Trajectory-Consistent Dropout for Uncertainty Decomposition in Hamiltonian Neural Networks

Yuki Tachibana
Department of Engineering Science
University of Oxford
OXFIDA Ltd
Oxford, United Kingdom
yuki.tachibana@jesus.ox.ac.uk

Stephen J. Roberts
Department of Engineering Science
University of Oxford
Oxford, United Kingdom
stephen.roberts@eng.ox.ac.uk

Abstract

Hamiltonian Neural Networks (HNNs) produce stable long-horizon rollouts by integrating a learned Hamiltonian with a symplectic scheme. Monte Carlo dropout (Fresh-mask MC) conflicts with this per-trajectory interpretation because resampling the dropout mask inside a leapfrog integrator means successive sub-steps follow different subnetworks, so a single MC rollout no longer corresponds to any one sampled Hamiltonian. We propose *trajectory-consistent dropout*, implemented as *Fixed-mask MC*, where each stochastic rollout selects one mask from a finite bank and reuses it across the full integration path. This restores a coherent per-rollout Hamiltonian sample and makes uncertainty decomposition operational. Crossing the mask axis with an initial-condition ensemble separates epistemic uncertainty, the model's own, from aleatoric uncertainty, the sensitivity to the initial state. Experiments on classical dynamical systems show three benefits. First, sampled trajectories conserve energy better over long horizons, with 1.9–4.7 times less physical energy drift than fresh-mask rollouts. Second, the method achieves a better aleatoric-epistemic decomposition. Fixed-mask MC recovers the initial-condition noise closest to its true scale and has the best likelihood at that scale, whereas fresh-mask baselines fit a larger noise scale that absorbs model error into the aleatoric term. Third, because model error stays on the epistemic axis, epistemic uncertainty responds to distribution shift. Out of distribution on a coupled pendulum, the epistemic share of predictive variance rises by 0.32 for Fixed-mask MC but by only 0.03 for Fresh-mask MC, whose epistemic variance is about 19 times smaller. On a transverse-field Ising model the decomposition carries over to latent quantum dynamics, while predictive gains remain system-dependent. Fixed-mask MC thus gives trajectory-consistent, physically interpretable uncertainty from a single training run.

1 Introduction

Hamiltonian Neural Networks [Greydanus et al., 2019] parametrise the Hamiltonian $H_\theta(\mathbf{p}, \mathbf{q})$ and integrate Hamilton's equations over momentum $\mathbf{p}$ and position $\mathbf{q}$, encoding energy conservation as a hard inductive bias. Compared with unconstrained dynamics models, this bias improves long-horizon prediction in conservative systems because rollouts are bounded, reversible, and physically consistent by construction. For scientific simulation, engineering design, and risk modelling, however, rollout accuracy is not enough. Downstream decisions also require *calibrated uncertainty*, including the distinction between uncertainty about the learned dynamics and sensitivity to an imprecise initial state. This is critical outside of the training regime, where a reliable indication of model ignorance is desired.

40th Conference on Neural Information Processing Systems (NeurIPS 2026).

Why the decomposition matters in practice. Uncertainty quantification supports risk assessment, reliability, and safety decisions in many applications of ML [Nemani et al., 2023], yet a single predictive variance only says that a forecast is uncertain. It does not explain *why*, and therefore does not provide guidance on specific areas of improvement. When uncertainty traces back to noisy or incomplete knowledge of the current state, the response is better sensing, filtering, or state estimation. When instead it reflects the model's ignorance about operating conditions it was never trained on, sharper sensing of the current state cannot fix the problem, and the appropriate responses are to collect training data in the unfamiliar regime, revise the model, or abstain from action until confidence is restored [Kapteyn et al., 2022, DeChant and Moradkhani, 2014, Michelmore et al., 2020, Nguyen et al., 2019, 2022]. For example, consider a satellite whose trajectory is forecast by a learned surrogate of its orbital dynamics. If the forecast becomes unreliable, a useful uncertainty estimate should help distinguish whether the issue is imprecise knowledge of the current position and velocity (a state-estimation problem) or the spacecraft having entered a perturbation regime absent from training (a model-coverage problem). This is exactly the distinction our decomposition targets for Hamiltonian systems, separating what we do not know about *which state the system is in* from what we do not know about *how it evolves*.

Conflict with standard dropout. Monte Carlo dropout [Gal and Ghahramani, 2016a] is the standard single-model recipe for uncertainty in neural networks, as it requires no extra parameters, uses a single training pass, and has a well-defined variational interpretation. Plugging it into an HNN, however, creates a modelling mismatch that is easy to miss. A leapfrog integrator is symplectic *for a fixed Hamiltonian*. If the dropout mask is resampled at every integrator sub-step, each step evaluates a different Hamiltonian H_{m_t}. The composite map is still symplectic, but it no longer discretises any single H. Per-step energy errors no longer share one modified Hamiltonian around which to cancel. Drift can accumulate with rollout length, and the physics prior that motivated the HNN is weakened. The controlled experiments in Section 5 test whether this modelling failure also appears in MSE, uncertainty scoring, calibration, and energy or norm drift.

Trajectory-consistent dropout. We call the resulting construction *trajectory-consistent dropout*, and report it as Fixed-mask MC in the experiments. Fixing the dropout mask for an entire rollout makes each stochastic sample a coherent Hamiltonian system under the integrator assumptions stated in Corollary 1. Figure 1 illustrates the mechanism. Predictive uncertainty then arises from an ensemble over coherent sampled rollouts rather than from step-wise noise injected into a single trajectory.

This masking rule is related to locked or variational dropout in sequence models [Gal and Ghahramani, 2016b, Merity et al., 2018] and to finite-mask ensemble methods such as Masksembles [Durasov et al., 2021] and Controlled Dropout [Hasan et al., 2023]. The contribution is not the fixed mask alone, but the Hamiltonian rollout contract. In a symplectic integration loop, holding the stochastic realisation fixed is what lets a sample retain the semantics of one coherent Hamiltonian trajectory. We therefore frame the contribution as diagnosing an overlooked incompatibility between per-evaluation dropout and geometric integration. Section 3 sets out how this differs from locked dropout.

Uncertainty should respond to physics. Because each mask's rollout is conditionally deterministic given an initial condition, we can construct an IC noise ensemble with N perturbed initial conditions and M masks per initial condition. Without IC replicas, rollout variance would conflate mask disagreement with sensitivity to the starting state. The IC ensemble lets us hold the starting state fixed while varying the mask, and vice versa. The law of total variance then separates $\mathrm{Var}_{\mathrm{epi}}$, the mask disagreement at a fixed initial condition, from $\mathrm{Var}_{\mathrm{alea}}$, the sensitivity of the mean prediction to initial-condition perturbations. This decomposition addresses a question that fresh-mask dropout answers less cleanly. Does the model-sample component increase when test-time physics moves away from the training support? Fresh-mask dropout can still be decomposed over rollout randomness and IC perturbations, but its stochastic axis is not a coherent Hamiltonian sample. The Coupled Pendulum stress test in Section 5.3 answers it for both contracts at matched dropout rate and rollout horizon. Out of distribution the epistemic share of predictive variance rises by 0.32 under fixed masks but by only 0.03 under fresh masks, which attribute model error to the initial-condition axis and keep about $19\times$ less variance on the model axis.

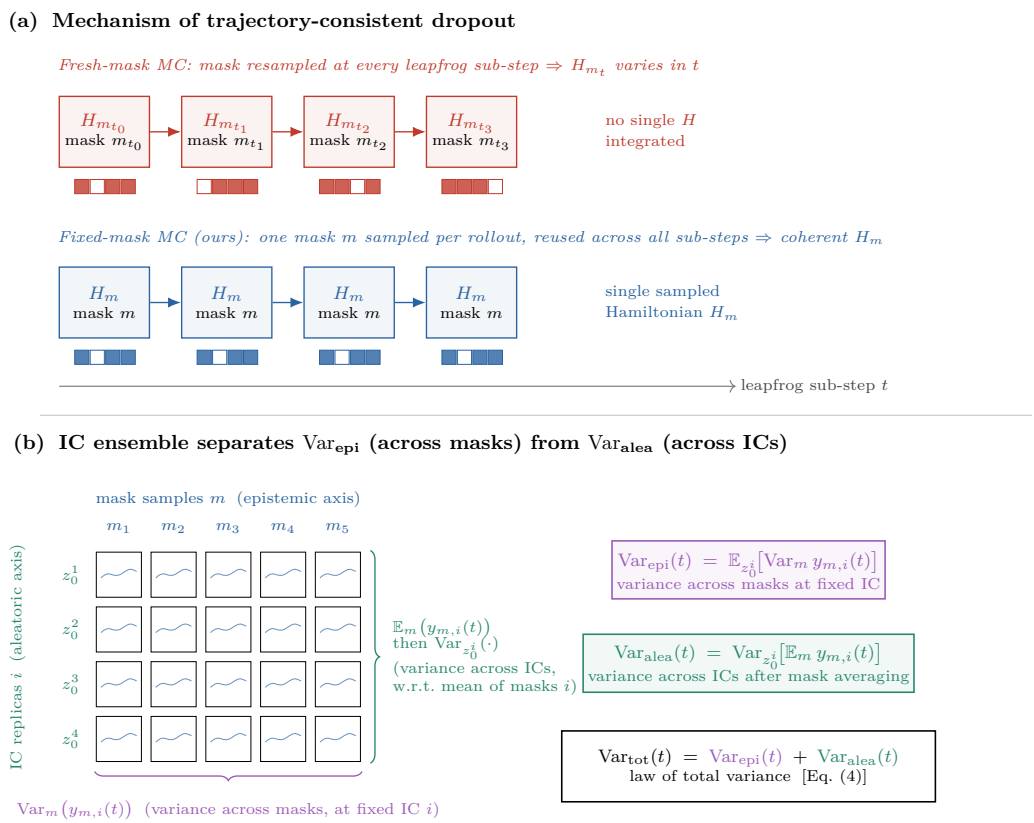


Figure 1: Mechanism of trajectory-consistent dropout. Fresh-mask MC dropout changes the mask inside a rollout, so the integrator no longer follows a single sampled Hamiltonian. Fixed-mask dropout samples one mask per rollout, producing a coherent sampled-Hamiltonian trajectory. The bottom panel shows how the initial condition (IC) ensemble separates epistemic variance $\mathrm{Var}_{\mathrm{epi}}$ (across masks at fixed initial condition) from aleatoric variance $\mathrm{Var}_{\mathrm{alea}}$ (across initial conditions after mask averaging). See Equation (4).

The value of this construction is the combination, not either axis alone. A finite mask bank gives a reusable model-uncertainty axis, and the IC ensemble gives a reusable state-preparation axis. When the same mask is held fixed for the whole rollout, the tensor retains the structure "which Hamiltonian sample?" crossed with "which initial condition?". This makes the IC-vs-model split a practical axis-specific estimator rather than a retrospective explanation.

Contributions. First, we introduce Fixed-mask MC for HNN rollouts. A finite mask bank is sampled once and each trajectory reuses one mask at every leapfrog sub-step, giving a per-mask Hamiltonian interpretation under the separable-integrator assumptions of Corollary 1. Second, crossing this mask axis with an initial-condition (IC) ensemble yields the IC-vs-model decomposition in Equation (4). Third, we evaluate the contract across a controlled complexity ladder[1], with ten training seeds per single-model method and likelihoods reported both under a fitted and under the true initial-condition noise scale. Coupled Pendulum provides the main physics-distance OOD evidence, Spring Chain tests structured pairwise interactions, and TFIM probes a harder latent quantum setting. Deep Ensembles remain compute-heavy accuracy references. Fixed-mask MC provides the single-training-run mechanism for trajectory-consistent uncertainty. Finally, an exploratory hidden-parameter benchmark (Appendix H.9) shows that per-mask energy drift flags a hidden mass-ratio shift from short observed prefixes, whereas predictive variance detects it only weakly under either scheme.

[1] Code to reproduce all tables and figures: `https://github.com/yukitachibana-eth/hnn-trajectory-dropout`.

2 Background

Hamiltonian neural networks. A separable Hamiltonian system has state $z = (\mathbf{p}, \mathbf{q}) \in \mathbb{R}^{2d}$ and satisfies

$$H(\mathbf{p}, \mathbf{q}) = T(\mathbf{p}) + V(\mathbf{q}), \qquad \dot{\mathbf{p}} = -\nabla_{\mathbf{q}} H, \quad \dot{\mathbf{q}} = \nabla_{\mathbf{p}} H. \tag{1}$$

An HNN parametrises H_θ with a neural network and integrates (1) with a symplectic scheme. For the formal symplecticity statement, we use the standard separable Störmer–Verlet (leapfrog) integrator

$$\mathbf{p}_{1/2} = \mathbf{p} - \tfrac{\Delta t}{2} \nabla_{\mathbf{q}} H(\mathbf{p}, \mathbf{q}), \quad \mathbf{q}' = \mathbf{q} + \Delta t \, \nabla_{\mathbf{p}} H(\mathbf{p}_{1/2}, \mathbf{q}), \quad \mathbf{p}' = \mathbf{p}_{1/2} - \tfrac{\Delta t}{2} \nabla_{\mathbf{q}} H(\mathbf{p}_{1/2}, \mathbf{q}'). \tag{2}$$

MC dropout as a Bayesian approximation. Gal and Ghahramani [2016a] show that a neural network with dropout and L_2 regularisation approximates a variational posterior over weights, and that predictive uncertainty can be recovered by averaging forward passes with fresh Bernoulli masks. Concrete dropout [Gal et al., 2017] makes the drop probability learnable, and Variational dropout [Kingma et al., 2015, Molchanov et al., 2017] adds Gaussian noise to the weights directly. All three families resample the injected noise at every forward pass in their original formulations.

Fixed-mask and Fresh-mask MC. We compare two ways of using dropout masks inside a rollout. Write $H(z; b)$ for the network Hamiltonian under dropout mask b. *Fresh-mask MC* draws a new dropout mask at every Hamiltonian evaluation. *Fixed-mask MC* draws one mask $b^{(m)}$ from a finite bank for each rollout and reuses it at every leapfrog sub-step, so the stochastic sample is the mask index m, resampled only across rollouts.

Why dropout and leapfrog conflict. For separable Hamiltonians, leapfrog is a symplectic composition of canonical subflows. Standard backward-error analysis then gives near-conservation of a modified Hamiltonian over bounded time intervals under the usual smoothness and step-size assumptions. Let H_k denote the Hamiltonian evaluated at sub-step k of one rollout. The two mask contracts differ only in this sequence:

$$\begin{aligned} &\text{Fresh-mask MC:} \quad H_k(z) = H(z; b_k), && b_k \text{ drawn anew at every sub-step } k, \\ &\text{Fixed-mask MC:} \quad H_k(z) = H(z; b^{(m)}), && \text{the same mask for all } k. \end{aligned} \tag{3}$$

Under fixed masks the rollout discretises one time-independent Hamiltonian, and Corollary 1 applies. Under fresh masks every sub-step is still the exact flow of a separable masked Hamiltonian, so the composite map remains symplectic. What fails is the interpretation of the whole trajectory. The generator is step-varying and random, which breaks the time-translation symmetry behind energy conservation, so no single modified Hamiltonian is nearly conserved and the sampled energy can drift diffusively (stochastic heating). On the Spring Chain system (Section 5.4), the contrast between Fresh-mask and Fixed-mask drift of the sampled Hamiltonian (H-drift) is among the largest observed across systems.

3 Related work

Structure-preserving neural dynamics. Hamiltonian Neural Networks (HNNs) [Greydanus et al., 2019], Symplectic Recurrent Neural Networks [Chen et al., 2020], SympNets [Jin et al., 2020], Lagrangian Neural Networks [Cranmer et al., 2020], Hamiltonian graph networks [Sanchez-Gonzalez et al., 2019], and port-Hamiltonian neural extensions [Desai et al., 2021] encode physical structure through Hamiltonian or Lagrangian parametrisations, symplectic integration, graph structure, or passivity. Adaptable HNNs [Han et al., 2021] also condition on varying system parameters. These works motivate our backbone, but their primary objective is accurate structure-preserving dynamics rather than uncertainty quantification that remains valid inside a symplectic rollout. We ask a narrower question. Given a learned Hamiltonian model with dropout-style uncertainty, how should stochastic samples be used so that each rollout still corresponds to one coherent Hamiltonian?

Dropout and ensemble uncertainty. MC dropout [Gal and Ghahramani, 2016a], locked dropout for recurrent networks [Gal and Ghahramani, 2016b], Concrete dropout [Gal et al., 2017], and Variational dropout [Kingma et al., 2015, Molchanov et al., 2017] provide single-model uncertainty estimates by injecting mask or weight noise. In a generic feed-forward predictor, resampling this

noise at inference is natural. In sequence models, locked dropout [Gal and Ghahramani, 2016b, Merity et al., 2018] already shows that reusing one stochastic realisation across time can be the right modelling choice. In a leapfrog HNN rollout, however, the consequence is specifically Hamiltonian. Fresh masks inside the integration loop mean that the effective Hamiltonian changes within a single trajectory. We therefore place the fixed-mask rule in the locked-dropout family and do not claim the masking mechanism as new. Three things differ. First, the failure mode is geometric. Under fresh masks the leapfrog composite no longer discretises a single Hamiltonian, so backward-error analysis no longer supplies one nearly conserved modified energy (Corollary 1). Second, a fixed mask is a coherent model-sample axis that can be crossed with an initial-condition ensemble, a decomposition with no counterpart in recurrent locked dropout. Third, the difference is measurable. Over ten seeds Fixed-mask MC lowers H-drift relative to Fresh-mask MC by 0.153 on the Coupled Pendulum and 1.64 on the Spring Chain (Holm-adjusted $p < 0.05$, Table 20). Deep Ensembles [Lakshminarayanan et al., 2017] avoid this particular problem because each ensemble member is a deterministic model, and they are strong accuracy and likelihood baselines throughout our experiments. Their trade-off is cost. They obtain uncertainty by training, storing, and evaluating multiple independent networks. SWAG [Maddox et al., 2019] offers a cheaper weight-space approximation, but it does not by itself specify the trajectory-level masking contract needed for single-Hamiltonian rollouts.

Finite masks and controlled dropout. The closest methodological relatives are Masksembles [Durasov et al., 2021] and Controlled Dropout [Hasan et al., 2023], which introduce finite or structured mask sets as an efficient alternative to independent ensembles. Our distinction is not the existence of a finite mask bank alone. The distinction is to treat the mask index as a trajectory-level Hamiltonian sample. Once a rollout begins, the same mask is held fixed through every leapfrog sub-step. This converts the finite mask bank into a stable model-sample axis and makes it possible to cross that axis with an initial-condition ensemble. The resulting decomposition is therefore tied to a Hamiltonian trajectory interpretation instead of only to calibrated predictive intervals.

Bayesian physics-informed and stochastic Hamiltonian models. Bayesian PINNs [Yang et al., 2021] place probabilistic inference inside physics-constrained learning, often with substantially higher compute cost. Hamiltonian and port-Hamiltonian Gaussian processes [Beckers et al., 2022, Ewering et al., 2025, Leung and Paré, 2025] and related GP formulations provide posterior uncertainty with strong structure but inherit GP scaling constraints. Recent stochastic port-Hamiltonian approaches, such as stochastic port-Hamiltonian neural networks [Di Persio et al., 2026], address a different modelling question. In those models, stochasticity is part of the physical dynamics being learned. In contrast, our systems are deterministic conditional on the control and initial state. The stochasticity represents model and IC/preparation uncertainty around that deterministic Hamiltonian rollout.

Calibration and graph uncertainty. Post-hoc calibration and graph uncertainty methods are complementary (Appendix C, with a comparison table).

4 Method

4.1 Notation

We write a classical phase-space state as $z = (\mathbf{p}, \mathbf{q})$ and the TFIM latent state as $z = \mathrm{concat}(q, p)$ for $\psi = q + ip$. A selected finite-bank mask is indexed by m, and initial-condition (IC) or state-preparation replicas are indexed by i. The decomposition is operational rather than a claim of exact Bayesian identifiability. Here, $\mathrm{Var}_{\mathrm{epi}}$ denotes a model-sample variance proxy across sampled Hamiltonians, while $\mathrm{Var}_{\mathrm{alea}}$ denotes initial-condition, preparation, or reconstruction sensitivity at fixed dynamics. The full notation table and axis conventions are given in Appendix D.

4.2 Fixed-mask dropout

At initialisation we sample a bank $\{b^{(1)}, \ldots, b^{(M)}\}$ of M Bernoulli masks per `MaskedLinear` layer and store them as buffers. Each rollout selects one mask index $m \in \{1, \ldots, M\}$ and reuses it through all leapfrog sub-steps. The index is resampled only during training or across MC samples. This is the modelling contract audited in Appendix K and detailed algorithmically in Appendix C.

Corollary 1 (Per-mask modified energy)**.** *Fix a mask $b^{(m)}$ from the finite bank and define $H_m(p,q) = H_\theta(p,q,b^{(m)})$. Assume that (A1) H_m is separable, $H_m = T_m(p) + V_m(q)$, (A2) H_m is smooth, with derivatives bounded uniformly over the bank on a compact set containing the rollout, and (A3) the step size Δt is below a threshold set by these bounds and the horizon $N\Delta t$ is bounded. Then (i) because the mask is fixed, every sub-step of the rollout evaluates the same Hamiltonian H_m, (ii) the leapfrog map for this fixed H_m is symplectic, (iii) backward-error analysis [Hairer et al., 2006] gives a modified Hamiltonian $\widetilde{H}_m = H_m + O(\Delta t^2)$, and (iv) $\widetilde{H}_m$ is nearly conserved along the rollout, so $|H_m(z_N) - H_m(z_0)| = O(\Delta t^2)$, with constants uniform over the finite mask bank.*

The proof and full assumptions are given in Appendix A. Corollary 1 is standard geometric integration applied to one fixed parametrisation, and it holds for any fixed parametrisation of H_θ, including a deterministic HNN. Our contribution is the contrast with fresh masks, for which step (i) fails (Equation (3)). The corollary does not claim that H_m is the true physical Hamiltonian or that predictive error is small. It only states that each stochastic sample has a coherent Hamiltonian rollout interpretation.

Extensions beyond MC dropout. The same construction applies to other stochastic regularisers provided the random draw is fixed per rollout. Concrete dropout [Gal et al., 2017] becomes trajectory-consistent by pre-sampling M hard Gumbel mask realisations, and Variational dropout [Kingma et al., 2015, Molchanov et al., 2017] by pre-sampling M realisations of the Gaussian weight noise. In both cases the inference-time forward pass selects a single pre-drawn sample and reuses it for the entire rollout.

4.3 Initial-condition (IC) noise ensemble and IC-vs-model variance decomposition

Because each fixed-mask rollout is conditionally deterministic given an initial condition (IC), we can draw a clean variance decomposition. For each test initial condition z_0, we construct N perturbed state-preparation replicas $z_0^{(i)}$ and launch Algorithm 1 from each. Classical systems use Gaussian phase-space perturbations, while TFIM uses the latent wavefunction perturbation described in Appendix G.2. The law of total variance then gives

$$\begin{aligned} \mathrm{Var}_{\text{total}}(t) &= \underbrace{\mathbb{E}_{z_0^{(i)}} \mathrm{Var}_m[y_{m,i}(t)]}_{\mathrm{Var}_{\text{epi}}(t)} + \underbrace{\mathrm{Var}_{z_0^{(i)}} \mathbb{E}_m[y_{m,i}(t)]}_{\mathrm{Var}_{\text{alea}}(t)}, \qquad \mathrm{Var}_{\text{total}} = \langle \mathrm{Var}_{\text{total}}(t) \rangle_{t,d}, \\ \mathrm{Var}_{\text{epi}} &= \langle \mathrm{Var}_{\text{epi}}(t) \rangle_{t,d}, \qquad \mathrm{Var}_{\text{alea}} = \langle \mathrm{Var}_{\text{alea}}(t) \rangle_{t,d}, \qquad \alpha := \frac{\mathrm{Var}_{\text{epi}}}{\mathrm{Var}_{\text{total}}}. \end{aligned} \tag{4}$$

Here $\langle \cdot \rangle_{t,d}$ denotes averaging over the evaluation time grid and state dimensions. Thus the decomposition is pointwise in time, while the reported $\mathrm{Var}_{\text{epi}}$, $\mathrm{Var}_{\text{alea}}$, $\mathrm{Var}_{\text{total}}$, and α in tables and figures are trajectory-level summaries. The $\mathrm{Var}_{\text{epi}}$ term measures disagreement among masks while holding the initial condition fixed. We interpret it as a model-sample variance proxy rather than as an exact posterior epistemic term. The $\mathrm{Var}_{\text{alea}}$ term measures sensitivity to initial-condition perturbation while holding the effective dynamics fixed. In this paper, "aleatoric" therefore means state-preparation or reconstruction sensitivity rather than generic observation noise. Thus larger α means that a larger share of predictive variance is model-sample driven. Fresh-mask MC also permits a formal law-of-total-variance decomposition over rollout randomness and IC perturbations. Its stochastic rollout axis, however, does not correspond to coherent sampled Hamiltonians, because masks are resampled within the rollout and couple effects inside every integrator call.

The practical point is that the decomposition is read from crossed mask and initial-condition axes rather than from a flattened pool of rollouts. Fixed masks make the model-sample index stable for an entire trajectory, so averaging over masks estimates dynamics uncertainty and averaging over initial-condition replicas estimates state-preparation sensitivity. Appendix D gives the full axis-preserving reduction used for implementation and TFIM regrouping.

We report IC-ensemble likelihoods under two conventions. In the fitted convention the IC-noise scale $\hat{\sigma}$ is inferred on validation data. In the true-scale convention it is fixed at the data-generating value $\sigma = 0.1$, adding the $\mathcal{N}(0, 0.1^2)$ measurement-noise term ($\epsilon = 0.10$ for TFIM, without that term). Because a fitted scale can absorb model error into the aleatoric term, we report $\hat{\sigma}$ next to every likelihood. Appendix G.1 details both conventions.

5 Experiments

5.1 Controlled complexity ladder

The experiments form a controlled complexity ladder. Coupled Pendulum is the cleanest OOD setting because the Hamiltonian depends on one interpretable control, the mass ratio $r = m_2/m_1$. Spring Chain adds pairwise interactions through a stiffness/connectivity matrix K (connection-matrix backbone shared by all methods, Appendix C.1), making it a structured stress test rather than a one-dimensional physics-distance claim. TFIM then moves to latent quantum wavefunction dynamics, with $\psi = q + ip$ and transverse-field control g, to test whether the same mask-vs-IC decomposition remains auditable in a harder dynamical system.

Methods and metrics. Baselines include a deterministic HNN, a five-member Deep Ensemble, Fresh-mask MC (Section 2), Concrete dropout, and Variational dropout. Deep Ensembles are treated as expensive reference rows rather than as the primary single-model competitor. Fresh-mask MC is the primary comparator because per-evaluation resampling is the default behaviour of off-the-shelf MC-dropout implementations, and hence the failure mode we diagnose. The comparison is controlled. On each system, every single-model arm shares the architecture, dropout rate, integrator, data and training budget, and each model is evaluated with the masks it was trained with. Concrete dropout [Gal et al., 2017], Variational dropout [Kingma et al., 2015, Molchanov et al., 2017] and Deep Ensembles [Lakshminarayanan et al., 2017] widen the comparison. Our central claim is that the mask contract changes trajectory-level behaviour within MC dropout, the most widely used single-model uncertainty family. Every single-model arm is trained with ten seeds (42–51). Table 1 reports means and Table 9 mean $\pm$ sd. Bold indicates the best stochastic single-model arm in a column when its paired tests, all of which involve Fixed-mask MC, are significant. We compare Fixed-mask MC with each other stochastic single-model arm by paired effect sizes with 95% percentile bootstrap confidence intervals, and test them with two-sided paired t-tests, Holm-corrected within three predefined families per system (Tables 20 and 21, protocol in Appendix K). "Significantly" refers to these tests at Holm-adjusted $p < 0.05$ unless stated otherwise. Likelihoods follow Section 4.3. Sample counts are matched across the stochastic arms of a system (Appendix K). We use MSE for mean squared trajectory-state error over the evaluated rollout. We report $\mathrm{Var}_{\mathrm{epi}}$, $\mathrm{Var}_{\mathrm{alea}}$, and the epistemic fraction α defined in Equation (4). These are trajectory-level summaries, the pointwise variances averaged over rollout time ($t \geq 1$) and state dimensions before forming α. Thus high α means that predictive variance is driven more by mask/model uncertainty than by initial-condition sensitivity.

5.2 Main quantitative comparison

Table 1 summarises the controlled runs for the two classical systems. Fixed-mask MC is competitive with other single-model stochastic baselines, but we do not claim a universal advantage on every predictive metric. On both systems, Fixed-mask MC has a significantly lower drift of its own sampled Hamiltonian (H-drift) than each other stochastic method, at the level of the deterministic network (0.00432 vs. 0.00428 on the Coupled Pendulum, 0.108 vs. 0.106 on the Spring Chain, against 0.157 and 1.751 for Fresh-mask MC). On the Spring Chain it also has a significantly lower OOD MSE (4.010) than each other stochastic method. On the Coupled Pendulum its OOD MSE (1.032) is on par with Concrete (1.026) and Fresh-mask MC (1.097) (Table 20). Fixed masks matter because they preserve a trajectory-level Hamiltonian interpretation and enable the decomposition, even when predictive advantages are system- and metric-dependent.

Fitted IC noise absorbs model error. The data carry IC noise $\sigma = 0.1$, whereas the IC-ensemble likelihood fits $\hat{\sigma}$ on validation data. Fixed-mask MC fits the scale closest to the truth ($\hat{\sigma} = 0.250$ on the Coupled Pendulum, 0.130 on the Spring Chain). Fresh-mask MC fits 0.331 and 0.483, and the deterministic, Concrete and Variational models fit 0.252–0.351. On these systems an inflated $\hat{\sigma}$ lets the aleatoric (IC) term absorb model error. Under the fitted convention Fixed-mask MC is still significantly better in 8 of the 12 paired NLL comparisons (Table 20), but the fresh-mask baselines attribute almost all predictive variance to the initial state (OOD $\alpha = 0.0326$ and 0.0464 for Fresh-mask MC, 0.0448 and 0.0661 for Concrete). This is why Fresh-mask MC has a significantly lower Spring Chain OOD G-NLL under $\hat{\sigma}$ than Fixed-mask MC (4.779 vs. 11.78). At the true $\sigma = 0.1$ the ranking reverses. Fixed-mask MC has a significantly lower ID and OOD G-NLL than each other

Table 1: Main comparison on the Coupled Pendulum and the Spring Chain. G-NLL is the Gaussian NLL of the IC $\times$ mask ensemble under two IC-noise conventions, where $\hat{\sigma}$ is the IC-noise scale fitted per run on validation data, and σ=0.1 fixes it at the data's true value and adds the $\mathcal{N}(0, 0.1^2)$ measurement-noise term. H-drift is the drift of the model's own sampled Hamiltonian along the rollout (a self-consistency measure, not physical energy, with physical drift in Table 19). Cov90 errors are in Table 7.

	MSE			G-NLL ($\hat{\sigma}$)		G-NLL (σ=0.1)			
Method	ID	OOD	$\hat{\sigma}$	ID	OOD	ID	OOD	OOD α	H-drift
Coupled pendulum									
Fresh-mask MC	0.117	1.097	0.331	0.274	3.298	1.064	14.45	0.0326	0.157
Concrete	0.115	1.026	0.294	0.267	3.837	1.091	13.35	0.0448	0.124
Variational	0.209	1.111	0.351	0.517	2.871	1.340	7.770	0.213	0.530
Fixed-mask MC	0.113	1.032	0.250	**0.224**	2.774	**0.477**	**4.739**	0.524	**0.00432**
Deterministic	0.119	1.189	0.290	0.273	4.255	1.332	20.08	n/a	0.00428
Deep ens. $\times$5	0.115	1.036	0.291	0.261	2.948	1.083	8.106	0.461	0.00431
Spring chain									
Fresh-mask MC	0.654	4.424	0.483	0.980	**4.779**	5.221	33.84	0.0464	1.751
Concrete	0.667	4.370	0.252	1.729	15.38	5.013	40.45	0.0661	1.732
Variational	1.721	4.430	0.310	1.776	7.711	3.878	17.42	0.426	4.111
Fixed-mask MC	**0.613**	**4.010**	0.130	0.989	11.78	**0.974**	**11.55**	0.833	**0.108**
Deterministic	0.529	4.441	0.253	1.370	18.64	5.055	66.14	n/a	0.106
Deep ens. $\times$5	0.538	4.105	0.216	1.576	17.24	3.812	35.62	0.271	0.106

stochastic method on both systems (0.477/4.739 on the Coupled Pendulum and 0.974/11.55 on the Spring Chain, against 1.064/14.45 and 5.221/33.84 for Fresh-mask MC, Table 20).

5.3 Coupled Pendulum OOD decomposition

The Coupled Pendulum is the paper's principal evidence because it has a physically interpretable OOD axis. The mass ratio $r = m_2/m_1$ is swept beyond the training interval, and the question is whether the uncertainty decomposition changes in the right direction. In distribution, uncertainty should be dominated by initial-condition sensitivity. Better state preparation or measurement would reduce the uncertainty. As the mass ratio moves away from the training physics, uncertainty should shift toward the epistemic axis because the learned Hamiltonian itself is being queried outside its support.

Figure 2 gives the main OOD decomposition evidence. The α panel asks whether the epistemic share increases OOD, and the physics-distance panel asks whether absolute epistemic uncertainty grows with distance from the training support. Fixed-mask MC carries about $19\times$ more variance on the mask axis than Fresh-mask MC, both in distribution ($\mathrm{Var}_{\mathrm{epi}} = 0.0203$ vs. 0.00108) and OOD (0.119 vs. 0.00637, Tables 16 and 20), while retaining competitive MSE and H-drift in Table 1. Its epistemic fraction rises from 0.203 to 0.524 ($+32.1$ percentage points), against 0.00753 to 0.0326 ($+2.5$) for Fresh-mask MC. The sharper rise reflects the attribution. Fresh masks absorb model error into the IC term (fitted $\hat{\sigma} = 0.331$ vs. 0.250), which holds their α near zero, and the likelihood at the true $\sigma = 0.1$ favours Fixed-mask MC (Appendix I).

5.4 Spring Chain structured stress test

Spring Chain is the intermediate complexity step. It tests whether the decomposition remains informative when the OOD shift changes pairwise interactions rather than a single scalar mass ratio. Fixed-mask MC gives the strongest conservation profile among the stochastic methods in Table 1 and remains competitive on OOD MSE. Over 1000-step rollouts, the per-rollout physical energy error of Fresh-mask MC grows to $4.71\times$ (in distribution) and $3.46\times$ (OOD) that of Fixed-mask MC, and to $1.94\times$ and $1.88\times$ on the Coupled Pendulum (Appendix J). The appendix OOD diagnostic further shows that epistemic variance is sensitive to support shifts and grows with distance from the training support. This evidence is less clean than the Coupled Pendulum result, because a scalar

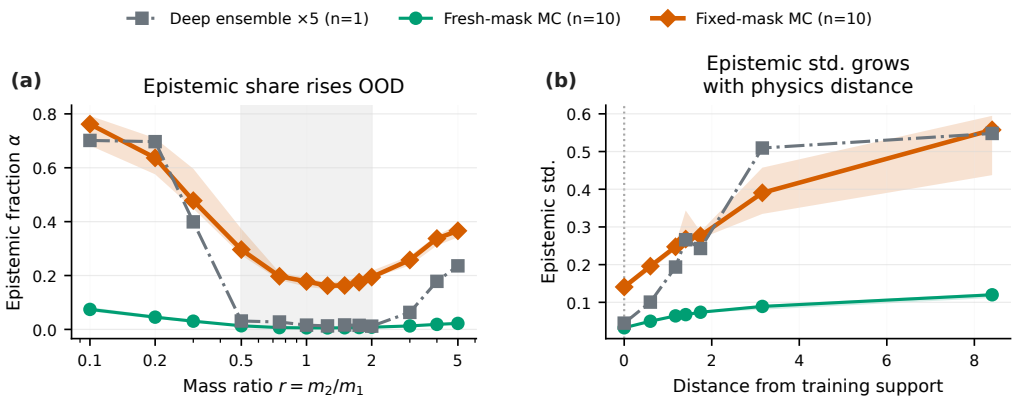


Figure 2: Main Coupled Pendulum OOD evidence. The left panel shows the epistemic fraction α (Equation (4)) over the mass-ratio sweep, and the right panel shows epistemic standard deviation versus distance from the training support. The shaded band marks the training support in mass ratio. Fixed-mask MC gives a stronger single-model epistemic OOD response than Fresh-mask MC. Deep Ensemble is included as a compute-heavy reference curve.

stiffness proxy cannot fully capture topology-dependent changes in the Hamiltonian mode structure. Still, it supports the method on a structured interacting system rather than serving only as a negative boundary case.

5.5 TFIM latent quantum dynamics

TFIM tests whether the same mask-vs-state-preparation decomposition can be applied to latent wavefunction dynamics. The model operates on $z = \mathrm{concat}(q, p)$ for $\psi = q + ip$, with $g = h/J$ supplied as an external control. We treat this as a supporting diagnostic, not as central evidence for fixed-mask dominance. Appendix G.2 defines the TFIM version of the IC ensemble.

Figure 3 shows that the mask-vs-IC decomposition remains visible in a harder latent system. It also states the main limitation. The quantitative TFIM table is not a fixed-mask performance win. At the true $\epsilon = 0.10$, however, Fixed-mask MC has a significantly lower OOD G-NLL (-1.250 vs. 3.343 for Fresh-mask MC) and OOD coverage error than each other stochastic arm (Tables 8 and 21). Fixed-mask dropout provides trajectory-consistent decomposition, not a guarantee of best predictive performance on every Hamiltonian system.

Appendix material. The detailed Coupled Pendulum OOD table, full decomposition table, TFIM quantitative table, efficiency table, Spring Chain OOD plots, and additional ablations are reported in Appendix H, Appendix I, and Appendix J.

6 Discussion

Relationship to Deep Ensembles. A Deep Ensemble produces independently learned Hamiltonians and is a strong accuracy reference. Its training, storage, and evaluation costs scale with the number of members. Fixed-mask dropout keeps training and storage single-model while paying mainly for additional inference samples. Appendix H.8 (Table 14) reports this cost profile. Its crossed mask and initial-condition axes also make the law-of-total-variance split a direct tensor reduction rather than a post-hoc aggregate statistic.

Decision value and a rule of thumb. The two axes call for different actions. Large or growing $\mathrm{Var}_{\mathrm{epi}}$ signals that the model is queried where it has not learned the dynamics, so the response is to collect data in that regime, revise the model, or abstain. Large $\mathrm{Var}_{\mathrm{alea}}$ signals sensitivity to the initial state, so the response is better sensing or state estimation. As a rule of thumb, hold the mask fixed for any quantity read along a trajectory, such as energy drift or per-rollout scores, because only then does each sample follow one Hamiltonian. At equal M this needs the same number of rollouts and

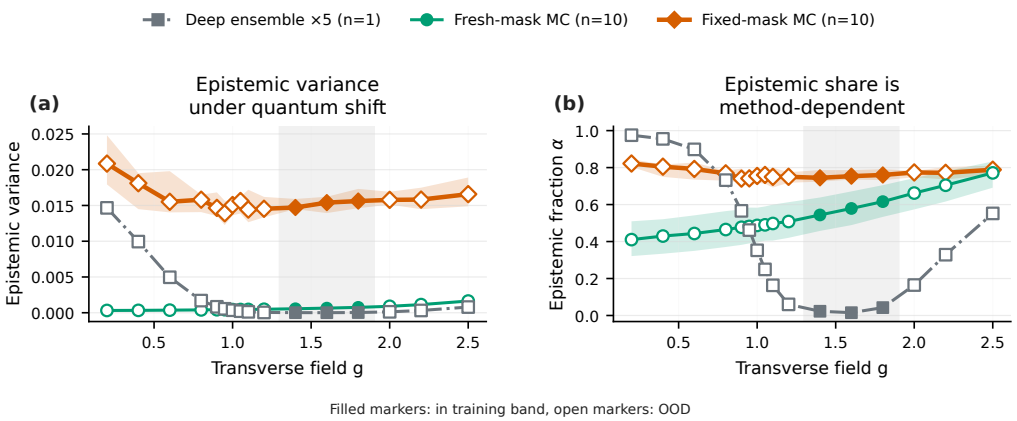


Figure 3: TFIM latent-dynamics stress test. The shaded region marks the training band, filled markers denote in-band controls, and open markers denote OOD controls. Fixed-mask MC has the largest epistemic variance and highest average epistemic fraction α in this diagnostic, showing that the axis-preserving decomposition transfers to latent quantum dynamics. The same experiment also illustrates the limitation. Fixed-mask MC is not the universal winner on point prediction or conservation metrics in this more complex system.

network evaluations as fresh-mask sampling. We have not validated a closed-loop decision policy built on this split.

Limitations and broader impacts. Corollary 1 covers conservative, separable systems integrated with leapfrog only. Likelihood rankings depend on the IC-noise convention. On the Coupled Pendulum the physical-energy benefit appears only at long horizons. The controlled, low-dimensional systems omit trajectory noise, partial observability, forcing and model mismatch. The split is a diagnostic next to validation, not a safety guarantee. Appendix B gives the full limitations, dissipative systems, structured OOD distances, invariant architectures and broader impacts.

7 Conclusion

Fresh-mask MC weakens the trajectory-level Hamiltonian semantics of HNNs by resampling dropout masks inside the integrator, whereas Fixed-mask MC samples one mask per rollout and reuses it at every leapfrog sub-step so that each stochastic trajectory follows a single masked Hamiltonian H_m. On both classical systems Fixed-mask MC fits the initial-condition noise closest to its true scale and gives the best likelihood at that scale, whereas the fresh-mask baselines reach competitive fitted-scale likelihood by inflating $\hat{\sigma}$ and attributing model error to the initial state. The held-out mass-ratio study shows that fixed masks keep model uncertainty on the model axis, where fresh masks count it as initial-state noise, and over long rollouts each sampled trajectory keeps its physical energy far better than under Fresh-mask MC. Spring Chain extends this decomposition to structured pairwise interactions, while TFIM shows that predictive gains remain system-dependent in harder latent dynamics. Fixed-mask MC is therefore a practical single-training-run route to trajectory-consistent uncertainty estimates whose IC-vs-model split can guide whether to improve state estimation, collect new training data, revise the model, or abstain.

Acknowledgments and Disclosure of Funding

YT's DPhil is funded by OXFIDA Ltd, which covers tuition fees; YT is also affiliated with OXFIDA Ltd. We thank Juliusz Ziomek, Masaki Adachi and Masahiro Fujisawa for helpful discussions and feedback on earlier drafts, and the anonymous reviewers for their comments. Experiments were run on NVIDIA RTX A6000 GPUs provided by the Machine Learning Research Group, University of Oxford.

References

Thomas Beckers, Jacob H. Seidman, Paris Perdikaris, and George J. Pappas. Gaussian Process Port-Hamiltonian Systems: Bayesian Learning with Physics Prior. In *Proceedings of the IEEE 61st Conference on Decision and Control*, pages 1447–1453, 2022. doi: 10.1109/CDC51059.2022.9992733. URL https://doi.org/10.1109/CDC51059.2022.9992733.

Zhengdao Chen, Jianyu Zhang, Martin Arjovsky, and Léon Bottou. Symplectic Recurrent Neural Networks. In *International Conference on Learning Representations*, 2020. URL https://openreview.net/forum?id=BkgYPREtPr.

Miles Cranmer, Sam Greydanus, Stephan Hoyer, Peter Battaglia, David Spergel, and Shirley Ho. Lagrangian Neural Networks. In *ICLR 2020 Workshop on Integration of Deep Neural Models and Differential Equations*, 2020. URL https://arxiv.org/abs/2003.04630.

Alessandro De Luca, Alin Albu-Schäffer, Sami Haddadin, and Gerd Hirzinger. Collision Detection and Safe Reaction with the DLR-III Lightweight Manipulator Arm. In *Proceedings of the IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*, pages 1623–1630, 2006. doi: 10.1109/IROS.2006.282053. URL https://doi.org/10.1109/IROS.2006.282053.

C. M. DeChant and H. Moradkhani. Toward a Reliable Prediction of Seasonal Forecast Uncertainty: Addressing Model and Initial Condition Uncertainty with Ensemble Data Assimilation and Sequential Bayesian Combination. *Journal of Hydrology*, 519:2967–2977, 2014. doi: 10.1016/j.jhydrol.2014.05.045. URL https://doi.org/10.1016/j.jhydrol.2014.05.045.

Shaan Desai, Marios Mattheakis, David Sondak, Pavlos Protopapas, and Stephen Roberts. Port-Hamiltonian Neural Networks for Learning Explicit Time-Dependent Dynamical Systems. *Physical Review E*, 104(3):034312, 2021. doi: 10.1103/PhysRevE.104.034312. URL https://doi.org/10.1103/PhysRevE.104.034312.

Luca Di Persio, Matthias Ehrhardt, and Youness Outaleb. Stochastic Port-Hamiltonian Neural Networks: Universal Approximation with Passivity Guarantees. *arXiv preprint arXiv:2603.10078*, 2026. doi: 10.48550/arXiv.2603.10078. URL https://arxiv.org/abs/2603.10078.

Nikita Durasov, Timur Bagautdinov, Pierre Baque, and Pascal Fua. Masksembles for Uncertainty Estimation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 13534–13543, 2021. doi: 10.1109/CVPR46437.2021.01333. URL https://doi.org/10.1109/CVPR46437.2021.01333.

Jan-Hendrik Ewering, Robin E. Herrmann, Niklas Wahlström, Thomas B. Schön, and Thomas Seel. Learning Dynamics from Input-Output Data with Hamiltonian Gaussian Processes. *arXiv preprint arXiv:2511.05330*, 2025. doi: 10.48550/arXiv.2511.05330. URL https://arxiv.org/abs/2511.05330.

Yarin Gal and Zoubin Ghahramani. Dropout as a Bayesian Approximation: Representing Model Uncertainty in Deep Learning. In *Proceedings of the 33rd International Conference on Machine Learning*, volume 48 of *Proceedings of Machine Learning Research*, pages 1050–1059. PMLR, 2016a.

Yarin Gal and Zoubin Ghahramani. A Theoretically Grounded Application of Dropout in Recurrent Neural Networks. In *Advances in Neural Information Processing Systems*, volume 29, 2016b.

Yarin Gal, Jiri Hron, and Alex Kendall. Concrete Dropout. In *Advances in Neural Information Processing Systems*, volume 30, 2017.

Sam Greydanus, Misko Dzamba, and Jason Yosinski. Hamiltonian Neural Networks. In *Advances in Neural Information Processing Systems*, volume 32, 2019.

Chuan Guo, Geoff Pleiss, Yu Sun, and Kilian Q. Weinberger. On Calibration of Modern Neural Networks. In *Proceedings of the 34th International Conference on Machine Learning*, volume 70 of *Proceedings of Machine Learning Research*, pages 1321–1330. PMLR, 2017.

Ernst Hairer, Christian Lubich, and Gerhard Wanner. *Geometric Numerical Integration: Structure-Preserving Algorithms for Ordinary Differential Equations*. Springer, 2nd edition, 2006.

Chen-Di Han, Bryan Glaz, Mulugeta Haile, and Ying-Cheng Lai. Adaptable Hamiltonian Neural Networks. *Physical Review Research*, 3(2):023156, 2021. doi: 10.1103/PhysRevResearch.3.023156. URL https://doi.org/10.1103/PhysRevResearch.3.023156.

Mehedi Hasan, Ibrahim Hossain, Ashikur Rahman, and Saeid Nahavandi. Controlled Dropout for Uncertainty Estimation. In *Proceedings of the IEEE International Conference on Systems, Man, and Cybernetics*, pages 973–980, 2023. doi: 10.1109/SMC53992.2023.10394101. URL https://doi.org/10.1109/SMC53992.2023.10394101.

Arman Hasanzadeh, Ehsan Hajiramezanali, Shahin Boluki, Mingyuan Zhou, Nick Duffield, Krishna Narayanan, and Xiaoning Qian. Bayesian Graph Neural Networks with Adaptive Connection Sampling. In *Proceedings of the 37th International Conference on Machine Learning*, volume 119 of *Proceedings of Machine Learning Research*, pages 4094–4104. PMLR, 2020. URL https://proceedings.mlr.press/v119/hasanzadeh20a.html.

Pengzhan Jin, Zhen Zhang, Aiqing Zhu, Yifa Tang, and George Em Karniadakis. SympNets: Intrinsic Structure-Preserving Symplectic Networks for Identifying Hamiltonian Systems. *Neural Networks*, 132:166–179, 2020. doi: 10.1016/j.neunet.2020.08.017. URL https://doi.org/10.1016/j.neunet.2020.08.017.

Michael G. Kapteyn, David J. Knezevic, D. B. P. Huynh, M. Tran, and Karen E. Willcox. Data-Driven Physics-Based Digital Twins via a Library of Component-Based Reduced-Order Models. *International Journal for Numerical Methods in Engineering*, 123(13):2986–3003, 2022. doi: 10.1002/nme.6423. URL https://doi.org/10.1002/nme.6423.

Diederik P. Kingma, Tim Salimans, and Max Welling. Variational Dropout and the Local Reparameterization Trick. In *Advances in Neural Information Processing Systems*, volume 28, 2015.

Volodymyr Kuleshov, Nathan Fenner, and Stefano Ermon. Accurate Uncertainties for Deep Learning Using Calibrated Regression. In *Proceedings of the 35th International Conference on Machine Learning*, volume 80 of *Proceedings of Machine Learning Research*, pages 2796–2804. PMLR, 2018.

Balaji Lakshminarayanan, Alexander Pritzel, and Charles Blundell. Simple and Scalable Predictive Uncertainty Estimation using Deep Ensembles. In *Advances in Neural Information Processing Systems*, volume 30, 2017.

Chi Ho Leung and Philip E. Paré. Learning Passive Continuous-Time Dynamics with Multistep Port-Hamiltonian Gaussian Processes. *arXiv preprint arXiv:2510.00384*, 2025. doi: 10.48550/arXiv.2510.00384. URL https://arxiv.org/abs/2510.00384.

Wesley J. Maddox, Pavel Izmailov, Timur Garipov, Dmitry Vetrov, and Andrew Gordon Wilson. A Simple Baseline for Bayesian Uncertainty in Deep Learning. In *Advances in Neural Information Processing Systems*, volume 32, 2019.

Stephen Merity, Nitish Shirish Keskar, and Richard Socher. Regularizing and Optimizing LSTM Language Models. In *International Conference on Learning Representations*, 2018. URL https://openreview.net/forum?id=SyyGPP0TZ.

Rhiannon Michelmore, Matthew Wicker, Luca Laurenti, Luca Cardelli, Yarin Gal, and Marta Kwiatkowska. Uncertainty Quantification with Statistical Guarantees in End-to-End Autonomous Driving Control. In *Proceedings of the IEEE International Conference on Robotics and Automation*, pages 7344–7350, 2020. doi: 10.1109/ICRA40945.2020.9196844. URL https://doi.org/10.1109/ICRA40945.2020.9196844.

Dmitry Molchanov, Arsenii Ashukha, and Dmitry Vetrov. Variational Dropout Sparsifies Deep Neural Networks. In *Proceedings of the 34th International Conference on Machine Learning*, volume 70 of *Proceedings of Machine Learning Research*, pages 2498–2507. PMLR, 2017.

Venkat Nemani, Luca Biggio, Xun Huan, Zhen Hu, Olga Fink, Anh Tran, Yan Wang, Xiaoge Zhang, and Chao Hu. Uncertainty Quantification in Machine Learning for Engineering Design and Health Prognostics: A Tutorial. *Mechanical Systems and Signal Processing*, 205:110796, 2023. doi: 10.1016/j.ymssp.2023.110796. URL https://doi.org/10.1016/j.ymssp.2023.110796.

Vu-Linh Nguyen, Sébastien Destercke, and Eyke Hüllermeier. Epistemic Uncertainty Sampling. In *Discovery Science*, volume 11828 of *Lecture Notes in Computer Science*, pages 72–86. Springer, 2019. doi: 10.1007/978-3-030-33778-0_7. URL https://doi.org/10.1007/978-3-030-33778-0_7.

Vu-Linh Nguyen, Mohammad Hossein Shaker, and Eyke Hüllermeier. How to Measure Uncertainty in Uncertainty Sampling for Active Learning. *Machine Learning*, 111(1):89–122, 2022. doi: 10.1007/s10994-021-06003-9. URL https://doi.org/10.1007/s10994-021-06003-9.

Alvaro Sanchez-Gonzalez, Victor Bapst, Kyle Cranmer, and Peter Battaglia. Hamiltonian Graph Networks with ODE Integrators. In *NeurIPS Workshop on Machine Learning and the Physical Sciences*, 2019. URL https://arxiv.org/abs/1909.12790.

Andrew Gordon Wilson, Zhiting Hu, Ruslan Salakhutdinov, and Eric P. Xing. Deep Kernel Learning. In *Proceedings of the 19th International Conference on Artificial Intelligence and Statistics*, volume 51 of *Proceedings of Machine Learning Research*, pages 370–378. PMLR, 2016.

Liu Yang, Xuhui Meng, and George Em Karniadakis. B-PINNs: Bayesian Physics-Informed Neural Networks for Forward and Inverse PDE Problems with Noisy Data. *Journal of Computational Physics*, 425:109913, 2021. doi: 10.1016/j.jcp.2020.109913. URL https://doi.org/10.1016/j.jcp.2020.109913.

A Proof of Corollary 1

For completeness, we state the assumptions used by Corollary 1. Let $H_m(p,q) := H_\theta(p,q,b^{(m)})$ for a fixed mask $b^{(m)}$ in the finite bank $\mathcal{B} = \{b^{(1)}, \dots, b^{(M)}\}$. Assume:

(A1) **Separability.** $H_m(p,q) = T_m(p) + V_m(q)$ for every $m \in \{1, \dots, M\}$.

(A2) **Smoothness on a compact domain.** There exists a compact set $\Omega \subset \mathbb{R}^{2d}$ containing the numerical rollout such that $H_m \in C^\infty(\Omega)$ with

$$\sup_{m \le M} \sup_{|\alpha| \le k} \left\| \partial^\alpha H_m \right\|_{L^\infty(\Omega)} \le C_\Omega,$$

for some $k \ge 4$ sufficient for the $O(\Delta t^2)$ truncation of the modified Hamiltonian.

(A3) **Step-size and horizon conditions** [Hairer et al., 2006]. The Störmer–Verlet step size satisfies $\Delta t \le h_0$ with $h_0 \left\| \nabla^2 H_m \right\|_{L^\infty(\Omega)} < c_0$ for the universal constant c_0 of symplectic backward-error analysis, and the rollout horizon obeys $N\Delta t \le T_{\max}$ with the numerical trajectory remaining in Ω.

Given a fixed binary mask $b^{(m)} \in \mathcal{B}$, the functions T_m and V_m in $H_m(\mathbf{p},\mathbf{q}) = T_m(\mathbf{p}) + V_m(\mathbf{q})$ are fixed smooth scalar functions on the rollout domain. The Störmer–Verlet step in Equation (2), with $H = H_m$, is the composition of the exact canonical subflows generated by T_m and V_m. Each subflow preserves the canonical symplectic form, and therefore the composition has Jacobian J satisfying $J^\top \mathbb{J} J = \mathbb{J}$ for the canonical symplectic matrix $\mathbb{J}$. This is the symplecticity statement. Determinant one or volume preservation alone would not be sufficient.

Backward-error analysis (see Chapter IX of Hairer et al. [2006]) then yields a modified Hamiltonian $\tilde{H}_m = H_m + \Delta t^2 H_2 + \Delta t^4 H_4 + \cdots$ that the numerical trajectory conserves in the sense of an asymptotic series over bounded time intervals under the usual smoothness and step-size assumptions. The learned masked-Hamiltonian error is therefore $|H_m(z_N) - H_m(z_0)| = O(\Delta t^2)$ over $N\Delta t \le T_{\max}$, with constants that depend on derivative norms of H_m and are uniform over the finite mask bank when those norms are uniformly bounded. This is a conditional statement. Once the mask is fixed, the sampled neural Hamiltonian defines the vector field being integrated. The

argument does not require the sampled Hamiltonian to equal the ground-truth physical Hamiltonian, and it does not make a prediction-error claim.

When the mask is resampled at every step, the step-wise Hamiltonian H_{m_t} varies with t. Each step is still symplectic, so the composite map is symplectic, but no single $\tilde{H}$ is conserved by the numerical trajectory. The modified-Hamiltonian argument breaks because the backward-error expansion assumes a single smooth vector field over the rollout. The issue is therefore not merely a larger variance estimate. The stochastic sample no longer has the interpretation of one Hamiltonian trajectory.

Remark 1 (Scope and necessity).

- *What Corollary 1 claims.* Each stochastic sample is a well-posed Hamiltonian rollout, a trajectory of one coherent sampled dynamics under a symplectic integrator with controlled $O(\Delta t^2)$ energy error.
- *What it does not claim.* Corollary 1 does not assert that H_m approximates the true physical Hamiltonian, nor that the predictive error of the rollout is small.
- *Necessity of fixed masks.* If the mask is resampled at any sub-step of the rollout, the effective generator becomes step-dependent, $H_{m_0}, H_{m_1}, \ldots, H_{m_{N-1}}$, and the conclusions above fail. The composite map remains symplectic but is not the discretisation of any single sampled Hamiltonian, and the modified-Hamiltonian argument no longer applies.

B Extended discussion, limitations and broader impacts

Conservative-system assumption. Corollary 1 assumes a fixed separable Hamiltonian and Störmer–Verlet integration. Dissipative, forced or controlled systems require port-Hamiltonian extensions [Desai et al., 2021]. The fixed-mask idea still transfers as a trajectory-consistency principle, but the formal symplecticity statement must be replaced by the appropriate passivity or conservation law. We do not derive a passivity guarantee for fixed masks. This is future work.

System-dependent OOD distance. Spring Chain shows both the promise and the difficulty of structured OOD evaluation. Epistemic variance responds to support shifts and grows with pooled physics distance, but a scalar stiffness proxy averages over topology changes that alter the Hamiltonian mode structure. The practical recommendation is to select the stochastic family and OOD distance per system while preferring fixed masks when per-trajectory consistency or IC-vs-model decomposition is required.

Invariant architectures and norm drift. The TFIM norm drift does not originate in the dropout scheme. The deterministic HNN, which has no masks, also drifts in norm (0.0123, Table 11), because the MLP backbone does not preserve the norm by construction. The masks change only the size of the drift. Fixed-mask sampling is complementary to norm-preserving or equivariant architectures. With the mask fixed, such a network still defines one Hamiltonian per rollout while keeping its invariant, so the IC-vs-model decomposition applies unchanged. To keep equivariance, the mask must be tied across symmetry-related channels. An untied mask breaks the symmetry of each sampled subnetwork. This is a design hypothesis that still needs empirical checks.

Limitations and broader impacts. Inference cost scales with the number of retained mask samples, as shown by the $M = 100$ and $M = 200$ efficiency rows. The formal symplecticity statement assumes a separable Hamiltonian. We also do not prove that finite-mask dropout converges to the MC dropout variational approximation as $M \to \infty$. The paper instead gives empirical evidence that finite mask banks are sufficient for the reported decompositions. Likelihood rankings depend on the IC-noise convention, so we report both conventions and $\hat{\sigma}$ itself. A better estimate of the IC noise would move the fitted-scale likelihood towards the true-scale one, where Fixed-mask MC has the lowest NLL of every stochastic method on both classical systems. On the Coupled Pendulum the physical-energy benefit of fixed masks appears only at long horizons (Appendix J). The systems are controlled simulations chosen to isolate the mechanism. State uncertainty enters the decomposition only through the initial condition, and the per-frame measurement noise on the classical targets enters only the likelihood. We do not model measurement or process noise along the trajectory, partial observability, external forcing, or structural model mismatch (a true system outside the learned

Hamiltonian class). Each adds uncertainty that the two-axis split does not separate. The systems are also low-dimensional (at most 32 state dimensions). Growth of the state space and Monte Carlo error are separate questions. The mask count is a sampling question (Table 13), while the behaviour of the decomposition in high-dimensional or macroscopic systems is untested, and we make no claim there. As for broader impacts, the method targets scientific and engineering surrogates.

C Method and related-work details

Algorithm 1 Fixed-mask HNN rollout.

Require: HNN H_θ with mask bank $\{b^{(1)}, \ldots, b^{(M)}\}$, initial state $z_0 = (\mathbf{p}_0, \mathbf{q}_0)$, step size Δt, number of steps N, sample count $M_{\text{eval}} \leq M$.

for $m = 1, \ldots, M_{\text{eval}}$ **do**
  Activate mask m in every `MaskedLinear` layer.
  $(\mathbf{p}, \mathbf{q}) \leftarrow (\mathbf{p}_0, \mathbf{q}_0)$.
  for $t = 1, \ldots, N$ **do**
    $\mathbf{p}_{1/2} \leftarrow \mathbf{p} - \frac{\Delta t}{2} \nabla_{\mathbf{q}} H_\theta(\mathbf{p}, \mathbf{q}, b^{(m)})$
    $\mathbf{q}' \leftarrow \mathbf{q} + \Delta t \, \nabla_{\mathbf{p}} H_\theta(\mathbf{p}_{1/2}, \mathbf{q}, b^{(m)})$
    $\mathbf{p}' \leftarrow \mathbf{p}_{1/2} - \frac{\Delta t}{2} \nabla_{\mathbf{q}} H_\theta(\mathbf{p}_{1/2}, \mathbf{q}', b^{(m)})$
    $(\mathbf{p}, \mathbf{q}) \leftarrow (\mathbf{p}', \mathbf{q}')$
  end for
  Store trajectory $z_{0:N}^{(m)}$.
end for
return ensemble $\{z_{0:N}^{(m)}\}_{m=1}^{M_{\text{eval}}}$.

Table 2: Comparison with related uncertainty families. The target niche is the combination of single-model cost, trajectory-level Hamiltonian interpretation, and an explicit initial-condition-vs-model decomposition axis.

Method	Single model	Fixed rollout sample	Hamiltonian interpretation	IC/model split	Physics OOD
MC dropout [Gal and Ghahramani, 2016a]	✓	✗	✗	partial	✗
Concrete dropout [Gal et al., 2017]	✓	✗	✗	partial	✗
Variational dropout [Kingma et al., 2015, Molchanov et al., 2017]	✓	✗	✗	partial	✗
Deep Ensembles [Lakshminarayanan et al., 2017]	✗	✓	✓	aggregate	partial
Masksembles [Durasov et al., 2021]	✓	partial	✗	partial	✗
Controlled Dropout [Hasan et al., 2023]	✓	partial	✗	partial	✗
Bayesian PINNs [Yang et al., 2021]	✗	✓	PDE-specific	✓	partial
Hamiltonian/Port-Hamiltonian GP [Beckers et al., 2022, Ewering et al., 2025, Leung and Paré, 2025]	✗	✓	✓	GP-native	partial
Stochastic port-Hamiltonian models [Di Persio et al., 2026]	varies	✓	✓	physical noise	partial
Fixed-mask dropout (ours)	✓	✓	✓	✓	✓

Calibration and graph uncertainty. Post-hoc calibration methods [Guo et al., 2017, Kuleshov et al., 2018] and graph and GP-based uncertainty methods [Hasanzadeh et al., 2020, Wilson et al., 2016] are complementary to our approach. Calibration can improve interval coverage after predictions are generated, and graph UQ can be combined with structured simulators. Neither addresses the issue studied here, which is to preserve a coherent sampled Hamiltonian across the numerical integration path while keeping model and IC/preparation uncertainty separable. Table 2 gives a compact comparison.

C.1 Supporting architecture with a connection matrix

For the Spring Chain we use a connection-matrix architecture

$$H(\mathbf{p}, \mathbf{q}, K) = T(\mathbf{p}) + \sum_{(i,j)\in \mathrm{edges}(K)} V_\theta(q_i,\, q_j,\, k_{ij}), \tag{5}$$

where K is a sparse spring-stiffness matrix and V_θ has input dimension three. The architecture is chain-length agnostic. A model trained on $N{=}4$ masses can be evaluated on larger N by applying the same edge network to additional edges. Fixed-mask dropout is orthogonal to this architectural choice, and all methods compared on a given system share the same backbone.

D Notation, uncertainty axes, and OOD definitions

This appendix records the detailed notation and distance definitions used by the main text.

D.1 Notation and uncertainty axes

We write a phase-space state as $z = (\mathbf{p}, \mathbf{q})$ for the classical systems and use $z = \mathrm{concat}(q, p)$ for the TFIM latent state. A system-level control variable is denoted by c. It is the mass ratio r in the Coupled Pendulum, the stiffness/topology pair (k, c) in the Spring Chain, and the transverse field g in TFIM. The finite mask bank has M entries and a selected mask index is denoted by m. Initial-condition or preparation replicas are indexed by i, and latent/bootstrap replicas inside a given intended initial condition are indexed by l.

The terminology is operational. We use *epistemic* uncertainty for variation across sampled Hamiltonians, represented by the mask/model axis m. We use *aleatoric* uncertainty for variation induced by the initial-condition, preparation, or reconstruction ensemble, represented by the initial-condition (IC) axis i and, in TFIM, the latent/bootstrap axis l. Thus "aleatoric" means initial-condition/preparation uncertainty in this paper rather than direct observation noise on every coordinate.

Table 3: Notation used for stochastic and physical axes. The central decomposition requires the model-sample axis and the initial-condition/preparation axis to remain separately identifiable through the rollout.

Symbol	Meaning	Typical role
m	mask/model sample index	epistemic axis
M	finite mask-bank size	number of sampled Hamiltonians
i	intended initial condition or initial-condition perturbation index	initial-condition/preparation axis
l	latent/bootstrap replica index	TFIM reconstruction axis
t	time index	rollout horizon
r	Coupled Pendulum mass ratio m_2/m_1	OOD control
(k, c)	spring stiffness and topology/stratum	structured OOD proxy
g	TFIM transverse field h/J	OOD control
$\mathrm{Var}_{\mathrm{epi}}$	variance across mask/model samples	model uncertainty
$\mathrm{Var}_{\mathrm{alea}}$	variance across initial-condition/preparation samples	initial-condition uncertainty
α	$\mathrm{Var}_{\mathrm{epi}}/\mathrm{Var}_{\mathrm{total}}$	epistemic fraction

Algorithm 2 Axis-preserving uncertainty decomposition.

Require: Predictions $y_{m,i,l}(t)$ with mask/model index m, intended initial-condition index i, optional latent/bootstrap index l, and time index t.

1: Preserve the mask axis through the full rollout. Do not flatten m into the initial-condition or time axis.
2: Preserve initial-condition/preparation replicas as a separate axis. In TFIM, regroup flat samples to (i, l) using metadata before reductions.
3: Estimate model uncertainty by variance over m at fixed initial-condition/preparation sample, then average over initial-condition/preparation samples.
4: Estimate initial-condition/preparation uncertainty by variance over initial-condition/preparation samples after averaging over m.
5: Report $\mathrm{Var}_{\mathrm{epi}}$, $\mathrm{Var}_{\mathrm{alea}}$, and α (Equation (4)) with the aggregation axes listed in the table or appendix.

D.2 Formal OOD definitions

For all experiments, in-distribution means that the system-level control variable lies inside the support used for training and model selection. OOD evaluation changes that control while leaving the evaluation horizon, mask rule, checkpoint-selection rule, and metric definitions unchanged.

Coupled Pendulum. The control is the mass ratio $r = m_2/m_1$. The in-distribution set is $\mathcal{R}_{\mathrm{ID}} = [0.5, 2.0]$. A rollout is OOD when $r \notin \mathcal{R}_{\mathrm{ID}}$, and its physics distance is

$$d^{\mathrm{DP}}_{\mathrm{phys}}(r) = \sqrt{C_{\mathrm{kin}} \left(1/r - 1/r^\star\right)^2 + C_{\mathrm{force}} \left(r - r^\star\right)^2},$$

where $r^\star$ is the point of $\mathcal{R}_{\mathrm{ID}}$ closest to r, and $C_{\mathrm{kin}} = 1.10$ and $C_{\mathrm{force}} = 0.327$ are the training-set means of p_2^2 and $\sin^2 q_2$. The two terms are the root-mean-square changes in the only r-dependent parts of the vector field, $\dot{q}_2 = p_2/r$ and the gravity force $r \sin q_2$ on the second pendulum. Therefore every in-distribution point has zero distance, while OOD points are measured by how much their dynamics differ from those at the nearest support boundary.

Spring Chain. The controls are the spring-stiffness parameter k and the connection-matrix family or topology stratum c. For each stratum, let $\mathcal{K}_{\mathrm{ID}}(c)$ denote the training-support interval for k. A rollout is OOD when either its topology stratum is held out or $k \notin \mathcal{K}_{\mathrm{ID}}(c)$. For the plotted stiffness sweep, the reported distance is

$$d^{\mathrm{SC}}_{\mathrm{phys}}(k) = s \min_{k^\star \in \mathcal{K}_{\mathrm{ID}}} |k - k^\star| \,,$$

with $\mathcal{K}_{\mathrm{ID}} = [1, 5]$ for every stratum and $s = 2.39$ the root-mean-square spring force per unit stiffness over the training states, so that $d^{\mathrm{SC}}_{\mathrm{phys}}$ is the expected change in spring force. This distance is a controlled one-dimensional stiffness proxy. It is physically meaningful for the stiffness sweep, but it is not a complete Hamiltonian distance because topology and mode structure can change even when a scalar stiffness distance is small. This is why the Spring Chain is treated as a harder structured-system stress test rather than as evidence for a universal monotone OOD-distance claim.

TFIM. The control is the transverse field $g = h/J$. The in-distribution set is the training control set $\mathcal{G}_{\mathrm{ID}}$ or its closed training interval when the grid is dense. A rollout is OOD when $g \notin \mathcal{G}_{\mathrm{ID}}$. For strictly positive field grids, the OOD distance is

$$d^{\mathrm{TFIM}}_{\mathrm{phys}}(g) = \min_{g^\star \in \mathcal{G}_{\mathrm{ID}}} |\log(g/g^\star)| \,.$$

If a TFIM sweep includes $g = 0$, the same slot is filled by the normalised absolute distance to the training interval instead of a log-ratio distance. In all cases g is a control parameter, not a dynamical coordinate.

E Training details and hyperparameters

Unless otherwise stated, methods compared within a system use the same data split, HNN backbone, optimiser family, and training budget. The method-specific differences are the uncertainty mechanism and rollout sampling convention. $T_{\mathrm{train}} < T_{\mathrm{eval}}$ because long training rollouts destabilise

Table 4: System-level training and rollout settings. Width and depth refer to the hidden network inside the HNN backbone. All rows use the leapfrog integrator.

System	Backbone and split	Training settings	Evaluation settings
Coupled Pendulum	`MassRatioSimpleHNN`, conditioned on mass ratio $r = m_2/m_1$. Training covers $r \in [0.5, 2.0]$ and OOD evaluates lower and higher held-out ratios.	AdamW, learning rate 10^{-3}, batch 128, 600 epochs, width 500, depth 1, `intsin2x` activation.	$T_{\text{train}} = 10$, $T_{\text{eval}} = 60$, $\Delta t = 0.1$. OOD uses 10 MC samples and IC-ensemble metrics use 50 MC samples.
Spring Chain	`SeparableHNN`, conditioned on connection matrix K. Training covers $k \in [1, 5]$ over ten topologies and OOD evaluates softer and stiffer chains.	AdamW, learning rate 10^{-3}, batch 256, 600 epochs, width 400, depth 1, `intsin2x` activation.	$T_{\text{train}} = 10$, $T_{\text{eval}} = 100$ for OOD metrics, 60 for IC-ensemble NLL, $\Delta t = 0.1$. Evaluation uses 50 MC samples.
TFIM	`TFIMSeparableHNN`, conditioned on operator $A(g)$ and control g. Training covers $g \in [1.3, 1.9]$ for $L = 4$, with held-out controls extending to both sides of this band.	AdamW, learning rate 10^{-3}, batch 128, 400 epochs, width 100, depth 1, `intsin2x` activation. Horizon curriculum from $T_{\text{train}} = 10$ to 15, with the full horizon in the last 100 epochs.	$T_{\text{train}} = 15$, $T_{\text{eval}} = 41$, $\Delta t = 0.05$. State error and drift use 10 MC samples. IC-ensemble metrics use 50 MC samples and 20 hidden state-preparation replicas.

HNN training. The classical arms use a fixed T_{train} and TFIM a short-to-long curriculum ending at $T_{\text{train}} = 15$. The `intsin2x` activation is $\sigma(x) = x/2 - \sin(2x)/4 + x^2/2 = \int_0^x (\sin^2 u + u)\, du$.

Classical data generation. The Coupled Pendulum has $H = p_1^2/2 + p_2^2/(2r) + (1 - \cos q_1) + r(1 - \cos q_2) + (q_1 - q_2)^2/2$ with $m_1 = 1$. Training uses the seven mass ratios $r \in \{0.5, 0.75, \ldots, 2.0\}$ and OOD evaluation $r \in \{0.1, 0.2, 0.3, 3, 4, 5\}$. Initial states are drawn as $q_i \sim \mathcal{U}(-\pi/2, \pi/2)$ and $p_i \sim \mathcal{U}(-0.8, 0.8)$. The Spring Chain has $N = 4$ unit masses and $H = \sum_i p_i^2/2 + \frac{1}{2}\sum_{i<j} K_{ij}(q_i - q_j)^2 + \frac{1}{2}k(q_1^2 + q_N^2)$. Each connection matrix is one stiffness $k \sim \mathcal{U}(1, 5)$ times a random symmetric adjacency with edge probability 0.7. Training uses 20 matrices with ten distinct adjacency patterns, and testing 10 newly drawn matrices. Initial states are drawn as $p_i, q_i \sim \mathcal{U}(-2.5, 2.5)$, with both end masses starting at rest at zero. Both systems are integrated with leapfrog at $\Delta t = 0.1$, and every stored frame, including the initial one, carries independent $\mathcal{N}(0, 0.1^2)$ measurement noise. The IC ensemble perturbs this observed initial state (Appendix G.1), and the $\mathcal{N}(0, 0.1^2)$ term enters only the likelihood, not the variance decomposition.

For the classical systems, IC-ensemble NLL and decomposition metrics use 20 Gaussian phase-space replicas per initial state, with the state-preparation noise scale inferred at evaluation as described in Appendix G. For TFIM, the nested decomposition uses 20 hidden state-preparation replicas per intended initial state, with the latent wavefunction convention described in Appendix G.2. All TFIM arms were trained for the full 400-epoch budget (early stopping, with a 300-epoch minimum, never triggered). The TFIM Variational row uses KL weight 10^{-4}. Controlled stochastic runs record mean per-seed runtimes of approximately 1.5–2.5 hours for Coupled Pendulum, 5–7 hours for Spring Chain and 2.1–2.5 hours for TFIM, depending on the method. Jobs were run on shared NVIDIA RTX A6000 GPUs, so these are upper bounds. Runtimes of the classical Deep Ensemble members were not recorded consistently, so we do not report them.

Table 5: Method-level stochastic settings. All single-model rows use seeds 42–51 on every system, while Deep Ensemble rows use five independently trained members.

Method	Stochastic convention	Count	Seeds or members
Fixed-mask MC	Dropout probability $p = 0.1$, with one finite-bank mask index fixed for each rollout.	Mask bank $M = 200$.	Seeds 42–51.
Fresh-mask MC	Dropout probability $p = 0.1$, fresh-mask MC dropout baseline.	Same MC sample counts as Fixed-mask MC within each system.	Seeds 42–51.
Concrete	Concrete dropout baseline with stochastic single-model samples. Initialised with p=0.1.	Same MC sample counts as Fixed-mask MC within each system.	Seeds 42–51.
Variational	Variational dropout baseline with multiplicative weight-noise samples. Initialised with p=0.1.	Same MC sample counts as Fixed-mask MC within each system.	Seeds 42–51.
Deep Ensemble	Independent deterministic HNN members. The member index supplies the model-sample axis.	Five members.	Members $0, \ldots, 4$.
Deterministic HNN	No stochastic mask or weight noise.	One model.	Seeds 42–51.

F TFIM protocol details

The closed-system TFIM experiment uses the one-dimensional Hamiltonian

$$H(g) = -J \sum_i \sigma_i^z \sigma_{i+1}^z - gJ \sum_i \sigma_i^x,$$

with $J = 1$ and $L = 4$ throughout. The dataset is generated by exact dense unitary evolution. For each control value g, the dense Hamiltonian is built and diagonalised, the initial-state bank is evolved on the selected time grid, and canonical latent trajectories are stored for training.

State and split structure. Trajectories $z(t) \in \mathbb{R}^{32}$ are stored with metadata for intended ICs, latent replicas, g values, and the time grid, in the grouped layout (k, i, l, t, d), where k indexes g, i indexes intended initial conditions, l indexes latent replicas or bootstrap reconstructions, t indexes time, and d is the latent dimension.

Uncertainty scenarios. Clean exact data contains no initial-condition perturbation. The initial-state preparation setting perturbs only the $t = 0$ latent state and exact-evolves each perturbed state. The shadow-bootstrap reconstruction setting samples random local-Pauli measurements at $t = 0$, bootstrap-resamples the shot record, reconstructs plausible pure-state representatives, fixes gauge deterministically, and exact-evolves each reconstruction.

G Gaussian IC-ensemble NLL reporting convention

The main tables use a single likelihood convention, Gaussian NLL computed from the initial-condition (IC) ensemble by moment matching the predictive mean and variance. This choice is deliberately narrow. The paper's uncertainty decomposition already crosses a mask axis with an IC axis, so the likelihood score should be computed from the same object rather than from a separately fitted calibration family.

Heavier-tailed and mixture predictive families answer a different question, how well a post-hoc marginal calibrator can fit the empirical predictive samples. That comparison can be useful for model diagnosis, but it injects calibrator-choice noise into the result ranking. In particular, a method

Table 6: Effect of using the IC ensemble for in-distribution Gaussian NLL on the Coupled Pendulum. The dropout-only row evaluates predictive variance from mask samples alone. The IC-ensemble rows use the validation-inferred IC-noise scale and moment-match the crossed IC-by-mask predictive samples used for the paper's decomposition. The reduction column is the decrease from the Fresh-mask MC dropout-only NLL. Larger reductions indicate a larger improvement.

Method	NLL convention	ID G-NLL	Reduction vs dropout-only	$\hat{\sigma}$
Fresh-mask MC (dropout-only)	dropout-only	104.82 $\pm$4.6	n/a	n/a
Fresh-mask MC (IC ensemble)	IC ensemble	0.274 $\pm$0.012	104.55 $\pm$4.6	0.331 $\pm$0.0050
Fixed-mask MC (IC ensemble)	IC ensemble	0.224 $\pm$0.011	104.60 $\pm$4.6	0.250 $\pm$0.0029

can look better or worse depending on whether the diagnostic rewards Gaussian bulk fit, tail fit, or mixture flexibility.

For this reason, Table 1 and Table 11 report only IC-ensemble Gaussian NLL, both with the validation-inferred IC-noise scale described below and with the true scale. Coverage error under both conventions is retained as a separate interval diagnostic (Tables 7 and 8), and the decomposition tables report $\mathrm{Var}_{\mathrm{epi}}$, $\mathrm{Var}_{\mathrm{alea}}$, and α (Equation (4)) directly. This keeps the main evidence aligned with the modelling claim that fixed masks make the mask and IC axes separately observable during a Hamiltonian rollout.

Table 6 quantifies why the IC ensemble is the appropriate in-distribution likelihood object for this paper. On the Coupled Pendulum, Fresh-mask MC's dropout-only Gaussian NLL is very large because mask variation alone does not represent the observation and state-preparation spread used by the decomposition. Moment matching the crossed IC ensemble reduces Fresh-mask MC's ID Gaussian NLL from 104.82 to 0.274. Fixed-mask MC under the same IC-ensemble convention gives the lowest ID Gaussian NLL.

G.1 IC-ensemble construction and noise-scale inference

All reported IC-ensemble NLLs use the same outer evaluation structure for the Coupled Pendulum, Spring Chain, and TFIM experiments. For each evaluation case, we construct an $N \times M$ predictive tensor. The N axis contains perturbed initial conditions and represents state-preparation sensitivity. The M axis contains stochastic model samples and represents model-sample uncertainty. The model-sample axis is the mask index for Fixed-mask MC, the dropout sample index for the other stochastic single-model baselines, and the ensemble-member index for Deep Ensemble where that reference is reported.

Classical IC replicas. For the two classical systems, the state is a phase-space vector $z = (p, q) \in \mathbb{R}^{2d}$. The IC ensemble is formed by perturbing this observed initial state,

$$z_0^{(i)} = z_0 + \epsilon_i, \qquad \epsilon_i \sim \mathcal{N}(0, \sigma_{\mathrm{IC}}^2 I), \qquad i = 1, \ldots, N.$$

This models repeated noisy measurements or state estimates of the same underlying classical state. In the reported tables, the scalar perturbation scale is inferred from validation data as described below.

Inferring the IC-noise scale. The perturbation scale σ_{IC} is inferred on validation data rather than fixed by hand. For each evaluated classical run, a bounded one-dimensional search chooses the value of σ_{IC} that minimises the IC-ensemble Gaussian NLL on the validation split. The selected scale is then held fixed for the reported in-distribution and OOD evaluations.

Two conventions. A fitted scale can absorb model error. If the mean rollout is biased, a larger $\hat{\sigma}$ widens the predictive distribution and lowers the NLL, at the price of attributing model error to the initial state. We therefore also report every likelihood and coverage with σ fixed at the data-generating value 0.1, adding the $\mathcal{N}(0, 0.1^2)$ per-frame measurement noise to the predictive samples (Table 7). At the true scale Fixed-mask MC has a significantly lower G-NLL and Cov90 error than each other stochastic arm in every column, whereas under the fitted scale Fresh-mask MC has a significantly lower Spring Chain Cov90 error and OOD G-NLL than Fixed-mask MC (Table 20).

Table 7: Likelihood and coverage under the two IC-noise conventions. $\hat{\sigma}$ denotes IC noise fitted per run on validation data, and σ=0.1 denotes IC noise fixed at the data's true value, with independent $\mathcal{N}(0, 0.1^2)$ measurement noise added to the predictive samples. Cov90 err is the absolute deviation of the empirical 90% coverage from 0.9. The NLL masks-only columns give the NLL of the mask ensemble alone plus measurement noise (no IC-noise term).

		G-NLL ($\hat{\sigma}$)		G-NLL (σ=0.1)	
Method	$\hat{\sigma}$	ID	OOD	ID	OOD
Coupled pendulum					
Fresh-mask MC	0.331 ±0.0050	0.274 ±0.012	3.298 ±0.17	1.064 ±0.028	14.45 ±0.88
Concrete	0.294 ±0.0041	0.267 ±0.0091	3.837 ±0.13	1.091 ±0.026	13.35 ±0.36
Variational	0.351 ±0.026	0.517 ±0.083	2.871 ±0.32	1.340 ±0.21	7.770 ±1.5
Fixed-mask MC	0.250 ±0.0029	**0.224** ±0.011	2.774 ±0.17	**0.477** ±0.039	**4.739** ±0.39
Deterministic	0.290 ±0.0030	0.273 ±0.0062	4.255 ±0.30	1.332 ±0.043	20.08 ±1.5
Deep ens. ×5	0.291 (n=1)	0.261 (n=1)	2.948 (n=1)	1.083 (n=1)	8.106 (n=1)
Spring chain					
Fresh-mask MC	0.483 ±0.0083	0.980 ±0.055	**4.779** ±0.19	5.221 ±0.55	33.84 ±0.77
Concrete	0.252 ±0.018	1.729 ±0.12	15.38 ±2.3	5.013 ±0.33	40.45 ±2.1
Variational	0.310 ±0.058	1.776 ±0.28	7.711 ±1.3	3.878 ±0.51	17.42 ±2.5
Fixed-mask MC	0.130 ±0.0022	0.989 ±0.14	11.78 ±1.1	**0.974** ±0.13	**11.55** ±1.1
Deterministic	0.253 ±0.0069	1.370 ±0.051	18.64 ±1.1	5.055 ±0.10	66.14 ±3.1
Deep ens. ×5	0.216 (n=1)	1.576 (n=1)	17.24 (n=1)	3.812 (n=1)	35.62 (n=1)

Seeds 42–51 (n=10), mean ± sd over seeds. Deep ens. ×5 is a single 5-member ensemble (n=1). n/a = not defined for that model (e.g. no mask ensemble). Bold indicates the best stochastic single-model arm in a column when its paired tests in Tables 20 and 21, all of which involve Fixed-mask MC, are significant (Holm-adjusted $p < 0.05$). Paired effect sizes with 95% confidence intervals are in Table 20 (classical) and Table 21 (TFIM). Deterministic and deep-ensemble rows (italic, below the dotted rule) are references and are not compared for bold.

Table 7: Likelihood and coverage under the two IC-noise conventions (continued).

	Cov90 err ($\hat{\sigma}$)		Cov90 err (σ=0.1)		NLL masks-only	
Method	ID	OOD	ID	OOD	ID	OOD
Coupled pendulum						
Fresh-mask MC	0.0119 ±0.0030	0.321 ±0.011	0.235 ±0.0022	0.568 ±0.0092	4.038 ±0.11	25.82 ±1.7
Concrete	0.0152 ±0.0032	0.352 ±0.0082	0.239 ±0.0031	0.559 ±0.0052	3.209 ±0.061	27.04 ±1.0
Variational	0.0246 ±0.0098	0.271 ±0.030	0.287 ±0.023	0.477 ±0.037	2.224 ±0.31	10.23 ±2.4
Fixed-mask MC	0.0131 ±0.0057	0.300 ±0.023	**0.157** ±0.0063	**0.407** ±0.029	1.001 ±0.063	6.208 ±0.60
Deterministic	0.0431 ±0.0025	0.402 ±0.018	0.305 ±0.0055	0.620 ±0.0071	n/a	n/a
Deep ens. ×5	0.0214 (n=1)	0.284 (n=1)	0.240 (n=1)	0.438 (n=1)	6.938 (n=1)	21.42 (n=1)
Spring chain						
Fresh-mask MC	**0.0103** ±0.0072	**0.352** ±0.0097	0.227 ±0.018	0.634 ±0.0066	15.33 ±0.88	54.22 ±1.2
Concrete	0.0854 ±0.0096	0.539 ±0.019	0.235 ±0.014	0.654 ±0.0037	12.83 ±0.72	85.30 ±5.8
Variational	0.125 ±0.031	0.448 ±0.027	0.286 ±0.020	0.570 ±0.026	5.529 ±0.99	24.06 ±4.3
Fixed-mask MC	0.0748 ±0.0091	0.527 ±0.016	**0.104** ±0.0054	**0.520** ±0.016	1.395 ±0.20	16.28 ±1.8
Deterministic	0.0821 ±0.0030	0.587 ±0.0076	0.278 ±0.0066	0.709 ±0.0065	n/a	n/a
Deep ens. ×5	0.0844 (n=1)	0.574 (n=1)	0.199 (n=1)	0.651 (n=1)	21.25 (n=1)	129.43 (n=1)

Gaussian NLL and coverage. For likelihood reporting, the $N \times M$ rollouts are flattened into one predictive sample set at each time and state coordinate. We moment-match a Gaussian predictive distribution from that sample set and evaluate the held-out target under this Gaussian. The reported IC-ensemble Gaussian NLL averages this score over evaluation cases, time, and state dimensions. The Cov90 diagnostic uses the same crossed ensemble but reports empirical central 90% coverage from the sample quantiles. The coverage-error columns in Tables 7 and 8 are $|\mathrm{Cov90} - 0.90|$.

Relation to the variance decomposition. The same crossed tensor also defines the axis-preserving decomposition. At each time point, the implementation computes mask variation at fixed IC and IC variation after mask averaging, as in Equation (4). The scalar $\mathrm{Var}_{\mathrm{epi}}$, $\mathrm{Var}_{\mathrm{alea}}$, and $\mathrm{Var}_{\mathrm{total}}$ values reported in tables and figures are then averaged over the relevant evaluation cases, times, and state dimensions. Inferring σ_{IC} changes the scale of the IC replicas but does not alter the mask-sampling protocol or the definition of α (Equation (4)).

G.2 TFIM IC-ensemble convention

The TFIM experiment uses the same $N \times M$ reduction as the classical systems, but the aleatoric axis has a different physical meaning. The model state is the real latent vector $z = \mathrm{concat}(q, p)$ representing the complex wavefunction $\psi = q + ip$, and the transverse-field value $g = h/J$ is treated as an external control rather than as a noisy dynamical coordinate.

Latent state-preparation replicas. For an intended normalised quantum initial state $|\psi\rangle \in \mathbb{C}^{2^L}$, TFIM replicas are generated in Hilbert space rather than by adding unconstrained Euclidean noise to the real latent coordinates. We draw an isotropic circular complex Gaussian direction

$$\eta \sim \mathcal{CN}(0, I_{2^L})$$

and remove the component parallel to the intended state,

$$\eta_\perp = \eta - \langle\psi|\eta\rangle|\psi\rangle.$$

For $\eta_\perp \neq 0$, which holds almost surely under the continuous Gaussian draw, we define

$$t = \frac{\eta_\perp}{\|\eta_\perp\|}.$$

Thus $\langle\psi|t\rangle = 0$ and $\|t\| = 1$. This projection samples a random direction in the complex-orthogonal, phase-horizontal subspace at $|\psi\rangle$. It removes both the radial component and the physically irrelevant global-phase direction, rather than perturbing along the full real tangent space of the unit sphere.

The latent IC replica is then

$$|\psi^{(i)}\rangle = \frac{|\psi\rangle + \epsilon t_i}{\||\psi\rangle + \epsilon t_i\|}.$$

Because $t_i \perp \psi$ and $\|t_i\| = 1$, the normalisation factor is

$$\||\psi\rangle + \epsilon t_i\| = \sqrt{1 + \epsilon^2}.$$

Equivalently,

$$|\psi^{(i)}\rangle = \frac{|\psi\rangle + \epsilon t_i}{\sqrt{1 + \epsilon^2}}.$$

This construction keeps the perturbed state on the unit-norm pure-state manifold. It also gives ϵ a direct geometric interpretation. The fidelity to the intended state is

$$|\langle\psi|\psi^{(i)}\rangle|^2 = \frac{1}{1 + \epsilon^2},$$

so the corresponding Fubini–Study angle is

$$\theta_{\mathrm{FS}} = \arctan(\epsilon).$$

The ensemble therefore models independent imperfect preparations, or independent latent reconstructions, of the same intended quantum state at fixed control value.

Table 8: TFIM (operator-structured HNN backbone) with a nested IC × mask ensemble scored with the state-preparation noise ϵ fitted per run on a validation grid ($\hat{\epsilon}$) and with ϵ fixed at the data's true 0.10. Columns are G-NLL, Cov90 error and the mask-ensemble variance fraction α, in and out of distribution. TFIM has no per-step observation noise, so no measurement-noise term is added. $\hat{\epsilon}$ is selected from a discrete grid with step 0.05.

		G-NLL		Cov90 err		α	
Method	ϵ	ID	OOD	ID	OOD	ID	OOD
Fitted scale $\hat{\epsilon}$							
Fresh-mask MC	0.140 ± 0.052	**−2.276** ± 0.13	2.720 ± 2.5	0.0829 ± 0.012	0.222 ± 0.084	0.551 ± 0.15	0.527 ± 0.16
Concrete	0.265 ± 0.058	−1.939 ± 0.13	0.584 ± 1.9	0.0859 ± 0.014	0.148 ± 0.074	0.211 ± 0.050	0.201 ± 0.055
Variational	0.595 ± 0.016	−0.611 ± 0.026	−0.287 ± 0.040	**0.0164** ± 0.0076	0.0925 ± 0.012	0.599 ± 0.013	0.567 ± 0.011
Fixed-mask MC	0.585 ± 0.047	−0.994 ± 0.047	−0.937 ± 0.048	0.0985 ± 6.8e-4	0.0897 ± 0.0064	0.757 ± 0.039	0.768 ± 0.034
Deterministic	0.150 ± 0	−2.456 ± 0.021	2.395 ± 5.5	0.0282 ± 0.0068	0.197 ± 0.091	0 ± 0	0 ± 0
True scale ϵ=0.10							
Fresh-mask MC	0.10	**−2.363** ± 0.023	3.343 ± 2.5	0.0696 ± 0.0042	0.260 ± 0.041	0.677 ± 0.011	0.657 ± 0.013
Concrete	0.10	−2.211 ± 0.084	5.339 ± 5.0	**0.0388** ± 0.015	0.327 ± 0.049	0.624 ± 0.049	0.607 ± 0.057
Variational	0.10	−0.618 ± 0.065	0.416 ± 0.098	0.115 ± 0.020	0.258 ± 0.013	0.977 ± 0.0015	0.973 ± 0.0014
Fixed-mask MC	0.10	−1.350 ± 0.055	**−1.250** ± 0.11	0.0858 ± 0.0026	**0.0778** ± 0.011	0.988 ± 0.0020	0.988 ± 0.0018
Deterministic	0.10	−2.495 ± 0.046	8.327 ± 12	0.102 ± 0.013	0.321 ± 0.090	0 ± 0	0 ± 0

Thus $\mathrm{Var}_{\mathrm{alea}}$ in TFIM should be read as sensitivity to latent state-preparation or reconstruction uncertainty at fixed g, not as process noise in the Schrödinger dynamics and not as uncertainty in the control field. The mask axis, Gaussian NLL convention, and definition of α (Equation (4)) are otherwise unchanged. The TFIM perturbation strength ϵ is selected before held-out reporting by minimising validation NLL over a discrete grid of candidate perturbation strengths. Table 8 also reports every metric at the true $\epsilon = 0.10$. This differs from the continuous classical search, but it serves the same role of choosing the IC-noise scale before reporting held-out NLL, coverage, and decomposition statistics.

Scope. This convention calibrates only the perturbation scale used to construct the IC ensemble. It does not alter the learned model, the mask-sampling protocol, or the evaluation targets.

H Additional experimental checks

This appendix collects supporting experimental details that are useful for reproducibility and interpretation, but too specific for the main paper narrative.

Map of results. The detailed Coupled Pendulum OOD table is in Appendix H.1. The Spring Chain OOD figure is in Appendix H.3. The all-method Coupled Pendulum decomposition figure is in Appendix H.2. The all-method TFIM decomposition figure and quantitative table are in Appendix H.4. The mask-bank size study is in Appendix H.6 and the cost profile in Appendix H.8. The full epistemic/aleatoric decomposition table is in Appendix I, the expanded drift table and the long-horizon physical energy drift are in Appendix J, and likelihoods under both noise conventions are in Appendix G. An exploratory hidden-parameter shift-detection benchmark is in Appendix H.9. A Poisson-bracket symmetry diagnostic on TFIM is in Appendix H.5.

Table 9: Full version of Table 1 with mean ± sd over seeds.

	MSE			G-NLL ($\hat{\sigma}$)		G-NLL (σ=0.1)			
Method	ID	OOD	$\hat{\sigma}$	ID	OOD	ID	OOD	OOD α	H-drift
Coupled pendulum									
Fresh-mask MC	0.117 ± 0.0058	1.097 ± 0.068	0.331 ± 0.0050	0.274 ± 0.012	3.298 ± 0.17	1.064 ± 0.028	14.45 ± 0.88	0.0326 ± 0.0033	0.157 ± 0.0076
Concrete	0.115 ± 0.0035	1.026 ± 0.031	0.294 ± 0.0041	0.267 ± 0.0091	3.837 ± 0.13	1.091 ± 0.026	13.35 ± 0.36	0.0448 ± 0.0031	0.124 ± 0.0043
Variational	0.209 ± 0.032	1.111 ± 0.046	0.351 ± 0.026	0.517 ± 0.083	2.871 ± 0.32	1.340 ± 0.21	7.770 ± 1.5	0.213 ± 0.051	0.530 ± 0.12
Fixed-mask MC	0.113 ± 0.0046	1.032 ± 0.045	0.250 ± 0.0029	**0.224** ± 0.011	2.774 ± 0.17	**0.477** ± 0.039	**4.739** ± 0.39	0.524 ± 0.066	**0.00432** ± 5.3e-5
Deterministic	0.119 ± 0.0026	1.189 ± 0.11	0.290 ± 0.0030	0.273 ± 0.0062	4.255 ± 0.30	1.332 ± 0.043	20.08 ± 1.5	n/a	0.00428 ± 4.6e-5
Deep ens. ×5	0.115 (n=1)	1.036 (n=1)	0.291 (n=1)	0.261 (n=1)	2.948 (n=1)	1.083 (n=1)	8.106 (n=1)	0.461 (n=1)	0.00431 (n=1)
Spring chain									
Fresh-mask MC	0.654 ± 0.051	4.424 ± 0.049	0.483 ± 0.0083	0.980 ± 0.055	**4.779** ± 0.19	5.221 ± 0.55	33.84 ± 0.77	0.0464 ± 0.0022	1.751 ± 0.11
Concrete	0.667 ± 0.037	4.370 ± 0.062	0.252 ± 0.018	1.729 ± 0.12	15.38 ± 2.3	5.013 ± 0.33	40.45 ± 2.1	0.0661 ± 0.011	1.732 ± 0.21
Variational	1.721 ± 0.11	4.430 ± 0.063	0.310 ± 0.058	1.776 ± 0.28	7.711 ± 1.3	3.878 ± 0.51	17.42 ± 2.5	0.426 ± 0.13	4.111 ± 0.58
Fixed-mask MC	**0.613** ± 0.026	**4.010** ± 0.053	0.130 ± 0.0022	0.989 ± 0.14	11.78 ± 1.1	**0.974** ± 0.13	**11.55** ± 1.1	0.833 ± 0.028	**0.108** ± 0.0014
Deterministic	0.529 ± 0.0062	4.441 ± 0.10	0.253 ± 0.0069	1.370 ± 0.051	18.64 ± 1.1	5.055 ± 0.10	66.14 ± 3.1	n/a	0.106 ± 3.3e-4
Deep ens. ×5	0.538 (n=1)	4.105 (n=1)	0.216 (n=1)	1.576 (n=1)	17.24 (n=1)	3.812 (n=1)	35.62 (n=1)	0.271 (n=1)	0.106 (n=1)

H.1 Detailed Coupled Pendulum OOD table

Table 10 gives the detailed Coupled Pendulum OOD diagnostics that support the main-body combined OOD figure.

H.2 All-method Coupled Pendulum OOD decomposition

Figure 4 repeats the main Coupled Pendulum OOD decomposition with the additional stochastic dropout families.

H.3 Spring Chain OOD diagnostic

Figure 5 records the Spring Chain OOD diagnostic. This plot supports the intermediate-complexity claim. Epistemic variance changes across OOD stiffness values and increases with pooled distance from the training support. It is less central than the Coupled Pendulum because a single stiffness-distance proxy does not capture all topology-dependent interaction changes.

H.4 TFIM quantitative summary

Table 11 gives the TFIM quantitative summary. Fixed-mask MC is not the universal winner on point prediction, norm drift, energy drift, or Gaussian NLL in this more complex latent-dynamics setting.

H.5 Poisson-bracket symmetry diagnostic on TFIM

The TFIM Hamiltonian conserves $G(q,p) = \sum_i (q_i^2 + p_i^2)$, so the Poisson bracket $\{G, H\}$ vanishes for the exact H. For a learned H we report the normalised residual $|\{G,H\}|/(\|\nabla G\|\,\|\nabla H\|)$ of the operator-structured backbone, with gradients by automatic differentiation. It is evaluated at all

Table 10: Coupled-pendulum OOD stress test with mean epistemic variance in and out of distribution, OOD MSE, the Spearman correlation between MC predictive variance and the log distance of the OOD mass ratio from the training range (six ratios, hence coarse), and the sampled-Hamiltonian drift averaged over the rollout and at the final frame (Table 18). $\mathrm{Var}_{\mathrm{epi}}$ is computed from 10 mask samples without IC replicas, so it differs from Table 16. Variance columns have no preferred direction. The drift columns are self-consistency measures of the model's own Hamiltonian.

Method	$\mathrm{Var}_{\mathrm{epi}}$ ID	$\mathrm{Var}_{\mathrm{epi}}$ OOD	OOD MSE	OOD Spearman ρ	Drift mean	Drift final
Fresh-mask MC	8.6e-4 ± 4.1e-5	0.00564 ± 7.4e-4	1.097 ± 0.068	0.909 ± 0.030	0.157 ± 0.0076	0.179 ± 0.0084
Concrete	7.9e-4 ± 7.7e-5	0.00598 ± 0.0011	1.026 ± 0.031	0.851 ± 0.11	0.124 ± 0.0043	0.147 ± 0.0060
Variational	0.0133 ± 0.0046	0.0339 ± 0.011	1.111 ± 0.046	0.851 ± 0.16	0.530 ± 0.12	0.645 ± 0.16
Fixed-mask MC	0.0176 ± 0.0053	0.121 ± 0.043	1.032 ± 0.045	0.760 ± 0.096	**0.00432** ± 5.3e-5	0.00423 ± 3.0e-4
Deterministic	0 ± 0	0 ± 0	1.189 ± 0.11	n/a	0.00428 ± 4.6e-5	0.00354 ± 2.4e-4
Deep ens. ×5	0.00148 (n=1)	0.100 (n=1)	1.036 (n=1)	0.714 (n=1)	0.00431 (n=1)	0.00359 (n=1)

Seeds 42–51 (n=10), mean ± sd over seeds. Deep ens. ×5 is a single 5-member ensemble (n=1). n/a = not defined for that model (e.g. no mask ensemble). Bold indicates the best stochastic single-model arm in a column when its paired tests in Tables 20 and 21, all of which involve Fixed-mask MC, are significant (Holm-adjusted $p < 0.05$). Paired effect sizes with 95% confidence intervals are in Table 20 (classical) and Table 21 (TFIM). Deterministic and deep-ensemble rows (italic, below the dotted rule) are references and are not compared for bold.

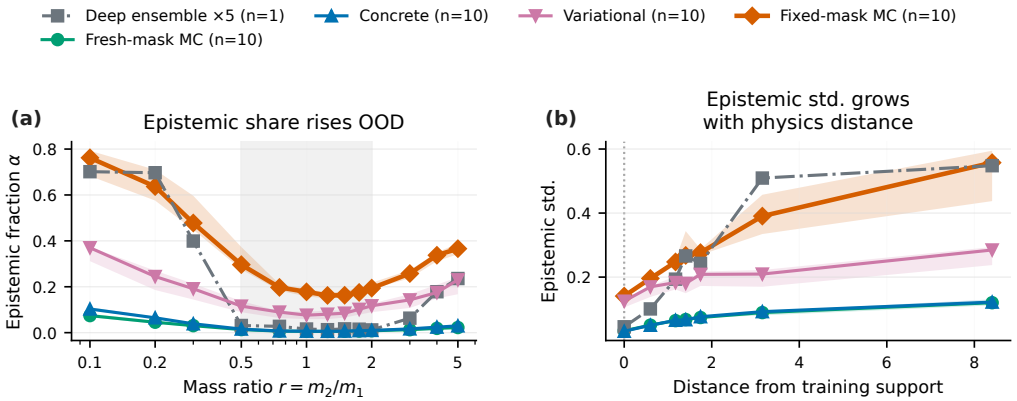


Figure 4: All-method Coupled Pendulum OOD decomposition. The left panel shows the epistemic fraction α (Equation (4)) across the mass-ratio sweep, and the right panel shows epistemic standard deviation versus distance from the training support. Deep Ensemble is retained as a compute-heavy reference. Concrete and Variational dropout are included here as additional single-model stochastic baselines.

23 test controls g, five time indices and the first 10 intended initial conditions (1150 states, identical for every model), for seeds 42–51. Fixed-mask MC uses 10 masks of its bank, Fresh-mask MC 10 fresh draws and the deterministic HNN a single pass. Two Fixed-mask banks are compared. The bank used throughout the paper redraws its mask index at every leapfrog sub-step during training. The second bank is trained with one index per rollout and is otherwise identical. At evaluation both hold one index for the whole rollout, so Corollary 1 applies to both and their contrast isolates the training-time part of the effect. Per seed we average the log residual over draws and states, and Table 12 reports geometric means. The exact Hamiltonian reaches 6×10^{-8} on rotated states, close

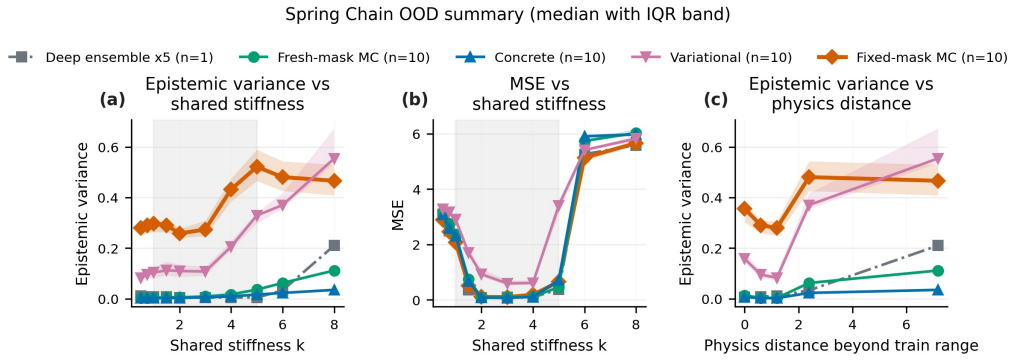


Figure 5: Spring-chain OOD diagnostic. Epistemic variance and MSE versus shared stiffness k, and epistemic variance versus physics distance. Shaded bands mark the training support. The positive distance response supports the structured-system stress test, while the remaining topology-dependent scatter shows that stiffness alone is an incomplete OOD coordinate.

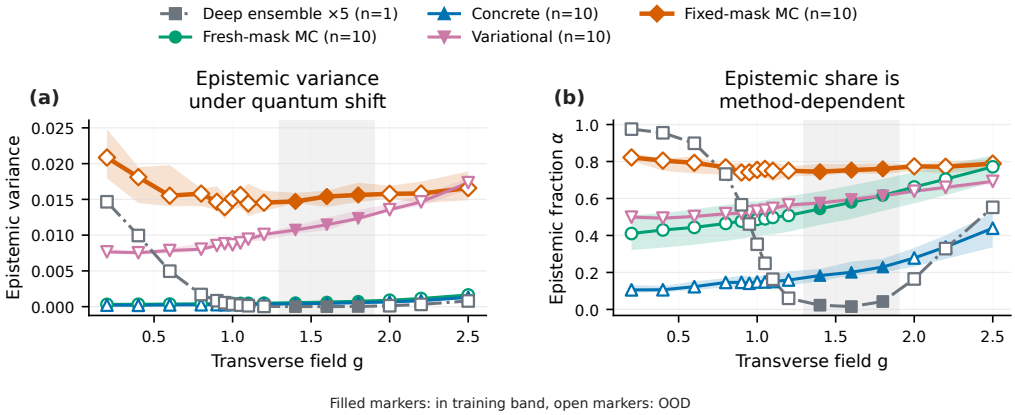


Figure 6: All-method TFIM decomposition diagnostic. The figure repeats the main TFIM decomposition plot with Concrete and Variational dropout included alongside Fixed-mask MC, Fresh-mask MC, and the Deep Ensemble reference. The TFIM setting remains a method-dependent stress test rather than a clean dominance result.

to float32 machine precision ($\epsilon \approx 1.2 \times 10^{-7}$), so the residuals of the learned models are not a numerical floor.

The deterministic HNN has the smallest residual, below every stochastic arm in all ten seeds. The dropout mask does not act on G, so a difference between stochastic arms can arise only indirectly through training. Holding the index per rollout during training lowers the residual by 22% relative to per-sub-step training (eight of ten seeds) and by 24% relative to Fresh-mask MC, while the per-sub-step bank sits 3% below Fresh-mask MC, not significantly. This is consistent with per-sub-step training producing a more diverse bank whose sampled Hamiltonians lie further from the symmetric mean, and with per-rollout training keeping each sampled Hamiltonian closer to it.

H.6 Mask-bank size on the Spring Chain

Table 13 varies the mask-bank size M for Fixed-mask MC on the Spring Chain (three seeds per M, descriptive). The fitted IC-noise scale falls from $\hat{\sigma} = 0.277$ at $M = 1$ to 0.131 at $M = 50$ and stays at 0.127–0.129 for $M = 100$–200. The OOD G-NLL at $\sigma = 0.1$ falls from 63.73 to 12.33 and then plateaus (11.74–11.77). OOD MSE changes little beyond $M = 20$ (4.126 to 4.006). A bank of $M = 200$, used throughout, is therefore past the point where the decomposition and likelihood change. On the Coupled Pendulum (ten seeds), Fixed-mask MC evaluated with M of its masks has

Table 11: Transverse-field Ising model (operator-structured HNN backbone with energy built from learned local operator terms, L=4, g-scan, M=200). State RMSE of the predicted wavefunction in and out of the training band of g. The IC $\times$ mask ensemble is scored with the state-preparation noise ϵ fitted per run on a validation grid ($\hat{\epsilon}$, a grid value) and with ϵ fixed at the true 0.10. Norm and energy drift are measured on the predicted wavefunction. α is the mask-ensemble fraction of predictive variance.

	State RMSE			G-NLL ($\hat{\epsilon}$)	
Method	ID	OOD	$\hat{\epsilon}$	ID	OOD
Fresh-mask MC	**0.0202** ±9.9e-4	0.0694 ±0.013	0.140 ±0.052	**−2.276** ±0.13	2.720 ±2.5
Concrete	0.0264 ±0.0021	0.0769 ±0.021	0.265 ±0.058	−1.939 ±0.13	0.584 ±1.9
Variational	0.147 ±0.0039	0.173 ±0.0014	0.595 ±0.016	−0.611 ±0.026	−0.287 ±0.040
Fixed-mask MC	0.0575 ±0.0065	0.0768 ±0.013	0.585 ±0.047	−0.994 ±0.047	−0.937 ±0.048
Deterministic	0.0189 ±5.7e-4	0.0487 ±0.023	0.150 ±0	−2.456 ±0.021	2.395 ±5.5

	G-NLL (ϵ=0.10)				
Method	ID	OOD	Norm drift	Energy drift	OOD α
Fresh-mask MC	**−2.363** ±0.023	3.343 ±2.5	0.0371 ±0.0014	**0.0994** ±0.0019	0.527 ±0.16
Concrete	−2.211 ±0.084	5.339 ±5.0	**0.0328** ±0.0057	0.107 ±0.017	0.201 ±0.055
Variational	−0.618 ±0.065	0.416 ±0.098	0.291 ±0.051	0.726 ±0.10	0.567 ±0.011
Fixed-mask MC	−1.350 ±0.055	**−1.250** ±0.11	0.146 ±0.041	0.299 ±0.086	0.768 ±0.034
Deterministic	−2.495 ±0.046	8.327 ±12	0.0123 ±0.0048	0.0318 ±0.014	0 ±0

Table 12: Poisson-bracket residual on TFIM ($L = 4$, operator-structured backbone, seeds 42–51). Top: geometric mean over seeds of the normalised residual, with the range over seeds. Bottom: seed-paired differences of the log residual (first minus second) with 95% t intervals and two-sided paired t-test p-values, unadjusted. Per-sub-step is the Fixed-mask MC bank used throughout the paper, per-rollout the bank trained with one mask index per rollout.

Method	Geometric mean	Range over seeds
Deterministic	0.0036	0.0021–0.0060
Fresh-mask MC	0.0199	0.0188–0.0207
Fixed-mask MC, per-sub-step training	0.0193	0.0140–0.0291
Fixed-mask MC, per-rollout training	0.0151	0.0107–0.0214

Comparison	Mean log difference	95% CI	p
Per-rollout − Per-sub-step	−0.243	$[-0.424, -0.062]$	0.014
Per-rollout − Fresh	−0.273	$[-0.403, -0.143]$	0.001
Per-sub-step − Fresh	−0.030	$[-0.171, +0.112]$	0.648
Per-rollout − Deterministic	+1.428	$[+1.098, +1.758]$	4×10^{-6}
Per-sub-step − Deterministic	+1.671	$[+1.363, +1.979]$	6×10^{-7}
Fresh − Deterministic	+1.701	$[+1.451, +1.951]$	9×10^{-8}

a median OOD MSE 4.4%, 0.8% and 0.7% above the $M = 50$ value at $M = 5$, 10 and 20, against at most 0.12% for Fresh-mask MC. Both are converged well below the $M = 200$ used throughout.

H.7 Per-mask and mean-trajectory energy views

A methodological caveat accompanies Corollary 1. The corollary gives a trajectory interpretation and symplecticity conditional on a fixed mask. It does not guarantee that the mean of mask trajectories conserves energy, because averaging trajectories that each live on a different energy shell is not itself a symplectic operation. In the 1000-step Spring Chain rollouts of Appendix J, the true-energy drift of the Fixed-mask MC ensemble-mean trajectory reaches 0.89 at $t = 1000$ in distribution, whereas its individual rollouts drift by 0.0552. We therefore evaluate energy only per sampled rollout, never on the ensemble-mean trajectory, and discuss the long-horizon results in Appendix J.

Table 13: Effect of the mask-bank size M on the Spring Chain, scored as in Table 7 (IC ensemble with $\hat{\sigma}$ fitted and with σ=0.1, masks-only NLL includes measurement noise, $n_{\mathrm{mc}} = \min(M, 50)$). Descriptive only, with three seeds (42–44) per M and no significance test. M=1 has a single mask, so mask-ensemble quantities are n/a.

	MSE			G-NLL ($\hat{\sigma}$)		G-NLL (σ=0.1)	
Mask bank	ID	OOD	$\hat{\sigma}$	ID	OOD	ID	OOD
M=1	0.652 ±0.036	4.369 ±0.068	0.277 ±0.015	1.450 ±0.10	14.99 ±1.7	6.184 ±0.50	63.73 ±0.98
M=5	0.589 ±0.026	4.181 ±0.088	0.158 ±0.0059	1.513 ±0.36	18.55 ±3.5	1.887 ±0.57	23.04 ±5.2
M=20	0.582 ±0.016	4.126 ±0.076	0.139 ±0.0062	1.006 ±0.18	13.17 ±0.34	1.034 ±0.17	13.77 ±0.28
M=50	0.592 ±0.013	4.095 ±0.070	0.131 ±0.0028	1.007 ±0.17	12.50 ±1.3	0.997 ±0.17	12.33 ±1.3
M=100	0.598 ±0.0048	4.044 ±0.055	0.127 ±0.0031	0.998 ±0.17	12.12 ±1.1	0.978 ±0.17	11.74 ±0.96
M=200	0.619 ±0.040	4.006 ±0.067	0.129 ±0.0034	0.983 ±0.15	12.07 ±0.79	0.967 ±0.15	11.77 ±0.52

	NLL masks-only		Cov90 err ($\hat{\sigma}$)		Cov90 err (σ=0.1)	
Mask bank	ID	OOD	ID	OOD	ID	OOD
M=1	n/a	n/a	0.124 ±0.0054	0.560 ±0.015	0.307 ±0.015	0.704 ±0.0019
M=5	5.155 ±2.1	56.44 ±18	0.120 ±0.011	0.591 ±0.016	0.136 ±0.021	0.611 ±0.019
M=20	1.619 ±0.25	20.58 ±0.41	0.108 ±0.0031	0.544 ±0.0024	0.109 ±0.0018	0.545 ±0.0072
M=50	1.449 ±0.26	17.57 ±2.2	0.104 ±0.0066	0.533 ±0.013	0.104 ±0.0059	0.527 ±0.014
M=100	1.405 ±0.27	16.65 ±1.7	0.104 ±0.0035	0.525 ±0.0066	0.104 ±0.0026	0.516 ±0.0047
M=200	1.389 ±0.19	16.66 ±0.74	0.103 ±0.0094	0.533 ±0.0044	0.103 ±0.0094	0.525 ±0.0024

			Phys. E-drift		
Mask bank	OOD α	$\mathrm{Var}_{\mathrm{epi}}/\mathrm{Var}_{\mathrm{alea}}$ (OOD)	ID	OOD	H-drift
M=1	n/a	n/a	0.657 ±0.034	3.815 ±0.15	0.103 ±0.010
M=5	0.590 ±0.14	1.620 ±0.79	0.704 ±0.052	3.894 ±0.046	0.104 ±0.0050
M=20	0.766 ±0.0075	3.270 ±0.14	0.740 ±0.0026	4.026 ±0.070	0.108 ±0.0026
M=50	0.802 ±0.026	4.108 ±0.68	0.766 ±0.017	3.966 ±0.026	0.106 ±0.0013
M=100	0.836 ±0.0056	5.087 ±0.21	0.784 ±0.058	4.098 ±0.17	0.107 ±0.0012
M=200	0.833 ±0.014	5.002 ±0.52	0.760 ±0.027	4.047 ±0.15	0.108 ±0.0019

H.8 Efficiency cost profile

Table 14 summarises the cost profile. Deep Ensembles are expensive accuracy references, while fixed-mask dropout is the single-model trajectory-consistent uncertainty mechanism.

H.9 Hidden-parameter shift detection

We ask whether a model can notice that the system it observes has silently changed. A Coupled Pendulum model is conditioned on a nominal mass ratio r_0, while the observed trajectory is generated at a true ratio r_{true}. Cases with $r_{\mathrm{true}} = r_0$ are negatives, mismatches are positives, and we report AUROC. The model sees only a revealed prefix of k observed steps. All r_{true} share one pool of 50 initial conditions from a ratio-independent sampler, so at $k = 0$ the model input is identical across conditions. Both scores use $M = 50$ samples per method. The predictive-variance score is the variance of the forecasts launched from the state at step k and conditioned on r_0, averaged over the remaining steps of a $T = 60$ horizon and the state dimensions. The prefix-drift score evaluates each sampled Hamiltonian H_m, conditioned on r_0, on the observed states $t = 0, \ldots, k$ and takes its standard deviation over t relative to its mean, averaged over samples. No integration is involved. Fixed-mask MC keeps one mask per sample along the prefix. Fresh-mask MC is scored under its native sampling, a fresh mask at every evaluation. We use 10 seeds (42–51), paired two-sided t-tests on per-seed AUROC and Holm correction within each family (Table 15).

As a leakage control, at $k = 0$ the pooled predictive-variance AUROC is 0.499, inside the chance band $[0.43, 0.57]$. Predictive variance detects the shift only weakly for both schemes. For held-out ratios at $r_0 = 1.0$ its seed-mean AUROC peaks at 0.641 (Fixed-mask MC) and 0.660 (Fresh-mask MC), both at $k = 10$. At the training-grid anchors $r_0 \in \{0.75, 1.0, 1.5\}$ with $k = 2$–10 it

Table 14: Efficiency comparison on the Coupled Pendulum benchmark. Relative values are normalised to the deterministic HNN. The Deep Ensemble row is extrapolated from a directly measured deterministic member under a five-member parallel assumption and is separated as a reference row. The sample counts in this table serve the inference-cost profile only. The matched sample counts behind the reported accuracy, likelihood and decomposition metrics are given in Appendix K and Table 4.

Method	Models	Samples	Train time	Infer time	Storage	Decomp. axes
Fresh-mask MC	1	10	1.044	1.195	1.000	no
Fixed-mask MC (M=100)	1	100	1.050	1.295	1.000	yes
Fixed-mask MC (M=200)	1	200	1.049	1.528	1.000	yes
Concrete	1	10	1.274	1.330	1.031	no
Variational	1	10	1.077	1.260	1.573	no
Det. HNN	1	1	1.000	1.000	1.000	no
Deep ens. ×5	5	5	5.000	5.000	5.000	yes

Table 15: Exploratory hidden-parameter shift detection on the Coupled Pendulum, reported as the AUROC for detecting that the true mass ratio r_{true} differs from the conditioning ratio r_0, given a revealed prefix of k observed steps.

		Revealed prefix length k				
r_0	Method	2	3	5	8	10
0.75	Fixed-mask MC	0.794	0.803	0.813	0.821	0.829
	Fresh-mask MC	0.640	0.701	0.765	0.808	0.827
	Δ (Fixed − Fresh)	$0.154^{\dagger}$	$0.103^{\dagger}$	$0.049^{\dagger}$	0.013	0.003
1.0	Fixed-mask MC	0.856	0.838	0.817	0.797	0.811
	Fresh-mask MC	0.651	0.705	0.767	0.810	0.830
	Δ (Fixed − Fresh)	$0.204^{\dagger}$	$0.133^{\dagger}$	$0.050^{\dagger}$	−0.013	−0.019
1.5	Fixed-mask MC	0.631	0.634	0.640	0.668	0.693
	Fresh-mask MC	0.609	0.650	0.695	0.728	0.743
	Δ (Fixed − Fresh)	$0.023^{\dagger}$	−0.016	$-0.055^{\dagger}$	$-0.060^{\dagger}$	$-0.049^{\dagger}$

Predictive-variance score	Fixed-mask MC	Fresh-mask MC
Leakage control, $k = 0$, $r_0 = 1.0$ (pooled over both methods)	0.499	
Peak over $k \in \{1, 2, 5, 10\}$, $r_0 = 1.0$ (held-out r_{true})	0.641 ($k = 10$)	0.660 ($k = 10$)
Range over $r_0 \in \{0.75, 1.0, 1.5\}$ (training-grid r_{true}), $k \in \{2, 3, 5, 8, 10\}$	0.468–0.572	0.483–0.573

The top block scores the per-sample Hamiltonian drift along the revealed prefix, with r_{true} over the six other training mass ratios. Δ is the mean paired per-seed difference Fixed-mask MC − Fresh-mask MC, and † marks Holm-adjusted $p < 0.05$ (Holm over the five k per anchor r_0).

stays between 0.468 and 0.573, and no paired difference survives Holm. We regard detection from predictive variance as an open problem.

The prefix-drift score detects the shift better, but the comparison is mixed. At $r_0 = 0.75$ and 1.0, Fixed-mask MC is significantly higher for $k = 2, 3, 5$ (at $k = 2$, 0.794 vs. 0.640 and 0.856 vs. 0.651). The gap vanishes around $k = 8$–10. At $r_0 = 1.5$ Fixed-mask MC is higher only at $k = 2$, and Fresh-mask MC is higher for $k \geq 5$. The benefit is thus an early-detection effect on short prefixes, not a uniform advantage.

Flagging anomalies or model change through violations of energy conservation is not new [De Luca et al., 2006]. Our claim is limited to the link with the sampling scheme. A fixed mask gives each sample one coherent H_m whose drift along the observed prefix can be read off directly. This benchmark is exploratory and independent of the α claim of the main text.

I Full variance decomposition notes

Table 16 reports the decomposition values used in the controlled comparison. The decomposition evidence is carried by the variance components themselves, $\mathrm{Var}_{\mathrm{epi}}$ for mask-conditioned model uncertainty, $\mathrm{Var}_{\mathrm{alea}}$ for initial-condition sensitivity, and α (Equation (4)) for the epistemic fraction.

The important qualitative pattern is that Fixed-mask MC makes the model-uncertainty axis visible on the Coupled Pendulum and Spring Chain without retraining an ensemble. On the Coupled Pendulum, its OOD epistemic component is larger than in the fresh-mask dropout rows, which prevents the predictive variance from collapsing into an almost purely aleatoric explanation. On the Spring Chain, Fixed-mask MC and Variational dropout both allocate more variance to epistemic uncertainty than Fresh-mask MC or Concrete dropout, while Fixed-mask MC also retains the strongest conservation profile among the stochastic methods.

The ID values in Table 16 are included to make the ID-to-OOD shift explicit. On the Coupled Pendulum, Fixed-mask MC moves from $\alpha = 0.203$ in distribution to 0.524 OOD, while Fresh-mask MC moves from 0.00753 to 0.0326, with about $19\times$ less variance on the mask axis (Section 5.3). On the Spring Chain, Fixed-mask MC has a higher in-distribution epistemic fraction, $\alpha = 0.803$, and still increases to 0.833 OOD, against 0.0153 to 0.0464 for Fresh-mask MC. How far α moves with a shift thus depends on where the decomposition attributes model error (below), and we do not test it as an OOD detector.

Where the model uncertainty is attributed. The difference between the two contracts is one of attribution. Under fresh masks little variance remains on the mask axis, and the validation fit widens the IC ensemble instead. At the fitted scale, Fresh-mask MC on the Coupled Pendulum has $\mathrm{Var}_{\mathrm{alea}} = 0.142$ (ID) and 0.189 (OOD), against 0.0798 and 0.104 for Fixed-mask MC, in line with its larger $\hat{\sigma}$ (0.331 vs. 0.250). At the true $\sigma = 0.1$ the IC term is nearly the same for both ($\mathrm{Var}_{\mathrm{alea}} = 0.0131$ vs. 0.0128 ID, 0.0182 vs. 0.0170 OOD), while the mask axis still carries $\mathrm{Var}_{\mathrm{epi}} = 0.00096$ vs. 0.0186 ID and 0.00627 vs. 0.118 OOD (seed means over the evaluation runs behind the $\sigma = 0.1$ columns of Table 7). Model error that a fixed mask exposes on the mask axis is thus, under fresh masks, absorbed into the replica ensemble and counted as aleatoric. The likelihood at the true scale does not depend on this attribution and favours Fixed-mask MC on both systems (Table 1).

The TFIM extension verifies that the nested mask-vs-initial-condition decomposition can be applied in latent wavefunction dynamics. We use it as a conservative audit rather than as the main decomposition evidence, because the Coupled Pendulum setting provides the cleaner interpretable OOD axis.

I.1 Sensitivity to the IC-noise scale

Table 17 rescales the validation-fitted IC-noise scale $\hat{\sigma}$ by 0, 0.25, 0.5, 1, 2 and 4 on the Coupled Pendulum (seeds 42–44). At $\sigma = 0$, $\mathrm{Var}_{\mathrm{alea}}$ is zero for Fixed-mask MC and at most 3.5×10^{-4} per mass ratio for Fresh-mask MC, whose replicas draw independent fresh masks, so their mask means differ by sampling noise ($\mathrm{Var}_{\mathrm{alea}} \approx \mathrm{Var}_{\mathrm{epi}}/50$, hence $\alpha = 0.980$). $\mathrm{Var}_{\mathrm{alea}}$ is non-decreasing in σ for every seed.

Out of distribution $\mathrm{Var}_{\mathrm{epi}}$ changes little with σ, from 0.1285 to 0.1561 ($1.21\times$) for Fixed-mask MC between $0\times$ and $4\times$, and in distribution it rises from 0.0197 to 0.0582. Fixed-mask MC exceeds Fresh-mask MC at every scale, split and seed, by 17.7–$20.3\times$ in seed means. The fraction α, by contrast, falls from 1 at $\sigma = 0$ to 0.095 (Fixed-mask MC, OOD, $4\times$). The decomposition is therefore operational and conditioned on the chosen IC perturbation model. Absolute $\mathrm{Var}_{\mathrm{epi}}$ is less sensitive to the IC-noise scale in these experiments.

The sweep fixes the mask and IC random seeds at 42 for every run, whereas Table 16 uses each run's own seed. This changes which 50 of the 200 masks are evaluated, the IC draws and $\hat{\sigma}$, so its values differ from Table 16. For seed 42, where the seeds coincide, the $\mathrm{Var}_{\mathrm{epi}}$ values agree exactly.

Ten seeds at the fitted scale. With seeds 42–51 at $1 \times \hat{\sigma}$, Fixed-mask MC has higher $\mathrm{Var}_{\mathrm{epi}}$ and lower $\mathrm{Var}_{\mathrm{alea}}$ than Fresh-mask MC in both splits (Table 20, family (iii): $\mathrm{Var}_{\mathrm{epi}}$ $+0.0192$ $[0.0179, 0.0205]$ ID and $+0.113$ $[0.0936, 0.132]$ OOD, $\mathrm{Var}_{\mathrm{alea}}$ -0.0623 $[-0.0646, -0.0598]$ ID and

Table 16: Variance decomposition of the nested IC × mask predictive ensemble into aleatoric (IC-conditional) and epistemic (mask-conditional at fixed IC) components at the fitted IC-noise scale ($\hat{\sigma}$, $\hat{\epsilon}$) in and out of distribution, their OOD ratio, and the epistemic fraction α. Deterministic models have no mask ensemble (epistemic terms are 0 or n/a by construction). Variance components have no preferred direction.

	$\mathrm{Var}_{\mathrm{alea}}$		$\mathrm{Var}_{\mathrm{epi}}$			α	
Method	ID	OOD	ID	OOD	$\mathrm{Var}_{\mathrm{epi}}/\mathrm{Var}_{\mathrm{alea}}$ (OOD)	ID	OOD
Coupled pendulum							
Fresh-mask MC	0.142 ± 0.0041	0.189 ± 0.012	0.00108 ± 5.1e-5	0.00637 ± 8.0e-4	0.0337 ± 0.0035	0.00753 ± 4.0e-4	0.0326 ± 0.0033
Concrete	0.112 ± 0.0032	0.145 ± 0.0071	9.9e-4 ± 6.6e-5	0.00681 ± 5.2e-4	0.0469 ± 0.0034	0.00880 ± 6.5e-4	0.0448 ± 0.0031
Variational	0.156 ± 0.023	0.155 ± 0.022	0.0161 ± 0.0059	0.0425 ± 0.014	0.275 ± 0.080	0.0922 ± 0.025	0.213 ± 0.051
Fixed-mask MC	0.0798 ± 0.0021	0.104 ± 0.0044	0.0203 ± 0.0022	0.119 ± 0.033	1.135 ± 0.29	0.203 ± 0.018	0.524 ± 0.066
Deterministic	0.110 ± 0.0021	0.180 ± 0.032	n/a	n/a	n/a	n/a	n/a
Deep ens. ×5	0.109 (n=1)	0.143 (n=1)	0.00206 (n=1)	0.123 (n=1)	0.856 (n=1)	0.0185 (n=1)	0.461 (n=1)
Spring chain							
Fresh-mask MC	0.561 ± 0.021	0.571 ± 0.023	0.00871 ± 5.5e-4	0.0278 ± 0.0016	0.0487 ± 0.0024	0.0153 ± 8.7e-4	0.0464 ± 0.0022
Concrete	0.150 ± 0.022	0.150 ± 0.022	0.00482 ± 8.3e-4	0.0106 ± 0.0024	0.0710 ± 0.013	0.0312 ± 0.0043	0.0661 ± 0.011
Variational	0.271 ± 0.086	0.243 ± 0.080	0.0985 ± 0.022	0.173 ± 0.041	0.876 ± 0.71	0.281 ± 0.11	0.426 ± 0.13
Fixed-mask MC	0.0373 ± 0.0011	0.0374 ± 0.0012	0.155 ± 0.024	0.192 ± 0.035	5.150 ± 1.0	0.803 ± 0.029	0.833 ± 0.028
Deterministic	0.153 ± 0.0085	0.152 ± 0.0085	n/a	n/a	n/a	n/a	n/a
Deep ens. ×5	0.111 (n=1)	0.107 (n=1)	0.00593 (n=1)	0.0400 (n=1)	0.372 (n=1)	0.0507 (n=1)	0.271 (n=1)
TFIM (operator-structured backbone)							
Fresh-mask MC	6.4e-4 ± 4.8e-4	6.5e-4 ± 4.8e-4	6.5e-4 ± 3.0e-5	5.9e-4 ± 3.7e-5	1.319 ± 0.69	0.551 ± 0.15	0.527 ± 0.16
Concrete	0.00201 ± 8.1e-4	0.00202 ± 8.1e-4	5.0e-4 ± 1.0e-4	4.7e-4 ± 1.3e-4	0.257 ± 0.085	0.211 ± 0.050	0.201 ± 0.055
Variational	0.00785 ± 3.3e-4	0.00780 ± 3.1e-4	0.0118 ± 7.8e-4	0.0102 ± 4.0e-4	1.313 ± 0.056	0.599 ± 0.013	0.567 ± 0.011
Fixed-mask MC	0.00494 ± 6.5e-4	0.00480 ± 6.4e-4	0.0155 ± 0.0017	0.0160 ± 0.0019	3.402 ± 0.68	0.757 ± 0.039	0.768 ± 0.034
Deterministic	6.5e-4 ± 3.8e-7	6.5e-4 ± 1.9e-6	0 ± 0	0 ± 0	0 ± 0	0 ± 0	0 ± 0

-0.0846 $[-0.0910, -0.0757]$ OOD, all $p_{\mathrm{Holm}} < 0.05$). $\mathrm{Var}_{\mathrm{total}}$ is lower in distribution (-0.0431 $[-0.0456, -0.0402]$), while the OOD difference ($+0.0281$ $[0.00475, 0.0545]$) is not significant after Holm. The effect is thus one of attribution, not a uniformly smaller predictive variance.

Table 17: Sensitivity of the variance decomposition to the IC-noise scale on the Coupled Pendulum, showing epistemic and aleatoric components and the epistemic fraction for Fixed-mask MC and Fresh-mask MC, in and out of distribution, at 0–4× the validation-fitted scale. Seeds 42–44, $t = 0$ excluded, descriptive (no tests).

			IC-noise scale (multiple of the fitted $\hat{\sigma}$)					
Method	Split	Quantity	0×	0.25×	0.5×	1×	2×	4×
Fixed-mask MC	ID	$\mathrm{Var}_{\mathrm{epi}}$	0.0197	0.0198	0.0202	0.0219	0.0290	0.0582
		$\mathrm{Var}_{\mathrm{alea}}$	0	0.0050	0.0202	0.0803	0.3167	1.228
		α	1.000	0.797	0.501	0.215	0.084	0.045
	OOD	$\mathrm{Var}_{\mathrm{epi}}$	0.1285	0.1286	0.1289	0.1302	0.1349	0.1561
		$\mathrm{Var}_{\mathrm{alea}}$	0	0.0067	0.0267	0.1055	0.4064	1.484
		α	1.000	0.950	0.827	0.551	0.249	0.095
Fresh-mask MC	ID	$\mathrm{Var}_{\mathrm{epi}}$	9.7e-4	9.8e-4	0.0010	0.0011	0.0015	0.0032
		$\mathrm{Var}_{\mathrm{alea}}$	1.9e-5	0.0090	0.0358	0.1421	0.5564	2.143
		α	0.980	0.098	0.027	0.008	0.003	0.001
	OOD	$\mathrm{Var}_{\mathrm{epi}}$	0.0071	0.0071	0.0071	0.0072	0.0074	0.0088
		$\mathrm{Var}_{\mathrm{alea}}$	1.4e-4	0.0132	0.0516	0.1971	0.7152	2.463
		α	0.980	0.350	0.121	0.035	0.010	0.004

J Energy conservation across averaging modes

For stochastic Hamiltonian models, the relevant conservation question depends on the averaging mode. Each fixed mask defines its own Hamiltonian flow, whereas the mean of several mask-conditioned trajectories need not lie on any single energy shell.

How to read the drift column. The H-drift entry should be read as a self-consistency measure of the sampled Hamiltonian, not as physical energy conservation and not as a replacement for OOD MSE or Gaussian NLL. Fixed-mask rows are designed to conserve a sampled Hamiltonian along each rollout. Fresh-mask rows can have competitive mean trajectories, but the Hamiltonian sampled by the network changes inside a rollout, which is the modelling-contract failure that motivates the method. Table 20 gives the paired differences in physical energy drift at the evaluation horizon (Phys. E-drift, the absolute drift $|E(z_T) - E(z_1)|$ of the true Hamiltonian from the first predicted frame to the last step, averaged over masks and the full test set).

Long-horizon physical energy. H-drift measures whether a rollout follows the Hamiltonian the network sampled. It says nothing about the true energy. We therefore evaluate the true Hamiltonian along each sampled rollout, relative to that rollout's own first predicted frame and divided by its magnitude (self-referenced and target-free), for $T = 1000$ steps with 10 masks per stochastic arm on both classical systems (Table 19 and Figure 8). On the Spring Chain (1200 test cases per split) the per-rollout physical energy error of Fresh-mask MC relative to Fixed-mask MC grows from $1.46\times$ at $t = 100$ to $4.71\times$ at $t = 1000$ in distribution, and from $1.18\times$ to $3.46\times$ out of distribution (ten seeds, and at $T = 1000$ Fixed-mask MC is lower on 10/10 seeds, unadjusted paired t-test $p = 3 \times 10^{-11}$ and 4×10^{-11}). On the Coupled Pendulum (all 1400 in-distribution and 1200 OOD test cases) the same ratio grows from $0.82\times$ to $1.94\times$ in distribution and from $0.89\times$ to $1.88\times$ out of distribution (10/10 seeds at $T = 1000$, $p = 1 \times 10^{-13}$ and 4×10^{-9}). Fixed-mask MC is slightly worse at $t = 100$ and overtakes Fresh-mask MC after about $t = 200$ in distribution and $t = 400$ out of distribution. Four caveats apply. In distribution the deterministic network still drifts least (Fixed-mask MC has $1.15\times$ its drift at $T = 1000$ on the Spring Chain and $1.28\times$ on the Coupled Pendulum). Out of distribution Fixed-mask MC is lower on the Spring Chain ($0.79\times$) and not distinguishable from it on the Coupled Pendulum ($p = 0.12$). At the usual evaluation horizon the benefit is absent or reversed on the Coupled Pendulum. Finally, this is not an accuracy claim. From $t \approx 400$ the ensemble-mean MSE of every arm is dominated by dephasing, and energy is meaningful only along each sampled rollout, not along the ensemble-mean trajectory (Appendix H.7).

Table 18: Drift along the rollout. For the Coupled Pendulum and Spring Chain, drift of the model's own sampled Hamiltonian (self-consistency, not physical energy, see Table 19) averaged over the rollout steps and at the final frame, rescored with the mask that produced each rollout. For TFIM, norm and energy drift of the predicted wavefunction (from Table 11).

Method	Mean H-drift	Final H-drift
Coupled pendulum		
Fresh-mask MC	0.157 ±0.0076	0.179 ±0.0084
Concrete	0.124 ±0.0043	0.147 ±0.0060
Variational	0.530 ±0.12	0.645 ±0.16
Fixed-mask MC	**0.00432** ±5.3e-5	0.00423 ±3.0e-4
Deterministic	0.00428 ±4.6e-5	0.00354 ±2.4e-4
Deep ens. ×5	0.00431 (n=1)	0.00359 (n=1)
Spring chain		
Fresh-mask MC	1.751 ±0.11	1.854 ±0.11
Concrete	1.732 ±0.21	1.761 ±0.27
Variational	4.111 ±0.58	4.739 ±0.64
Fixed-mask MC	**0.108** ±0.0014	0.104 ±0.0047
Deterministic	0.106 ±3.3e-4	0.0791 ±0.0020
Deep ens. ×5	0.106 (n=1)	0.0774 (n=1)

Method	Norm drift	Energy drift
TFIM (operator-structured backbone)		
Fresh-mask MC	0.0371 ±0.0014	**0.0994** ±0.0019
Concrete	**0.0328** ±0.0057	0.107 ±0.017
Variational	0.291 ±0.051	0.726 ±0.10
Fixed-mask MC	0.146 ±0.041	0.299 ±0.086
Deterministic	0.0123 ±0.0048	0.0318 ±0.014

Deep Ensembles. Deep Ensemble rows are included as reference estimates in Table 18. They are useful as a compute-heavy benchmark, but they are not part of the paired tests because the paper's main comparison is between single-model stochastic mechanisms.

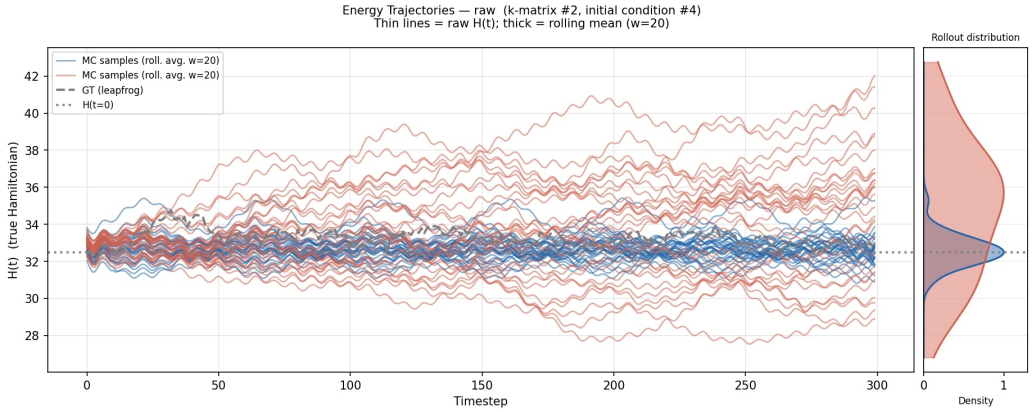


Figure 7: Example Spring Chain energy trajectories for stochastic rollouts. Thin curves show raw Hamiltonian traces and thick curves show rolling averages (blue Fixed-mask MC, red Fresh-mask MC). Fixed-mask MC keeps the sampled trajectories concentrated near the initial energy shell, while fresh-mask sampling produces visibly broader energy drift over the same rollout. This single-case visualisation complements the aggregate drift values in Table 18.

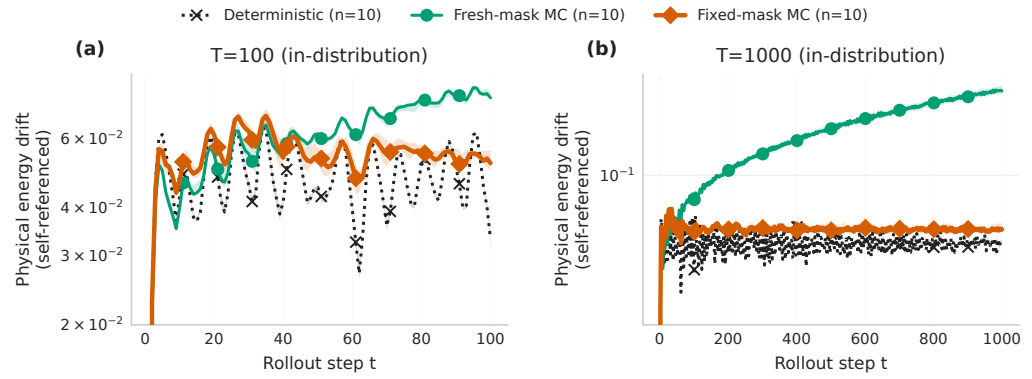


Figure 8: Spring-chain physical energy drift versus rollout step, in distribution. Drift is the true Hamiltonian along each predicted rollout relative to its first predicted frame, evaluated per sampled rollout and never on the ensemble-mean trajectory. Median and interquartile range over ten seeds. (a) $T = 100$. (b) $T = 1000$. Values at $t = 100$ and $t = 1000$ are in Table 19.

Table 19: Long-horizon physical energy drift. Relative drift $|E(z_t) - E(z_1)|/|E(z_1)|$ of the true Hamiltonian along each sampled rollout at t=100 and t=1000 of 1000-step rollouts with 10 masks, mean $\pm$ sd over ten seeds. Ratios quoted in the text are means of per-seed ratios.

	T=100		T=1000	
Method	ID	OOD	ID	OOD
Long horizon, Spring Chain (self-referenced physical drift, ten seeds)				
Fresh-mask MC	0.0770 $\pm$0.0025	0.482 $\pm$0.10	0.259 $\pm$0.015	1.500 $\pm$0.11
Fixed-mask MC	0.0531 $\pm$0.0042	0.406 $\pm$0.044	0.0552 $\pm$0.0032	0.433 $\pm$0.025
Deterministic	0.0329 $\pm$0.0019	0.533 $\pm$0.14	0.0479 $\pm$0.0018	0.550 $\pm$0.12
Long horizon, Coupled Pendulum (self-referenced physical drift, ten seeds)				
Fresh-mask MC	0.0811 $\pm$0.0020	0.463 $\pm$0.048	0.186 $\pm$0.0040	0.911 $\pm$0.039
Fixed-mask MC	0.0995 $\pm$0.0066	0.527 $\pm$0.061	0.0963 $\pm$0.0037	0.491 $\pm$0.053
Deterministic	0.0656 $\pm$0.0044	0.412 $\pm$0.044	0.0752 $\pm$0.0027	0.434 $\pm$0.091

K Protocol lock and sanity checks

Validation and test horizons. Model selection uses the same rollout length as the reported evaluation, so checkpoint choice and learning-rate scheduling are aligned with the reported long-horizon metric.

Mask interpretation. The fixed-mask condition means one sampled mask index per trajectory, held fixed across every leapfrog sub-step of that trajectory. Fresh-mask baselines resample inside the rollout and are therefore not interpreted as samples from a single Hamiltonian. Concrete and Variational variants are only labelled fixed-mask when their inference-time stochastic realisation is drawn once and reused across the whole rollout. This consistency check is independent of the dropout family and is the reason the main claim is about trajectory consistency rather than a universal MC-dropout leaderboard.

TFIM coordinate convention. TFIM latent states are $z = \mathrm{concat}(q, p)$ for $\psi = q + ip$, and the HNN rollout takes them in $[p, q]$ order. The transverse field g is a control input, not part of the dynamical state.

TFIM initial-condition grouping. When intended initial-condition (IC) metadata is present, train/validation/test selection is performed by intended initial-state blocks before latent replicas are flattened into the sample axis. This prevents leakage in which two latent reconstructions of the same intended state appear in different splits.

Energy and norm aggregation. Classical drift is reported in two forms, the drift of the model's own sampled Hamiltonian (H-drift, a self-consistency measure) and the physical energy drift of the true Hamiltonian along each sampled rollout. The main text uses the H-drift column for readability and sends physical and long-horizon drift to Appendix J. TFIM reports both norm drift and energy drift, because preserving the latent pure-state norm is part of the modelling contract.

Matched sample counts. Sample counts are matched across the stochastic arms of a system. On the Coupled Pendulum, OOD rollouts use 10 MC samples and the IC-ensemble metrics 50 MC samples crossed with 20 state-preparation replicas. On the Spring Chain, OOD metrics use 50 MC samples and IC-ensemble metrics 50×20. On TFIM, state error and drift use 10 MC samples and IC-ensemble metrics 50×20 (Appendix E).

Statistical protocol. For every tested column we report the paired effect size of Fixed-mask MC minus each comparator (Fresh-mask MC, Concrete, Variational), the mean over the ten matched seeds 42–51 of the difference of seed-level means, with a 95% percentile bootstrap confidence interval over seeds (10,000 resamples). Each difference is tested with a two-sided paired t-test on the seed-level means, and p-values are Holm-corrected within predefined families, separately for each system: (i) accuracy and drift (MSE or state error, H-drift, physical, norm and energy drift), (ii) likelihood and coverage (G-NLL under both noise conventions and Cov90 error), and (iii) variance decomposition ($\mathrm{Var}_{\mathrm{epi}}$, $\mathrm{Var}_{\mathrm{alea}}$ and $\mathrm{Var}_{\mathrm{total}} = \mathrm{Var}_{\mathrm{epi}} + \mathrm{Var}_{\mathrm{alea}}$, in and out of distribution). A family contains every comparator $\times$ column test in it. The Coupled Pendulum, Spring Chain and TFIM are separate systems. Seed pools and systems are never combined in one correction. The fitted noise scales and α are descriptive and untested. The deterministic HNN is a reference row with an effect size but no test, and the Deep Ensemble ($n = 1$) is not tested. The ten-seed sample size and this protocol were fixed in advance, and the analysis was run once (Tables 20 and 21). Results with three seeds (Tables 13 and 17) are descriptive and carry no significance claims. The long-horizon energy p-values in Appendix J come from a separate paired analysis outside these families and are reported unadjusted.

Table 20: Paired effect sizes for the classical systems (Coupled Pendulum and Spring Chain), computed as Fixed-mask MC minus each comparator, with 95% bootstrap CIs and Holm-adjusted significance marks. Columns are those of Tables 1, 7 and 16, and Phys. E-drift is defined in Appendix J.

	MSE		
Comparator	ID	OOD	H-drift
Coupled pendulum, family (i) accuracy and drift			
Fresh-mask MC	-0.00344 [-0.00768, 0.00173]	-0.0650 [-0.107, -0.0317]	-0.153^{*} [-0.157, -0.148]
Concrete	-0.00201 [-0.00570, 0.00197]	0.00515 [-0.0229, 0.0323]	-0.120^{*} [-0.122, -0.117]
Variational	-0.0958^{*} [-0.113, -0.0792]	-0.0798^{*} [-0.113, -0.0459]	-0.525^{*} [-0.593, -0.458]
Deterministic (ref.)	-0.00520 [-0.00828, -0.00206]	-0.158 [-0.228, -0.0977]	3.7e-5 [2.7e-6, 7.4e-5]
Spring chain, family (i) accuracy and drift			
Fresh-mask MC	-0.0404^{*} [-0.0686, -0.0134]	-0.413^{*} [-0.459, -0.366]	-1.643^{*} [-1.705, -1.581]
Concrete	-0.0534^{*} [-0.0844, -0.0272]	-0.360^{*} [-0.398, -0.321]	-1.625^{*} [-1.761, -1.512]
Variational	-1.108^{*} [-1.185, -1.039]	-0.420^{*} [-0.446, -0.393]	-4.003^{*} [-4.375, -3.706]
Deterministic (ref.)	0.0841 [0.0687, 0.101]	-0.431 [-0.490, -0.371]	0.00112 [2.2e-4, 0.00197]

	Phys. E-drift	
Comparator	ID	OOD
Coupled pendulum, family (i) accuracy and drift		
Fresh-mask MC	0.0143^{*} [0.00659, 0.0215]	0.0271 [-0.0586, 0.105]
Concrete	0.0191^{*} [0.0143, 0.0237]	0.0884 [0.0379, 0.140]
Variational	-0.184^{*} [-0.219, -0.151]	-0.581^{*} [-0.714, -0.477]
Deterministic (ref.)	0.0301 [0.0241, 0.0368]	0.0161 [-0.0804, 0.0933]
Spring chain, family (i) accuracy and drift		
Fresh-mask MC	-0.332^{*} [-0.354, -0.307]	-0.529^{*} [-0.628, -0.424]
Concrete	-0.119^{*} [-0.146, -0.0970]	0.129^{*} [0.0712, 0.200]
Variational	-2.678^{*} [-2.879, -2.505]	-4.124^{*} [-4.761, -3.580]
Deterministic (ref.)	0.129 [0.112, 0.145]	-0.882 [-1.350, -0.504]

Cells give the mean paired difference Fixed-mask MC $-$ comparator and the 95% bootstrap CI, and * marks Holm-adjusted $p < 0.05$. The confidence intervals are not adjusted for multiplicity, so an interval can exclude zero while the Holm-adjusted test is not significant. Negative = Fixed-mask MC lower. $\mathrm{Var}_{\mathrm{total}}$ is defined in Equation (4).

Table 20: Paired effect sizes for the classical systems (continued).

	G-NLL ($\hat{\sigma}$)	
Comparator	ID	OOD
Coupled pendulum, family (ii) likelihood: G-NLL		
Fresh-mask MC	-0.0501^{*} [−0.0596, −0.0399]	-0.524^{*} [−0.678, −0.384]
Concrete	-0.0426^{*} [−0.0485, −0.0366]	-1.063^{*} [−1.175, −0.944]
Variational	-0.293^{*} [−0.342, −0.249]	-0.0973 [−0.311, 0.116]
Deterministic (ref.)	-0.0487 [−0.0537, −0.0441]	-1.481 [−1.697, −1.269]
Spring chain, family (ii) likelihood: G-NLL		
Fresh-mask MC	0.00911 [−0.0668, 0.0950]	7.001^{*} [6.388, 7.630]
Concrete	-0.740^{*} [−0.843, −0.625]	-3.605^{*} [−5.218, −2.062]
Variational	-0.787^{*} [−1.021, −0.590]	4.069^{*} [3.044, 5.103]
Deterministic (ref.)	-0.382 [−0.473, −0.287]	-6.865 [−7.862, −5.950]

	G-NLL (σ=0.1)	
Comparator	ID	OOD
Coupled pendulum, family (ii) likelihood: G-NLL		
Fresh-mask MC	-0.587^{*} [−0.611, −0.560]	-9.714^{*} [−10.30, −9.210]
Concrete	-0.614^{*} [−0.636, −0.593]	-8.614^{*} [−8.874, −8.331]
Variational	-0.863^{*} [−0.975, −0.746]	-3.031^{*} [−3.975, −2.164]
Deterministic (ref.)	-0.855 [−0.868, −0.839]	-15.34 [−16.32, −14.53]
Spring chain, family (ii) likelihood: G-NLL		
Fresh-mask MC	-4.247^{*} [−4.565, −3.929]	-22.29^{*} [−22.91, −21.71]
Concrete	-4.039^{*} [−4.201, −3.857]	-28.90^{*} [−29.90, −27.71]
Variational	-2.904^{*} [−3.155, −2.646]	-5.873^{*} [−7.050, −4.591]
Deterministic (ref.)	-4.081 [−4.159, −4.004]	-54.59 [−56.24, −52.90]

Table 20: Paired effect sizes for the classical systems (continued).

Comparator	Cov90 err ($\hat{\sigma}$) ID	Cov90 err ($\hat{\sigma}$) OOD
Coupled pendulum, family (ii) likelihood: Cov90 error		
Fresh-mask MC	0.00121 [−0.00260, 0.00496]	−0.0207 [−0.0362, −0.00481]
Concrete	−0.00207 [−0.00510, 0.00142]	−0.0520* [−0.0662, −0.0363]
Variational	−0.0115* [−0.0177, −0.00555]	0.0292 [0.00791, 0.0504]
Deterministic (ref.)	−0.0300 [−0.0338, −0.0268]	−0.101 [−0.118, −0.0828]
Spring chain, family (ii) likelihood: Cov90 error		
Fresh-mask MC	0.0646* [0.0581, 0.0710]	0.175* [0.163, 0.185]
Concrete	−0.0106 [−0.0185, −0.00196]	−0.0113 [−0.0260, 0.00411]
Variational	−0.0499* [−0.0715, −0.0306]	0.0793* [0.0647, 0.0941]
Deterministic (ref.)	−0.00723 [−0.0128, −5.9e-4]	−0.0602 [−0.0698, −0.0512]

Comparator	Cov90 err (σ=0.1) ID	Cov90 err (σ=0.1) OOD
Coupled pendulum, family (ii) likelihood: Cov90 error		
Fresh-mask MC	−0.0783* [−0.0818, −0.0748]	−0.160* [−0.178, −0.142]
Concrete	−0.0826* [−0.0869, −0.0789]	−0.151* [−0.168, −0.133]
Variational	−0.130* [−0.143, −0.117]	−0.0700* [−0.0956, −0.0441]
Deterministic (ref.)	−0.149 [−0.152, −0.145]	−0.213 [−0.230, −0.194]
Spring chain, family (ii) likelihood: Cov90 error		
Fresh-mask MC	−0.123* [−0.134, −0.111]	−0.114* [−0.125, −0.105]
Concrete	−0.131* [−0.138, −0.123]	−0.134* [−0.143, −0.126]
Variational	−0.182* [−0.191, −0.172]	−0.0497* [−0.0615, −0.0355]
Deterministic (ref.)	−0.174 [−0.178, −0.170]	−0.189 [−0.197, −0.181]

Table 20: Paired effect sizes for the classical systems (continued).

Comparator	$\mathrm{Var}_{\mathrm{epi}}$ (ID)	$\mathrm{Var}_{\mathrm{alea}}$ (ID)	$\mathrm{Var}_{\mathrm{total}}$ (ID)
Coupled pendulum, family (iii) variance decomposition, ID			
Fresh-mask MC	0.0192^{*} [0.0179, 0.0205]	-0.0623^{*} [−0.0646, −0.0598]	-0.0431^{*} [−0.0456, −0.0402]
Concrete	0.0193^{*} [0.0180, 0.0206]	-0.0319^{*} [−0.0338, −0.0298]	-0.0126^{*} [−0.0150, −0.0102]
Variational	0.00421 [8.0e-7, 0.00798]	-0.0764^{*} [−0.0902, −0.0630]	-0.0722^{*} [−0.0884, −0.0565]
Deterministic (ref.)	n/a	−0.0306 [−0.0323, −0.0288]	n/a
Spring chain, family (iii) variance decomposition, ID			
Fresh-mask MC	0.146^{*} [0.132, 0.160]	-0.524^{*} [−0.536, −0.511]	-0.378^{*} [−0.393, −0.363]
Concrete	0.150^{*} [0.136, 0.164]	-0.113^{*} [−0.125, −0.101]	0.0370^{*} [0.0137, 0.0584]
Variational	0.0564^{*} [0.0415, 0.0732]	-0.234^{*} [−0.284, −0.183]	-0.178^{*} [−0.229, −0.127]
Deterministic (ref.)	n/a	−0.116 [−0.121, −0.111]	n/a

Comparator	$\mathrm{Var}_{\mathrm{epi}}$ (OOD)	$\mathrm{Var}_{\mathrm{alea}}$ (OOD)	$\mathrm{Var}_{\mathrm{total}}$ (OOD)
Coupled pendulum, family (iii) variance decomposition, OOD			
Fresh-mask MC	0.113^{*} [0.0936, 0.132]	-0.0846^{*} [−0.0910, −0.0757]	0.0281 [0.00475, 0.0545]
Concrete	0.112^{*} [0.0933, 0.132]	-0.0411^{*} [−0.0449, −0.0376]	0.0712^{*} [0.0500, 0.0927]
Variational	0.0766^{*} [0.0565, 0.0968]	-0.0504^{*} [−0.0645, −0.0367]	0.0261 [−2.4e-4, 0.0557]
Deterministic (ref.)	n/a	−0.0755 [−0.0942, −0.0582]	n/a
Spring chain, family (iii) variance decomposition, OOD			
Fresh-mask MC	0.164^{*} [0.145, 0.184]	-0.534^{*} [−0.547, −0.520]	-0.369^{*} [−0.390, −0.348]
Concrete	0.182^{*} [0.161, 0.202]	-0.113^{*} [−0.126, −0.0999]	0.0685^{*} [0.0403, 0.0942]
Variational	0.0192 [−0.00508, 0.0414]	-0.206^{*} [−0.251, −0.158]	-0.186^{*} [−0.229, −0.147]
Deterministic (ref.)	n/a	−0.114 [−0.120, −0.109]	n/a

Table 21: Paired effect sizes for the transverse-field Ising model, computed as Fixed-mask MC minus each comparator, with 95% bootstrap CIs and Holm-adjusted significance marks. Columns are those of Tables 11, 8 and 16 (TFIM block).

	State RMSE	
Comparator	ID	OOD
TFIM, family (i) accuracy and drift		
Fresh-mask MC	0.0373^{*} [0.0334, 0.0415]	0.00739 [2.9e-5, 0.0147]
Concrete	0.0311^{*} [0.0267, 0.0355]	−7.3e-5 [−0.0155, 0.0160]
Variational	-0.0893^{*} [−0.0942, −0.0839]	-0.0963^{*} [−0.103, −0.0889]
Deterministic (ref.)	0.0386 [0.0349, 0.0425]	0.0281 [0.0198, 0.0367]

Comparator	Norm drift	Energy drift
TFIM, family (i) accuracy and drift		
Fresh-mask MC	0.109^{*} [0.0856, 0.134]	0.199^{*} [0.150, 0.250]
Concrete	0.113^{*} [0.0890, 0.140]	0.192^{*} [0.146, 0.238]
Variational	-0.145^{*} [−0.185, −0.111]	-0.427^{*} [−0.522, −0.340]
Deterministic (ref.)	0.134 [0.109, 0.160]	0.267 [0.213, 0.322]

Cells as in Table 20.

Table 21: Paired effect sizes for the transverse-field Ising model (continued).

	G-NLL ($\hat{\epsilon}$)	
Comparator	ID	OOD
TFIM, family (ii) likelihood: G-NLL		
Fresh-mask MC	1.281^{*} [1.187, 1.361]	-3.657^{*} [−5.042, −2.073]
Concrete	0.944^{*} [0.876, 1.009]	−1.521 [−2.709, −0.443]
Variational	-0.383^{*} [−0.416, −0.348]	-0.649^{*} [−0.679, −0.616]
Deterministic (ref.)	1.461 [1.436, 1.486]	−3.332 [−6.778, −0.450]

	G-NLL (ϵ=0.10)	
Comparator	ID	OOD
TFIM, family (ii) likelihood: G-NLL		
Fresh-mask MC	1.013^{*} [0.975, 1.049]	-4.594^{*} [−5.914, −3.002]
Concrete	0.861^{*} [0.794, 0.927]	-6.589^{*} [−9.599, −3.625]
Variational	-0.731^{*} [−0.781, −0.679]	-1.666^{*} [−1.726, −1.606]
Deterministic (ref.)	1.145 [1.105, 1.181]	−9.578 [−17.26, −3.147]

Table 21: Paired effect sizes for the transverse-field Ising model (continued).

	Cov90 err ($\hat{\epsilon}$)	
Comparator	ID	OOD
TFIM, family (ii) likelihood: Cov90 error		
Fresh-mask MC	0.0156^{*} [0.00853, 0.0225]	-0.132^{*} [−0.177, −0.0774]
Concrete	0.0126 [0.00482, 0.0217]	−0.0581 [−0.100, −0.0157]
Variational	0.0821^{*} [0.0771, 0.0860]	−0.00276 [−0.00995, 0.00341]
Deterministic (ref.)	0.0703 [0.0665, 0.0744]	−0.107 [−0.165, −0.0560]

	Cov90 err (ϵ=0.10)	
Comparator	ID	OOD
TFIM, family (ii) likelihood: Cov90 error		
Fresh-mask MC	0.0162^{*} [0.0130, 0.0195]	-0.182^{*} [−0.201, −0.159]
Concrete	0.0470^{*} [0.0389, 0.0550]	-0.250^{*} [−0.280, −0.219]
Variational	-0.0293^{*} [−0.0420, −0.0172]	-0.180^{*} [−0.185, −0.174]
Deterministic (ref.)	−0.0160 [−0.0240, −0.00872]	−0.243 [−0.294, −0.197]

Table 21: Paired effect sizes for the transverse-field Ising model (continued).

Comparator	$\mathrm{Var}_{\mathrm{epi}}$ (ID)	$\mathrm{Var}_{\mathrm{alea}}$ (ID)	$\mathrm{Var}_{\mathrm{total}}$ (ID)
TFIM, family (iii) variance decomposition, ID			
Fresh-mask MC	0.0149^{*} [0.0139, 0.0159]	0.00430^{*} [0.00394, 0.00462]	0.0192^{*} [0.0181, 0.0202]
Concrete	0.0150^{*} [0.0141, 0.0160]	0.00293^{*} [0.00225, 0.00362]	0.0180^{*} [0.0172, 0.0187]
Variational	0.00377^{*} [0.00239, 0.00500]	-0.00291^{*} [−0.00335, −0.00256]	8.7e-4 [−4.8e-4, 0.00218]
Deterministic (ref.)	0.0155 [0.0145, 0.0165]	0.00429 [0.00386, 0.00461]	0.0198 [0.0189, 0.0207]

Comparator	$\mathrm{Var}_{\mathrm{epi}}$ (OOD)	$\mathrm{Var}_{\mathrm{alea}}$ (OOD)	$\mathrm{Var}_{\mathrm{total}}$ (OOD)
TFIM, family (iii) variance decomposition, OOD			
Fresh-mask MC	0.0154^{*} [0.0144, 0.0166]	0.00415^{*} [0.00377, 0.00451]	0.0196^{*} [0.0184, 0.0208]
Concrete	0.0155^{*} [0.0145, 0.0167]	0.00278^{*} [0.00215, 0.00343]	0.0183^{*} [0.0172, 0.0194]
Variational	0.00575^{*} [0.00462, 0.00706]	-0.00300^{*} [−0.00343, −0.00270]	0.00275^{*} [0.00149, 0.00418]
Deterministic (ref.)	0.0160 [0.0150, 0.0171]	0.00415 [0.00373, 0.00447]	0.0201 [0.0191, 0.0212]

NeurIPS Paper Checklist

1. **Claims**

   Question: Do the main claims made in the abstract and introduction accurately reflect the paper's contributions and scope?

   Answer: [Yes]

   Justification: The abstract and Section 1 state the central contribution as trajectory-consistent dropout for Hamiltonian rollouts and an IC-vs-model uncertainty decomposition. The scope is limited throughout the paper. Section 5.2 and Section 7 explicitly avoid claiming a universal predictive-performance advantage.

   Guidelines:

   - The answer [N/A] means that the abstract and introduction do not include the claims made in the paper.
   - The abstract and/or introduction should clearly state the claims made, including the contributions made in the paper and important assumptions and limitations. A [No] or [N/A] answer to this question will not be perceived well by the reviewers.
   - The claims made should match theoretical and experimental results, and reflect how much the results can be expected to generalize to other settings.
   - It is fine to include aspirational goals as motivation as long as it is clear that these goals are not attained by the paper.

2. **Limitations**

   Question: Does the paper discuss the limitations of the work performed by the authors?

   Answer: [Yes]

   Justification: Section 6 and Appendix B discuss the conservative-system assumption, system-dependent OOD distances, and the TFIM stress-test result. The appendix further separates robustness, energy, and calibration details from the main claim.

   Guidelines:

   - The answer [N/A] means that the paper has no limitation while the answer [No] means that the paper has limitations, but those are not discussed in the paper.
   - The authors are encouraged to create a separate "Limitations" section in their paper.
   - The paper should point out any strong assumptions and how robust the results are to violations of these assumptions (e.g., independence assumptions, noiseless settings, model well-specification, asymptotic approximations only holding locally). The authors should reflect on how these assumptions might be violated in practice and what the implications would be.
   - The authors should reflect on the scope of the claims made, e.g., if the approach was only tested on a few datasets or with a few runs. In general, empirical results often depend on implicit assumptions, which should be articulated.
   - The authors should reflect on the factors that influence the performance of the approach. For example, a facial recognition algorithm may perform poorly when image resolution is low or images are taken in low lighting. Or a speech-to-text system might not be used reliably to provide closed captions for online lectures because it fails to handle technical jargon.
   - The authors should discuss the computational efficiency of the proposed algorithms and how they scale with dataset size.
   - If applicable, the authors should discuss possible limitations of their approach to address problems of privacy and fairness.
   - While the authors might fear that complete honesty about limitations might be used by reviewers as grounds for rejection, a worse outcome might be that reviewers discover limitations that aren't acknowledged in the paper. The authors should use their best judgment and recognize that individual actions in favor of transparency play an important role in developing norms that preserve the integrity of the community. Reviewers will be specifically instructed to not penalize honesty concerning limitations.

3. **Theory assumptions and proofs**

Question: For each theoretical result, does the paper provide the full set of assumptions and a complete (and correct) proof?

Answer: [Yes]

Justification: Corollary 1 states the separability, smoothness, step-size, and horizon assumptions needed for the per-mask trajectory interpretation. The proof is given in Appendix A, and Remark 1 clarifies that the corollary does not claim low prediction error or recovery of the true physical Hamiltonian.

Guidelines:

- The answer [N/A] means that the paper does not include theoretical results.
- All the theorems, formulas, and proofs in the paper should be numbered and cross-referenced.
- All assumptions should be clearly stated or referenced in the statement of any theorems.
- The proofs can either appear in the main paper or the supplemental material, but if they appear in the supplemental material, the authors are encouraged to provide a short proof sketch to provide intuition.
- Inversely, any informal proof provided in the core of the paper should be complemented by formal proofs provided in appendix or supplemental material.
- Theorems and Lemmas that the proof relies upon should be properly referenced.

4. **Experimental result reproducibility**

Question: Does the paper fully disclose all the information needed to reproduce the main experimental results of the paper to the extent that it affects the main claims and/or conclusions of the paper (regardless of whether the code and data are provided or not)?

Answer: [Yes]

Justification: Section 5 defines the controlled complexity ladder, methods, and metrics, while Appendix E gives training details and Appendix K documents protocol checks. The code release (`https://github.com/yukitachibana-eth/hnn-trajectory-dropout`) includes the fixed-mask implementation, evaluation scripts, artifact-export path, and precomputed summaries used to regenerate the reported tables and figures.

Guidelines:

- The answer [N/A] means that the paper does not include experiments.
- If the paper includes experiments, a [No] answer to this question will not be perceived well by the reviewers: Making the paper reproducible is important, regardless of whether the code and data are provided or not.
- If the contribution is a dataset and/or model, the authors should describe the steps taken to make their results reproducible or verifiable.
- Depending on the contribution, reproducibility can be accomplished in various ways. For example, if the contribution is a novel architecture, describing the architecture fully might suffice, or if the contribution is a specific model and empirical evaluation, it may be necessary to either make it possible for others to replicate the model with the same dataset, or provide access to the model. In general. releasing code and data is often one good way to accomplish this, but reproducibility can also be provided via detailed instructions for how to replicate the results, access to a hosted model (e.g., in the case of a large language model), releasing of a model checkpoint, or other means that are appropriate to the research performed.
- While NeurIPS does not require releasing code, the conference does require all submissions to provide some reasonable avenue for reproducibility, which may depend on the nature of the contribution. For example
  (a) If the contribution is primarily a new algorithm, the paper should make it clear how to reproduce that algorithm.
  (b) If the contribution is primarily a new model architecture, the paper should describe the architecture clearly and fully.

(c) If the contribution is a new model (e.g., a large language model), then there should either be a way to access this model for reproducing the results or a way to reproduce the model (e.g., with an open-source dataset or instructions for how to construct the dataset).

(d) We recognize that reproducibility may be tricky in some cases, in which case authors are welcome to describe the particular way they provide for reproducibility. In the case of closed-source models, it may be that access to the model is limited in some way (e.g., to registered users), but it should be possible for other researchers to have some path to reproducing or verifying the results.

5. **Open access to data and code**

   Question: Does the paper provide open access to the data and code, with sufficient instructions to faithfully reproduce the main experimental results, as described in supplemental material?

   Answer: [Yes]

   Justification: The experiments use simulated classical and quantum dynamical systems rather than private datasets. The code release (`https://github.com/yukitachibana-eth/hnn-trajectory-dropout`) documents data generation, precomputed result summaries, table/figure export, and smoke-test commands. Appendix E and Appendix D.2 specify the training and OOD evaluation settings needed to interpret those scripts.

   Guidelines:

   - The answer [N/A] means that paper does not include experiments requiring code.
   - Please see the NeurIPS code and data submission guidelines (`https://neurips.cc/public/guides/CodeSubmissionPolicy`) for more details.
   - While we encourage the release of code and data, we understand that this might not be possible, so [No] is an acceptable answer. Papers cannot be rejected simply for not including code, unless this is central to the contribution (e.g., for a new open-source benchmark).
   - The instructions should contain the exact command and environment needed to run to reproduce the results. See the NeurIPS code and data submission guidelines (`https://neurips.cc/public/guides/CodeSubmissionPolicy`) for more details.
   - The authors should provide instructions on data access and preparation, including how to access the raw data, preprocessed data, intermediate data, and generated data, etc.
   - The authors should provide scripts to reproduce all experimental results for the new proposed method and baselines. If only a subset of experiments are reproducible, they should state which ones are omitted from the script and why.
   - At submission time, to preserve anonymity, the authors should release anonymized versions (if applicable).
   - Providing as much information as possible in supplemental material (appended to the paper) is recommended, but including URLs to data and code is permitted.

6. **Experimental setting/details**

   Question: Does the paper specify all the training and test details (e.g., data splits, hyperparameters, how they were chosen, type of optimizer) necessary to understand the results?

   Answer: [Yes]

   Justification: Section 5 describes the systems, baselines, and evaluation metrics. Appendix E reports optimiser, learning rate, width, epochs, curriculum, rollout, mask-bank, and TFIM training details, and Appendix D.2 defines the OOD splits for each system.

   Guidelines:

   - The answer [N/A] means that the paper does not include experiments.
   - The experimental setting should be presented in the core of the paper to a level of detail that is necessary to appreciate the results and make sense of them.
   - The full details can be provided either with the code, in appendix, or as supplemental material.

7. **Experiment statistical significance**

   Question: Does the paper report error bars suitably and correctly defined or other appropriate information about the statistical significance of the experiments?

   Answer: [Yes]

   Justification: Table 9 reports mean$\pm$standard deviation for stochastic dropout rows where multiple seeds are available and explicitly marks Deep Ensemble as a single reference estimate. Appendix tables and figures, including Tables 16, 11, and 18, follow the same convention where applicable. Tables 20 and 21 report paired effect sizes with 95% bootstrap confidence intervals and Holm-corrected paired t-tests (Appendix K).

   Guidelines:

   - The answer [N/A] means that the paper does not include experiments.
   - The authors should answer [Yes] if the results are accompanied by error bars, confidence intervals, or statistical significance tests, at least for the experiments that support the main claims of the paper.
   - The factors of variability that the error bars are capturing should be clearly stated (for example, train/test split, initialization, random drawing of some parameter, or overall run with given experimental conditions).
   - The method for calculating the error bars should be explained (closed form formula, call to a library function, bootstrap, etc.)
   - The assumptions made should be given (e.g., Normally distributed errors).
   - It should be clear whether the error bar is the standard deviation or the standard error of the mean.
   - It is OK to report 1-sigma error bars, but one should state it. The authors should preferably report a 2-sigma error bar than state that they have a 96% CI, if the hypothesis of Normality of errors is not verified.
   - For asymmetric distributions, the authors should be careful not to show in tables or figures symmetric error bars that would yield results that are out of range (e.g., negative error rates).
   - If error bars are reported in tables or plots, the authors should explain in the text how they were calculated and reference the corresponding figures or tables in the text.

8. **Experiments compute resources**

   Question: For each experiment, does the paper provide sufficient information on the computer resources (type of compute workers, memory, time of execution) needed to reproduce the experiments?

   Answer: [Yes]

   Justification: Appendix E reports the main hardware and training setup, and Appendix H.8 gives a method-level efficiency comparison. The code release provides the practical reproduction path and expected runtime guidance for regenerating artifacts from precomputed summaries versus rerunning training.

   Guidelines:

   - The answer [N/A] means that the paper does not include experiments.
   - The paper should indicate the type of compute workers CPU or GPU, internal cluster, or cloud provider, including relevant memory and storage.
   - The paper should provide the amount of compute required for each of the individual experimental runs as well as estimate the total compute.
   - The paper should disclose whether the full research project required more compute than the experiments reported in the paper (e.g., preliminary or failed experiments that didn't make it into the paper).

9. **Code of ethics**

   Question: Does the research conducted in the paper conform, in every respect, with the NeurIPS Code of Ethics https://neurips.cc/public/EthicsGuidelines?

   Answer: [Yes]

   Justification: The work uses simulated dynamical systems, does not involve human subjects or private data, and reports limitations and system-dependent failure modes. The paper

frames uncertainty estimates as diagnostics for scientific and engineering models, not as standalone safety guarantees.

Guidelines:

- The answer [N/A] means that the authors have not reviewed the NeurIPS Code of Ethics.
- If the authors answer [No], they should explain the special circumstances that require a deviation from the Code of Ethics.
- The authors should make sure to preserve anonymity (e.g., if there is a special consideration due to laws or regulations in their jurisdiction).

10. **Broader impacts**

Question: Does the paper discuss both potential positive societal impacts and negative societal impacts of the work performed?

Answer: [Yes]

Justification: The introduction motivates uncertainty decomposition for reliability-sensitive scientific and engineering decisions, and Section 6 and Appendix B state limitations that prevent overinterpreting the method as a universal safety guarantee. The experiments are simulation-based and do not introduce a deployed decision system or human-facing model.

Guidelines:

- The answer [N/A] means that there is no societal impact of the work performed.
- If the authors answer [N/A] or [No], they should explain why their work has no societal impact or why the paper does not address societal impact.
- Examples of negative societal impacts include potential malicious or unintended uses (e.g., disinformation, generating fake profiles, surveillance), fairness considerations (e.g., deployment of technologies that could make decisions that unfairly impact specific groups), privacy considerations, and security considerations.
- The conference expects that many papers will be foundational research and not tied to particular applications, let alone deployments. However, if there is a direct path to any negative applications, the authors should point it out. For example, it is legitimate to point out that an improvement in the quality of generative models could be used to generate Deepfakes for disinformation. On the other hand, it is not needed to point out that a generic algorithm for optimizing neural networks could enable people to train models that generate Deepfakes faster.
- The authors should consider possible harms that could arise when the technology is being used as intended and functioning correctly, harms that could arise when the technology is being used as intended but gives incorrect results, and harms following from (intentional or unintentional) misuse of the technology.
- If there are negative societal impacts, the authors could also discuss possible mitigation strategies (e.g., gated release of models, providing defenses in addition to attacks, mechanisms for monitoring misuse, mechanisms to monitor how a system learns from feedback over time, improving the efficiency and accessibility of ML).

11. **Safeguards**

Question: Does the paper describe safeguards that have been put in place for responsible release of data or models that have a high risk for misuse (e.g., pre-trained language models, image generators, or scraped datasets)?

Answer: [N/A]

Justification: The paper does not release a high-capability generative model, a pretrained language model, or a model intended for direct deployment in high-risk decision making. The released assets are simulation, training, and evaluation code for controlled dynamical-system experiments.

Guidelines:

- The answer [N/A] means that the paper poses no such risks.
- Released models that have a high risk for misuse or dual-use should be released with necessary safeguards to allow for controlled use of the model, for example by requiring that users adhere to usage guidelines or restrictions to access the model or implementing safety filters.

- Datasets that have been scraped from the Internet could pose safety risks. The authors should describe how they avoided releasing unsafe images.
- We recognize that providing effective safeguards is challenging, and many papers do not require this, but we encourage authors to take this into account and make a best faith effort.

12. **Licenses for existing assets**

    Question: Are the creators or original owners of assets (e.g., code, data, models), used in the paper, properly credited and are the license and terms of use explicitly mentioned and properly respected?

    Answer: [Yes]

    Justification: The paper cites prior methods, datasets are generated synthetically, and the code release (MIT licence) documents software and license information for the provided artifacts. No third-party private datasets or restricted human-subject assets are used.

    Guidelines:

    - The answer [N/A] means that the paper does not use existing assets.
    - The authors should cite the original paper that produced the code package or dataset.
    - The authors should state which version of the asset is used and, if possible, include a URL.
    - The name of the license (e.g., CC-BY 4.0) should be included for each asset.
    - For scraped data from a particular source (e.g., website), the copyright and terms of service of that source should be provided.
    - If assets are released, the license, copyright information, and terms of use in the package should be provided. For popular datasets, `paperswithcode.com/datasets` has curated licenses for some datasets. Their licensing guide can help determine the license of a dataset.
    - For existing datasets that are re-packaged, both the original license and the license of the derived asset (if it has changed) should be provided.
    - If this information is not available online, the authors are encouraged to reach out to the asset's creators.

13. **New assets**

    Question: Are new assets introduced in the paper well documented and is the documentation provided alongside the assets?

    Answer: [Yes]

    Justification: The new assets are code, simulation protocols, and precomputed summaries for reproducing the reported tables and figures. The code release documents repository layout, data generation, artifact mapping, seeds, and the distinction between precomputed-result export and full retraining.

    Guidelines:

    - The answer [N/A] means that the paper does not release new assets.
    - Researchers should communicate the details of the dataset/code/model as part of their submissions via structured templates. This includes details about training, license, limitations, etc.
    - The paper should discuss whether and how consent was obtained from people whose asset is used.
    - At submission time, remember to anonymize your assets (if applicable). You can either create an anonymized URL or include an anonymized zip file.

14. **Crowdsourcing and research with human subjects**

    Question: For crowdsourcing experiments and research with human subjects, does the paper include the full text of instructions given to participants and screenshots, if applicable, as well as details about compensation (if any)?

    Answer: [N/A]

    Justification: The work uses simulated Coupled Pendulum, Spring Chain, and TFIM systems and does not involve crowdsourcing, user studies, or human-subject data.

    Guidelines:

- The answer [N/A] means that the paper does not involve crowdsourcing nor research with human subjects.
- Including this information in the supplemental material is fine, but if the main contribution of the paper involves human subjects, then as much detail as possible should be included in the main paper.
- According to the NeurIPS Code of Ethics, workers involved in data collection, curation, or other labor should be paid at least the minimum wage in the country of the data collector.

15. **Institutional review board (IRB) approvals or equivalent for research with human subjects**

    Question: Does the paper describe potential risks incurred by study participants, whether such risks were disclosed to the subjects, and whether Institutional Review Board (IRB) approvals (or an equivalent approval/review based on the requirements of your country or institution) were obtained?

    Answer: [N/A]

    Justification: No human participants, crowdsourced workers, medical records, personal data, or behavioural data are used, so IRB review is not applicable.

    Guidelines:

    - The answer [N/A] means that the paper does not involve crowdsourcing nor research with human subjects.
    - Depending on the country in which research is conducted, IRB approval (or equivalent) may be required for any human subjects research. If you obtained IRB approval, you should clearly state this in the paper.
    - We recognize that the procedures for this may vary significantly between institutions and locations, and we expect authors to adhere to the NeurIPS Code of Ethics and the guidelines for their institution.
    - For initial submissions, do not include any information that would break anonymity (if applicable), such as the institution conducting the review.

16. **Declaration of LLM usage**

    Question: Does the paper describe the usage of LLMs if it is an important, original, or non-standard component of the core methods in this research? Note that if the LLM is used only for writing, editing, or formatting purposes and does not impact the core methodology, scientific rigor, or originality of the research, declaration is not required.

    Answer: [N/A]

    Justification: LLM-based assistance was used for grammar checks/refinement and code/documentation support. LLMs are not part of the proposed method, evaluation protocol, data generation process, or the reported scientific results. All claims, experiments, and associated text were reviewed and verified by the authors.

    Guidelines:

    - The answer [N/A] means that the core method development in this research does not involve LLMs as any important, original, or non-standard components.
    - Please refer to our LLM policy in the NeurIPS handbook for what should or should not be described.