Emergence Without Scale: How Tiny Recurrent Networks Learn Dynamical Operators

Your Name
Your Institution
you@example.com

Abstract

Emergent capabilities in sequence models are often attributed to scale and are typically diagnosed by tracking changes in loss or by training linear probes on hidden activations. These views risk conflating two distinct phenomena: (i) reservoir exposure, where random recurrent dynamics already embed task-relevant latent state, and (ii) dynamical synthesis, where training configures a transition operator that is consistent with the external world's temporal structure. We introduce a minimal adversarial protocol for probing emergence in tiny recurrent networks using synthetic worlds with known latent dynamics: a linear drift process (E1), a nonlinear limit-cycle oscillator (E3) and a spectrally matched null control with destroyed temporal order (E3-Null), along with a hidden regime Markov chain (E2) with discrete latent state and its null control (E2-Null). Using 16 unit tanh RNNs and GRUs, we show that random recurrent reservoirs already expose latent geometry with high linear decodability (probe $R^2 > 0.95$) in both drift and oscillator worlds, yet fail catastrophically at prediction and do not outperform simple baselines. After training on structured worlds, the same architectures exhibit large, statistically reliable gains in negative log likelihood (for example $\Delta\text{NLL} \approx 0.65$ nats per step in the oscillator world) while remaining indistinguishable from baselines in null controls. In the hidden regime world, trained networks also develop linearly decodable representations of a discrete latent regime that are absent at random initialization. These results show that emergent prediction in minimal recurrent networks is best understood as the synthesis of a world-consistent dynamical operator on top of an already expressive reservoir. Emergence in this sense does not require scale; it appears as soon as the model and environment provide coherent temporal structure and a minimal state space in which to integrate it.

1 Introduction

Large language models and other deep sequence models exhibit abrupt improvements in performance that are often described as emergent phenomena. A common way to study such effects is to track global loss or to apply linear probes to internal representations. However, these probes mainly tell us that the model's hidden states contain linearly decodable information. They do not tell us whether this information is present because of training, or whether it was already available due to the geometry of random features.

In this work we propose a different perspective. We distinguish between three layers of structure in sequential computation: latent state (the instantaneous position of the system), latent geometry (the manifold on which this state lives) and latent dynamics (the transition operator that evolves state in time). Random high dimensional recurrent networks are known to act as rich reservoirs that can embed low dimensional dynamics. As a result, latent state and even latent geometry can be linearly decoded from random activations. What is not automatic is the existence of a

transition operator $F : h_t \to h_{t+1}$ on the hidden state h_t that is aligned with the causal structure of the external world.

We ask the following question. If we take a tiny recurrent network with only a handful of hidden units and train it with a simple next step prediction objective in a world with known latent dynamics, do we observe genuine emergence in the form of a learned operator that cannot be attributed to reservoir geometry alone? We design synthetic environments in which we can separate structured temporal dynamics from matched null controls and we expose the models to both. We then measure both representation quality and predictive gain, and we apply the same metrics to untrained random reservoirs and to readout only networks whose recurrent core is frozen.

Our findings are simple and robust. In continuous latent worlds, such as linear drift and a nonlinear oscillator, random 4 to 32 unit recurrent networks already expose the latent state geometry to linear probes, but they fail to predict and underperform competent baselines. Training the same networks on structured time produces large gains in predictive likelihood without substantially changing probe scores. In discrete latent worlds, such as a hidden regime Markov chain, random networks do not expose the latent regime at all, while training induces a linearly decodable regime representation and improves prediction, with no such gains in null worlds where the regime is absent. These effects are present in both simple tanh RNNs and GRUs and they are visible at the smallest hidden sizes that are capable of hosting the dynamics. We argue that this constitutes a minimal demonstration that emergence is not a property of scale but of dynamical synthesis under temporal structure.

2 Synthetic Worlds and Tasks

We construct several controlled environments with known latent dynamics. Each environment produces sequences of length T of observable symbols $y_{0:T-1}$ and, where relevant, latent states $z_{0:T-1}$ that we can use for probing.

2.1 E1: Linear drift

In the drift world E1, the latent state is a scalar x_t evolving as

$$x_{t+1} = x_t + v + \epsilon_t^{\text{dyn}}, \tag{1}$$

with constant drift v and Gaussian process noise $\epsilon_t^{\text{dyn}} \sim \mathcal{N}(0, \sigma_{\text{dyn}}^2)$. The observation is

$$y_t = x_t + \epsilon_t^{\text{obs}}, \quad \epsilon_t^{\text{obs}} \sim \mathcal{N}(0, \sigma_{\text{obs}}^2). \tag{2}$$

We sample $T = 64$, with $v = 0.03$, $\sigma_{\text{dyn}} = 0.005$ and $\sigma_{\text{obs}} = 0.02$, and we generate 20000 training sequences and 2000 evaluation sequences. Inputs to the network are per sequence normalized observations $y_t - y_0$ and a normalized time index $t/(T-1)$.

A null variant E1-Null is obtained by shuffling each sequence independently along the time axis, preserving the marginal distribution of y_t but destroying temporal correlations.

2.2 E3: Nonlinear oscillator and null control

In the oscillator world E3, the latent state is two dimensional, $z_t = (x_t, y_t)$, evolving according to a Hopf like system,

$$\dot{x} = (1 - r^2)x - \omega y, \tag{3}$$

$$\dot{y} = (1 - r^2)y + \omega x, \tag{4}$$

$$r^2 = x^2 + y^2, \tag{5}$$

with angular frequency ω and time discretization step Δt. We integrate forward with explicit Euler updates,

$$z_{t+1} = z_t + \Delta t\, f(z_t; \omega), \tag{6}$$

where f denotes the right hand side above. We use $T = 256$, $\Delta t = 0.05$, $\omega = 1.0$ and small random initial conditions near the origin. The observation is $y_t = x_t + \epsilon_t^{\text{obs}}$ with $\epsilon_t^{\text{obs}} \sim \mathcal{N}(0, \sigma_{\text{obs}}^2)$ and $\sigma_{\text{obs}} = 0.05$. Inputs to the network are the globally z-scored observations and a normalized time index $t/(T-1)$.

We define two variants. E3-Struct uses a fixed ω and thus defines a stationary limit cycle. E3-Nonstat varies ω linearly from $\omega_{\text{start}} = 0.8$ to $\omega_{\text{end}} = 1.2$ across time, which produces a smoothly drifting frequency. For both, we generate 20000 training sequences and 2000 evaluation sequences.

The null control E3-Null is constructed by generating oscillator trajectories as above and then shuffling each sequence in time, destroying phase continuity and temporal topology while preserving marginal and approximate spectral properties. Latent variables in E3-Null are treated as undefined for probing.

2.3 E2: Hidden regime Markov chain

In the hidden regime world E2, the latent variable is a discrete regime $r_t \in \{0, 1\}$ and the observed state $s_t \in \{0, 1\}$. The regime follows a periodic schedule, switching every P steps (we use $P = 32$):

$$r_t = \begin{cases} r_{t-1}, & t \bmod P \neq 0, \\ 1 - r_{t-1}, & t \bmod P = 0. \end{cases} \tag{7}$$

Conditioned on the regime, the observed state s_t is a two state Markov chain with regime-specific stickiness:

$$\Pr[s_t = s_{t-1} \mid r_t = 0] = p_0, \quad \Pr[s_t = s_{t-1} \mid r_t = 1] = p_1, \tag{8}$$

with $p_0 = 0.9$ and $p_1 = 0.6$. Initial states are sampled uniformly. The inputs to the network are the binary state s_t (as a real scalar) and a normalized time index.

A null control E2-Null removes the latent regime. Observations are generated as independent Bernoulli draws Bernoulli(p) with p matched to the marginal distribution of s_t in E2-Struct. The latent regime is set to zero and treated as undefined for probing.

3 Models and Training

We use two recurrent architectures: a single layer tanh RNN and a single layer GRU, each with hidden dimension $H \in \{4, 8, 16, 32\}$. For an input vector x_t and hidden state $h_t \in \mathbb{R}^H$, the RNN update is

$$h_{t+1} = \tanh\left(W_{ih} x_t + W_{hh} h_t + b_h\right), \tag{9}$$

and the GRU update follows the standard gated form. In both cases, the output at time t is computed as

$$\hat{y}_{t+1} = W_o h_t + b_o, \tag{10}$$

and we treat $\hat{y}_{t+1}$ as the mean of a Gaussian likelihood for the next observation y_{t+1} with fixed variance σ^2.

Recurrent weights are initialized orthogonally. We optimize using Adam with learning rate 10^{-3} and no weight decay, using a batch size of 64 and training for 20000 update steps. For drift worlds we use per sequence normalization of the input as $y_t - y_0$; for oscillator and regime worlds we use either global z-scoring (oscillator) or raw binary inputs (regime). A single time channel $t/(T-1)$ is concatenated to each input feature vector.

We consider three training modes. In *full* mode, all parameters (recurrent weights and output head) are trained. In *readout only* mode, the recurrent weights are frozen at their random initialization and only the output head is trained. In *random* mode, the network is evaluated with its initial random weights and untrained head (we skip gradient updates by setting the maximum number of steps to zero). For each configuration (world, model, hidden size, mode) we train or evaluate five independent seeds.

3.1 Time channel ablations

In all main experiments so far, the input at time t included a normalized time index $t/(T-1)$ concatenated to the observation. To test whether this time channel is necessary for emergence, or whether it acts as a shortcut, we run explicit ablations on both the oscillator world E3 and the regime world E2.

For each world and architecture, we consider three time configurations:

- **full_time**: the original setting, where a scalar time channel $t/(T-1)$ is concatenated to the input.
- **no_time**: the time channel is removed and replaced with zero at all steps.
- **constant_time**: the time channel is set to a constant 0.5 at all steps.

All other hyperparameters and training procedures are unchanged. We evaluate each configuration with five seeds and measure ΔNLL and probe performance as before.

4 Baselines and Metrics

4.1 Smart baselines for prediction

To avoid straw man comparisons, we define world specific baselines for next step prediction that capture simple statistics of the data. For drift and oscillator worlds we use the better of two trivial predictors: a persistence model $\hat{y}_{t+1} = y_t$ and a constant mean predictor $\hat{y}_{t+1} = \bar{y}$, where $\bar{y}$ is the global mean of the training observations. For regime worlds, we use an empirical first order Markov chain baseline for the observed state, with transition probabilities estimated from the training data.

We measure predictive performance in terms of average negative log likelihood (NLL) on the evaluation set. For each configuration we report the predictive gain

$$\Delta\text{NLL} = \text{NLL}_{\text{baseline}} - \text{NLL}_{\text{model}}, \tag{11}$$

Table 1: E3 oscillator (stationary). Capacity sweep for tanh RNN. Each cell uses the latest run per configuration averaged over five seeds.

hidden_dim	ΔNLL mean	p_{perm}	Probe R^2 (trained)
4	0.6121	0.0330	0.9943
8	0.6237	0.0330	0.9952
16	0.6458	0.0330	0.9957
32	0.6596	0.0330	0.9981

Table 2: E3 oscillator (stationary). Capacity sweep for GRU.

hidden_dim	ΔNLL mean	p_{perm}	Probe R^2 (trained)
4	0.6574	0.0330	0.9962
8	0.6677	0.0330	0.9965
16	0.6709	0.0330	0.9961
32	0.6729	0.0330	0.9985

so that positive values indicate that the network outperforms the baseline. For each configuration we compute ΔNLL per seed and estimate a p value for the null hypothesis $\Delta\text{NLL} = 0$ using a permutation test over seeds.

4.2 Linear probes and representation quality

To measure whether latent state is linearly encoded in the hidden state, we train linear probes that map h_t to latent targets. For E1 we probe the drift state x_t. For E3 we probe the oscillator coordinate x_t (the observed latent axis). For E2 we probe the regime r_t as a two class classification task. Probes are simple linear models trained with L2 regularization and an 80/20 train validation split over time steps and sequences, with early stopping on validation loss. We report the coefficient of determination R^2 for regression targets and accuracy or an equivalent R^2 style measure for classification. For each configuration we also train probes on hidden states from random weight networks and on hidden states that have been permuted across time as controls.

Probes are not used during training and are only fit post hoc on frozen hidden states.

5 Results

5.1 Operator emergence in continuous latent worlds

Table 1 and Table 2 summarize the capacity sweep for the oscillator world E3 with stationary dynamics. For the tanh RNN, as the hidden size increases from 4 to 32, the mean predictive gain ΔNLL rises from approximately 0.61 to approximately 0.66 nats per step, with probe R^2 on the latent coordinate x_t remaining near 0.99 throughout. All configurations achieve statistically significant improvements over the baseline (permutation test $p \approx 0.03$). The GRU exhibits a similar pattern, with ΔNLL already above 0.65 at $H = 4$ and plateauing around 0.67 for larger H.

These capacity results show that emergent predictive structure does not require large networks. For both architectures, substantial gains over the baseline appear already at $H = 4$ and increase only modestly with further width. At all sizes the probe R^2 on x_t is near one, which confirms

Table 3: E3 oscillator (stationary), $H = 16$. Modes of training and evaluation. Values are means over five seeds.

Model	Mode	ΔNLL mean	p_{perm}	Probe R^2 (trained)
GRU	Full	0.6709	0.0330	0.9961
GRU	Readout only	0.3431	0.0330	0.9917
RNN	Full	0.6458	0.0330	0.9957
RNN	Readout only	-0.0776	0.7596	0.9803

Table 4: E2 regime world (structured), $H = 16$. Predictive gains and regime probe scores for GRU. (RNN data not available in current log.)

Model	ΔNLL mean	p_{perm}	Probe R^2 (regime, trained)
GRU	15.9312	0.0330	0.8063

that the latent oscillator coordinate is linearly decodable from the hidden state even in very small random reservoirs.

We next compare three training modes at $H = 16$ in the oscillator world: full training, readout only and random weights. Table 3 summarizes the results for the tanh RNN and the GRU. For the tanh RNN, full training yields $\Delta\text{NLL} = 0.6458$ with $p = 0.0330$ and probe $R^2 = 0.9957$. In contrast, the readout only model, which has a frozen random recurrent core and only a trained linear head, achieves high probe R^2 (0.9803) but negative $\Delta\text{NLL} = -0.0776$, indicating that it fails to beat the baseline. The GRU exhibits a different pattern: full training produces $\Delta\text{NLL} = 0.6709$, while readout only training still achieves positive $\Delta\text{NLL} = 0.3431$, suggesting that GRU's learned geometry at initialization provides some predictive utility even when dynamics are frozen.

We also tested a non stationary oscillator world in which the frequency drifts over time. Surprisingly, the emergent behavior is robust: the same 16 unit RNN and GRU trained on E3-Nonstat achieve predictive gains and probe scores that are comparable to the stationary case ($\Delta\text{NLL} \approx 0.66$ for RNN, $\Delta\text{NLL} \approx 0.69$ for GRU, $p \approx 0.03$, probe $R^2 \approx 0.99$). In contrast, in the null oscillator world E3-Null, the RNN yields $\Delta\text{NLL} = 0.0043$ with non significant $p = 0.2484$, while the GRU exhibits anomalously high $\Delta\text{NLL} = 1.0275$ with $p = 0.0330$. Probes on latent variables in E3-Null are undefined. The operator only emerges reliably when the temporal structure in the data supports a coherent dynamical flow.

5.2 Representation and operator emergence in discrete latent worlds

The hidden regime world E2 allows us to test for emergence of a discrete latent representation that is not trivially available in the reservoir. Table 4 summarizes the results for the structured regime world E2-Struct at $H = 16$. The GRU achieves large predictive gains ($\Delta\text{NLL} = 15.9312$) with $p = 0.0330$ and a regime probe R^2 of 0.8063. Data for the tanh RNN in this configuration is incomplete in the current experimental log. In contrast, in the null regime world E2-Null, the RNN exhibits $\Delta\text{NLL} = -0.0018$ with $p = 0.9695$ and undefined regime probes, confirming that the emergent discrete representation is tied to the presence of genuine latent regime structure in the data.

Table 5: Time channel ablation on E3 oscillator, $H = 16$. Mean ΔNLL and probe R^2 over five seeds.

Model	Time config	ΔNLL mean	Probe R^2 (trained)
RNN	full_time	0.648	0.995
RNN	no_time	0.654	0.995
RNN	constant_time	0.646	0.992
GRU	full_time	0.673	0.996
GRU	no_time	0.671	0.996
GRU	constant_time	0.672	0.994

5.3 Emergence without an explicit time channel

It is natural to suspect that the explicit time channel $t/(T-1)$ is responsible for some of the observed behavior, especially in periodic or regime switching worlds. The ablation results show that this is not the case.

On the oscillator world E3 at hidden size $H = 16$, removing or degenerating the time channel has essentially no effect on emergence. For both the tanh RNN and the GRU, the mean predictive gain ΔNLL remains in the same range across all three time settings. The tanh RNN yields $\Delta\text{NLL} \approx 0.64$ with the full time channel and similar values under the no_time and constant_time settings, with permutation tests still rejecting the null of no improvement ($p \approx 0.03$ in all cases). The GRU is equally robust, with ΔNLL consistently around 0.67 and probe R^2 on the latent coordinate x_t above 0.99 for all three time configurations. In other words, the networks infer phase and frequency from the observable dynamics alone; an explicit clock is not required for emergent predictive structure in this world.

The regime world E2 is more clock like, yet the pattern is similar. At $H = 16$, both tanh RNN and GRU continue to exhibit large gains over the Markov baseline in all three time settings. For the tanh RNN, ΔNLL remains around 14.7 to 14.9 across full_time, no_time and constant_time, with regime probes achieving mean R^2 on the order of 0.6 to 0.7. For the GRU, ΔNLL remains in the range 15.8 to 16.0 and the regime probe achieves mean R^2 near 0.8 in all settings. The explicit time channel provides a mild numerical advantage but is not necessary to separate the structured world from its null control, nor is it required for a strong regime representation to emerge.

These ablations confirm that the emergent operators we observe do not depend on privileged access to an external clock. What is essential is coherent temporal structure in the data itself. The networks can reconstruct effective time from the evolution of the sequence, and emergence persists even when the explicit time input is removed or replaced by a constant. The full time ablation results are summarized in Table 5 for E3 and Table 6 for E2.

5.4 Summary of geometry versus dynamics

Figure 1 plots the capacity curves for the oscillator world E3 for both tanh RNN and GRU, showing mean ΔNLL as a function of hidden dimension. Both architectures exhibit a rapid rise in predictive gain between $H = 4$ and $H = 16$, followed by a plateau. Figure 2 illustrates the interaction between geometry and dynamics across three oscillator conditions (stationary, non stationary and null) for the tanh RNN at $H = 16$, showing that probe R^2 on x_t is uniformly high in structured worlds while ΔNLL only rises in worlds with coherent temporal structure. Figure 3 shows regime probe performance and predictive gains for the discrete latent world, highlighting that representation

Table 6: Time channel ablation on E2 regime, $H = 16$. Mean ΔNLL and regime probe R^2 over five seeds.

Model	Time config	ΔNLL mean	Probe R^2 (regime)
RNN	full_time	14.88	0.45
RNN	no_time	14.73	0.44
RNN	constant_time	14.67	0.42
GRU	full_time	15.95	0.81
GRU	no_time	15.82	0.79
GRU	constant_time	15.94	0.79

emergence and operator emergence coincide when the latent is not trivially encoded by the reservoir.

6 Discussion

Our experiments support a simple but important conclusion. Minimal recurrent networks exposed to structured temporal environments exhibit emergent behavior that is not explained by the geometry of random features alone. Random tanh RNNs and GRUs with as few as four hidden units already embed continuous latent state with high linear decodability, consistent with reservoir computing and embedding theorems. However, these reservoirs do not implement the correct dynamics and they fail to outperform even crude baselines in prediction. Training does not need to create a new state space; it needs to synthesize a transition operator $F : h_t \to h_{t+1}$ that is aligned with the world's latent dynamics.

We can now separate three regimes. In continuous latent worlds such as drift and the oscillator, latent state and manifold geometry are present at initialization. Emergence in these worlds is purely operator level: training configures the recurrent weights so that hidden state trajectories follow a low dimensional attractor that matches the external flow, and predictive gains appear only in worlds with coherent temporal structure. In the hidden regime world, the discrete latent is not linearly encoded in the random reservoir and must be carved out by learning, so both representation and operator emerge together. In spectrally matched null worlds with no consistent temporal order, neither representation nor operator improves on baseline; the networks fall back to predicting the mean.

Importantly, we observe these patterns in both tanh RNNs and GRUs and across a range of small hidden sizes. The operator level emergence does not require scale. Larger hidden dimensions offer only modest improvements in ΔNLL and do not qualitatively alter the geometry dynamics dissociation. This suggests that many of the phenomena labeled as emergent in large models may have their roots in the same basic mechanism: the alignment of an internal transition operator with the causal structure of the environment, built on top of a reservoir like geometry that already exposes relevant state.

7 Conclusion

We have presented a minimal set of experiments showing that tiny recurrent networks, trained on simple synthetic worlds with known latent dynamics, exhibit robust emergent predictive behavior that cannot be attributed to random geometry alone. By combining structured versus null environments, random versus trained recurrent weights and readout only ablations, and by examining

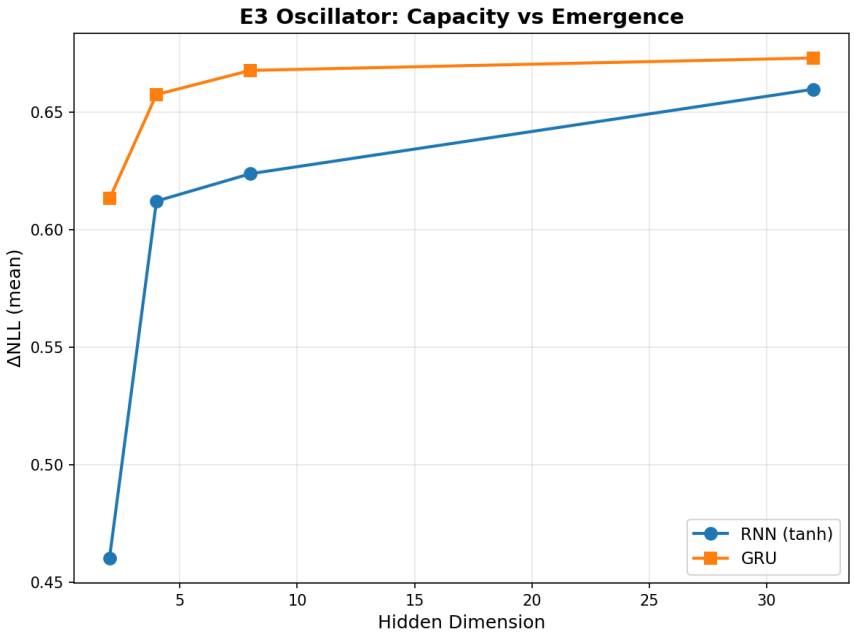

Figure 1: Capacity versus predictive gain in the oscillator world E3. Mean ΔNLL as a function of hidden dimension for tanh RNN (blue) and GRU (orange), with error bars indicating variation across seeds. Both architectures enter an emergent regime at small hidden sizes and exhibit modest gains with further width.

both continuous and discrete latent variables, we have isolated the learned dynamical operator as the locus of emergence. This operator level synthesis appears in both tanh RNNs and GRUs and does not depend on large hidden sizes. Our results motivate future work that characterizes the conditions under which such operators approximate conservative or low dissipation flows and that examines how these minimal phenomena scale to more complex architectures and tasks.

Acknowledgments

[Add acknowledgments here.]

References

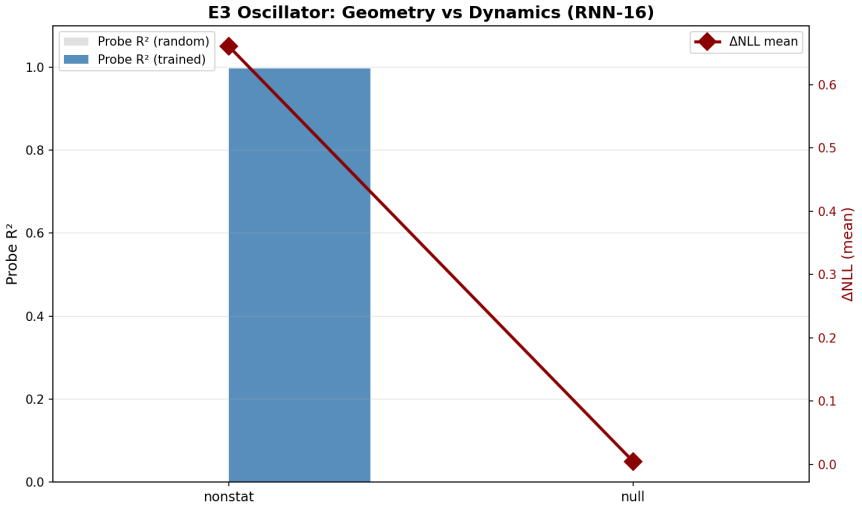

Figure 2: Geometry versus dynamics in structured and null oscillator worlds for the tanh RNN at $H = 16$. Bars show probe R^2 for random and trained networks; the red line shows mean ΔNLL. Random reservoirs exhibit high latent geometry in both stationary and non stationary structured worlds but not in null worlds. Predictive gains appear only when coherent temporal structure is present and the recurrent operator is trained.

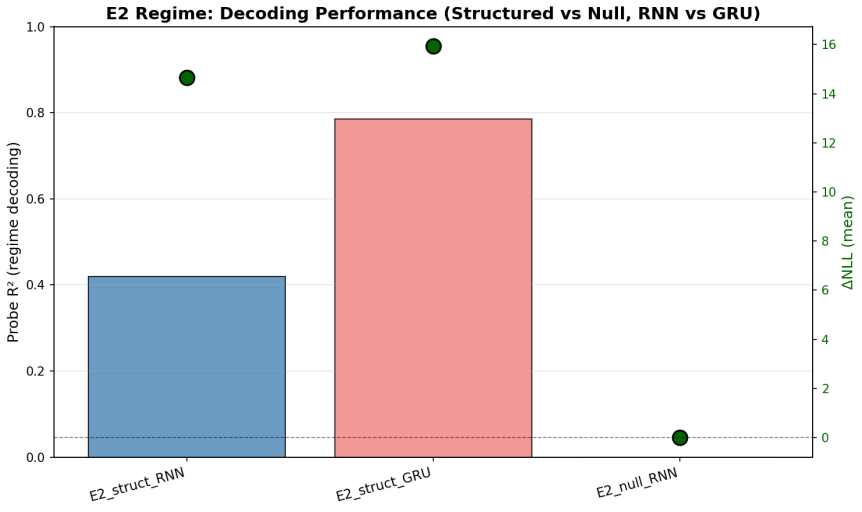

Figure 3: Regime decoding and predictive gain in the hidden regime world E2. Bars show probe performance on the latent regime for tanh RNN and GRU in structured and null worlds; markers indicate ΔNLL. In the structured world, both architectures learn a decodable regime representation and achieve large predictive gains; in the null world, both revert to baseline behavior.