable post hoc, but actively constrains learning toward functionally relevant dynamical solutions.

Finally, we provide a biologically plausible synaptic mechanism for the emergence of mixed-symmetry connectivity. We show that a modified STDP rule combining predominantly symmetric plasticity with a weak antisymmetric component naturally produces tilted Mexican-hat connectivity profiles. This learning rule generates stable traveling-wave dynamics whose speed

and direction depend continuously on the degree of symmetry breaking (Figure 3), linking classical STDP, recent experimental evidence for symmetric CA3-CA3 plasticity, and the emergence of directed hippocampal sequences.

Together, these results support a view of the hippocampus as a flow-stabilized recurrent system operating in an intermediate dynamic regime. In this regime, symmetric recurrence stabilizes internal representations while structured antisymmetry breaks time-reversal symmetry to generate controlled temporal progression. Traveling waves are therefore not epiphenomenal, but a necessary dynamical consequence of how hippocampal circuitry balances stability and flow. *More broadly, our findings suggest that hippocampal computation reflects a combination of task-driven learning and circuit-level inductive biases, with symmetry breaking serving as a fundamental organizing principle for replay, prediction and wave dynamics*.

2. Methods

(a) Task and network architecture

To examine the impact of structural inductive bias in hippocampal recurrent circuitry, we developed a predictive RNN model inspired by place-cell navigation. The task consisted of a simulated rat traversing a continuous circular track at constant velocity. Neurons in the recurrent layer, representing CA3, received location-specific bell-shaped inputs corresponding to place fields distributed uniformly around the ring (Fig. 1A). The network was trained to perform one-step-ahead prediction [2], mapping population activity at time t to activity at time $t+1$.

The model was implemented as an Elman RNN in PyTorch (v2.6.0) and trained using gradient descent with backpropagation through time and a fixed learning rate. Recurrent units used a $\tanh$ nonlinearity and output units used a sigmoid nonlinearity. Input-to-hidden and hidden-to-output weights were frozen to isolate the role of recurrent dynamics. Training terminated when either a maximum epoch count was reached or the loss dropped below 1% of its initial value and changed by less than $10^{-3}\%$ over ten consecutive epochs. Network dimensions and hyperparameters are summarized in Table S1. Training sequences consisted of repeated laps around the circular track with fixed direction and velocity.

(b) Recurrent connectivity families and inductive biases

We examined multiple recurrent connectivity families encoding distinct inductive biases: random, identity, shift, and Mexican-hat connectivity. Random dense initialization served as a baseline, with other families normalized to match. Identity connectivity isolates temporal persistence without lateral interactions. The shift and Mexican-hat matrices are circulant, implementing translation-invariant connectivity on the ring. The shift matrix represents a maximally asymmetric extreme that induces directional propagation, while Mexican-hat connectivity implements distance-dependent excitation and inhibition supporting continuous attractor dynamics.

(c) Symmetry-antisymmetry decomposition and control parameter

In addition to the baseline shift and Mexican-hat initializations, we also implemented variants with specifically tuned symmetry components. Specifically, the recurrent matrices W were decomposed into symmetric S and antisymmetric components A

$$S = \frac{1}{2}\left(W + W^T\right) \quad A = \frac{1}{2}\left(W - W^T\right) \tag{2.1}$$

and recombined as

$$W(\alpha_0) = (\alpha_0)S + (1-\alpha_0)A \quad \text{with } \alpha_0 \in [0,1] \tag{2.2}$$

This parameter α_0 controls the initial balance between reciprocal (symmetric, $\alpha_0 = 1$) and directional (antisymmetric, $\alpha_0 = 0$) interactions. To track how symmetry evolved during training, we defined a time-dependent symmetry index using Frobenius norms as

$$\alpha(t) = \frac{||S||_F}{||S||_F + ||A||_F} \quad \alpha \in [0,1] \tag{2.3}$$

which provides a low-dimensional summary of recurrent structure, explicitly parameterizing the degree of symmetry breaking in the network. α increases as the symmetric component becomes dominant, where $\alpha = 0$ represents the purely anti-symmetric extreme and $\alpha = 1$ denotes the purely symmetric state.

(d) Training, replay, and prediction protocols

Networks were trained using teacher forcing [56], with the true input provided at each time step. Two evaluation regimes (prediction and replay) were used. For prediction, networks received a short prefix of true input followed by noise and were assessed on their ability to autonomously continue the learned sequence. For replay, networks were driven entirely by low-amplitude noise to access internally generated dynamics.

(e) Mexican-hat connectivity construction

Mexican-hat connectivity was constructed by defining a one-dimensional difference-of-Gaussians kernel on a ring of size N. For offsets from the diagonal $d \in 0, 1, ..., H-1$, we define the signed circular distance as

$$d_{signed} = ((d + \lfloor H/2 \rfloor) \mod H) - \lfloor H/2 \rfloor, \quad r(d) = |d_{signed}| \tag{2.4}$$

and construct kernel

$$\mathrm{DoG}(d) = a \exp\left(-\frac{r(d)^2}{2\sigma_e^2}\right) - b \exp\left(-\frac{r(d)^2}{2\sigma_i^2}\right) \tag{2.5}$$

where σ_e and σ_i are the excitation and inhibition widths, and a and b are the excitatory and inhibitory amplitude respectively. b was chosen to approximately balance total excitation and inhibition. The kernel was scaled to match a target spectral radius ρ_{target} using the maximum magnitude of its Fourier coefficients. In our experiments, $\sigma_i > \sigma_e$ to produce a local excitatory bump with broader surround inhibition.

The final recurrent weight matrix was obtained by embedding the kernel into a circulant matrix

$$W_{ij} = \mathrm{DoG}([(i - j - \Delta) \bmod N]) \tag{2.6}$$

where Δ optionally introduces a directional shift. For shifted Mexican-hat matrices, the symmetry-antisymmetry balance was further controlled via the α decomposition above.

3. Results

(a) Task-driven symmetry breaking in recurrent connectivity supports replay and prediction

Training an Elman recurrent neural network from random initialization was sufficient to generate hippocampal-like replay and one-step-ahead prediction on a continuous circular navigation task (Fig. 1A-C). These phenomena are consistent with prior work showing that generic recurrent networks can support sequential dynamics under appropriate task constraints [2,35]. We therefore focus on the structure of the learned recurrent connectivity that gives rise to these dynamics.

Despite starting from unstructured random weights, learning consistently reshaped the recurrent matrix into a highly organized form, indicating that this structure is selected by the task rather than imposed architecturally. When neurons were sorted by their peak activation time during the training sequence, the final recurrent connectivity exhibited a clear banded organization (Fig. 1D). Averaging weights along the diagonals revealed a Mexican-hat-like interaction profile, characterized by strong local excitation flanked by longer-range inhibition. Such profiles are canonical in continuous attractor models and are known to support low-dimensional ring manifolds capable of stable pattern completion [17,21–24,26–28,57].

Crucially, however, the learned connectivity was not purely symmetric. The eigenspectrum of the full recurrent matrix contained complex-valued modes (Fig. 1D, right), indicating the presence of antisymmetric structure. While most eigenvalues clustered near the origin, a small number of dominant modes emerged. In particular, the leading eigenvalue had a real component of 2.16 and a nonzero imaginary component of $\pm$ 0.15, indicating a strongly amplifying mode with weak rotational structure. This mixed real-imaginary signature distinguishes the learned connectivity from a purely symmetric Mexican-hat kernel, which would support static attractors but not sustained temporal progression.

To isolate the contributions of these components, we decomposed the learned recurrent matrix into symmetric and antisymmetric parts (Fig. 1E-F). The symmetric component dominated the overall structure ($||W_{sym}||_F = 3.65$) and reproduced both the Mexican-hat trace profile and a purely real eigenspectrum (Fig. 1E), consistent with stable fixed-point and continuous attractor dynamics [17,21,22,57]. In contrast, the antisymmetric component, though smaller in magnitude ($||W_{anti}||_F = 1.94$), accounted for approximately 35% of the total recurrent norm and exhibited imaginary eigenmodes within the unit circle (Fig. 1F), consistent with rotational or flow-like dynamics that bias activity propagation along the ring [34].

The spatial structure of the antisymmetric component closely resembles a noisy first spatial derivative of the symmetric Mexican-hat profile, with odd symmetry around zero offset and localization near the attractor peak (Fig. 1F, middle). Such derivative-like antisymmetric perturbations have been shown to convert static continuous attractors into traveling-wave solutions by introducing a controlled direction bias [34].

This decomposition shows how replay and prediction emerge from the learned connectivity. The symmetric component stabilizes representations on a continuous attractor manifold, while the antisymmetric component breaks time-reversal symmetry and selects flow direction along that manifold. The resulting dynamics correspond to traveling waves on a ring, producing smooth phase advance in polar coordinates and coherent sequential structure during both replay and prediction (Fig. 1B). These trajectories do not perfectly overlap, indicating that the learned flow is flexible and not exactly aligned with the task manifold. Notably, the dominant outlier eigenmode combines strong symmetric gain with a weak antisymmetric contribution, suggesting that sequence generation arises from a primarily amplifying mode that is gently rotated in state space rather than purely oscillatory dynamics [34,35,38,39].

These results demonstrate that optimization for one-step-ahead prediction from random connectivity reliably induces a mixed-symmetry recurrent organization, rather than converging to a purely symmetric attractor network. This provides evidence that

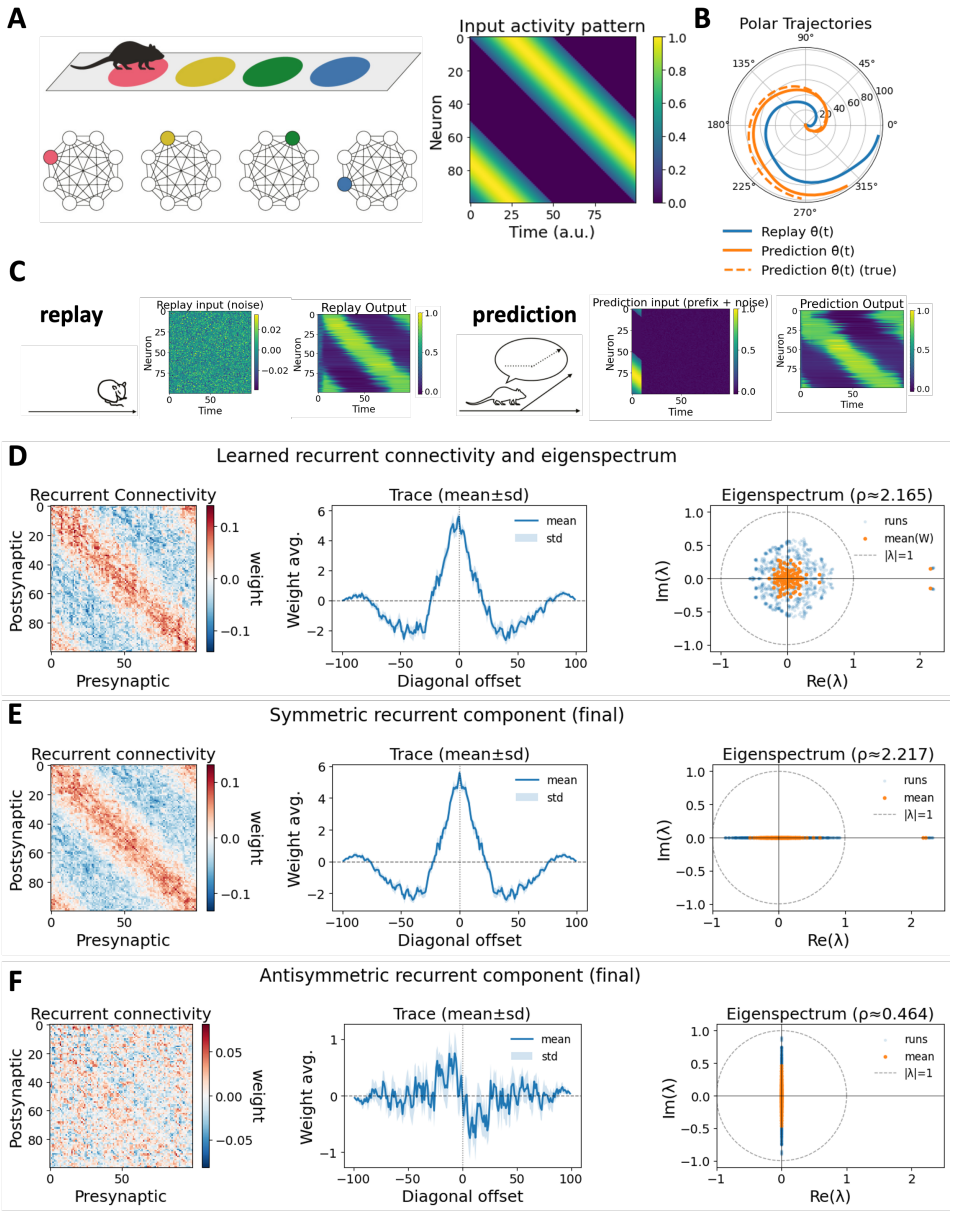

Figure 1: **Learned symmetry breaking in recurrent connectivity supports replay and prediction. (A)** Task schematic: a rat traverses a circular track and receives bell-shaped, location-specific input. **(B)** Polar trajectories during replay (blue) and prediction (orange), showing coherent sequential dynamics. **(C)** Evaluation paradigms: replay driven by noise (left) and prediction following a brief input prefix (right). **(D)** Learned recurrent weight matrix, averaged across runs and sorted by neurons' peak activation time, revealing a banded structure aligned with sequence order. **(E)** Symmetric component of the learned connectivity, exhibiting a Mexican-hat-like profile and real-valued eigenmodes. **(F)** Antisymmetric component, showing off-diagonal structure and imaginary eigenmodes consistent with directional or flow-like dynamics.

symmetry breaking can be learned from task demands, while the residual mismatch between replay and prediction dynamics suggests that learning alone does not fully constrain the recurrent structure. These findings show that mixed-symmetry can emerge from task optimization alone, but the resulting dynamics remain variably aligned with replay and prediction, motivating a systematic examination of how structured recurrent inductive biases shape learning and performance.

(b) CA3-like recurrent inductive bias improves replay, prediction, and learning dynamics

To assess how recurrent inductive bias shapes hippocampal-like computation, we compared networks initialized with different CA3-inspired recurrent weight motifs (identity, shift, Mexican-hat, see Methods) and evaluated autonomous replay and one-step-ahead prediction. Replay probes internally generated dynamics in the absence of informative input, whereas prediction tests whether the learned temporal structure is maintained after withdrawal of external drive.

Performance across both replay and prediction depended strongly on the initial recurrent structure (Tables 1-2; Fig. 2). Networks initialized with Mexican-hat (center-surround) connectivity consistently achieved the strongest performance across all metrics. During replay, Mexican-hat initialization yielded the highest ring-decode R^2 and near-maximal residual lag-1 autocorrelation (Table 1), indicating that internally generated activity remains confined to the learned ring manifold while evolving smoothly over time. More information on the metrics used can be found in the Supplemental Material.

Table 1: Replay metrics

metric	random	identity	pure shift	mixed shift	Mexican hat ($\alpha_0 = 0.7$)
ring decode R^2 (↑)	0.963 ± 0.012	0.211 ± 0.27	0.782 ± 0.013	0.929 ± 0.027	**0.975**± 0.0029
residual lag-1 autocorr (↑)	0.99 ± 0.0032	0.176 ± 0.27	0.99 ± 0.0029	0.99 ± 0.002	**0.99**± 0.0061

Table 2: Prediction metrics

metric	random	identity	pure shift	mixed shift	Mexican hat ($\alpha_0 = 0.7$)
mean resultant length R (↑)	0.956 ± 0.0071	0.939 ± 0.015	0.836 ± 0.0083	0.922 ± 0.024	**0.987**± 0.0044
ring decode R^2 (↑)	0.968 ± 0.011	0.957 ± 0.014	0.778 ± 0.012	0.921 ± 0.036	**0.99**± 0.0039
MSE (↓)	0.0646 ± 0.0066	0.079 ± 0.0076	0.114 ± 0.0052	0.0814 ± 0.013	**0.0499**± 0.0046

Random initialization also supported replay, demonstrating that task-driven learning alone can induce continuous-attractor-like dynamics. However, replay trajectories were less consistent across runs and showed weaker alignment between replay and prediction than in Mexican-hat networks 2D). In contrast, identity initialization, which is maximally symmetric but and lacking lateral interactions, failed to support replay, exhibiting low ring-decode R^2 and poor temporal coherence. These results indicate that temporal persistence alone is insufficient for autonomous sequence generation. Further discussion on identity and shift matrices can be found in the Supplemental Material.

Prediction performance further separated the connectivity families (Table 2). Mexican-hat networks simultaneously maximized angular accuracy (ring-decode R^2 and mean resultant length R) while minimizing mean squared error. Shift connectivity, which hardwires directionality, improved prediction and replay relative to identity initialization but underperformed Mexican-hat connectivity and exhibited replay breakdown at higher symmetry levels (Supplementary Fig. S4). Thus, directionality alone is also insufficient without a stabilizing recurrent scaffold.

Structured inductive bias also altered optimization dynamics. Mexican-hat initialization often began at lower loss and exhibited a pronounced early reduction in error compared to random connectivity (Fig. 2A,B). For specific initial symmetry values, training showed a two-phase learning profile that was absent under random initialization, with an early rapid loss decrease followed by slower refinement This indicates that structured recurrence not only improves final performance but places the network in a more favorable dynamical regime for learning. These performance and learning advantages raise the question of what recurrent structure is ultimately selected by optimization under different initial conditions. We therefore next examine how learning reshapes recurrent connectivity and whether it converges to a common dynamical regime across structured initializations.

(c) Learning converges to a mixed-symmetry regime supporting traveling-wave dynamics

Across structured initializations, learning consistently drove recurrent connectivity toward an intermediate balance between symmetric and antisymmetric components, quantified by the symmetry index $\alpha(t)$ (Fig. 2C). Despite wide variation in initial symmetry, $\alpha(t)$ converged to a predominantly symmetric regime with a persistent antisymmetric component, indicating that intermediate symmetry is a task-selected property rather than an incidental outcome of training. Across all conditions, the best performance was obtained for off-centered Mexican-hat connectivity initialized with an intermediate symmetry level ($\alpha_0 = 0.70$), which places the network near the task-selected symmetry regime from the outset (Fig. 2A-C).

The learned recurrent weights preserve a Mexican-hat-like center-surround profile characterized by strong local excitation and broader inhibition (Fig. 2E-F). Decomposition of the final connectivity revealed that the symmetric component dominated the overall structure and eigenspectrum (Fig. 2G, $||W_{sym}||_F = 3.7$), consistent with stabilization of a continuous attractor manifold. In contrast, the antisymmetric component was smaller in magnitude (Fig. 2H; $||W_{anti}||_F = 0.42$, $||A||/(||S||+||A||) = 0.10$) but introduced imaginary eigenmodes that bias activity propagation. As in the randomly initialized networks, the learned antisymmetric component closely resembles a noisy spatial derivative of the symmetric Mexican-hat profile, with odd symmetry around zero and localization near the attractor peak (Fig. 2H, middle). This derivative-like structure is consistent with theoretical predictions that weak antisymmetric perturbations of continuous attractors induce directional drift without destabilizing the underlying manifold.

Functionally, this mixed-symmetry regime supports dynamics that are neither purely static nor purely feedforward. Replay and prediction trajectories were more closely aligned for Mexican-hat initializations than for random connectivity (Fig. 2D), indicating that internally generated dynamics reflect the learned predictive structure. Spectral analyses further showed that the best-performing initializations already contained weak imaginary components (Fig. 2E), which were preserved and refined during training (Fig. 2F-H).

These results show that learning does not converge to purely symmetric attractor connectivity or strongly antisymmetric sequential architectures. Instead, task optimization reliably selects an intermediate symmetry regime in which symmetric recurrence stabilizes representations while structured antisymmetry enables controlled temporal progression. Task learning can

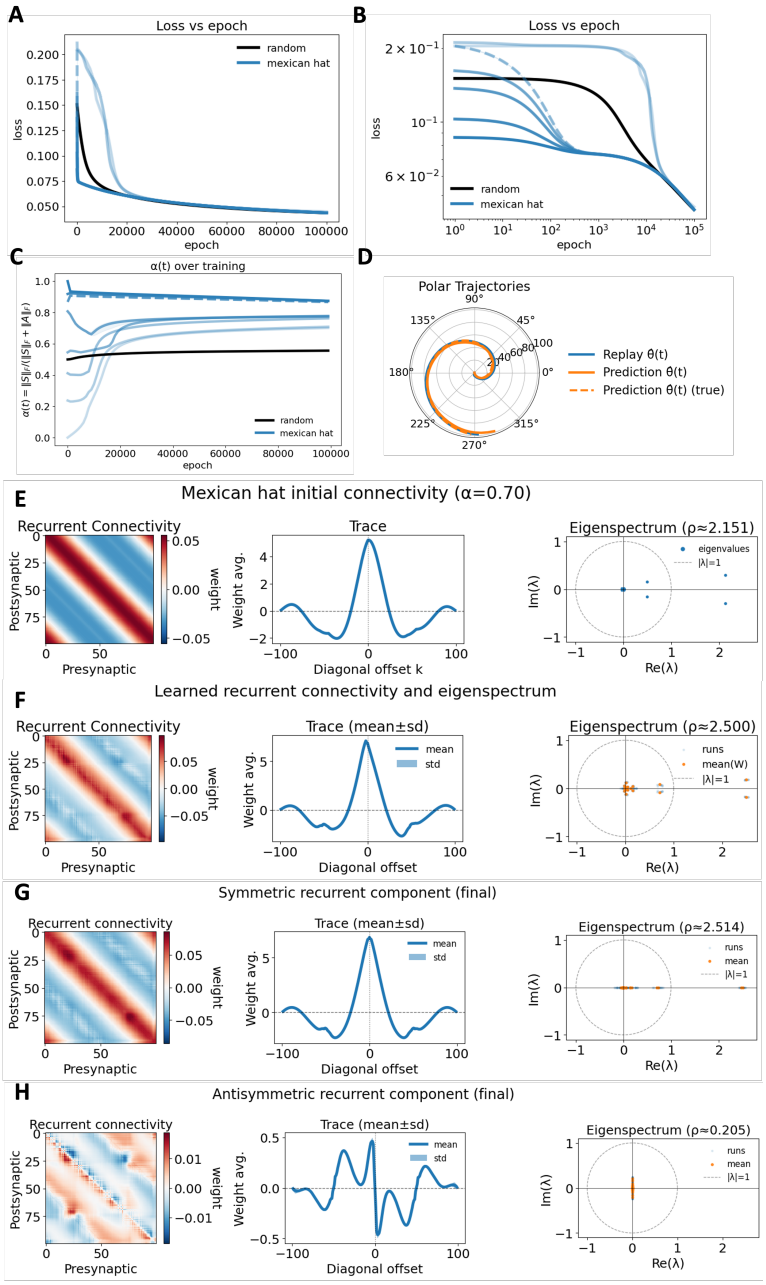

Figure 2: **Structured CA3-like recurrent connectivity accelerates learning and stabilizes traveling-wave dynamics. (A,B)** Training loss for networks initialized with circulant Mexican-hat connectivity (blue; varying initial symmetry α_0, dashed: $\alpha_0 = 0.70$) compared with random connectivity (black), shown on linear and log-log axes. **(C)** Evolution of the symmetry index $\alpha(t)$, showing convergence to an intermediate, symmetry-dominated regime. **(D)** Polar trajectories of replay and one-step-ahead prediction for $\alpha_0 = 0.70$. **(E)** Initial Mexican-hat connectivity ($\alpha_0 = 0.70$), its diagonal trace, and eigenspectrum. **(F)** Learned recurrent connectivity after training. **(G,H)** Symmetric (G) and antisymmetric (H) components of the learned connectivity, showing a dominant attractor-like scaffold with a weaker directional component.

recover mixed-symmetry recurrent structure from random connectivity, but biologically constrained, structured inductive bias reliably accelerates learning, stabilizes dynamics, and yield the most robust replay and prediction. In the next section, we examine how biologically plausible plasticity mechanisms, specifically modified STDP rules, can give rise to this structured symmetry breaking.

(d) Heterogeneous STDP induces symmetry breaking in Mexican-hat connectivity

To connect the mixed-symmetry recurrent structure observed in the trained RNNs to biologically plausible learning mechanisms, we analyzed a modified STDP rule that interpolates between antisymmetric and symmetric plasticity. Classical STDP is temporally asymmetric, strengthening synapses when presynaptic activity precedes postsynaptic firing and weakening the reverse. While

such causal learning supports sequence formation, it does not by itself produce the stabilized center-surround structure observed in the learned recurrent connectivity (Figure 2). Conversely, purely symmetric Hebbian plasticity produces stable attractors but lacks directional bias [21].

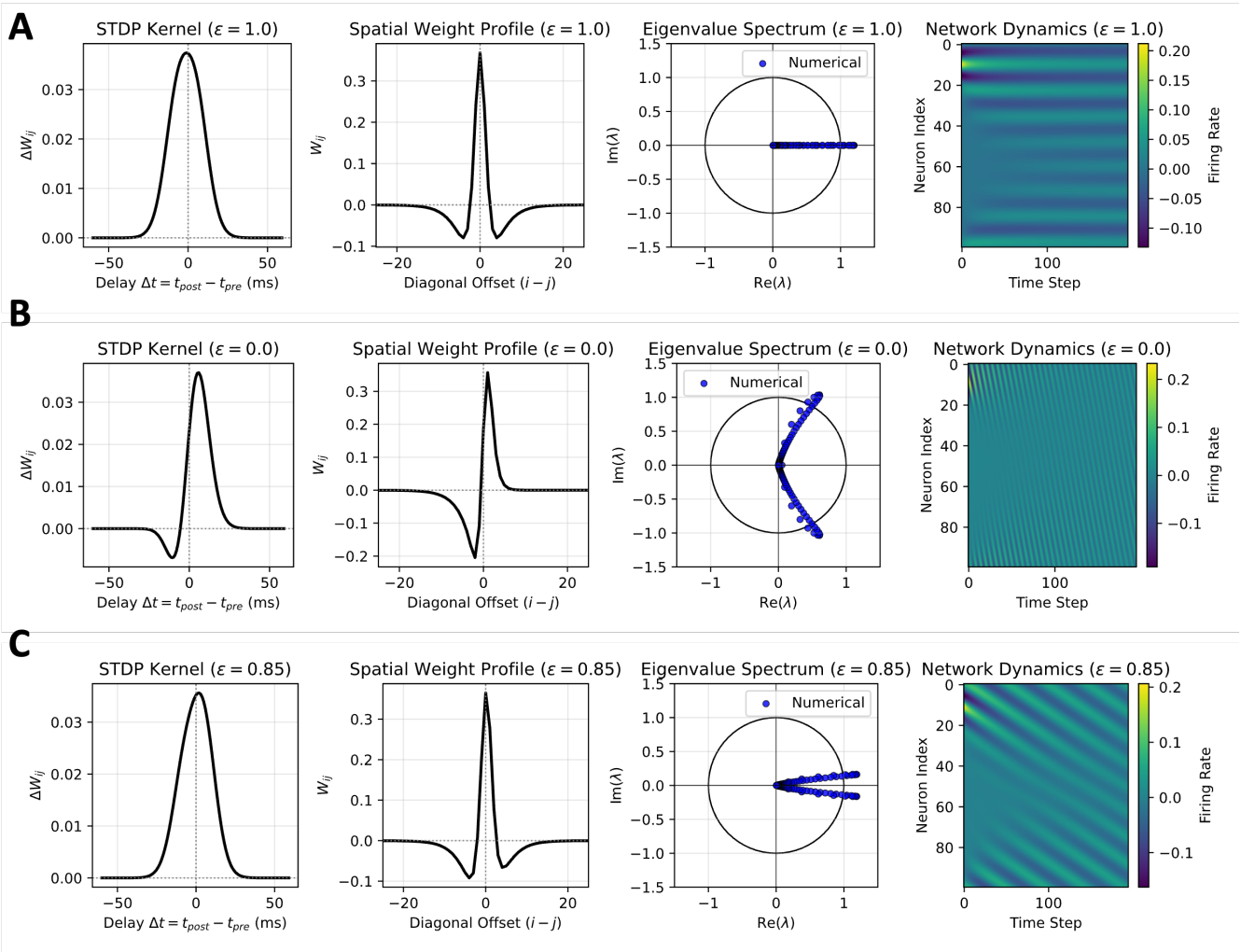

Figure 3: **Symmetry breaking in STDP generates flow-stabilized Mexican-hat dynamics.** Each panel shows the learned synaptic weight profile for different settings of the symmetry parameter (ϵ). The first column shows the STDP learning kernel, the second column the weight profile after learning, the third column the computed eigenvalue spectrum of the learned matrix and the fourth column the resulting network activity dynamics. We use STDP with parameters $A_+ = 1.0, A_- = 0.5, \tau_+ = 1.0, \tau_- = 3.0$. **(A) Pure symmetry ($\epsilon = 1$):** symmetric Mexican-hat profile with real eigenvalues and a stationary attractor. Activity bump is stationary (zero velocity). **(B) Pure antisymmetry ($\epsilon = 0$):** standard STDP gives highly asymmetric weight profile with fast activity drift. **(C) Symmetry breaking ($\epsilon = 0.85$):** weakly tilted Mexican hat with mixed real-imaginary eigenvalues, producing a flow-stabilized attractor with a slowly propagating activity bump.

We therefore considered a modified learning rule that linearly combines a standard antisymmetric STDP kernel ($K(s)$) with a retroactive, symmetry-inducing component controlled by a scalar parameter $\epsilon \in [0, 1]$. The resulting synaptic update from neuron j to neuron i is:

$$W_{ij} = \int_{-\infty}^{\infty} K(s) \int_{-\infty}^{\infty} u_i(t)\, u_j(t-s)\, dt ds + \epsilon \int_{-\infty}^{\infty} K(s) \int_{-\infty}^{\infty} u_j(t)\, u_i(t-s)\, dt ds \tag{3.1}$$

where $u(t)$ represents the neural activity, ϵ controls the degree of symmetry, and $K(s)$ is a biphasic STDP kernel:

$$K(s) = \begin{cases} A_+ e^{-s/\tau_+} & \text{if } s > 0 \quad \text{(LTP)} \\ -A_- e^{s/\tau_-} & \text{if } s < 0 \quad \text{(LTD)} \end{cases} \tag{3.2}$$

where A_+, A_- are the maximum learning rates and τ_+, τ_- are the time constants for potentiation and depression, respectively.

Driven by the above traveling Gaussian input, this learning rule leads to a family of synaptic weights whose spatial interaction pattern is determined by the degree of symmetry breaking, parameterized by ϵ(analytical derivation in Supplementary Section

7; Fig. 3). For $\epsilon = 0$, the rule reduces to standard STDP and yields a strongly antisymmetric interaction profile, dominated by imaginary eigenvalues and fast drift of the activity bump (Fig. 3a). For $\epsilon = 1$, a symmetric Mexican-hat profile is produced with a purely real eigenspectrum as a stationary continuous attractor (Fig. 3b). For intermediate values ($\epsilon = 0.7$), the learned synaptic weight profile becomes a weakly tilted Mexican hat, preserving the stabilizing center-surround structure while introducing a small antisymmetric component (Fig. 3c). The corresponding eigenspectrum contains dominant real modes with only small imaginary components, yielding a stable activity bump that propagates at a constant velocity.

Although ϵ plays a role analogous to the symmetry index α used to characterize learned recurrent weights, the two quantities are not identical. ϵ parametrizes symmetry at the level of synaptic plasticity, whereas α quantifies the symmetry of the overall connectivity. Nevertheless, intermediate values of ϵ naturally give rise to recurrent connectivity in the same mixed-symmetry regime favored by task optimization in the RNNs.

These results show that the flow-stabilized attractor regime identified in the trained networks is accessible through biologically plausible, local learning rules, without requiring nonlocal supervision or explicit architectural constraints.

The formation of the center-surround structure itself—short-range excitation flanked by long-range inhibition—relies on a temporal antisymmetry between the plasticity windows defined in Eq. 3.2. When the causal LTP window is narrow and strong, and the anti-causal LTD window is broad and weak, the summation of the forward and backward kernels results in constructive interference at short latencies. The overlap of the causal and retroactive kernels creates a strong, symmetric peak at the center (excitation), while the broader LTD tails dominate at longer delays to carve out the surround inhibition necessary for stability. For $\epsilon < 1$, the resulting profile retains a slight directional bias, effectively shifting the center of the interaction kernel.

This derivation demonstrates that the "flow-stabilized attractor" regime is not merely an arbitrary outcome of backpropagation but is accessible to biological circuits via simple tuning of the retrograde plasticity signal (ϵ). In biological terms, ϵ may correspond to the strength of "retroactive" potentiation, such as that observed in recent STDP experiments in CA3 [58], or the modulation of back-propagating action potentials by acetylcholine, allowing the hippocampus to dynamically interpolate between storage (attractor) and replay (flow) modes through purely local changes.

4. Discussion

(a) Symmetry breaking as a mechanism for hippocampal sequence generation

A central conclusion of this work is that hippocampal-like recurrent networks operate most effectively in an intermediate regime that balances stability and directed flow. In this regime, recurrent dynamics support persistent internal representations while enabling controlled motion along those representations. Such flow provides a unified dynamical account of replay, prediction, and traveling waves in hippocampal populations [3].

Purely symmetric recurrent connectivity stabilizes attractor states and supports pattern completion but suppresses temporal structure. Conversely, strongly asymmetric connectivity introduces directionality at the expense of stability. A long line of theoretical work has shown that symmetry breaking in recurrent networks is essential for generating ordered temporal dynamics [9,16,29,34,41]. In hippocampal models, asymmetric interactions have been invoked to explain path integration, sequential recall, and internally generated trajectories [10,17,35,38,40].

Our results extend this framework by showing that replay and prediction emerge most robustly when symmetric and antisymmetric components coexist in a controlled ratio. Pure symmetry yields static attractors incapable of temporal progression, while excessive antisymmetry destabilizes representations. Importantly, networks trained from various initial conditions reliably converge toward this mixed-symmetry regime, suggesting that partial symmetry breaking is not an arbitrary modeling assumption but a task-selected dynamical solution that is controlled.

(b) Circulant connectivity and flow-equivariant dynamics in hippocampal networks

An important theoretical insight emerging from this work is the role of translation-invariant recurrent structure, formalized as Toeplitz or, more specifically, circulant connectivity [36,59]. In such networks, synaptic weights depend primarily on relative position, yielding spatial Fourier eigenmodes. Symmetric Toeplitz structure selectively stabilizes low-frequency modes, producing smooth continuous attractors, whereas antisymmetric components introduce imaginary eigenvalues that generate coherent drift along the attractor manifold.

From this perspective, the balance between symmetric and antisymmetric structure directly determines whether dominant modes are stationary or propagating. Our findings show that learning consistently drives recurrent connectivity toward this structured regime, even when starting from dense random matrices. This suggests that circulant-like structure is not merely imposed for analytical convenience but represents an efficient solution for predictive sequence learning on continuous spaces.

Approximate translation invariance is biologically plausible in hippocampal CA3, where recurrent connectivity is dense and lacks obvious spatial anchoring. Experience-dependent plasticity would naturally reinforce relative rather than absolute coding, while deviations from perfect symmetry provide a principled mechanism for generating directed flow without disrupting representational stability. In this way, partial symmetry breaking transforms a static attractor into a flow-stabilized internal representation.

Within this framework, traveling waves arise as a natural dynamical consequence of flow along an internal representational manifold. Such waves have been observed across hippocampal and cortical systems, from theta waves propagating through the

hippocampus [15,60] to high-frequency voltage dynamics revealed by modern imaging [61]. Recent theoretical work emphasizes their computational role in binding memory variables and encoding temporal context [36,62,63].

Our model provides a concrete instantiation of this principle: partial symmetry breaking within a circulant recurrent scaffold yields traveling-wave dynamics that unify replay, prediction, and internally generated sequences under a single mechanism—flow along an internal representational manifold. In this sense, hippocampal recurrent dynamics implement a form of flow-equivariant computation [59], in which task-relevant structure is preserved under continuous transformations while temporal progression is encoded through controlled motion along that structure.

(c) Relationship to prior models

Zhang's seminal work explicitly decomposed recurrent connectivity into even (symmetric) and odd (antisymmetric) components, and demonstrated that the odd component generates rotational flow in head-direction networks [34]. He further showed that optimal dynamics arise in an intermediate regime dominated by symmetry but containing a weaker antisymmetric component, and that the antisymmetric structure approximates the spatial derivative of the symmetric component.

Our results closely parallel this framework but extend it in several key ways. First, whereas Zhang prescribed the relative strength of even and odd components to implement velocity integration, we show that this balance can emerge through learning under a predictive objective. Second, Zhang's model relies on externally driven velocity inputs, whereas our networks generate replay and prediction internally, in the absence of explicit motion signals. Finally, although the learned antisymmetric component closely resembles the spatial derivative of the symmetric Mexican-hat profile, this relationship arises naturally from optimization rather than being imposed. Together, these findings extend symmetry-based theories from hand-designed integrators to learned, internally generative hippocampal dynamics.

Recent theoretical syntheses emphasize that neural circuits often operate between pure attractor and pure integrator regimes [64]. Our results identify partial symmetry breaking as a concrete mechanism for realizing such intermediate dynamics in hippocampal networks.

(d) Structured versus random recurrent connectivity

Random recurrent networks can exhibit rich dynamics, particularly near the edge of chaos [33], and training can harness this richness for computation [53,65]. However, such networks often require fine-tuning to suppress instability and may lack robustness across tasks.

In contrast, we find that structured inductive biases, specifically Mexican-hat-like connectivity with partial symmetry breaking, provide a stable substrate for learning (see the discussion on pretraining as motivation in the Supplemental Material). Networks initialized near this regime learn faster, converge more reliably, and exhibit more interpretable dynamics than those initialized randomly. This aligns with work showing that population structure constrains and stabilizes computation in recurrent circuits [52,66]. From a biological standpoint, these results support the idea that hippocampal circuitry may be developmentally biased toward certain connectivity motifs, with learning refining rather than discovering them *de novo*.

(e) Biological plausibility

The coexistence of symmetric and asymmetric connectivity need not imply multiple specialized learning rules. Synapses exhibit substantial heterogeneity in timing, plasticity, and short-term dynamics [38,49,58,67]. Our heterogeneous STDP results demonstrate that realistic synaptic diversity can naturally give rise to the structured symmetry breaking required for flow, providing a plausible biological substrate for the dynamics observed here.

Our findings support a biologically grounded view of hippocampal CA3 as a flow-stabilized recurrent circuit in which memory, replay, and prediction emerge from a balance between symmetric stabilization and directed flow. Conceptual frameworks emphasize that the hippocampus is fundamentally a sequence generator, with internally generated, ordered dynamics forming the basis of memory and prediction rather than static representations [3]. Consistent with this view, we find that recurrent networks trained on predictive sequence tasks converge to a regime dominated by symmetric, Mexican-hat-like connectivity with a smaller but essential antisymmetric component. This mixed-symmetry organization closely mirrors experimental observations in CA3, where recurrent synapses exhibit largely symmetric plasticity optimized for autoassociative storage [58], alongside structural and functional asymmetries between pyramidal subtypes that support sequential activation [49–51,55]. Although learning from random connectivity can recover this regime, structured CA3-like inductive biases accelerate convergence and stabilize dynamics. This suggests that hippocampal circuitry constrains learning toward a family of dynamical regimes well suited for sequential memory tasks, with experience selecting task-specific operating points within this constrained space.

This framework yields clear experimental predictions, such as selectively perturbing directional coupling should impair replay speed or directionality without abolishing completion (e.g. targeted optogenetic manipulation of CA3 ensembles [68], while state-dependent modulation of CA3 input pathways should bias internal glow without disrupting the underlying attractor geometry [67]. Importantly, while the exact symmetry level, propagation speed, and gain depend on task demands, the principle of dominant symmetry augmented by weak derivative-like antisymmetry appears to be task-general. Together, these results argue that hippocampal sequence generation reflects an interaction between biological inductive biases and task-driven learning, rather than learning alone.

Taken together, these results suggest that hippocampal replay and prediction emerge from a delicate balance between stability and flow, implemented through partial symmetry breaking in recurrent connectivity. Circulant structure defines the internal representational manifold and antisymmetry determines how that manifold is traversed. Flow is therefore a functional requirement for hippocampal computation, enabling internal models to be dynamically explore rather than statically recalled. More broadly, these results suggest that CA3-like inductive bias does not prescribe a specific computation, but instead constrains learning toward a family of dynamical regimes well suited for sequential memory tasks—a tilted Mexican hat. Structured connectivity restricts the search space of learning, while task demands select a particular operating point within that space. The convergence of mixed-symmetry structure across random initialization, structured bias, and biologically plausible plasticity argues that this organization if not a task artifact. Although the precise symmetry level, propagation timescale, and antisymmetric gain depend on task demands, the organizing principle of dominant symmetry combined with weak, derivative-like antisymmetry appears to be general.

5. Conclusion

This work identifies partial symmetry breaking in recurrent connectivity as a circuit-level principle that enables hippocampal networks to jointly support stable memory representations, internally generated replay, and predictive sequence dynamics. Using recurrent neural networks trained on a continuous prediction task, we show that these functions emerge most robustly in an intermediate regime dominated by symmetric structure but augmented by a weaker antisymmetric component. This mixed-symmetry organization stabilizes internal representations while enabling directed flow along them, yielding traveling-wave dynamics that unify replay, prediction, and sequence generation within a single recurrent circuit.

Importantly, this regime is not imposed but selected by learning. Across random and structured initializations, networks reliably converge toward a tilted Mexican-hat-like connectivity, suggesting that symmetry breaking acts as a low-dimensional control parameter governing the balance between stability and motion. These findings support a view in which hippocampal circuitry is shaped by biological inductive biases that constrain learning toward a family of dynamical regimes well suits for sequential memory, with experience refining task-specific operating points within that space rather than discovering structure *de novo*.

Several limitations of the present study point to important directions for future work. We focused on a single continuous prediction task and abstracted away cell-type diversity, neuromodulation, and short-term synaptic dynamics, all of which are known to influence hippocampal computation. Whether the same mixed-symmetry regime generalizes to non-spatial memories, branching sequences, or goal-dependent dynamics remains an open question. Extending this framework to multi-area models incorporating CA1, entorhinal cortex, and state-dependent modulation will be essential for linking circuit dynamics more directly to behavior.

The framework developed here yields clear experimental predictions. Selective perturbations of directional coupling within CA3 should impair replay speed or directionality without abolishing pattern completion, while state-dependent modulation of CA3 inputs should bias internal flow while preserving the underlying attractor geometry. Advances in large-scale electrophysiology and imaging now make it possible to directly test whether hippocampal population dynamics exhibit the predicted weak rotational components associated with partial symmetry breaking.

In conclusion, these results support a view of hippocampal CA3 as a flow-stabilized recurrent circuit in which memory, replay, and prediction arise from the interaction between biologically constrained inductive biases and experience-dependent learning. Partial symmetry breaking provides a principled mechanism for balancing stability and motion, offering a unifying circuit-level account of hippocampal sequence generation across behavioral states.

Acknowledgements. This research was supported by an NIH Director's Pioneer Award (DP1NS149613) and by a grant from the Office of Naval Research (N00014-23-1-2069).

References

1. O'Keefe J, Dostrovsky J. 1971 The hippocampus as a spatial map. Preliminary evidence from unit activity in the freely-moving rat. *Brain Research* **34**, 171–175. (10.1016/0006-8993(71)90358-1)
2. Chen Y, Zhang H, Cameron M, Sejnowski T. 2024 Predictive Sequence Learning in the Hippocampal Formation. *Neuron* **112**. (https://doi.org/10.1016/j.neuron.2024.05.024)
3. Buzsáki G, Tingley D. 2018 Space and Time: The Hippocampus as a Sequence Generator. *Trends Cogn Sci* **22**, 853–869. (10.1016/j.tics.2018.07.006)
4. Treves A, Rolls ET. 1994 Computational analysis of the role of the hippocampus in memory. *Hippocampus* **4**, 374–391. (10.1002/hipo.450040319)
5. McNaughton BL, Morris RGM. 1987 Hippocampal synaptic enhancement and information storage within a distributed memory system. *Trends Neurosci.* **10**, 408–415. (10.1016/0166-2236(87)90011-7)
6. Skaggs WE, McNaughton BL. 1992 Computational approaches to hippocampal function. *Curr. Opin. Neurobiol.* **2**, 209–211. (10.1016/0959-4388(92)90014-C)
7. O'Keefe J, Recce ML. 1993 Phase relationship between hippocampal place units and the EEG theta rhythm. *Hippocampus* **3**, 317–330. (10.1002/hipo.450030307)

8. Skaggs WE, McNaughton BL, Wilson MA, Barnes CA. 1996 Theta phase precession in hippocampal neuronal populations and the compression of temporal sequences. *Hippocampus* **6**, 149–172. (10.1002/(SICI)1098-1063(1996)6:2<149::AID-HIPO6>3.0.CO;2-K)
9. Tsodyks MV, Skaggs WE, Sejnowski TJ, McNaughton BL. 1996 Population Dynamics and Theta Rhythm Phase Precession of Hippocampal Place Cell Firing: a Spiking Neuron Model. *Hippocampus* **6**, 271–280. (10.1002/(SICI)1098-1063(1996)6:3<271::AID-HIPO5>3.0.CO;2-Q)
10. Lisman JE. 1999 Relating Hippocampal Circuitry to Function: Recall of Memory Sequences by Reciprocal Dentate–CA3 Interactions. *Neuron* **22**, 233–242. (10.1016/S0896-6273(00)81085-5)
11. Johnson A, Redish AD. 2007 Neural ensembles in CA3 transiently encode paths forward of the animal at a decision point. *J Neurosci* **27**, 12176–12189. (10.1523/JNEUROSCI.3761-07.2007)
12. Manns JR, Howard MW, Eichenbaum H. 2007 Gradual changes in hippocampal activity support remembering the order of events. *Neuron* **56**, 530–540. (10.1016/j.neuron.2007.08.017)
13. Wilson MA, McNaughton BL. 1994 Reactivation of Hippocampal Ensemble Memories During Sleep. *Science* **265**, 676–679. (10.1126/science.8036517)
14. Pfeiffer BE, Foster DJ. 2013 Hippocampal place cell sequences depict future paths to remembered goals. *Nature* **497**, 74–79. (10.1038/nature12112)
15. Lubenov EV, Siapas AG. 2009 Hippocampal theta oscillations are travelling waves. *Nature* **459**, 534–539. (10.1038/nature08010)
16. Abbott LF, Blum KI. 1996 Functional Significance of Long-Term Potentiation for Sequence Learning and Prediction. *Cereb Cortex* **6**, 406–416. (10.1093/cercor/6.3.406)
17. Samsonovich A, McNaughton BL. 1997 Path Integration and Cognitive Mapping in a Continuous Attractor Neural Network Model. *J Neurosci* **17**, 5900–5920. (10.1523/JNEUROSCI.17-15-05900.1997)
18. Buzsáki G, Moser EI. 2013 Memory, navigation and theta rhythm in the hippocampal-entorhinal system. *Nat Neurosci* **16**, 130–138. (10.1038/nn.3304)
19. Chen Z, Gomperts SN, Yamamoto J, Wilson MA. 2013 Neural Representation of Spatial Topology in the Rodent Hippocampus. *Neural computation* **26**, 1. (10.1162/NECO_a_00538)
20. O'Keefe J, Burgess N. 2005 Dual phase and rate coding in hippocampal place cells: theoretical significance and relationship to entorhinal grid cells. *Hippocampus* **15**, 853–866. (10.1002/hipo.20115)
21. Hopfield JJ. 1982 Neural networks and physical systems with emergent collective computational abilities. *Proc Natl Acad Sci U S A* **79**, 2554–2558. (10.1073/pnas.79.8.2554)
22. Amari Si. 1977 Dynamics of pattern formation in lateral-inhibition type neural fields. *Biol. Cybern.* **27**, 77–87. (10.1007/BF00337259)
23. Amit DJ, Tsodyks MV. 1991 Quantitative study of attractor neural network retrieving at low spike rates:: I. substrate—spikes, rates and neuronal gain. *Network: Computation in Neural Systems* **2**, 259–273. (10.1088/0954-898X_2_3_003)
24. Amit D, Tsodyks M. 1992 Effective neurons and attractor neural networks in cortical environment. *Network* **3**, 121–137. (10.1088/0954-898X/3/2/003)
25. Chandra S, Sharma S, Chaudhuri R, Fiete I. 2025 Episodic and associative memory from spatial scaffolds in the hippocampus. *Nature* **638**, 739–751. (10.1038/s41586-024-08392-y)
26. Griniasty M, Tsodyks MV, Amit DJ. 1993 Conversion of Temporal Correlations Between Stimuli to Spatial Correlations Between Attractors. *Neural Computation* **5**, 1–17. (10.1162/neco.1993.5.1.1)
27. Amit DJ, Brunel N, Tsodyks MV. 1994 Correlations of cortical Hebbian reverberations: theory versus experiment. *J Neurosci* **14**, 6435–6445. (10.1523/JNEUROSCI.14-11-06435.1994)
28. Cugliandolo LF, Tsodyks MV. 1994 Capacity of networks with correlated attractors. *J. Phys. A: Math. Gen.* **27**, 741. (10.1088/0305-4470/27/3/018)
29. Seung H. 1996 How the brain keeps the eyes still. *Proc. Natl. Acad. Sci. U.S.A.* **93**, 13339–13344. (10.1073/pnas.93.23.13339)
30. Eliasmith C. 2005 A Unified Approach to Building and Controlling Spiking Attractor Networks. *Neural Computation* **17**, 1276–1314. (10.1162/0899766053630332)
31. Kleinfeld D. 1986 Sequential state generation by model neural networks.. *Proceedings of the National Academy of Sciences* **83**, 9469–9473. (10.1073/pnas.83.24.9469)
32. Sompolinsky H, Kanter I. 1986 Temporal Association in Asymmetric Neural Networks. *Phys. Rev. Lett.* **57**, 2861–2864. (10.1103/PhysRevLett.57.2861)
33. Sompolinsky H, Crisanti A, Sommers HJ. 1988 Chaos in Random Neural Networks. *Phys. Rev. Lett.* **61**, 259–262. (10.1103/PhysRevLett.61.259)
34. Zhang K. 1996 Representation of spatial orientation by the intrinsic dynamics of the head-direction cell ensemble: a theory. *J. Neurosci.* **16**, 2112–2126. (10.1523/JNEUROSCI.16-06-02112.1996)
35. Rajan K, Harvey CD, Tank DW. 2016 Recurrent Network models of sequence generation and memory. *Neuron* **90**, 128–142. (10.1016/j.neuron.2016.02.009)
36. Keller TA, Muller L, Sejnowski T, Welling M. 2024 Traveling Waves Encode the Recent Past and Enhance Sequence Learning. ICLR. arXiv:2309.08045 [cs] (10.48550/arXiv.2309.08045)
37. Mohan UR, Zhang H, Jacobs J. 2022 The direction and timing of theta and alpha traveling waves modulate human memory processing. (10.1101/2022.02.07.479466)
38. Bush D, Philippides A, Husbands P, O'Shea M. 2010 Dual Coding with STDP in a Spiking Recurrent Neural Network Model of the Hippocampus. *PLoS Comput Biol* **6**, e1000839. (10.1371/journal.pcbi.1000839)
39. Romani S, Tsodyks M. 2015 Short-term plasticity based network model of place cells dynamics. *Hippocampus* **25**, 94–105. (10.1002/hipo.22355)
40. Roxin A, Brunel N, Hansel D. 2005 Role of Delays in Shaping Spatiotemporal Dynamics of Neuronal Activity in Large Networks. *Phys. Rev. Lett.* **94**, 238103. (10.1103/PhysRevLett.94.238103)
41. Pouget A, Sejnowski TJ. 1995 pp. 368–371. In *Dynamic Remapping*, pp. 368–371. Cambridge, MA: MIT Press 2 edition.

42. Kurikawa T, Kaneko K. 2021 Multiple-timescale neural networks: Generation of history-dependent sequences and inference through autonomous bifurcations. *Front. Comput. Neurosci.* **15**, 743537. (10.3389/fncom.2021.743537)
43. Betteti S, Baggio G, Bullo F, Zampieri S. 2025 Input-driven dynamics for robust memory retrieval in Hopfield networks. *Science Advances* **11**, eadu6991. (10.1126/sciadv.adu6991)
44. Karuvally A, Sejnowski T, Siegelmann HT. 2023 General Sequential Episodic Memory Model. In Krause A, Brunskill E, Cho K, Engelhardt B, Sabato S, Scarlett J, editors, *Proceedings of the 40th International Conference on Machine Learning* vol. 202*Proceedings of Machine Learning Research* pp. 15900–15910. PMLR.
45. Karuvally A, Lertsaroj P, Sejnowski T, Siegelmann HT. 2025 Exponential Dynamic Energy Network for High Capacity Sequence Memory. In *The Thirty-ninth Annual Conference on Neural Information Processing Systems*.
46. Markram H, Lübke J, Frotscher M, Sakmann B. 1997 Regulation of synaptic efficacy by coincidence of postsynaptic APs and EPSPs. *Science* **275**, 213–215. (10.1126/science.275.5297.213)
47. Bi GQ, Poo MM. 1998 Synaptic modifications in cultured hippocampal neurons: dependence on spike timing, synaptic strength, and postsynaptic cell type. *J Neurosci* **18**, 10464–10472. (10.1523/JNEUROSCI.18-24-10464.1998)
48. Gillett M, Brunel N. 2024 Dynamic control of sequential retrieval speed in networks with heterogeneous learning rules. *eLife* **12**, RP88805. (10.7554/eLife.88805)
49. Sammons RP, Masserini S, Moreno Velasquez L, Metodieva VD, Cano G, Sannio A, Orlando M, Maier N, Kempter R, Schmitz D. 2025 Sub-type specific connectivity between CA3 pyramidal neurons may underlie their sequential activation during sharp waves. *eLife* **13**, RP98653. (10.7554/eLife.98653)
50. Watson JF, Vargas-Barroso V, Jonas P. 2025 Cell-specific wiring routes information flow through hippocampal CA3. *Cell Rep.* **44**, 116080. (10.1016/j.celrep.2025.116080)
51. Layous R, Tamás B, Mike A, Sipos E, Arszovszki A, Brunner J, Szatai Á, Yaseen F, Andrási T, Szabadics J. 2025 Optical recordings of unitary synaptic connections reveal high and random local connectivity between CA3 pyramidal cells. *J. Neurosci.* **45**, e0102252025. (10.1523/JNEUROSCI.0102-25.2025)
52. Chenkov N, Sprekeler H, Kempter R. 2017 Memory replay in balanced recurrent networks. *PLoS Comput Biol* **13**, e1005359. (10.1371/journal.pcbi.1005359)
53. Laje R, Buonomano DV. 2013 Robust timing and motor patterns by taming chaos in recurrent neural networks. *Nat Neurosci* **16**, 925–933. (10.1038/nn.3405)
54. Buzsáki G. 2010 Neural syntax: cell assemblies, synapsembles and readers. *Neuron* **68**, 362–385. (10.1016/j.neuron.2010.09.023)
55. Watson JF, Vargas-Barroso V, Morse-Mora RJ, Navas-Olive A, Tavakoli MR, Danzl JG, Tomschik M, Rössler K, Jonas P. 2025 Human hippocampal CA3 uses specific functional connectivity rules for efficient associative memory. *Cell* **188**, 501–514.e18. (10.1016/j.cell.2024.11.022)
56. Williams RJ, Zipser D. 1989 A learning algorithm for continually running fully recurrent neural networks. *Neural Comput.* **1**, 270–280. (10.1162/neco.1989.1.2.270)
57. Ben-Yishai R, Bar-Or RL, Sompolinsky H. 1995 Theory of orientation tuning in visual cortex.. *Proc. Natl. Acad. Sci. U.S.A.* **92**, 3844–3848. (10.1073/pnas.92.9.3844)
58. Mishra RK, Kim S, Guzman SJ, Jonas P. 2016 Symmetric spike timing-dependent plasticity at CA3–CA3 synapses optimizes storage and recall in autoassociative networks. *Nat Commun* **7**, 11552. (10.1038/ncomms11552)
59. Keller TA. 2025 Flow equivariant recurrent neural networks. In *Proceedings of the 39th International Conference on Neural Information Processing Systems*.
60. Patel J, Fujisawa S, Berényi A, Royer S, Buzsáki G. 2012 Traveling theta waves along the entire septotemporal axis of the hippocampus. *Neuron* **75**, 410–417. (10.1016/j.neuron.2012.07.015)
61. Haziza S, Chrapkiewicz R, Zhang Y, Kruzhilin V, Li J, Li J, Delamare G, Swanson R, Buzsáki G, Kannan M, Vasan G, Lin MZ, Zeng H, Daigle TL, Schnitzer MJ. 2025 Imaging high-frequency voltage dynamics in multiple neuron classes of behaving mammals. *Cell* **188**, 4401–4423.e31. (10.1016/j.cell.2025.06.028)
62. Muller L, Churchland PS, Sejnowski TJ. 2024 Transformers and cortical waves: encoders for pulling in context across time. *Trends Neurosci* **47**, 788–802. (10.1016/j.tins.2024.08.006)
63. Karuvally A, Sejnowski TJ, Siegelmann HT. 2024 Hidden Traveling Waves bind Working Memory Variables in Recurrent Neural Networks. arXiv:2402.10163 [cs] (10.48550/arXiv.2402.10163)
64. Khona M, Fiete IR. 2022 Attractor and integrator networks in the brain. *Nat Rev Neurosci* **23**, 744–766. (10.1038/s41583-022-00642-0)
65. Sussillo D, Abbott LF. 2009 Generating Coherent Patterns of Activity from Chaotic Neural Networks. *Neuron* **63**, 544–557. (10.1016/j.neuron.2009.07.018)
66. Dubreuil A, Valente A, Beiran M, Mastrogiuseppe F, Ostojic S. 2022 The role of population structure in computations through neural dynamics. *Nat Neurosci* **25**, 783–794. (10.1038/s41593-022-01088-4)
67. Vandael D, Okamoto Y, Jonas P. 2021 Transsynaptic modulation of presynaptic short-term plasticity in hippocampal mossy fiber synapses. *Nat Commun* **12**, 2912. (10.1038/s41467-021-23153-5)
68. Oishi N, Nomoto M, Ohkawa N, Saitoh Y, Sano Y, Tsujimura S, Nishizono H, Matsuo M, Muramatsu Si, Inokuchi K. 2019 Artificial association of memory events by optogenetic stimulation of hippocampal CA3 cell ensembles. *Mol Brain* **12**, 2. (10.1186/s13041-018-0424-1)
69. Quentin R, King JR, Sallard E, Fishman N, Thompson R, Buch ER, Cohen LG. 2019 Differential brain mechanisms of selection and maintenance of information during working memory. *J. Neurosci.* **39**, 3728–3740. (10.1523/JNEUROSCI.2764-18.2019)
70. Durbin J, Watson GS. 1971 Testing for serial correlation in least squares regression. III. *Biometrika* **58**, 1. (10.2307/2334313)
71. Batschelet E. 1983 Circular statistics in biology. *J. R. Stat. Soc. Ser. A* **146**, 91. (10.2307/2981498)

6. Supplemental Material

(a) Pretraining as evidence for structural inductive bias

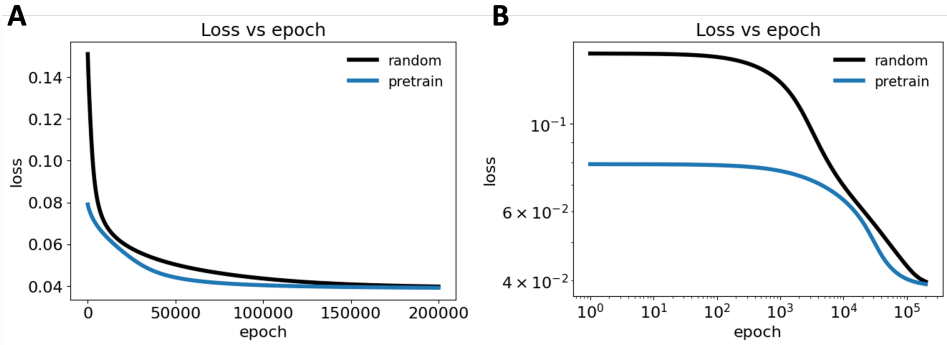

Figure S1: **Effect of pretraining on learning dynamics. (A)** Training loss as a function of epoch shown on linear axes. Networks initialized with pretrained recurrent weights (blue), obtained from the final weights of the randomly initialized network, begin at a lower loss and exhibit faster early optimization compared to networks trained from random initialization (black). **(B)** Same data plotted on log-log axes, highlighting differences in early and intermediate training regimes. Pretrained networks avoid the prolonged high-loss plateau observed under random initialization and transition earlier into the rapid loss-decrease regime. In both cases, training converges to a similar final loss.

Pretraining the recurrent connectivity using the final weights obtained from a randomly initialized network significantly alters the early learning dynamics without changing the eventual performance ceiling (Fig. S1). As shown in both linear (Fig. S1A) and log-log (Fig. S1B) axes, pretrained networks start at a markedly lower loss and bypass the extended high-loss plateau characteristic of random initialization. This leads to faster progress through early and intermediate training epochs. Despite these differences in learning trajectory, both conditions converge to nearly identical final loss values, indicating that pretraining primarily improves optimization efficiency rather than the final representational capacity of the model. These results suggest that structure acquired during prior training places the network in a favorable regime of parameter space, facilitating faster learning while reserving the same asymptotic solution.

(b) Model hyperparameters

Table S1: Replay metrics

hyperparameter	value
number of neurons	100
number of epochs	100000
sequence length	100
learning rate	0.01

(c) Identity and shift initializations: isolating persistence, directionality, and their limits in CA3-like circuits

To further dissect which structural features of CA3 recurrent connectivity are necessary for hippocampal-like replay and prediction, we examined two extreme and deliberately simplified structures: identity and cyclic shift connectivity. These architectures isolate the complementary computational primitives of temporal persistence and directional propagation (symmetric and antisymmetric aspects) that are each present in CA3, although not in isolation.

The identity matrix represents a maximally symmetric, self-coupled network in which each unit predominantly feeds back onto itself. We include this case as a conceptual control that isolates temporal persistence without lateral interactions. Such a network can maintain activity locally and support short-horizon, input-driven prediction. However, it fundamentally fails to capture CA3's role as an associative recurrent circuit. Consistent with this, identity-initialized networks reliably learned one-step prediction when driven by external input but failed at autonomous replay. The resulting weight profiles remained identity-dominated with only weak lateral structure, which was insufficient to sustain a self-organized sequence when driven by noise. Thus, persistence alone is not sufficient for traveling waves or replay. This reinforces the idea that CA3 replay requires not just stability, but structured interactions that organize population activity along a low-dimensional manifold.

In contrast, cyclic shift connectivity (Fig. S2) imposes a purely directional recurrent structure, in which each neuron excites its success along a ring. This architecture hardwires a translation operator into the recurrent matrix and naturally produces a

sequential propagation of activity. As such, it captures the CA3 function of directional progression through a learned sequence and is consistent with experimental and theoretical work implicating antisymmetric or non-reciprocal connectivity arising from temporally antisymmetric plasticity such as STDP. Indeed, antisymmetric components in CA3 may reflect experience-dependent strengthening of forward transitions during repeated trajectories, theta sequences, or sharp-wave ripple replay.

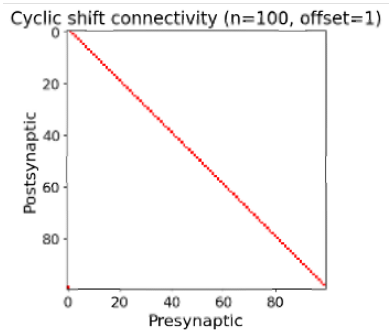

Figure S2: **Cyclic shift recurrent connectivity.** Visualization of a cyclic shift recurrent weight matrix ($n = 100$, offset$= 1$) in which each unit excites its immediate successor on the ring with wrap-around at the boundaries. This structure implements a purely directional, translation-invariant connectivity pattern that promotes sequential propagation of activity without symmetric recurrent stabilization.

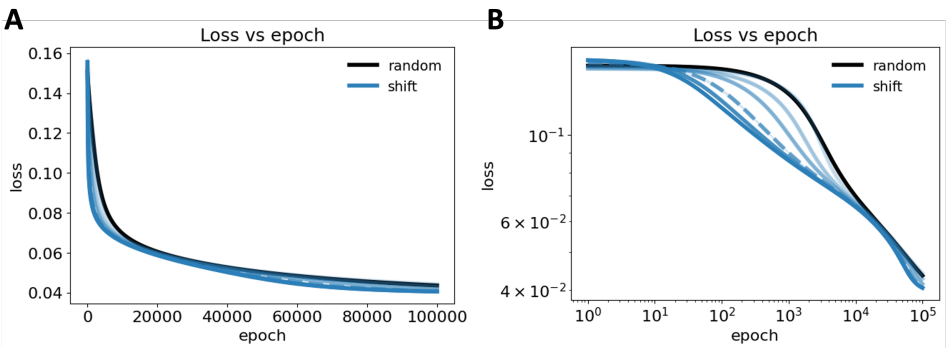

Figure S3: **Training loss for shift-initialized recurrent networks across symmetry levels.** Training loss as a function of epoch for shift initializations (blue) spanning initial symmetry ratios of $\alpha_0 \in [0, 1]$ compared to random initialization (black). The dashed blue curve is $\alpha_0 = 0.60$, the beginning of the regime where replay breaks down despite comparable training loss. **(A)** linear scale and **(B)** log-log scale. Overall, shift-initialized networks exhibit similar final and initial losses to random connectivity, while early training dynamics are altered. Stable replay here is not guaranteed.

However, pure shift connectivity lacks the bidirectional, stabilizing interactions required for attractor-like memory retrieval. Our results make this limitation explicit. While shift-initialized networks were able to outperform identity networks on both replay and prediction, their ability to support replay depended critically on the initial symmetry-antisymmetry balance. As the initial symmetry ratio α_0 increased, the structure became more similar to the identity matrix, with a complete sharp breakdown in replay behavior at $\alpha_0 = 0.60$ indicated by ring-decode R^2 collapse, residual temporal coherence deterioration, and polar trajectories which degenerated into fixed or noisy states (Fig. S4-5) Indeed, at $\alpha_0 = 0$, the recurrent matrix is a pure shift matrix, but at $\alpha_0 = 1$, the matrix has the same limitations as the identity matrix. Notably, prediction performance remained largely intact across this transition, indicating that locally accurate state updates can still be learned even when the global autonomous dynamics fail.

This highlights a key computational distinction. Prediction can be supported by a locally correct transition operator under external drive, whereas replay requires a globally consistent dynamical system with both a stable manifold and a directed flow along it. Pure shift connectivity provides direction without sufficient stabilization; identity provides stabilization without direction. Random connectivity can, through learning, self-organize both to some degree but does so less intuitively than biologically inspired structures.

Across structured initializations, training consistently drove the recurrent connectivity toward an intermediate symmetry regime, quantified by convergence of $\alpha(t)$ (Fig. S6). Fully symmetric networks (i.e. identity) rapidly gained antisymmetric structure. Shift networks, spanning fully antisymmetric to symmetric initial conditions, converged toward a mid-range α value. Of note, the intermediate range here is lower than that of Mexican-hat initialization. This is likely connected to the breakdown in replay behavior observed at $\alpha_0 = 0.60$. This convergence is again not an artifact of initialization, but instead an emergent,

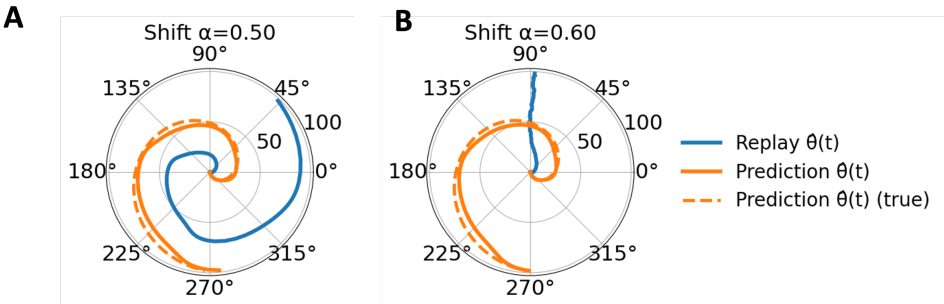

Figure S4: **Breakdown of replay for shift-initialized networks at higher symmetry**. Polar trajectories of learned replay (blue) and prediction (orange; dashed indicated ground truth) for shift-initialized recurrent connectivity with **(A)** $\alpha_0 = 0.50$ and **(B)** $\alpha_0 = 0.60$. At $\alpha_0 = 0.50$, replay follows a coherent progression around the ring, whereas increasing symmetry to $\alpha_0 = 0.60$ leads to a collapse of replay dynamics despite preserved prediction performance, indicating a sharp transition in the ability of shift connectivity to support stable autonomous sequence generation.

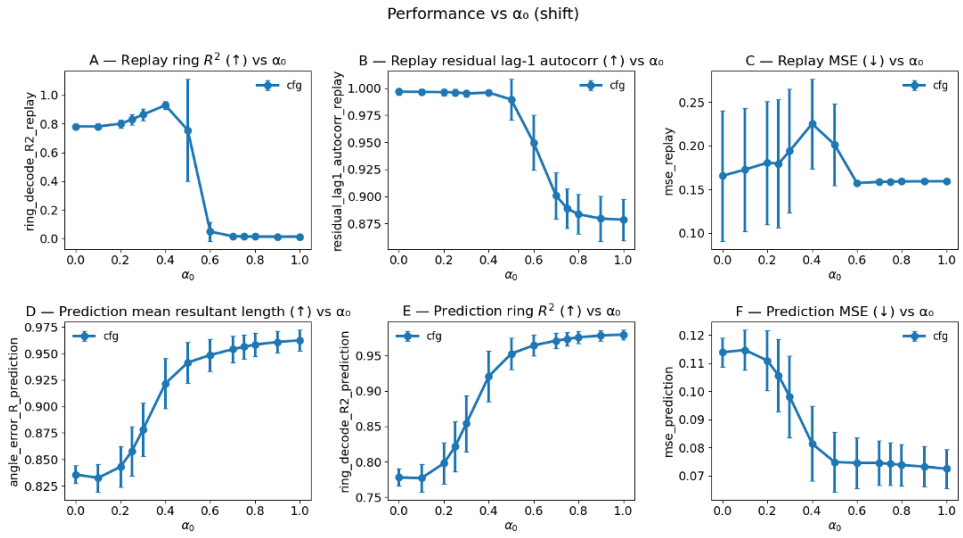

Figure S5: **Performance as a function of initial symmetry for shift connectivity.** Replay and prediction performance metrics plotted as a function of the initial symmetry ratio α_0 for shift-initialized recurrent connectivity. Top row: replay metrics including ring-decode R^2 **(A)**, residual lag-1 autocorrelation **(B)**, and replay MSE **(C)**. Bottom row: prediction metrics including mean resultant length of angular error **(D)**, ring-decode R^2 **(E)**, and prediction MSE **(F)**. Replay performance exhibits a sharp breakdown at $\alpha_0 = 0.60$, marked by the dramatic drop in ring-decode R^2 and reduced temporal coherence despite increased prediction performance. This difference indicates that increasing symmetry in shift connectivity impairs autonomous replay dynamics while leaving input-driven prediction intact.

task-selected property, reinforcing the conclusion that hippocampal-like computation requires a balance between symmetric stabilization and antisymmetric flow.

Importantly, a tilted Mexican-hat connectivity obtains this balance naturally. By providing a center-surround attractor scaffold from the outset, Mexican-hat initialization supports both stable memory representations and controlled propagation, yielding superior replay fidelity, tighter alignment between replay and prediction, and more consistent traveling-wave dynamics than either shift or identity connectivity. In this context, the shift and identity initializations serve as informative boundary cases that reveal that neither directionality nor persistence alone is sufficient. Instead, CA3-like replay emerges most robustly when directional biases are embedded with a stabilizing recurrent architecture. This is a regime that learning reliably discovers, but that structured inductive bias makes easier to reach and clearer to interpret.

Together, these results argue that CA3 is best understood not as a pure attractor (identity) network or a feedforward sequence generator (pure shift), but as a flow-stabilized recurrent system in which symmetric and antisymmetric interactions are jointly essential. Identity and shift initializations clarify the computational roles of these components, while Mexican-hat connectivity demonstrates how their integration supports traveling waves, replay, and prediction in a single recurrent circuit.

(d) Metrics computed

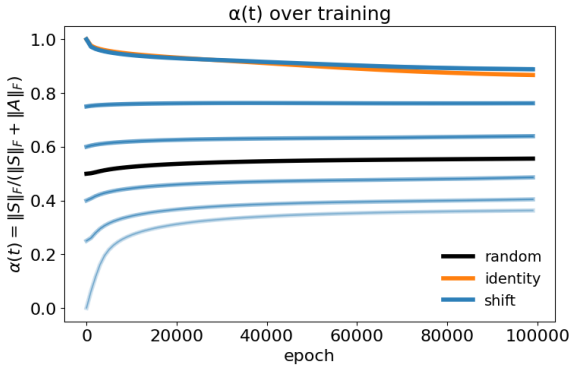

Figure S6: **Training drives recurrent connectivity toward an intermediate symmetry regime.** Evolution of the symmetry index $\alpha(t) = ||S||_F/(||S||_F + ||A||_F)$ during training for networks initialized with random (black), identity (orange), and shift (blue; multiple initial α_0) recurrent connectivity. Across initial conditions, learning consistently drives $\alpha(t)$ toward an intermediate range, indicating convergence to a mixed symmetric-antisymmetric recurrent regime.

(i) Ring decoding via cosine-sine regression

We evaluate decoding of a circular variable using linear (ridge) regression on cosine and sine targets, reporting performance via R^2 [69]. Decoding $\cos\theta$ and $\sin\theta$ avoids angular wrap-around discontinuities.

At each time step, the network produces an N-dimensional population activity vector interpreted as a ring code. Outputs are rectified and normalized to form a population distribution,

$$x_t = \text{normalize}(y_t) = \frac{\max(y_t, 0)}{\sum_{i=1}^{N} \max(y_{t,i}, 0) + \epsilon} \tag{6.1}$$

Each unit i is assigned a preferred angle $\phi_i = 2\pi i/N$. Given the target ring distribution p_t, the target angle is computed via its first circular moments

$$C_t = \sum_{i=1}^{N} p_{t,i}\cos(\phi_i), \quad S_t = \sum_{i=1}^{N} p_{t,i}\sin(\phi_i), \quad \theta_t = \text{atan2}(S_t, C_t) \tag{6.2}$$

and represented as $Y_t = [\cos\theta_t, \sin\theta_t]$

Stacking all time points yields $X \in \mathbb{R}^{T\times N}$ and $Y \in \mathbb{R}^{T\times 2}$. A ridge decoder is fit as

$$W = (X^\top X + \lambda I)^{-1} X^\top Y, \quad \hat{Y} = XW \tag{6.3}$$

Performance is quantified by computing R^2 separately for cosine and sine and averaging

$$\text{ring decode } R^2 = \frac{1}{2}\left(R^2_{\cos} + R^2_{\sin}\right) \tag{6.4}$$

This metric measures time-resolved fidelity to the trained target phase θ_t. It is not invariant to phase shifts, speed changes, or temporal offsets. For replay, it reflects spontaneous reproduction of the learned sequence under noise input. For prediction, it reflects accuracy in tracking the target trajectory given an initial drive. Values lie in $(-\infty, 1]$, with $R^2 = 1$ indicating a correctly positioned, stable activity bump at all time steps.

(ii) Lag-1 autocorrelation

Lag-1 autocorrelation quantifies temporal dependence in the magnitude of model residuals. For each time step, we compute the residual magnitude

$$r_t = ||\text{Target}_t - \text{Output}_t||_2 \tag{6.5}$$

Then demeaning the resulting time series

$$x_t = r_t - \text{mean}_t(r_t) \tag{6.6}$$

and compute the correlation between $(x_1, ..., x_{T-1})$ and $(x_2, ..., x_T)$ (standard lag-1 autocorrelation).

Values lie in $[-1, 1]$. Zero indicates temporally uncorrelated (white-noise-like) residuals, positive values indicate smooth temporal persistence of errors, and negative values indicate alternating structure. High lag-1 autocorrelation reflects slowly varying residual magnitudes consistent with coherent underlying dynamics [70].

(iii) Mean resultant length (angular error concentration)

This metric quantifies how concentrated the circular angular error $\Delta\theta(t)$ is over time between the true angle implied by the target ring distribution and the angle decoded from the model output [71]. It is also known as the mean resultant length, vector strength, or phase-locking value.

At each time step, target and model output activities are converted to angles, and the wrapped angular error $\Delta\theta_t$ is computed. Each error is represented as a unit vector

Let $\Delta\theta_{b,t}$ be the wrapped circular error between the predicted and ground truth angle. We can define unit vectors as

Then demean according to

$$u_t = (\cos\Delta\theta_t, \sin\Delta\theta_t) \tag{6.7}$$

which are averaged over time

$$\bar{u} = (\langle\cos\Delta\theta\rangle, \langle\sin\Delta\theta\rangle) \tag{6.8}$$

The mean resultant length is

$$R = ||\bar{u}|| = \sqrt{\langle\cos\Delta\theta\rangle^2 + \langle\sin\Delta\theta\rangle^2} \tag{6.9}$$

$R \in [0, 1]$, where $R = 1$ indicates highly consistent angular errors (strong concentration near a single phase, typically zero) and $R = 0$ indicates uniformly distributed, random angular errors.

7. Symmetric STDP Analysis

The analytic form of the final weights is derived by convolving the Effective STDP Kernel (which includes the symmetry correction ϵ) with the Autocorrelation of the traveling Gaussian input. This is derived from the STDP learning kernel defined as

$$W_{ij} = \int K(s) \int u_i^{(t)}\, u_j^{(t-s)}\, \mathrm{d}t\, \mathrm{d}s + \epsilon \int K(s) \int u_j^{(t)}\, u_i^{(t-s)}\, \mathrm{d}t\, \mathrm{d}s \tag{7.1}$$

where

$$K(s) = \begin{cases} A_+\, e^{-s/\tau_+} & s \geq 0 \\ -A_-\, e^{s/\tau_-} & s < 0 \end{cases} \quad \text{and} \quad u_i^{(t)} = \mathcal{N}(a, b, \mu)[i] = a e^{-b\,(i-(\mu+t))^2} \tag{7.2}$$

First computing the product $u_i^{(t)} u_j^{(t-s)}$

$$u_i^{(t)}\, u_j^{(t-s)} = \mathcal{N}(a, b, \mu + t)[i]\, \mathcal{N}(a, b, \mu + t - s)[j] \tag{7.3}$$

$$u_i^{(t)}\, u_j^{(t-s)} = \mathcal{N}(a, b, \mu + t)[i]\, \mathcal{N}(a, b, i - j + \mu + t - s)[i] \tag{7.4}$$

$$u_i^{(t)}\, u_j^{(t-s)} = \mathcal{N}(a, b, \mu - i)[-t]\, \mathcal{N}(a, b, \mu - (j + s))[-t] \tag{7.5}$$

Using Lemma 1,

$$u_i^{(t)}\, u_j^{(t-s)} = \mathcal{N}\left(a^2\, e^{-\frac{b(s+j-i)^2}{2}}, 2b, \mu - \frac{(i + j + s)}{2}\right)[-t] \tag{7.6}$$

Now the autocorrelation,

$$\int_{-\infty}^{\infty} u_i^{(t)}\, u_j^{(t-s)}\, \mathrm{d}t = \int_{-\infty}^{\infty} \mathcal{N}\left(a^2\, e^{-\frac{b(s+j-i)^2}{2}}, 2b, \mu - \frac{(i + j + s)}{2}\right)[-t]\, \mathrm{d}t \tag{7.7}$$

Using Lemma 2

$$= a^2 \sqrt{\frac{\pi}{2b}}\, e^{-\frac{b(s-(i-j))^2}{2}} = \mathcal{N}\left(a^2 \sqrt{\frac{\pi}{2b}}, \frac{b}{2}, (i - j)\right)[s] \tag{7.8}$$

Computing $W_{ij}^{(\epsilon=0)}$ (the weight matrix without any symmetric correction) using autocorrelations

$$W_{ij}^{(\epsilon=0)} = -A_- \int_{-\infty}^{0} e^{s/\tau_-}\, \mathcal{N}\left(a^2 \sqrt{\frac{\pi}{2b}}, \frac{b}{2}, (i - j)\right)[s]\, \mathrm{d}s + A_+ \int_{0}^{\infty} e^{-s/\tau_+}\, \mathcal{N}\left(a^2 \sqrt{\frac{\pi}{2b}}, \frac{b}{2}, (i - j)\right)[s]\, \mathrm{d}s \tag{7.9}$$

$$W_{ij}^{(\epsilon=0)} = -A_- a^2 \sqrt{\frac{\pi}{2b}} \int_{-\infty}^{0} e^{s/\tau_-}\, \mathcal{N}\left(1, \frac{b}{2}, (i - j)\right)[s]\, \mathrm{d}s + A_+ a^2 \sqrt{\frac{\pi}{2b}} \int_{0}^{\infty} e^{-s/\tau_+}\, \mathcal{N}\left(1, \frac{b}{2}, (i - j)\right)[s]\, \mathrm{d}s \tag{7.10}$$

using Lemma 4

$$W_{ij}^{(\epsilon=0)} = -A_- a^2 \sqrt{\frac{\pi}{2b}} \int_{-\infty}^{0} \mathcal{N}\left(\exp\left(\frac{i-j}{\tau_-} + \frac{1}{b\tau_-^2}\right), \frac{b}{2}, (i-j) + \frac{1}{b\tau_-}\right)[s]\,\mathrm{d}s$$
$$+ A_+ a^2 \sqrt{\frac{\pi}{2b}} \int_{0}^{\infty} \mathcal{N}\left(\exp\left(-\frac{i-j}{\tau_+} + \frac{1}{b\tau_+^2}\right), \frac{b}{2}, (i-j) - \frac{1}{b\tau_+}\right)[s]\,\mathrm{d}s \tag{7.11}$$

Let $\Delta = i - j$,

$$W_{ij}^{(\epsilon=0)} = \frac{\pi a^2}{2b}\left[A_+ e^{\left(\frac{1}{b\tau_+^2} - \frac{\Delta}{\tau_+}\right)} \operatorname{erfc}\left(\frac{1 - b\tau_+\Delta}{\tau_+\sqrt{2b}}\right) - A_- e^{\left(\frac{1}{b\tau_-^2} + \frac{\Delta}{\tau_-}\right)} \operatorname{erfc}\left(\frac{1 + b\tau_-\Delta}{\tau_-\sqrt{2b}}\right)\right] \tag{7.12}$$

When the STDP has the symmetric correction (ϵ), we have:

$$W_{ij} = \frac{\pi a^2}{2b}\left[A_+ e^{\left(\frac{1}{b\tau_+^2} - \frac{\Delta}{\tau_+}\right)} \operatorname{erfc}\left(\frac{1 - b\tau_+\Delta}{\tau_+\sqrt{2b}}\right) - A_- e^{\left(\frac{1}{b\tau_-^2} + \frac{\Delta}{\tau_-}\right)} \operatorname{erfc}\left(\frac{1 + b\tau_-\Delta}{\tau_-\sqrt{2b}}\right)\right]$$
$$+ \epsilon\frac{\pi a^2}{2b}\left[A_+ e^{\left(\frac{1}{b\tau_+^2} + \frac{\Delta}{\tau_+}\right)} \operatorname{erfc}\left(\frac{1 + b\tau_+\Delta}{\tau_+\sqrt{2b}}\right) - A_- e^{\left(\frac{1}{b\tau_-^2} - \frac{\Delta}{\tau_-}\right)} \operatorname{erfc}\left(\frac{1 - b\tau_-\Delta}{\tau_-\sqrt{2b}}\right)\right] \tag{7.13}$$

(a) Eigenvalues

Eigenvalues control the dynamics the RNN has; however, they are difficult to compute directly from the equations for W_{ij}. Instead, we compute it from the learning kernel in Equation 3.1 directly. Thus, the resultant W_{ij} is a Toeplitz matrix since W_{ij} is a function of $\Delta = i - j$, which are asymptotically diagonalized by the DFT matrix. We have

$$W_{ij} = \int K(s) \int u_i^{(t)}\, u_j^{(t-s)}\,\mathrm{d}t\,\mathrm{d}s + \epsilon \int K(s) \int u_j^{(t)}\, u_i^{(t-s)}\,\mathrm{d}t\,\mathrm{d}s \tag{7.14}$$

$$W_{ij} = a^2\sqrt{\frac{\pi}{2b}}\left(\int K(s)\mathcal{N}\left(1, \frac{b}{2}, \Delta\right)[s]\,\mathrm{d}s + \epsilon\int K(s)\mathcal{N}\left(1, \frac{b}{2}, -\Delta\right)[s]\,\mathrm{d}s\right) \tag{7.15}$$

$$W_{ij} = a^2\sqrt{\frac{\pi}{2b}}\left(\int K(s)\mathcal{N}\left(1, \frac{b}{2}, s - \Delta\right)\mathrm{d}s + \epsilon\int K(s)\mathcal{N}\left(1, \frac{b}{2}, s + \Delta\right)\mathrm{d}s\right) \tag{7.16}$$

Using the convolution theorem, we have

$$W_{ij} = a^2\sqrt{\frac{\pi}{2b}}\mathcal{F}^{-1}\left(\mathcal{F}[K(s)]\,\mathcal{F}\left[\mathcal{N}\left(1, \frac{b}{2}, s\right)\right] + \epsilon\mathcal{F}[K(s)]^\dagger\,\mathcal{F}\left[\mathcal{N}\left(1, \frac{b}{2}, s\right)\right]\right) \tag{7.17}$$

$$= a^2\sqrt{\frac{\pi}{2b}}\sqrt{\frac{2\pi}{b}}\mathcal{F}^{-1}\left[e^{\left(-\frac{\omega^2}{2b}\right)}\left(\mathcal{F}[K(s)] + \epsilon\mathcal{F}[K(s)]^\dagger\right)\right] \tag{7.18}$$

$$= a^2\sqrt{\frac{\pi}{2b}}\sqrt{\frac{2\pi}{b}}\mathcal{F}^{-1}\left[\left(\frac{A_+}{\frac{1}{\tau_+} + i\omega} - \frac{A_-}{\frac{1}{\tau_-} - i\omega}\right)e^{\left(-\frac{\omega^2}{2b}\right)}(1+\epsilon)\right] \tag{7.19}$$

$$= \frac{\pi a^2}{b}\mathcal{F}^{-1}\left[e^{\left(-\frac{\omega^2}{2b}\right)}\left((1+\epsilon)\left(\frac{A_+\tau_+}{1+\omega^2\tau_+^2} - \frac{A_-\tau_-}{1+\omega^2\tau_-^2}\right) - \mathbf{i}(1-\epsilon)\left(\frac{A_+\tau_+^2\omega}{1+\omega^2\tau_+^2} - \frac{A_-\tau_-^2\omega}{1+\omega^2\tau_-^2}\right)\right)\right] \tag{7.20}$$

From this, we obtain the eigenvalue expression

$$\lambda(\omega) = \frac{\pi a^2}{b}e^{\left(-\frac{\omega^2}{2b}\right)}\left((1+\epsilon)\left(\frac{A_+\tau_+}{1+\omega^2\tau_+^2} - \frac{A_-\tau_-}{1+\omega^2\tau_-^2}\right) - \mathbf{i}(1-\epsilon)\left(\frac{A_+\tau_+^2\omega}{1+\omega^2\tau_+^2} - \frac{A_-\tau_-^2\omega}{1+\omega^2\tau_-^2}\right)\right) \tag{7.21}$$

(b) Useful Lemmas

Lemma 1 (Product of Gaussian). *If* $\mathcal{N}(a, b, \mu) = ae^{-b(x-\mu)^2}$, *then*

$$\mathcal{N}(a_1, b_1, \mu_1)\mathcal{N}(a_2, b_2, \mu_1) = \mathcal{N}\left(a_1 a_2\, e^{-\frac{b_1 b_2(\mu_1-\mu_2)^2}{(b_1+b_2)}}, (b_1+b_2), \frac{b_1\mu_1 + b_2\mu_2}{b_1+b_2}\right) \tag{7.22}$$

Proof. Let the product of the two Gaussian functions be defined as $f(x) = \mathcal{N}(a_1, b_1, \mu_1)\,\mathcal{N}(a_2, b_2, \mu_2)$. By substituting the definition $\mathcal{N}(a, b, \mu) = ae^{-b(x-\mu)^2}$, we obtain:

$$f(x) = a_1 a_2 \exp\left(-\left[b_1(x-\mu_1)^2 + b_2(x-\mu_2)^2\right]\right) \tag{7.23}$$

Let $Q(x)$ denote the quadratic form in the exponent. Expanding the terms, we have:

$$\begin{aligned} Q(x) &= b_1(x^2 - 2x\mu_1 + \mu_1^2) + b_2(x^2 - 2x\mu_2 + \mu_2^2) \\ &= (b_1 + b_2)x^2 - 2x(b_1\mu_1 + b_2\mu_2) + (b_1\mu_1^2 + b_2\mu_2^2) \end{aligned} \tag{7.24}$$

To express $Q(x)$ in the standard form $b_c(x-\mu_c)^2 + K$, we define the combined precision b_c and combined mean μ_c as:

$$b_c = b_1 + b_2, \quad \mu_c = \frac{b_1\mu_1 + b_2\mu_2}{b_1 + b_2} \tag{7.25}$$

Rewriting the quadratic form by completing the square with respect to x:

$$Q(x) = (b_1 + b_2)\left(x - \frac{b_1\mu_1 + b_2\mu_2}{b_1 + b_2}\right)^2 + K \tag{7.26}$$

where the residual constant K is:

$$K = b_1\mu_1^2 + b_2\mu_2^2 - \frac{(b_1\mu_1 + b_2\mu_2)^2}{b_1 + b_2} \tag{7.27}$$

Simplifying K by finding a common denominator:

$$\begin{aligned} K &= \frac{(b_1\mu_1^2 + b_2\mu_2^2)(b_1 + b_2) - (b_1^2\mu_1^2 + 2b_1b_2\mu_1\mu_2 + b_2^2\mu_2^2)}{b_1 + b_2} \\ &= \frac{b_1b_2\mu_1^2 + b_1b_2\mu_2^2 - 2b_1b_2\mu_1\mu_2}{b_1 + b_2} \\ &= \frac{b_1b_2(\mu_1 - \mu_2)^2}{b_1 + b_2} \end{aligned} \tag{7.28}$$

Substituting $Q(x) = b_c(x-\mu_c)^2 + K$ back into the exponential function:

$$\begin{aligned} f(x) &= a_1a_2 \exp(-K) \exp\left(-b_c(x-\mu_c)^2\right) \\ &= \left[a_1a_2 \exp\left(-\frac{b_1b_2(\mu_1-\mu_2)^2}{b_1+b_2}\right)\right] \exp\left(-(b_1+b_2)(x-\mu_c)^2\right) \end{aligned} \tag{7.29}$$

By inspection, this result corresponds to the function $\mathcal{N}(a_c, b_c, \mu_c)$. □

Lemma 2 (Gaussian Integral)**.** *If* $\mathcal{N}(a, b, \mu) = ae^{-b(x-\mu)^2}$*, then*

$$\int_{-\infty}^{\infty} \mathcal{N}(a, b, \mu)\,\mathrm{d}x = a\sqrt{\frac{\pi}{b}} \tag{7.30}$$

Proof. Let the integral be denoted by I. We aim to evaluate:

$$I = \int_{-\infty}^{\infty} ae^{-b(x-\mu)^2}\,\mathrm{d}x \tag{7.31}$$

First, we factor out the constant amplitude a:

$$I = a\int_{-\infty}^{\infty} e^{-b(x-\mu)^2}\,\mathrm{d}x \tag{7.32}$$

We perform a change of variables to simplify the exponent. Let $u = \sqrt{b}(x-\mu)$. Consequently, the differential transforms as:

$$\mathrm{d}u = \sqrt{b}\,\mathrm{d}x \quad \Longrightarrow \quad \mathrm{d}x = \frac{1}{\sqrt{b}}\,\mathrm{d}u \tag{7.33}$$

The limits of integration remain from $-\infty$ to ∞. Substituting these into the integral:

$$I = a\int_{-\infty}^{\infty} e^{-u^2}\left(\frac{1}{\sqrt{b}}\right)\mathrm{d}u = \frac{a}{\sqrt{b}}\int_{-\infty}^{\infty} e^{-u^2}\,\mathrm{d}u \tag{7.34}$$

We recall the standard Gaussian integral identity:

$$\int_{-\infty}^{\infty} e^{-u^2}\,\mathrm{d}u = \sqrt{\pi} \tag{7.35}$$

Substituting this standard result back into our expression for I:

$$I = \frac{a}{\sqrt{b}}\left(\sqrt{\pi}\right) = a\sqrt{\frac{\pi}{b}} \tag{7.36}$$

This completes the proof. □

Lemma 3 (Gaussian Integral Bounds). *If* $\mathcal{N}(a,b,\mu) = ae^{-b(x-\mu)^2}$, *then*

$$\int_L^U \mathcal{N}(a,b,\mu)\,\mathrm{d}x = a\frac{1}{2}\sqrt{\frac{\pi}{b}}\left[erf\left(\sqrt{b}(U-\mu)\right) - erf\left(\sqrt{b}(L-\mu)\right)\right] \tag{7.37}$$

Proof. Let the integral be denoted by I. We aim to evaluate the definite integral:

$$I = \int_L^U ae^{-b(x-\mu)^2}\,\mathrm{d}x \tag{7.38}$$

First, we factor out the constant amplitude a:

$$I = a\int_L^U e^{-b(x-\mu)^2}\,\mathrm{d}x \tag{7.39}$$

We perform a change of variables to standardize the exponent. Let $t = \sqrt{b}(x-\mu)$. The differential transforms as $\mathrm{d}x = \frac{1}{\sqrt{b}}\,\mathrm{d}t$. We must also transform the limits of integration:

- When $x = L$, $t_L = \sqrt{b}(L-\mu)$.
- When $x = U$, $t_U = \sqrt{b}(U-\mu)$.

Substituting these into the integral:

$$I = \frac{a}{\sqrt{b}}\int_{t_L}^{t_U} e^{-t^2}\,\mathrm{d}t \tag{7.40}$$

We recall the definition of the error function, $\mathrm{erf}(z)$, which is defined as:

$$\mathrm{erf}(z) = \frac{2}{\sqrt{\pi}}\int_0^z e^{-t^2}\,\mathrm{d}t \tag{7.41}$$

Using the property of definite integrals that $\int_A^B f(t)\,\mathrm{d}t = \int_0^B f(t)\,\mathrm{d}t - \int_0^A f(t)\,\mathrm{d}t$, we can rewrite our integral in terms of the error function:

$$\int_{t_L}^{t_U} e^{-t^2}\,\mathrm{d}t = \left(\int_0^{t_U} e^{-t^2}\,\mathrm{d}t - \int_0^{t_L} e^{-t^2}\,\mathrm{d}t\right) \tag{7.42}$$

From the definition of $\mathrm{erf}(z)$, we have $\int_0^z e^{-t^2}\,\mathrm{d}t = \frac{\sqrt{\pi}}{2}\mathrm{erf}(z)$. Therefore:

$$\int_{t_L}^{t_U} e^{-t^2}\,\mathrm{d}t = \frac{\sqrt{\pi}}{2}\mathrm{erf}(t_U) - \frac{\sqrt{\pi}}{2}\mathrm{erf}(t_L) = \frac{\sqrt{\pi}}{2}\left[\mathrm{erf}(t_U) - \mathrm{erf}(t_L)\right] \tag{7.43}$$

Substituting this result back into the expression for I:

$$I = \frac{a}{\sqrt{b}}\left(\frac{\sqrt{\pi}}{2}\left[\mathrm{erf}(t_U) - \mathrm{erf}(t_L)\right]\right) \tag{7.44}$$

Rearranging terms and substituting the definitions of t_U and t_L:

$$I = a\frac{1}{2}\sqrt{\frac{\pi}{b}}\left[\mathrm{erf}\left(\sqrt{b}(U-\mu)\right) - \mathrm{erf}\left(\sqrt{b}(L-\mu)\right)\right] \tag{7.45}$$

□

Lemma 4 (Gaussian Exp Product). *If* $\mathcal{N}(a,b,\mu) = ae^{-b(x-\mu)^2}$, *then*

$$e^{cx}\,\mathcal{N}(a,b,\mu) = \mathcal{N}\left(a\exp\left(c\mu + \frac{c^2}{4b}\right), b, \mu + \frac{c}{2b}\right) \tag{7.46}$$

Proof. Let the product function be $f(x) = e^{cx}\,\mathcal{N}(a, b, \mu)$. Substituting the definition of the Gaussian function:

$$f(x) = e^{cx} \cdot ae^{-b(x-\mu)^2} = a \exp\left(cx - b(x-\mu)^2\right) \tag{7.47}$$

Let $E(x)$ denote the exponent. We expand the quadratic term to group coefficients of x:

$$\begin{aligned} E(x) &= cx - b(x^2 - 2x\mu + \mu^2) \\ &= -bx^2 + 2bx\mu + cx - b\mu^2 \\ &= -bx^2 + (2b\mu + c)x - b\mu^2 \end{aligned} \tag{7.48}$$

To restore the Gaussian form $-b(x-\mu_{\text{new}})^2 + K$, we factor out $-b$ from the terms involving x:

$$E(x) = -b\left(x^2 - \frac{2b\mu + c}{b}x\right) - b\mu^2 \tag{7.49}$$

We define the new mean μ_{new} as half the coefficient of the linear term inside the parenthesis:

$$\mu_{\text{new}} = \frac{1}{2}\left(\frac{2b\mu + c}{b}\right) = \mu + \frac{c}{2b} \tag{7.50}$$

Completing the square inside the parenthesis by adding and subtracting μ_{new}^2:

$$\begin{aligned} E(x) &= -b\left(x^2 - 2x\mu_{\text{new}} + \mu_{\text{new}}^2 - \mu_{\text{new}}^2\right) - b\mu^2 \\ &= -b(x-\mu_{\text{new}})^2 + b\mu_{\text{new}}^2 - b\mu^2 \end{aligned} \tag{7.51}$$

We simplify the constant terms (the residual) $K = b\mu_{\text{new}}^2 - b\mu^2$:

$$\begin{aligned} K &= b\left(\mu + \frac{c}{2b}\right)^2 - b\mu^2 \\ &= b\left(\mu^2 + \frac{2\mu c}{2b} + \frac{c^2}{4b^2}\right) - b\mu^2 \\ &= b\mu^2 + c\mu + \frac{c^2}{4b} - b\mu^2 \\ &= c\mu + \frac{c^2}{4b} \end{aligned} \tag{7.52}$$

Substituting $E(x) = -b(x-\mu_{\text{new}})^2 + K$ back into the original expression for $f(x)$:

$$f(x) = a\exp(K)\exp\left(-b(x-\mu_{\text{new}})^2\right) \tag{7.53}$$

Substituting the values of K and μ_{new}:

$$f(x) = \left[a\exp\left(c\mu + \frac{c^2}{4b}\right)\right]\exp\left(-b\left(x - \left(\mu + \frac{c}{2b}\right)\right)^2\right) \tag{7.54}$$

This matches the definition of $\mathcal{N}(a_{\text{new}}, b, \mu_{\text{new}})$. □