Comparing LSTM, GRU, Transformer, and Mamba for Statistics-Informed Neural Networks

Zihan Xu[1], Changho Kim[1], Roummel Marcia[1], Andy Nonaka[2], Yuanran Zhu[2]

[1] University of California, Merced [2] Lawrence Berkeley National Laboratory

Statistics-Informed Neural Network (SINN)

SINN [1] is a model originally based on recurrent neural networks (RNNs) that learns stochastic dynamics from target trajectories. Training matches selected statistical properties of trajectory ensembles rather than enforcing pathwise convergence. Once trained, it serves as a stochastic trajectory generator.

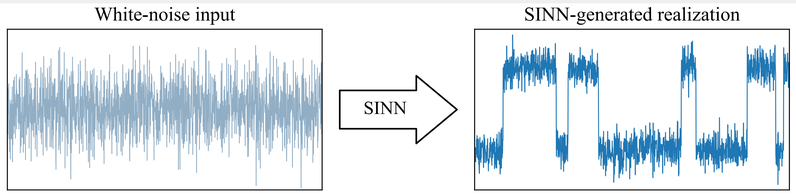


Figure 1. SINN as a stochastic trajectory generator.

Goal

Compare LSTM, GRU, causal Transformer and Mamba within SINN under a common protocol for each target.

Supplementary material

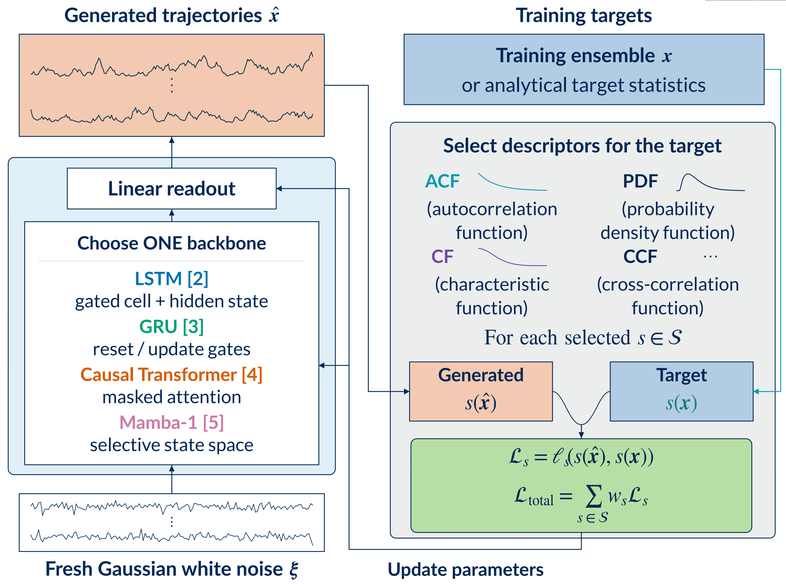


Figure 2. Choose one backbone and a target-specific set of statistics.

$$\text{Example (OU):} \quad \mathcal{L}_{\text{ACF}} = \text{MSE}\,(\text{ACF}(\hat{x}), \text{ACF}(x))$$

Protocol

Training

Similar parameter counts and matched seeds within each target. OU: 3 development seeds, 20,000 iterations. Other targets: 10 seeds, 50,000 iterations.

Fresh Gaussian noise is sampled at every iteration.

Evaluation

OU: 2,000 generated paths per fit, compared with analytical OU statistics. Development results only.

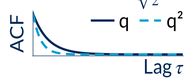
DW / Duffing: 400 generated vs 400 independent reference paths. Shuffle whole paths into two groups 999 times to test whether the observed discrepancy is unusually large.

Holm correction controls the chance of any false alarm across checks at 5% per fit. No detected difference does not prove a match.

Benchmark Selection

OU

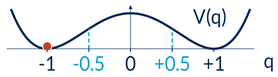
$$dq_t = -q_t\,dt + \frac{1}{\sqrt{2}}\,dW_t$$



In loss: ACF(q), ACF(q²), density.
Check: Amplitude and seed stability.

Double-well

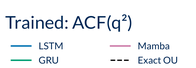
$$V(q) = 4(1-q^2)^2, \quad \beta\Delta V = 4$$



In loss: ACF(q), ACF(q²), density.
Additional: Switching rate, residence time.

Gamma-trawl

Marginal CF: $\varphi(u) = (1 - iu/2)^{-2}$



Planned: Positive-output training.
Match analytical CF families [7];
check distribution and dynamics.

Coupled Duffing (2D)

$$V = \frac{q_1^2+q_2^2}{2} + \frac{q_1^4}{4} + \frac{q_2^4}{8} + 2(q_1-q_2)^2$$



In loss: ACF, bidirectional CCF, q²-ACF, marginal + projected densities.
Additional: Full joint position density.

Figure 3. Three evaluated targets and Gamma-trawl as future work.

$$\text{DW/Duffing:} \quad \text{underdamped Langevin}, \quad \beta = 1, \quad \Gamma_{\text{DW}} = 1, \quad \Gamma_{\text{Duffing}} = \text{diag}(1, 0.5)$$

Ornstein–Uhlenbeck (OU)

Similar median errors; one fit has excessive fluctuations

Trained: ACF(q), ACF(q²), density. Median q²-ACF errors: 0.61–0.73 × 10^{-4}. One Transformer fit has excessive fluctuations: standard deviation 0.805 vs target 0.5.

Figure 4. OU development study: 3 seeds/backbone, 20k iterations, 2,000 paths/fit. Left: median curves. Dots: individual fits. Errors use 100 lags vs analytical OU.

Bistable Double-Well

Lowest position-distribution error, excess switching

Training targets: ACF(q), ACF(q²), and position density. The trained ACF(q²) still differs from reference in 38/40 fits. LSTM/GRU better reproduce switching and residence times; density fit alone is insufficient.

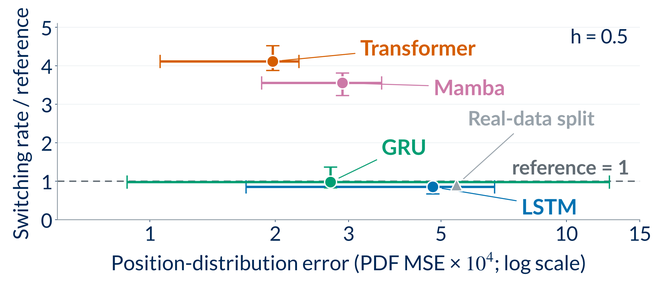
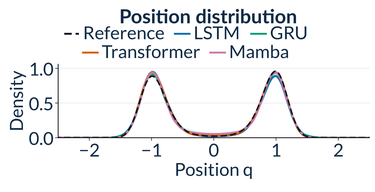


Figure 5. DW, 50k iterations. Median and IQR of 10 runs/backbone. Gray: real-data split. Switches go from ≤ −0.5 to ≥0.5 or back. Below: zoomed densities; black dashed: reference.



Backbone	Switching rate / reference	Time in one well* / reference
LSTM	0.85	1.05
GRU	0.98	0.90
Transformer	4.11	0.27
Mamba	3.55	0.29

Medians across runs. *Mean stay up to 40 time units, accounting for unfinished stays. Ratios of 1 match the reference. Residence: LSTM 9 runs, others 10.

Coupled Duffing (2D)

CCF mismatch in all Transformer/Mamba fits

In loss: ACFs, CCFs and densities. CCF differs from reference in 10/10 fits each for Transformer/Mamba; LSTM/GRU 0/10.

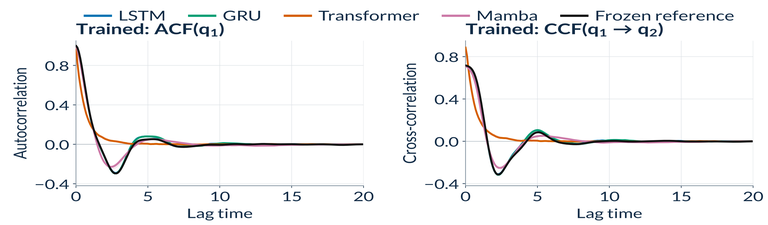


Figure 6. Duffing, 50k iterations, 10 runs/backbone. Lines and shading: median and IQR. Black: reference. Transformer misses the negative CCF dip. Training uses self-adaptive loss balancing [6].

Fit quality and computational cost

Backbone	CCF vs ref (%)	Joint PDF vs analytic (%)	Train 50k (min)	Generate (k paths/s)	Peak train (GiB)
LSTM	7.94	4.45	17.46	63.23	0.661
GRU	7.90	4.14	20.40	66.02	0.736
Transformer	60.86	4.02	41.55	28.42	3.960
Mamba	16.62	4.26	22.87	78.28	0.697

Medians of 10 runs/backbone. A100 80GB, 50k training iterations.
Generation: 400 paths, each with 200 burn-in + 200 retained steps.
Generator parameters (L/G/T/M): 13,122 / 13,063 / 12,867 / 13,282.
Joint density is outside the loss. Reference vs analytical density: 3.16% (one comparison).

Future Work: Gamma-Trawl

Enforce positivity during training

Gamma-trawl trajectories should be positive. Clipping negative outputs to zero creates a point mass absent from the target Gamma distribution.

$$z_t = Wh_t + b, \qquad \hat{x}_t = \text{softplus}(z_t) = \log(1 + e^{z_t}) > 0$$

Backbone state → Linear readout → Positive trajectory

Use the same Softplus readout for all four backbones. Compute statistical losses on the positive outputs and retrain each model.
Then evaluate both distributional fit and temporal dependence. Positivity alone does not guarantee either.

Conclusions

LSTM/GRU better match Duffing correlations and DW switching. DW q²-ACF remains poorly fitted.
OU: similar median errors, with one amplitude failure.
Validate dynamics explicitly. A good position distribution does not guarantee correct switching or residence times.

References

[1] Zhu et al., J. Comput. Phys. 474, 111819 (2023).
[2] Hochreiter & Schmidhuber, Neural Comput. 9, 1735–1780 (1997).
[3] Cho et al., EMNLP, 1724–1734 (2014).
[4] Vaswani et al., NeurIPS 30 (2017).
[5] Gu & Dao, Mamba, arXiv:2312.00752 (2023).
[6] Xiang et al., Neurocomputing 496, 11–34 (2022).
[7] Belkova & Mikhailov, Moscow Univ. Phys. Bull. 80 (Suppl. 3), S1077–S1084 (2025).

Acknowledgements

This material is based upon work supported by the U.S. Department of Energy (DOE), Office of Science, Office of Advanced Scientific Computing Research (ASCR) Funding for Accelerated, Inclusive Research (FAIR) program under Award Number DE-SC-0024240. Registration and travel support for this presentation was provided by the Society for Industrial and Applied Mathematics.