Testing the Associative Memory Interpretation of Protein Network Dynamics

Sreevatsa Nukala
AM226
Harvard University
snukala@hspsh.harvard.edu

1 Introduction

Understanding how biological networks process, store, and recover information is a central question in theoretical neuroscience and systems biology. Recent work has proposed interpreting protein interaction dynamics through the lens of associative memory, drawing an analogy between recurrent neural networks and biochemical regulatory systems: nonlinear recurrent models trained on protein expression dynamics have been claimed to encode stable biological states as attractors, enabling robust recovery of cell-cycle phases or expression patterns from partial or noisy observations [Samarasinghe et al., 2025].

Motivated by these claims, we ask whether learned recurrent models actually exhibit the global properties implied by the associative-memory hypothesis, such as stable attractors and basin-based recall. To address this question, we compare linear and nonlinear associative artificial neural networks (AANNs) trained under standard objectives and evaluate their behavior using long-horizon dynamical tests inspired by associative memory theory. We first formalize the dynamical assumptions underlying the associative-memory interpretation, then describe the surrogate generators, learned models, and evaluation procedures used in our study.

2 Problem Setup

The associative-memory interpretation implicitly treats protein interaction networks as autonomous dynamical systems with stable attractors and well-defined basins of attraction. To critically examine this claim, we first formalize associative memory in dynamical systems terms, providing the framework used throughout the paper to evaluate whether such attractor-based structure is recovered by learned recurrent models.

2.1 Dynamical systems and attractors

Let $x(t) \in \mathbb{R}^N$ denote the state of a recurrent system at time t. We consider autonomous continuous-time dynamics of the form

$$\dot{x}(t) = F(x(t)), \tag{1}$$

as well as discrete-time dynamical systems defined by an iterated map

$$x_{t+1} = f(x_t). \tag{2}$$

A point $x^* \in \mathbb{R}^N$ is a fixed point of the system if $F(x^*) = 0$ in continuous time, or equivalently $x^* = f(x^*)$ in discrete time. A fixed point is an attractor if trajectories initialized in a neighborhood of x^* converge to it asymptotically; the set of initial conditions that converge to a given attractor is its basin of attraction.

More generally, autonomous dynamical systems may exhibit non-point attractors, such as limit cycles or invariant manifolds, which similarly induce structured long-horizon behavior.

Preprint. Under review.

2.2 Associative memory as a dynamical property

We adopt a dynamical definition of associative memory [e.g. Hopfield, 1982, Izhikevich, 2007]: a system exhibits associative memory if it can reliably recover structured states from perturbed initial conditions through its intrinsic dynamics. Under this view, memory retrieval is not an instantaneous computation but a process of convergence under repeated application of the system dynamics.

Formally, let $\{x^{(1)}, \ldots, x^{(K)}\}$ denote a set of reference states. Associative recall requires that for each k, there exists a neighborhood U_k such that for any $\tilde{x}_0 \in U_k$, the iterated dynamics satisfy

$$\lim_{n \to \infty} f^n(\tilde{x}_0) = x^{(k)}, \tag{3}$$

or, in the case of non-point attractors, converge to the corresponding invariant set. This definition emphasizes robustness to perturbations and long-horizon stability rather than short-term prediction accuracy.

2.3 Recurrent neural networks as dynamical maps

We model learned recurrent systems as parameterized maps $f_\theta : \mathbb{R}^N \to \mathbb{R}^N$, where θ denotes the network parameters. In this work, we focus on affine and nonlinear update rules of the form

$$f_\theta(x) = Wx + b \quad \text{(linear)}, \tag{4}$$

and

$$f_\theta(x) = \sigma(Wx + b) \quad \text{(nonlinear)}, \tag{5}$$

where $W \in \mathbb{R}^{N \times N}$, $b \in \mathbb{R}^N$, and $\sigma(\cdot)$ is a bounded, monotone nonlinearity applied elementwise.

Although such architectures are expressive enough to represent multiple fixed points and attractors, the existence of these structures in the learned system depends not only on representational capacity but also on the learning objective and the information available during training.

2.4 Learning objectives and long-horizon behavior

Given observed trajectories $\{x_t\}_{t=0}^T$, a common learning objective for recurrent models is one-step prediction error,

$$\mathcal{L}_{\text{MSE}}(\theta) = \mathbb{E}_t \left[\|x_{t+1} - f_\theta(x_t)\|^2 \right]. \tag{6}$$

This objective enforces local consistency between successive states but does not directly constrain the asymptotic behavior of the iterated map $f_\theta^{(k)}$. Minimizing one-step error therefore does not guarantee preservation of attractors, basin geometry, or long-horizon stability, and a learned system may achieve low training error while exhibiting qualitatively different global dynamics from the data-generating process.

This distinction between local predictive accuracy and global dynamical structure motivates the evaluation of learned recurrent models through long-horizon rollouts and perturbation-based recall tests, which directly probe the existence and stability of attractors.

3 Ground-Truth Dynamical Systems

The associative-memory interpretation presupposes that the underlying biological system exhibits structured long-horizon behavior, such as stable fixed points or oscillatory attractors, that can in principle be recovered by a learned recurrent model. In the original work motivating this project, such structure arises from a detailed nonlinear system of ordinary differential equations (ODEs) describing regulatory interactions among cell-cycle proteins.

Our goal is not to reproduce this biochemical model, but to evaluate whether the associative-memory interpretation implied by such dynamics is recovered under standard learning objectives. Directly reproducing the full ODE system would introduce substantial mechanistic complexity and parameter sensitivity, making it difficult to disentangle failures of learning from idiosyncrasies of a particular biological model. Instead, we study a hierarchy of dynamical systems that abstract away biochemical detail while preserving dynamical properties essential to associative memory, enabling controlled analysis of whether learning recovers attractor structure independent of biological specificity.

3.1 Design principles for surrogate dynamics

Rather than attempting to reproduce biochemical mechanisms in detail, we focus on dynamical properties that are essential for an associative-memory interpretation:

- bounded state variables, reflecting normalized protein activity levels,
- nonlinear interactions capable of producing multistability or oscillations,
- autonomous dynamics with well-defined long-horizon behavior,
- robustness to perturbations, enabling meaningful recall tests.

Each dynamical system considered below satisfies these criteria, enabling isolation of the relationship between learning objectives and recovery of attractor structure.

3.2 Oscillatory dynamics via Hopf normal form

As a canonical example of nonlinear autonomous dynamics with a stable attractor, we consider networks of coupled Hopf (Stuart–Landau) oscillators [Strogatz, 2018]. For each unit i, the complex-valued state $z_i(t) \in \mathbb{C}$ evolves according to

$$\dot{z}_i(t) = (\mu - |z_i(t)|^2)z_i(t) + i\omega_i z_i(t) + \kappa \sum_j W_{ij} z_j(t), \tag{7}$$

where $\mu > 0$ controls the existence of a stable limit cycle, ω_i is the intrinsic frequency, κ is a coupling strength, and W encodes network connectivity.

For $\mu > 0$, the uncoupled system admits a stable limit cycle, while coupling induces phase-locked attractors corresponding to coherent oscillatory states. These attractors are robust to perturbations and depend on initial conditions, making the system a continuous-time analog of associative memory in which recall corresponds to convergence to a particular phase-locked regime.

To obtain real-valued observations comparable to protein activity measurements, we define the observed state as

$$x_i(t) = \sigma(\mathrm{Re}(z_i(t))), \tag{8}$$

where $\sigma(\cdot)$ is a bounded monotone nonlinearity applied elementwise. This observation map preserves the attractor structure while enforcing biologically plausible bounds.

3.3 Phase-based abstractions of cell-cycle dynamics

To further isolate attractor structure from mechanistic detail, we also consider a coarse-grained phase-based abstraction of cell-cycle dynamics, inspired by biochemical oscillator models [Novák and Tyson, 2008]. In this model, the system transitions among a small number of canonical phase states $\{x^{(1)}, \ldots, x^{(K)}\} \subset \mathbb{R}^N$, each representing a prototypical protein expression profile.

Dynamics are defined by relaxation toward the current phase state, with stochastic or deterministic transitions between phases. In discrete time, this can be written as

$$x_{t+1} = x_t + \alpha\big(x^{(k_t)} - x_t\big) + \xi_t, \tag{9}$$

where k_t denotes the active phase at time t, $\alpha > 0$ controls the strength of attraction, and ξ_t is small process noise.

Despite its simplicity, this abstraction explicitly encodes multiple attractors with well-defined basins and provides a controlled setting in which associative recall can be tested independently of oscillatory dynamics or detailed regulatory mechanisms.

3.4 Relationship to the original biochemical model

The oscillatory and phase-based systems studied here are not intended as biological replacements for the original biochemical ODE model. Rather, they serve as controlled surrogates that preserve the dynamical properties of nonlinearity, autonomy, and long-horizon structure required by the associative-memory interpretation. By examining learning behavior across this hierarchy of generators, we assess

whether recovery of attractor structure depends on biological specificity or instead reflects more fundamental constraints imposed by learning objectives and model class.

In the following sections, we train recurrent neural networks on trajectories generated by these systems and evaluate whether the learned dynamics preserve the attractor-based structure implied by the associative-memory hypothesis.

4 Learned Models and Training Objectives

Having defined a hierarchy of dynamical systems that exhibit structured long-horizon behavior, we now describe the class of recurrent models used to learn from generated trajectories and the objective under which learning is performed. All datasets considered in this work, regardless of their underlying generator, are used to train the same class of recurrent maps under an identical learning objective, allowing us to isolate the effect of the training criterion from properties specific to any individual dynamical system.

4.1 Associative artificial neural networks

We model learned dynamics using associative artificial neural networks (AANNs), implemented as recurrent maps $f_\theta : \mathbb{R}^N \to \mathbb{R}^N$ of the form

$$x_{t+1} = f_\theta(x_t). \tag{10}$$

We consider two commonly studied parameterizations:

$$f_\theta(x) = Wx + b \quad \text{(linear AANN)}, \tag{11}$$
$$f_\theta(x) = \sigma(Wx + b) \quad \text{(nonlinear AANN)}, \tag{12}$$

where $W \in \mathbb{R}^{N \times N}$, $b \in \mathbb{R}^N$, and $\sigma(\cdot)$ is a bounded monotone nonlinearity applied elementwise.

Both models define autonomous discrete-time dynamical systems whose long-horizon behavior is determined by the spectrum and geometry of the learned weight matrix. While nonlinear AANNs are capable of representing multiple fixed points and complex attractor structure in principle, such behavior is not guaranteed to emerge under arbitrary training objectives.

4.2 Connectivity constraints

To reflect biological sparsity and to remain consistent with the original work motivating this study [Samarasinghe et al., 2025], we impose structured connectivity constraints on the recurrent weight matrix W in most experiments. Specifically, W is masked by a fixed binary adjacency matrix encoding allowed interactions between state variables. This constraint limits the space of admissible dynamical maps and enforces locality of interactions, but does not otherwise alter the form of the update rule.

We emphasize that connectivity constraints affect expressivity but are independent of the learning objective itself. In later sections, we contrast masked and fully connected settings to disentangle the effects of architectural restrictions from those of optimization and objective design.

4.3 Training objective

Given a set of observed trajectories $\{x_t\}_{t=0}^T$ generated by an underlying dynamical system, all models are trained using a one-step prediction objective,

$$\mathcal{L}_{\text{MSE}}(\theta) = \mathbb{E}_t \left[\|x_{t+1} - f_\theta(x_t)\|^2 \right]. \tag{13}$$

This objective encourages the learned map to locally approximate the transition dynamics of the data-generating process in expectation.

Crucially, this objective does not directly constrain the global behavior of the iterated map $f_\theta^{(k)}$, which governs long-horizon stability, basin geometry, and attractor structure. As a result, minimizing one-step error may yield a learned system whose asymptotic dynamics differ qualitatively from those of the original generator, even when training error is low.

4.4 Scope and implications

By applying the same architecture class and training objective across a range of generators with known attractor structure, we can directly assess whether standard learning procedures recover the properties implied by the associative-memory interpretation and distinguish failures arising from insufficient model capacity from those arising from a mismatch between the learning objective and the dynamical property of interest. The following section evaluates the learned models using diagnostics that probe long-horizon behavior and perturbation-based recall, rather than relying solely on short-term prediction error.

5 Evaluation as a Dynamical Test

If associative memory is understood as a property of long-horizon dynamics, then evaluation of learned recurrent models must directly probe their asymptotic behavior rather than rely solely on short-term predictive accuracy, in contrast to many standard predictive objectives used for recurrent models [e.g. Jaeger, 2001, Sussillo and Abbott, 2009].

5.1 Open-loop rollouts

Given a learned recurrent map f_θ, we examine its autonomous behavior by iterating the system in open loop,

$$x_{t+1} = f_\theta(x_t), \tag{14}$$

starting from initial conditions drawn from observed trajectories. Open-loop rollouts reveal whether the learned dynamics remain stable, diverge, or collapse toward fixed points over long time horizons.

We report rollout error as a function of time,

$$\text{MSE}(t) = \mathbb{E}\left[\|x_t - \hat{x}_t\|^2\right], \tag{15}$$

where $\hat{x}_t$ denotes the predicted state under autonomous evolution. While this metric captures deviation from the ground-truth trajectory, it primarily serves as a diagnostic for stability rather than as a measure of associative recall.

5.2 Pattern completion and perturbation-based recall

To directly test associative memory, we evaluate pattern completion under perturbations. For a reference state $x^{(k)}$, we construct a perturbed initial condition

$$\tilde{x}_0 = x^{(k)} + \epsilon, \tag{16}$$

where ϵ is a small random perturbation. The system is then evolved autonomously for T steps,

$$\tilde{x}_{t+1} = f_\theta(\tilde{x}_t). \tag{17}$$

Successful recall is defined as convergence of the perturbed trajectory back to the reference attractor or its associated invariant set. This test directly probes the existence and stability of basins of attraction in the learned system, independent of the training objective.

5.3 Quantifying recall performance

We quantify recall performance using two complementary metrics. First, we define a phase recovery accuracy, which measures whether the final state $\tilde{x}_T$ is closest (under a suitable distance metric) to the correct reference state among a set of candidate patterns. Second, we define a recovery ratio that compares the distance from the perturbed initial condition to the reference state before and after autonomous evolution. Let $d(\cdot, \cdot)$ denote a distance measure on state space. The recovery ratio is given by

$$\rho = \frac{d(\tilde{x}_0, x^{(k)}) - d(\tilde{x}_T, x^{(k)})}{d(\tilde{x}_0, x^{(k)})}. \tag{18}$$

Positive values indicate convergence toward the reference state, while negative values indicate divergence or collapse away from the target. In practice, we ignore (or equivalently set $\rho = 0$ for) cases where $d(\tilde{x}_0, x^{(k)}) = 0$ to avoid division-by-zero.

(a) Rollout MSE over time.
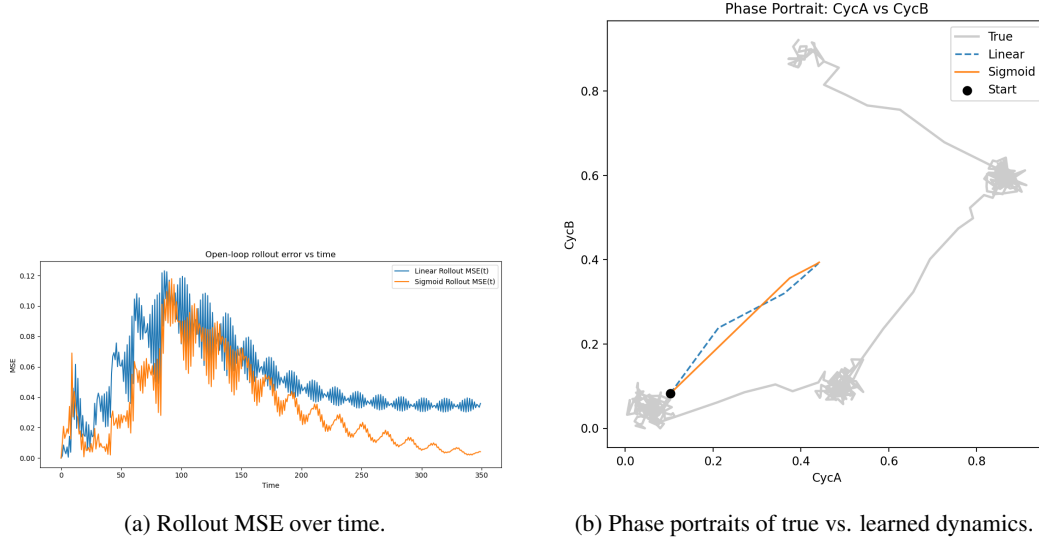
(b) Phase portraits of true vs. learned dynamics.

Figure 1: Hopf oscillator results. (a) Open-loop rollout MSE for linear and nonlinear AANNs trained on Hopf dynamics. (b) Representative phase portraits comparing ground-truth and learned trajectories, illustrating collapse toward a mean-field fixed point.

All evaluation metrics are applied uniformly across learned models and data generators. Importantly, these diagnostics assess properties of the learned dynamical system rather than performance on the training objective, allowing us to distinguish between models that achieve low one-step prediction error and those that preserve the long-horizon structure required for associative memory.

The following section presents empirical results using these evaluation criteria.

6 Results

We now evaluate whether recurrent models trained under a one-step prediction objective recover the attractor-based structure implied by the associative-memory interpretation. All results are reported using the dynamical diagnostics defined in Section 5.

6.1 Learning fails to recover attractor structure from oscillatory dynamics

We first evaluate learned dynamics on trajectories generated by coupled Hopf oscillators, which provide a canonical example of nonlinear autonomous dynamics with a stable oscillatory attractor. Figure 1 shows the rollout mean-squared error as a function of time for linear and nonlinear AANNs trained on Hopf-generated trajectories. While both models achieve low one-step prediction error, autonomous rollouts diverge rapidly from the ground-truth dynamics, indicating that local transition accuracy does not translate into long-horizon stability.

To understand the nature of this failure, we examine the geometry of learned trajectories in state space. Rather than preserving oscillatory structure, both linear and nonlinear models exhibit contraction toward a small region of state space, consistent with convergence to a single dominant fixed point corresponding approximately to the mean of the training data. This collapse occurs despite the presence of nonlinearities in the model architecture, suggesting that failure arises from the learning objective rather than from insufficient expressivity.

6.2 Pattern completion reveals absence of associative recall

We next evaluate learned models using perturbation-based pattern completion tests, which directly probe associative recall in the spirit of classical Hopfield networks [Hopfield, 1982]. For each reference state, small perturbations are applied and the system is evolved autonomously. Figure 2 summarizes performance: both linear and nonlinear AANNs perform at or near chance level in phase

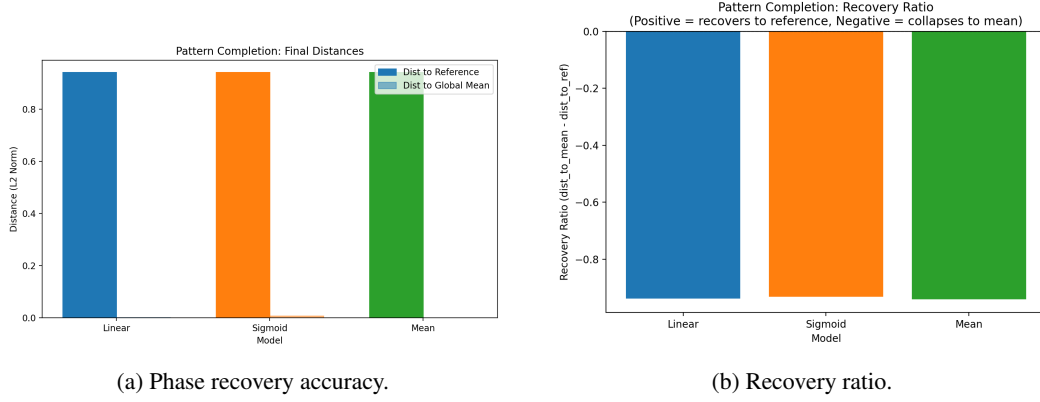(a) Phase recovery accuracy.

(b) Recovery ratio.

Figure 2: Perturbation-based recall for learned AANNs. (a) Phase recovery accuracy under perturbation-based recall, with performance near chance level. (b) Recovery ratio, with negative values indicating divergence away from reference states under autonomous evolution.

(a) Phase recovery accuracy.
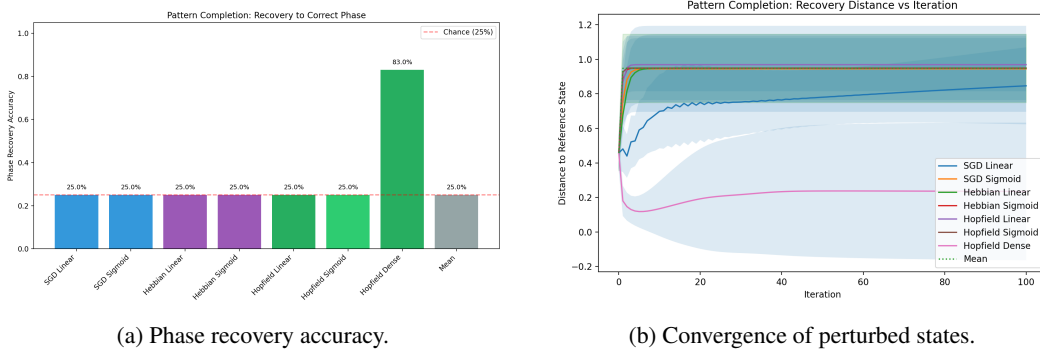

(b) Convergence of perturbed states.

Figure 3: Positive-control Hopfield-style nonlinear AANNs. (a) Phase recovery accuracy with explicit attractor encoding. (b) Distance to reference states as a function of iteration, showing rapid convergence of perturbed states.

recovery accuracy, and exhibit negative recovery ratios, meaning that autonomous dynamics move perturbed states farther from the reference rather than closer to it.

6.3 Explicit attractor encoding as a positive control

To distinguish between limitations of the learning objective and limitations of the model class, we next evaluate a positive control in which attractor structure is explicitly encoded in the network weights. Using an analytically constructed Hopfield-style nonlinear AANN with full connectivity [cf. Hopfield, 1982, Movellan, 1991], reference patterns are enforced as exact fixed points of the dynamics. Figure 3 shows that this construction yields high phase recovery accuracy and rapid convergence of perturbed states toward the correct attractor, establishing that the nonlinear AANN architecture is expressive enough to implement associative memory and that failure in learned models arises from the learning procedure rather than representational capacity.

7 Discussion

We tested whether learned recurrent models recover attractor-based structure from observed trajectories, as suggested by associative-memory interpretations of protein network dynamics. Across our dynamical generators and evaluation criteria, standard learning objectives fail to recover the long-horizon properties required for associative memory, even when the data-generating process exhibits clear attractor structure.

A central contribution of this study is the explicit operationalization of associative memory as a property of long-horizon dynamics rather than short-term prediction accuracy. Under a dynamical systems interpretation, associative recall requires that perturbed states converge back to reference attractors under autonomous evolution, emphasizing basin structure, stability, and asymptotic behavior. Viewed in this light, one-step prediction error is insufficient to establish associative memory: models can achieve low training error while exhibiting qualitatively incorrect global dynamics, including collapse to mean-field fixed points and loss of multistability.

The failure of both linear and nonlinear AANNs to recover associative memory across generators suggests that the limitation lies in the learning objective rather than representational capacity. The one-step prediction objective optimizes conditional expectations of the next state, encouraging local consistency rather than global stability and favoring contraction toward an average state in oscillatory or multistable systems. Positive control experiments make this distinction explicit: when attractors are analytically encoded into the network weights, nonlinear AANNs exhibit robust associative recall and stable basins of attraction, demonstrating that the architecture is expressive enough to implement associative memory in principle.

One natural hypothesis is that learning fails because attractor structure in the data is too weak or noisy. However, strengthening basin structure in the data-generating process does not rescue associative recall under the same objective: even when attractors are explicitly reinforced through phase-based relaxation dynamics, learned models converge to a reduced subset or a single dominant attractor. Thus the presence of attractors in the data alone is insufficient to guarantee their recovery; what matters is whether the learning objective constrains the global flow induced by the learned map.

Our findings suggest that the associative-memory interpretation proposed in prior work relies on implicit assumptions about how attractor structure is preserved under learning, such as observability or separation of timescales, that are not enforced by generic transition-based objectives. This does not invalidate the original interpretation, but clarifies its scope: associative memory may be present in biological dynamics, but its recovery by learned recurrent models is not guaranteed without objectives or architectures explicitly aligned to long-horizon properties. From this perspective, failures to exhibit associative recall reflect limitations of standard learning formulations rather than evidence against the underlying biological hypothesis.

More broadly, this work highlights a distinction between learning local predictive structure and learning global dynamical organization. Many questions in neural computation, from memory to decision-making and motor control, are inherently dynamical and depend on asymptotic behavior rather than instantaneous predictions, emphasizing the need for learning objectives that explicitly constrain long-horizon dynamics. Potential directions include objectives based on multi-step consistency, energy minimization, or contrasts between perturbed and unperturbed trajectories.

8   Limitations and Future Directions

This study focuses on a restricted class of recurrent architectures trained with a one-step prediction objective that does not explicitly constrain long-horizon dynamics. While this reflects common practice and enables controlled comparison across generators, it limits the types of dynamical structure that can be recovered.

Our surrogate dynamical systems also abstract away many biological details present in the original biochemical ODE model, so our results should be interpreted as statements about learnability of attractor structure under generic objectives rather than as claims about specific molecular mechanisms.

Future work could explore objectives explicitly aligned with dynamical properties, such as multi-step consistency losses, energy-based formulations, or contrasts between perturbed and unperturbed trajectories, and investigate how partial observability or architectural inductive biases affect recovery of associative memory.

9   Conclusion

We investigated whether learned recurrent models recover attractor-based associative memory from dynamical data. By evaluating linear and nonlinear associative artificial neural networks across a hierarchy of dynamical generators and using diagnostics aligned with dynamical systems theory, we

find that standard learning objectives fail to recover the long-horizon structure required for associative recall.

Our results demonstrate a clear separation between expressivity and learnability: while recurrent nonlinear models can represent associative memory when attractors are explicitly encoded, such structure is not reliably discovered through transition-based learning alone. These findings refine the associative-memory interpretation of learned protein network models and highlight the importance of aligning learning objectives with the dynamical properties of interest when modeling dynamical phenomena in neural computation.

References

John J. Hopfield. Neural networks and physical systems with emergent collective computational abilities. *Proceedings of the National Academy of Sciences*, 79(8):2554–2558, 1982. URL https://www.pnas.org/content/79/8/2554.

Eugene M. Izhikevich. *Dynamical Systems in Neuroscience: The Geometry of Excitability and Bursting*. MIT Press, 2007. ISBN 978-0-262-01257-9.

Herbert Jaeger. The "echo state" approach to analysing and training recurrent neural networks. *GMD Report*, 148, 2001. URL https://www.faculty.iu-bremen.de/hjaeger/pubs/EchoStatesTechRep.pdf.

Javier R. Movellan. Contrastive hebbian learning in the continuous hopfield model. In David S. Touretzky, Jeffrey L. Elman, Terrence J. Sejnowski, and Geoffrey E. Hinton, editors, *Connectionist Models*, pages 10–17. Morgan Kaufmann, 1991. doi: 10.1016/B978-1-4832-1448-1.50007-X. URL https://www.sciencedirect.com/science/article/pii/B978148321448150007X.

Béla Novák and John J. Tyson. Design principles of biochemical oscillators. *Nature Reviews Molecular Cell Biology*, 9:981–991, 2008. doi: 10.1038/nrm2502. URL https://www.nature.com/articles/nrm2502.

Sandhya Samarasinghe, Tran Nguyen Minh-Thai, Komal Sorthiya, and Don Kulasiri. Neurons and neural networks to model proteins and protein networks. *BioSystems*, 258:105613, 2025. ISSN 0303-2647. doi: 10.1016/j.biosystems.2025.105613. URL https://www.sciencedirect.com/science/article/pii/S0303264725002230.

Steven H. Strogatz. *Nonlinear Dynamics and Chaos*. CRC Press, 2018. ISBN 978-1-4987-4555-3.

David Sussillo and L. F. Abbott. Generating coherent patterns of activity from chaotic neural networks. *Neuron*, 63(4):544–557, 2009. doi: 10.1016/j.neuron.2009.07.018. URL https://www.cell.com/neuron/fulltext/S0896-6273(09)00453-9.

Hugh R. Wilson and Jack D. Cowan. Excitatory and inhibitory interactions in localized populations of model neurons. *Biophysical Journal*, 12(1):1–24, 1972. URL https://www.cell.com/biophysj/fulltext/0006-3495(72)86168-4.

A   Additional Experimental Details

In addition to the oscillatory and phase-based systems analyzed in the main text, we explored several other dynamical generators motivated by neural and biological population models, including discrete-time recurrent systems with sigmoid nonlinearities, modified Wilson–Cowan-type rate equations [Wilson and Cowan, 1972], and reduced variants of the biochemical ODE system proposed in prior work. These reduced ODE variants and population models reproduced oscillatory behavior under certain regimes but were highly sensitive to parameter choices and exhibited generator-specific behaviors that complicated controlled evaluation of associative recall. We therefore used them primarily for exploratory analysis and validation of simulation infrastructure and focus our main analysis on the surrogate systems that allow explicit control over attractor geometry, observability, and noise.
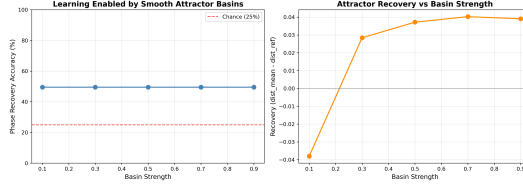
Figure 4: Phase recovery accuracy across increasing basin strength in phase-based dynamics. Strengthening attractor basins does not rescue associative recall under standard learning objectives.

B Additional Figures

B.1 Relaxation sweep: strengthening basins does not rescue learning

We tested whether strengthening attractor basins in the data-generating process improves recovery of associative memory. Using phase-based dynamics with varying relaxation strength, we trained nonlinear AANNs under the same one-step objective and evaluated recall performance. Figure 4 shows phase recovery accuracy as a function of basin strength. Despite substantial changes to the underlying dynamics, recall performance remains largely unchanged: learned models collapse to a subset of dominant states and fail to recover the full set of reference phases. These results indicate that stronger attractor structure in the data alone is insufficient to overcome the limitations imposed by the learning objective.

B.2 Implementation Notes

Code and experiment scripts are available at:
https://github.com/Sreevatsa03/protein-dynamics-aann.