

Uncovering and Fixing Collider Bias in Bayesian PINNs

Michael Obermayr
TU Graz

Robert Peharz
TU Graz

Abstract

Bayesian physics-informed neural networks (B-PINNs) are a popular framework for parameter and state inference from sparse or noisy observations. They are commonly formulated via a *collider structure*, in which physical and trajectory parameters are assumed to be a priori independent and become coupled through virtual likelihoods on differential-equation residuals that enforce physical consistency. We show that this modeling choice can induce severe systematic bias in the posterior over physical parameters: even when the prior is favorably centered on the ground-truth parameters, the resulting posterior can drift away and concentrate far from them. As a remedy, we advocate a hierarchical *chain model* in which physics generates trajectories, which in turn generate observations. The chain model does not suffer from this posterior bias, but it poses a harder, so-called doubly intractable, inference problem due to a physics-dependent normalization constant. This challenge can be resolved by discretizing the underlying stochastic dynamics, after which the chain posterior can be sampled exactly with particle MCMC. We identify two distinct mechanisms characterizing the collider bias, derive analytical approximations of their magnitudes, and establish diagnostic criteria for predicting when standard B-PINNs remain reliable. Experiments confirm the predicted bias and show that the chain formulation successfully avoids it.

1 INTRODUCTION

Inferring unknown physical parameters from dynamical systems is a fundamental challenge across science and engineering. Physics-informed neural networks (PINNs) [22] tackle this inverse problem by training neural trajectory surrogates to jointly minimize data

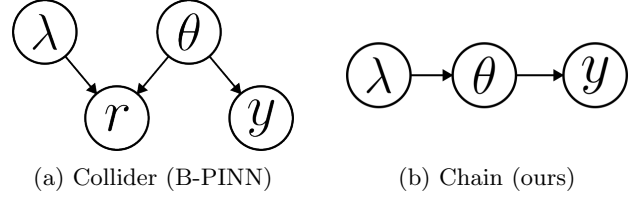


Figure 1: Modeling choices for physics-informed Bayesian inference. **(a) Collider (B-PINN):** Physics parameters λ and trajectory parameters θ are a priori independent, interacting only when conditioned on equation residuals r . **(b) Chain (ours):** Physics λ directly generates trajectories θ , which in turn generate data y .

mismatch and physical law violations. Under limited and noisy data, a Bayesian treatment naturally quantifies uncertainty over trajectories and parameters. Yet, their assumed dependency structure requires careful consideration, which is the focus of this paper.

Bayesian PINNs (B-PINNs) [29] assume the factorization

$$p_{\text{collider}}(\lambda, \theta, r, y) = \underbrace{p(\lambda)}_{\text{Physics Prior}} \underbrace{p(\theta)}_{\text{Traj. Prior}} \underbrace{p(r | \lambda, \theta)}_{\text{Residual Likelihood}} \underbrace{p(y | \theta)}_{\text{Observ. Likelihood}}. \quad (1)$$

Here $p(\lambda)$ and $p(\theta)$ are the priors over physics and trajectory parameters, respectively, and $p(y | \theta)$ is the observation likelihood. Residuals r serve as “virtual observations” of zero measuring physics violations at collocation points, typically under a narrow zero-mean Gaussian likelihood $p(r | \lambda, \theta)$. Thus, λ and θ are a priori independent and coupled only upon conditioning on $r = \mathbf{0}$. In graphical-model terms [13], this forms a *collider* at r (Fig. 1a), a structure widely adopted in the literature [14, 20, 25, 26].

Alternatively, one might use a hierarchical model, in which physics λ is equipped with a prior, trajectories θ depend on λ and observations y depend on θ :

$$p_{\text{chain}}(\lambda, \theta, y) \propto \underbrace{p(\lambda)}_{\text{Physics Prior}} \underbrace{p(\theta | \lambda)}_{\text{Trajectory Likelihood}} \underbrace{p(y | \theta)}_{\text{Observation Likelihood}}. \quad (2)$$

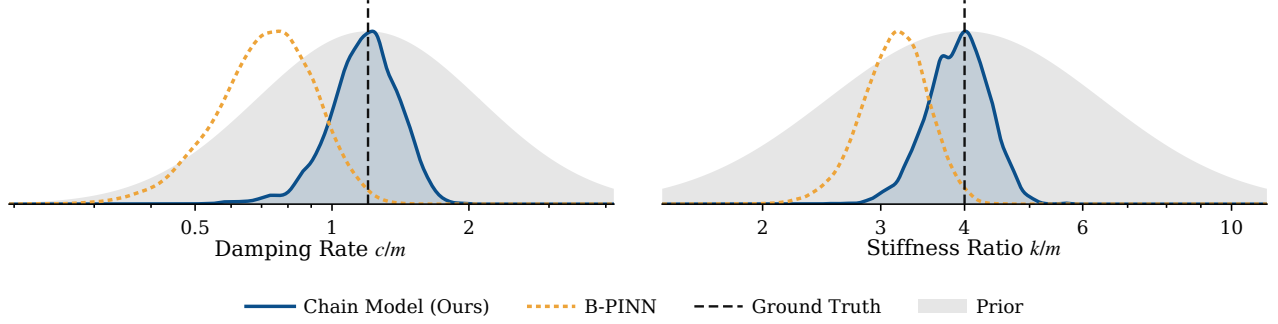


Figure 2: **Posterior marginals for the damped linear oscillator.** System governed by $\ddot{x} = -\frac{c}{m}\dot{x} - \frac{k}{m}x + \sigma_r \xi(t)$ with white noise $\xi(t)$, mass m , damping c and spring constant k , ground-truth parameters $\{m, c, k\} = \{1.0, 1.2, 4.0\}$, residual scale fixed at $\sigma_r = 0.3$, and 40 noisy observations ($\sigma_{\text{obs}} = 0.02$). Posteriors are sampled via NUTS (24,000 samples) for B-PINN and PMMH [2] (14,000 samples) for our proposed chain model. Although all three physical parameters are inferred jointly, only the ratios c/m and k/m are identifiable; we therefore display their corresponding marginal posteriors. Despite initialization at the ground truth and identical priors centered at the ground truth, only the chain model accurately recovers the true values. In contrast, B-PINN exhibits a systematic collider bias, underestimating dissipation (c/m) and stiffness (k/m). The bias is reproducible over resampled datasets. Full experimental details are provided in Section 5.

Here the dependency structure between λ and θ forms a *chain* (head-to-tail node) [13], as illustrated in Fig. 1b. Unlike the collider (1), θ depends directly on λ through the *trajectory likelihood* $p(\theta | \lambda)$ rather than indirectly through residual likelihood $p(r | \lambda, \theta)$.

Which of the two formulations is preferable?

The collider offers practical convenience: its factors are straightforward Gaussian priors and likelihoods. By contrast, specifying $p(\theta | \lambda)$ in the chain is less straightforward; it requires describing a distribution over trajectory parameters θ that may vary substantially with the physics λ . An intuitive construction is to start from the prior $p(\theta)$, similar as in the collider formulation, reweight it according to the exponential of the negative squared residual error, and renormalize the result. As detailed in Section 2, this renders the chain model (2) nearly identical to the collider model (1), up to a *physics-dependent* normalization constant Z_λ whose calculation requires an intractable marginalization over θ .

Despite the computational simplicity of the collider, the chain should be preferred. Conceptually, the residuals r in the collider formulation act merely as auxiliary variables to couple physics λ and trajectories θ , rather than quantities of substantive interest. The chain avoids this artificial construction by modeling trajectory dependence directly via $p(\theta | \lambda)$. Furthermore, forward-sampling the collider yields an implausible generative process—typically generating incompatible pairs of λ and θ that demand unlikely residuals. By contrast, the chain follows the natural

hierarchy in which physics λ generates trajectories θ , which in turn generate observations y .

Crucially, the collider systematically biases the posterior over physical parameters. To demonstrate this, we sample synthetic data from a damped linear oscillator with known ground truth, center prior modes at the true values, and sample the collider posterior with NUTS [12], using Fourier features to prevent spectral bias [27]. Despite these favorable conditions, the collider marginals *drift substantially from the ground truth* (Fig. 2), by about two posterior standard deviations in c/m . This is alarming, as the collider model can return confidently wrong estimates of the underlying physics. By contrast, our proposed chain inference exhibits the expected behavior: its posterior concentrates at the ground truth.

As shown in Section 2, this bias stems directly from the physics-dependent normalization constant Z_λ , which systematically favors less restrictive physics in the collider setting. While the chain formulation eliminates this bias, it yields a challenging, *doubly-intractable problem* [17]. Despite the prevalence of collider-based approaches in physics-informed neural inference, this structural bias remains largely unaddressed. To our knowledge, the only exceptions are Alberts, Hao, and Bilonis [1, 10, 11], who employ a similar chain structure and trace the posterior bias to Z_λ ; however, they address the doubly-intractable problem via computationally demanding nested inference without further analyzing the bias. Still, their formulation has gained

little traction in the community and the collider formulation remains the default in B-PINN practice [16, 24].

2 PROBLEM FORMULATION AND PROBABILISTIC MODELING

We consider parameterized second-order ordinary differential equations (ODEs) governing a trajectory $x(t) \in \mathbb{R}^d$:

$$M_\lambda(x, \dot{x}) \ddot{x} = F_\lambda(x, \dot{x}, t), \quad (3)$$

where $\dot{x}$ and $\ddot{x}$ denote the first and second time derivatives, $M_\lambda(x, \dot{x}) \in \mathbb{R}^{d \times d}$ is an invertible generalized mass matrix, and $F_\lambda(x, \dot{x}, t) \in \mathbb{R}^d$ is a force field driving the system, both parameterized by unknown physical parameters $\lambda \in \mathbb{R}^p$. Equation (3) is the general form of a broad class of second-order dynamical systems arising in classical mechanics, structural dynamics, and robotics [8, 18].

In practice, the trajectory $x(t)$ is rarely observed directly and is therefore treated as latent. Instead, we only have access to sparse, noisy observations $y = \{y_k\}_{k=1}^K$ at times t_k , governed by a Gaussian measurement model

$$y_k = x(t_k) + \epsilon_k, \quad \epsilon_k \sim \mathcal{N}(0, \Sigma_{\text{obs}}), \quad (4)$$

with known observation covariance Σ_{obs} .

Beyond the measurement, the dynamics themselves are typically uncertain: simplifications and unmodeled degrees of freedom inevitably perturb the ideal dynamics (3). We account for these discrepancies by introducing process noise on the velocity $v := \dot{x}$. With this, the phase-space state $z := (x^\top, v^\top)^\top \in \mathbb{R}^{2d}$ evolves according to an Itô stochastic differential equation (SDE)

$$dz_t = \underbrace{\begin{bmatrix} v_t \\ a_\lambda(z_t, t) \end{bmatrix}}_{=: f_\lambda(z_t, t)} dt + \begin{bmatrix} 0 \\ A_\lambda(z_t)^{-1} \Sigma_r^{1/2} \end{bmatrix} dW_t, \quad (5)$$

driven by standard Brownian motion $W_t \in \mathbb{R}^d$. Here, $a_\lambda(z, t) := M_\lambda(x, v)^{-1} F_\lambda(x, v, t)$ is the acceleration field and f_λ the drift. The physical disturbances are modeled as constant white noise with covariance Σ_r , and the invertible, possibly state- and parameter-dependent matrix $A_\lambda(z) \in \mathbb{R}^{d \times d}$ maps them into velocity increments.

The choice of A_λ determines how process noise enters the dynamics: setting $A_\lambda = I_d$ injects white noise directly into the accelerations, whereas setting $A_\lambda = M_\lambda$ models the perturbations as stochastic forces. The latter is more realistic for many mechanical systems, where physical disturbances act as fluctuating generalized forces or torques. As we show in Section 3, parameter dependence in A_λ is one of the factors contributing to collider bias in the inferred physical parameters.

With process noise present ($\Sigma_r \succ 0$), particular physics parameters λ induce a distribution $p(x | \lambda)$ over trajectories rather than a single deterministic trajectory. Our primary goal is to infer the marginal posterior distribution over the physical parameters given sparse observations of the trajectory:

$$p(\lambda | y) \propto p(\lambda) \int p(y | x) p(x | \lambda) dx. \quad (6)$$

Following standard B-PINN practice, we fix the residual scale Σ_r as a hyperparameter rather than inferring it alongside λ [14, 29].

2.1 Physics-Informed Residuals and Energy

To connect the trajectory distribution induced by the stochastic dynamics (5) with physics-informed inference, we express deviations from the governing dynamics through a residual energy. In physics-informed machine learning, the continuous trajectory is represented by a differentiable surrogate $x_\theta(t)$ parameterized by θ (e.g., neural network weights). At a set of collocation points $t_1 < \dots < t_N$, we define the physics residual

$$r_n(\theta, \lambda) := A_\lambda(z_\theta(t_n)) (\ddot{x}_\theta(t_n) - a_\lambda(z_\theta(t_n), t_n)), \quad (7)$$

where $z_\theta := (x_\theta^\top, \dot{x}_\theta^\top)^\top$. To relate these pointwise residuals to the process-noise model, suppose the collocation points are uniformly spaced with step size h . Over an interval of width h , the Brownian increment driving (5) has covariance $h\Sigma_r$. Equivalently, the corresponding average stochastic forcing over that interval has covariance Σ_r/h . This motivates modeling the residuals at the collocation points as Gaussian deviations with covariance Σ_r/h , yielding the quadratic residual energy

$$E[\theta, \lambda] = \frac{h}{2} \sum_{n=1}^N r_n(\theta, \lambda)^\top \Sigma_r^{-1} r_n(\theta, \lambda). \quad (8)$$

Thus, the stochastic dynamics induce a natural quadratic energy measuring the compatibility between a trajectory x_θ and physical parameters λ . This residual energy is the common ingredient underlying the probabilistic formulations considered next.

2.2 Probabilistic Models: Collider vs. Chain

We can now define the two probabilistic models—collider (1) and chain (2)—in a parallel manner and using similar components. Both factorizations share the same prior $p(\lambda)$ over physical parameters and the same observational likelihood,

$$p(y | \theta) = \prod_{k=1}^K \mathcal{N}(y_k | x_\theta(t_k), \Sigma_{\text{obs}}). \quad (9)$$

The key difference between them is how the residual energy (8) is used to couple the trajectory parameters θ and the physical parameters λ .

Collider formulation. Standard Bayesian PINNs [14, 25, 26, 29] place an independent prior $p(\theta)$ on the trajectory parameters and enforce the physical law through auxiliary residuals treated as pseudo-observations equipped with a narrow zero-mean Gaussian:

$$\begin{aligned} p(r = \mathbf{0} | \lambda, \theta) &= \prod_{n=1}^N \mathcal{N}(\mathbf{0} | r_n(\theta, \lambda), \frac{1}{h} \Sigma_r) \\ &= \frac{1}{Z_r} \exp(-E[\theta, \lambda]), \end{aligned} \quad (10)$$

where the Gaussian normalization constant $Z_r = \det(2\pi\Sigma_r/h)^{N/2}$ depends on Σ_r , but is invariant to λ and θ . Because λ and θ are assumed a priori independent and interact only via conditioning on $r = \mathbf{0}$, the residual acts as a collider node in the underlying directed acyclic graph (Fig. 1a).

Chain formulation. In contrast, the hierarchical chain formulation follows the generative direction $\lambda \rightarrow \theta \rightarrow y$ (Fig. 1b) by defining a conditional distribution over trajectories,

$$p(\theta | \lambda) = \frac{1}{Z_\lambda} \exp(-E[\theta, \lambda]) p(\theta). \quad (11)$$

Here, $p(\theta)$ is interpreted as a reference density over the trajectory parametrization, which can encode preferences for certain trajectories. Unlike in the collider formulation, it is not a prior over θ ; instead, the exponential factor tilts this reference density toward trajectories that are compatible with the λ . Normalizing the resulting conditional distribution requires the normalization constant

$$Z_\lambda = \int \exp(-E[\theta, \lambda]) p(\theta) d\theta, \quad (12)$$

which generally *depends on the physical parameters* λ .

Model Difference and The Collider Bias. The two formulations differ only through their normalization factors and we can write:

$$p_{\text{collider}}(\lambda, \theta, r = \mathbf{0}, y) = \frac{Z_\lambda}{Z_r} p_{\text{chain}}(\lambda, \theta, y). \quad (13)$$

Since Z_r is constant with respect to λ , this difference carries directly to the posterior over the physical parameters:

$$\Delta(\lambda) := \log \frac{p_{\text{collider}}(\lambda | y, r = \mathbf{0})}{p_{\text{chain}}(\lambda | y)} = \log Z_\lambda + C, \quad (14)$$

where we call the model difference $\Delta(\lambda)$ the *collider bias*. It is the mechanism underlying the posterior drift observed in Fig. 2. The term *bias* is deliberate: Δ displaces the posterior from the data-generating parameters, so that the collider’s credible intervals miss the ground truth, not merely the chain posterior (Section 5). The partition function Z_λ measures the amount of trajectory space, weighted by the reference density $p(\theta)$, that is compatible with the dynamics under λ . The collider formulation therefore reweights the desired posterior by Z_λ , favoring λ that admit a larger volume of low-residual trajectories, that is, less restrictive physics.

3 ANALYSIS OF THE COLLIDER BIAS

For B-PINNs with a neural trajectory surrogate, the collider bias $\Delta(\lambda)$ decomposes into two distinct mechanisms. In Section 4.2, we show that adopting a Markovian state-space latent model eliminates one mechanism and renders the other analytically tractable.

Laplace approximation of the collider bias. Because the partition function (12) is analytically intractable for neural surrogates, we place a Gaussian prior $p(\theta) = \mathcal{N}(0, P^{-1})$ on the weights and apply a Laplace expansion around the data-free mode $\theta^*(\lambda)$. The collider bias (14) evaluates to

$$\Delta(\lambda) \approx -\frac{1}{2} \log \det(G_\lambda + P) + C, \quad (15)$$

where G_λ is the Gauss–Newton curvature of the residual energy (8) at θ^* . The residual constrains $n_{\text{eff}} := \text{tr}[G_\lambda(G_\lambda + P)^{-1}] \leq \text{rank } G_\lambda \leq \min(Nd, \dim \theta)$ directions. Evaluating this curvature along the nominal trajectory $z(t) := z_{\theta^*}(t)$ over $t \in [0, \mathcal{T}]$ decomposes (15) into:

$$\begin{aligned} \Delta(\lambda) \approx & \underbrace{-\frac{n_{\text{eff}}}{Nd} \sum_{n=1}^N \log |\det A_\lambda(z(t_n))|}_{\text{(1) Residual metric bias}} \\ & \underbrace{-\frac{1}{2} \int_0^{\mathcal{T}} \sum_k |\text{Re } s_k(t)| dt}_{\text{(2) Dissipation bias}} + C, \end{aligned} \quad (16)$$

where $\{s_k(t)\}$ are the eigenvalues of the drift Jacobian Df_λ along $z(t)$ (full derivation in Appendix A.1). Eq. (16) explains the origin and direction of the bias.

Derivation Assumptions. This decomposition relies on four simplifying assumptions (Appendix A.1): (i) the surrogate closely satisfies the ODE at θ^* , with

the prior penalty depending only weakly on λ ; (ii) G_λ is evaluated along a single nominal trajectory; (iii) the residual loss dominates the prior along constrained directions ($n_{\text{eff}} \approx \text{rank } G_\lambda$ locally constant); and (iv) the residual metric and drift decouple additively. We also treat A_λ and Df_λ as piecewise constant in time, and assume the surrogate bandwidth Ω exceeds the system’s characteristic dynamical rates.

Mechanism 1: Residual metric bias. Because Z_r normalizes over residual space, it omits the Jacobian determinant $|\det A_\lambda|$ of the acceleration-to-residual mapping, whereas Z_λ accounts for it along all n_{eff} constrained directions. Consequently, acceleration residuals ($A_\lambda = I_d$) avoid this bias, whereas force residuals ($A_\lambda = M_\lambda$) add $-\frac{n_{\text{eff}}}{Nd} \sum_{n=1}^N \log|\det M_\lambda(z(t_n))|$, shifting posterior mass toward vanishing mass determinants. The residual metric is dictated by where the noise enters the dynamics (Section 2.1); switching to acceleration residuals for convenience changes the noise model and is therefore not a free choice.

This pathology worsens along drift-invariant parameter directions where jointly scaling masses and forces preserves drift. Because the bias grows with n_{eff} while the data only constrain scale via finite observations, the artifact can overwhelm the likelihood and collapse the posterior toward vanishing masses (Appendix A.2).

Mechanism 2: Dissipation bias. Lacking Markovian temporal structure, the neural surrogate couples perturbations through the operator $\frac{d}{dt} - Df_\lambda$ nonlocally over $[0, \mathcal{T}]$, generating the second term in (16).

When all perturbations decay ($\text{Re } s_k \leq 0$), the spectral sum reduces to phase-space divergence: $\sum_k |\text{Re } s_k| = -\nabla \cdot f_\lambda = -\text{tr}(\partial a_\lambda / \partial v)$. The dissipation bias simplifies to

$$\Delta_{\text{diss}}(\lambda) = \frac{1}{2} \int_0^\mathcal{T} \nabla \cdot f_\lambda(z(t), t) dt + \mathcal{O}(\mathcal{T}/\Omega) + C, \quad (17)$$

whose leading term matches the Onsager–Machlup path-action correction [19]. In dissipative systems ($\nabla \cdot f_\lambda < 0$), $e^{\Delta_{\text{diss}}}$ spuriously penalizes dissipation, underestimating friction, damping, or drag. Crucially, this penalty scales as $\Theta(\mathcal{T})$, so that the B-PINN’s error grows with the record length relative to its own posterior width (Section 5.2). In the presence of growing modes, it analogously penalizes stronger instability (Appendix A.1).

4 TRACTABLE INFERENCE IN THE CHAIN MODEL

With a global surrogate, the chain model yields a doubly intractable posterior [17], usually requiring expen-

sive inner-loop re-estimation of Z_λ via variational inference or Langevin dynamics that often diverges (Appendix A.4). Instead of attempting global integration, we discretize the generative SDE [23]. This factorizes Z_λ into analytic per-step normalizers (21), enabling asymptotically exact inference via particle MCMC.

4.1 The Chain Model as a State-Space Model

Discretizing $[0, \mathcal{T}]$ into T intervals of width h defines states $z_i = (x_i^\top, v_i^\top)^\top$ at $t_i = ih$ and applying a semi-implicit Euler–Maruyama scheme to (5) yields:

$$\begin{aligned} v_i &= v_{i-1} + h a_\lambda(z_{i-1}, t_{i-1}) + \sqrt{h} A_\lambda(z_{i-1})^{-1} \Sigma_r^{1/2} \xi_i, \\ x_i &= x_{i-1} + h v_i, \end{aligned} \quad (18)$$

with independent increments $\xi_i \sim \mathcal{N}(0, I_d)$. This discretization induces a Markovian trajectory prior over $\theta = z_{0:T}$:

$$\begin{aligned} p(z_{0:T} | \lambda) &= p(z_0) \prod_{i=1}^T K_\lambda(z_i | z_{i-1}), \\ K_\lambda(z_i | z_{i-1}) &= \delta(x_i - x_{i-1} - h v_i) \\ &\quad \cdot \mathcal{N}(v_i | \bar{v}_i, h A_\lambda^{-1} \Sigma_r A_\lambda^{-\top}), \end{aligned} \quad (19)$$

where $\bar{v}_i := v_{i-1} + h a_\lambda(z_{i-1}, t_{i-1})$ and A_λ is evaluated at z_{i-1} . Unlike (11), this prior admits analytic, per-step normalization. Combined with the prior $p(\lambda)$ and observations at $\{t_k\} \subseteq \{t_i\}_{i=1}^T$, this completes the chain model (2) over $(\lambda, z_{0:T})$ to target the exact marginal posterior $p(\lambda | y)$.

4.2 Exact Collider Bias on the State-Space Latent

On the Markovian latent state $\theta = z_{0:T}$, the collider bias can be evaluated in closed form. The velocity update in (18) corresponds to the discrete residual

$$r_i = A_\lambda(z_{i-1}) \left(\frac{v_i - v_{i-1}}{h} - a_\lambda(z_{i-1}, t_{i-1}) \right), \quad (20)$$

making the transition kernel (19) proportional to $\exp(-\frac{h}{2} r_i^\top \Sigma_r^{-1} r_i)$, which matches the summand of (8) for $N = T$. Because r_i is affine in v_i , each one-step transition normalizes analytically:

$$\begin{aligned} Z_\lambda(z_{i-1}) &:= \int \exp\left(-\frac{h}{2} r_i^\top \Sigma_r^{-1} r_i\right) dv_i \\ &= \frac{(2\pi h)^{d/2} \det(\Sigma_r)^{1/2}}{|\det A_\lambda(z_{i-1})|}. \end{aligned} \quad (21)$$

The drift a_λ acts merely as a mean shift and drops out of the integral, causing the dissipation bias to vanish.

Substituting $\prod_{i=1}^T Z_\lambda(z_{i-1})$ for Z_λ in (13) and subtracting $\log Z_r = \frac{T}{2} \log \det(2\pi\Sigma_r/h)$ gives the exact collider bias:

$$\Delta_{\text{Markov}}(\lambda, z_{0:T-1}) = -\sum_{i=1}^T \log|\det A_\lambda(z_{i-1})| + C. \quad (22)$$

Only the residual metric bias remains: under acceleration residuals ($A_\lambda = I_d$), Δ_{Markov} is independent of λ and both posteriors coincide; under force residuals ($A_\lambda = M_\lambda$), Δ_{Markov} biases the collider posterior toward vanishing mass determinants.

4.3 Exact Parameter Inference via Particle MCMC

For nonlinear dynamics, the marginal likelihood $p(y|\lambda) = \int p(z_{0:T}|\lambda)p(y|z_{0:T})dz_{0:T}$ lacks an analytic form. Rather than approximating it variationally [3], we exploit its Markovian structure, which admits sequential Monte Carlo estimation via a bootstrap particle filter [9]. We propagate P particles forward under (18), reweight by $\mathcal{N}(y_k|x_{t_k}, \Sigma_{\text{obs}})$ at observation times t_k , and adaptively resample [15]. The resulting estimator $\hat{p}(y|\lambda)$ costs $\mathcal{O}(PT)$ and is strictly unbiased. By the pseudo-marginal property [2], substituting $\hat{p}(y|\lambda)$ into the Metropolis–Hastings acceptance ratio guarantees Particle Marginal Metropolis–Hastings (PMMH) targets the exact posterior $p(\lambda|y)$ for any $P \geq 1$, where P governs only chain mixing. Sampling a trajectory alongside each accepted λ' additionally yields exact joint samples from $p(\lambda, z_{0:T}|y)$.

5 EXPERIMENTS

We test the analysis of Section 3 on three systems that isolate its bias mechanisms, and compare our inference scheme with nested inference.

5.1 Systems and Setup

Data are generated from the SDE (5) with isotropic process noise $\Sigma_r = \sigma_r^2 I_d$ fixed at ground truth; physics priors are log-normal with medians at the truth. Jointly scaling masses and forces by α leaves the drift invariant; this *drift-invariant direction* is only weakly identifiable via process noise.

Damped oscillator ($d = 1$): mass m , stiffness k , linear damping c , drift $m\ddot{x} = -c\dot{x} - kx$, truth $(m, c, k) = (1, 1.2, 4)$, and $\sigma_r = 0.3$. Because only c/m and k/m enter the drift, the invariant direction is $(m, c, k) \rightarrow \alpha(m, c, k)$. We observe position $K = 40$ times over $\mathcal{T} = 24$ ($\sigma_{\text{obs}} = 0.02$). Dissipation bias is active ($\nabla \cdot f_\lambda = -c/m < 0$).

Double pendulum ($d = 2$): masses m_1, m_2 on massless rods of lengths ℓ_1, ℓ_2 , with vertical angles $x = (\vartheta_1, \vartheta_2)$. The drift is $M_\lambda(x)\ddot{x} = F_\lambda(x, \dot{x})$ with state-dependent mass matrix M_λ and forces F_λ :

$$M_\lambda = \begin{bmatrix} (m_1+m_2)\ell_1^2 & \mu \cos \delta \\ \mu \cos \delta & m_2\ell_2^2 \end{bmatrix},$$

$$F_\lambda = - \begin{bmatrix} \mu\dot{\vartheta}_2^2 \sin \delta + (m_1+m_2)g\ell_1 \sin \vartheta_1 \\ -\mu\dot{\vartheta}_1^2 \sin \delta + m_2g\ell_2 \sin \vartheta_2 \end{bmatrix},$$

with $\delta := \vartheta_1 - \vartheta_2$, $\mu := m_2\ell_1\ell_2$, and $g = 9.81$. Truth: $(m_1, m_2, \ell_1, \ell_2) = (1, 0.5, 1, 0.8)$, $\sigma_r = 0.02$. Linearity in masses yields the invariant direction $(m_1, m_2) \rightarrow \alpha(m_1, m_2)$. Both angles are observed $K = 20$ times over $\mathcal{T} = 12$ ($\sigma_{\text{obs}} = 0.02 \text{ rad}$). The conservative divergence integrates to a bounded boundary term (38), but chaos yields a $\Theta(\mathcal{T})$ surrogate determinant term (Appendix A.3.6).

Relativistic particle ($d = 1$): rest mass m_0 , charge q , accelerated by $E(t) = 1.81 \cos(\pi t)$ up to $0.5c$. Drift: $m_0\gamma^3\ddot{x} = qE(t)$ with Lorentz factor $\gamma = (1 - \dot{x}^2/c^2)^{-1/2}$; truth: $(m_0, q, c) = (1, 1, 1)$, $\sigma_r = 0.02$. As only q/m_0 governs the drift, the invariant direction is $(m_0, q) \rightarrow \alpha(m_0, q)$. Position is observed $K = 20$ times over $\mathcal{T} = 12$ ($\sigma_{\text{obs}} = 0.005$). The divergence integrates to a bounded boundary term but alternates sign, so $\sum_k |\text{Re } s_k| = |\nabla \cdot f_\lambda|$ grows with $\mathcal{T}$ (Appendix A.3.6).

Evaluation grid. We evaluate two residual metrics—*acceleration* ($A_\lambda = I_d$, metric bias off) and *force/torque* ($A_\lambda = M_\lambda$, metric bias on)—across four inference schemes: **Chain (state space)**: our method (Section 4; PMMH, 1,000–2,000 particles, 24,000 iterations). **Collider (state space)**: differs from Chain solely by omitting the normalizer. **Collider B-PINN**: standard model (Section 2.2) with a Fourier-feature MLP ($\dim \theta \approx 1.4\text{k}$, $Nd \leq 720$), jointly sampled over (θ, λ) via NUTS (4 chains, 9,000 steps). **Chain (Fourier MLP)**: augments the B-PINN with the closed-form correction $-\Delta$ (16) to restore Z_λ . We run this only on the oscillator, where state-independent A_λ and $\nabla \cdot f_\lambda$ allow closed-form evaluation at each NUTS step; in the other systems, state dependence makes it prohibitive (Appendix A.4). Details and code: Appendix A.3 and <https://github.com/obermm/ColliderBias>.

5.2 Collider Bias and Its Removal

In Table 1, the chain model recovers the truth throughout: $|\log \alpha| \leq 0.18$ (≤ 1.1 posterior SDs) and identified parameters within 1.3 SDs (Appendix A.3.3). The collider fails as predicted.

Null condition. With acceleration residuals ($A_\lambda = I_d$) on the Markovian state-space latent, both biases

Table 1: **The chain model recovers the truth in every system; the collider fails where Section 3 predicts.** $\log \alpha$: posterior mean of the log scale factor along the weakly identified drift-invariant direction (Eq. (37); correct: ≈ 0). Other columns: posterior mean of an identified quantity divided by its true value (correct: 1.00). **Bold**: off by more than a factor of two. †: chains did not converge or stalled; median of the draws, for both columns of the system. ‡: exact Kalman likelihood, since PMMH cannot reach this posterior. All parameters in Appendix A.3.3.

Model (latent)	Oscillator		Double Pendulum		Relativistic Particle	
	$\log \alpha$	c/m	$\log \alpha$	m_2/m_1	$\log \alpha$	c
<i>Acceleration residuals ($A_\lambda = I_d$): metric bias inactive</i>						
Chain (state space)	-0.01	0.99	-0.03	0.99	-0.02	0.89
Collider (state space)	-0.01	0.99	-0.03	0.99	-0.02	0.89
Chain (Fourier MLP)	-0.00	1.00	—	—	—	—
Collider B-PINN (Fourier MLP)	-0.00	0.61	-0.03†	0.94	+0.00	1.78
<i>Force / torque residuals ($A_\lambda = M_\lambda$): metric bias active</i>						
Chain (state space)	-0.01	1.01	+0.04	1.00	-0.18	0.89
Collider (state space)	-42.42‡	27	-2.37†	0.30	-8.17†	4.1·10³
Chain (Fourier MLP)	+0.29	0.74	—	—	—	—
Collider B-PINN (Fourier MLP)	-0.93	2.05	-2.79	0.54	-1.88	1.61

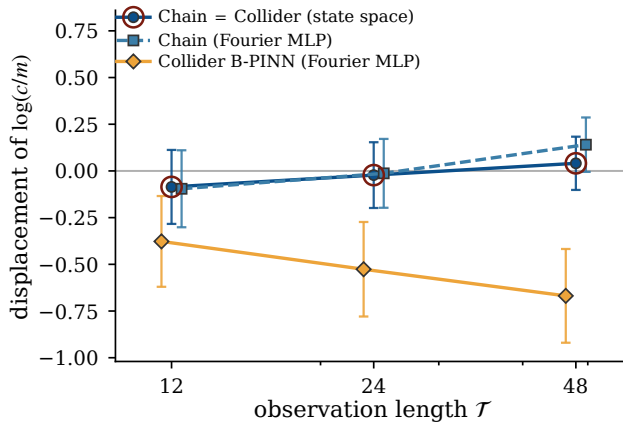


Figure 3: **Longer records amplify the dissipation bias** (oscillator, acceleration residuals). Posterior mean of $\log(c/m)$ minus its true value, with ± 1 posterior standard deviation. The chain stays near the truth on both latents, and its posterior contracts as $\mathcal{T}$ grows; on the state-space latent the collider coincides with it (open rings). The B-PINN drifts further from the truth while its posterior width stays constant, so its error grows from 1.55 to 2.67 standard deviations.

are inactive: $\Delta(\lambda)$ in Eq. (22) is constant, yielding bit-identical chain and collider draws.

Mechanism 1: residual metric bias. Under force residuals, the metric term in Eq. (16) rewards small masses by adding $-Td \log \alpha$ to the state-space log-density and $\approx -n_{\text{eff}} \log \alpha$ to the B-PINN, opposed along the invariant ray only by process noise (Appendix A.2). On state-space latents ($T = 1,248$ for the

oscillator), this reward dominates: the Kalman posterior collapses to $\log \alpha = -42.4$, and PMMH drifts toward zero mass until Euler integration fails. Measured $\Delta(\lambda)$ values match Eq. (22) to machine precision (Appendix A.3.8). On the B-PINN, a weaker reward ($n_{\text{eff}} \approx 110$) is checked by noise, yet mass scales are still underestimated by factors of 2.5 to 16. The damage is not confined to the drift-invariant direction: since $\det M_\lambda \propto \ell_1^2 \ell_2^2$ also rewards short rods, the pendulum B-PINN shrinks ℓ_1 and ℓ_2 to 0.14 and 0.27 of their true values, so removing the scale ambiguity by reparameterization would not remove the bias.

Mechanism 2: dissipation bias. Omitting divergence (17) biases B-PINNs toward underestimating dissipation. On the oscillator with acceleration residuals, B-PINN underestimates damping ($c/m = 0.73$ vs. true 1.20; Fig. 2). This is systematic: across ten dataset resamples, B-PINN underestimates $\log(c/m)$ by 0.42 ± 0.03 and lies below the truth on every dataset, with 90% credible intervals that cover the truth in only 4 of 10 cases, whereas our chain’s intervals always do (Appendix A.3.4). With σ_τ matched to the data, the bias persists for four tested $\sigma_\tau \in [0.05, 1]$ (Appendix A.3.5). For the relativistic particle, B-PINN overestimates the speed of light ($c = 1.78$ vs. 1.00): although the divergence integral vanishes, its absolute path integral does not, penalizing smaller c by 9.5 nats per e-fold (Appendix A.3.6). The pendulum is unassessed, as its B-PINN does not converge.

Longer records amplify dissipation bias. As the penalty scales as $\Theta(\mathcal{T})$, longer horizons ($\mathcal{T} = 12 \rightarrow 48$) push B-PINN damping estimates further

from the truth without posterior contraction (Fig. 3). Reweighting the chain posterior by $e^{\Delta_{\text{diss}}}$ predicts the bias’s direction, growth, and 60–75% of its size (Appendix A.3.8).

Restoring the normalizer on the surrogate.

Chain (Fourier MLP) adds $-\Delta$ to an otherwise identical B-PINN. Under acceleration residuals on the oscillator, it removes 97.5% of the $\log(c/m)$ bias. Under force residuals, where metric correction uses approximate n_{eff} , it eliminates two-thirds of the mass bias (overshooting slightly; Appendix A.3.7). Thus, the missing normalizer, not the neural surrogate, causes the pathology.

Comparison with nested inference. PMMH is up to $13\times$ faster than the B-PINN, depending on the system state size (Appendix A.3.1). We further benchmark PMMH against unmodified public implementations of NSVI [10] and NPSGLD [11] on Euler–Maruyama latents with closed-form Z_λ (Appendix A.4). PMMH converges across all settings (0.51–6.71 s/ESS). NPSGLD fails to converge, requiring $21\times$ – $2,900\times$ more compute per ESS while underestimating posterior SDs by up to $30\times$. NSVI converges to a biased damping ratio on the oscillator ($c/m \approx 0.80$ vs. 1.20) and raising its compute by a factor of 100 does not remove this bias.

6 CONCLUSIONS

B-PINNs suffer from an overlooked structural defect: conditioning on virtual residual observations turns physical parameters λ and trajectory surrogates into colliders. This implicitly neglects the parameter-dependent normalizing constant Z_λ , systematically biasing posterior inference. Even when initialized at the ground truth, standard B-PINNs underestimate damping and mass, and grow increasingly overconfident with longer observation records.

We approximated two bias mechanisms in closed form, confirmed them across three systems, yet showed that analytical corrections can remain insufficient. To eliminate collider bias entirely, we reformulate the trajectory as a Markovian state-space model. This enables an asymptotically exact particle MCMC scheme that avoids both the computational overhead of nested Langevin dynamics [11] and the variational gap of nested VI [10].

When are B-PINNs safe? Collider B-PINNs yield unbiased posteriors if two null conditions hold simultaneously (Section 3; Appendix Table 2): (i) the residual metric A_λ is parameter-independent (e.g., measuring acceleration rather than force); and (ii) the dissipation

bias (17) is negligible, with Δ_{diss} remaining roughly constant across posterior support. Otherwise, practitioners must adopt a Markovian chain formulation or, when A_λ and $\nabla \cdot f_\lambda$ are state-independent, apply the closed-form correction $-\Delta$ directly.

Limitations and Future Work. Our particle MCMC scheme targets only stochastic ODEs ($\Sigma_r \succ 0$) in low-to-moderate dimensions; particle filters degenerate as $\sigma_r \rightarrow 0$ and face weight collapse on finely discretized PDEs. Because collider bias persists in PDE systems, extending our framework to spatial domains will require scalable alternatives such as ensemble Kalman or score-based filtering. Furthermore, selecting Σ_r to prevent residual misspecification remains open. Ideally Σ_r would be inferred jointly with λ , but this introduces a third collider-bias mechanism that we leave to future work.

Implications. Although standard B-PINNs rely on collider structures, the resulting bias has gone largely unnoticed, as they appear well-behaved when the null conditions approximately hold. When violated, MCMC still converges cleanly to confidently wrong posteriors whose error is undetectable without ground truth. Crucially, this distortion peaks in weakly identified regimes and long horizons with sparse data—precisely when Bayesian rigor matters most. Practitioners should therefore check these null conditions and prefer a chain formulation.

ACKNOWLEDGEMENTS

This research was funded by the Austrian Research Promotion Agency (FFG) via the project VENTUS (910263) under the FFG AI for Green program.

References

- [1] Alex Alberts and Ilias Bilonis. Physics-informed information field theory for modeling physical systems with uncertainty quantification. *Journal of Computational Physics*, 486:112100, 2023. doi: 10.1016/j.jcp.2023.112100.
- [2] Christophe Andrieu, Arnaud Doucet, and Roman Holenstein. Particle Markov chain Monte Carlo methods. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 72(3):269–342, 2010. doi: 10.1111/j.1467-9868.2009.00736.x.
- [3] Cédric Archambeau, Manfred Oppel, Yuan Shen, Dan Cornford, and John Shawe-Taylor. Variational inference for diffusion processes. In J. Platt, D. Koller, Y. Singer, and S. Roweis, editors, *Advances in Neural Information Processing Systems*, volume 20. Curran Associates, Inc., 2007. URL https://proceedings.neurips.cc/paper_files/paper/2007/file/818f4654ed39a1c147d1e51a00ffb4cb-Paper.pdf.
- [4] Eli Bingham, Jonathan P. Chen, Martin Jankowiak, Fritz Obermeyer, Neeraj Pradhan, Theofanis Karaletsos, Rohit Singh, Paul A. Szerlip, Paul Horsfall, and Noah D. Goodman. Pyro: Deep universal probabilistic programming. *Journal of Machine Learning Research*, 20(28):1–6, 2019.
- [5] Maxwell Bolt. niff-replication: A JAX reimplementation of the Neural Information Field Filter. <https://github.com/BoltMaxwell/niff-replication>, 2026. Commit dab4403.
- [6] James Bradbury, Roy Frostig, Peter Hawkins, Matthew James Johnson, Chris Leary, Dougal Maclaurin, George Necoala, Adam Paszke, Jake VanderPlas, Skye Wanderman-Milne, and Qiao Zhang. JAX: composable transformations of Python+NumPy programs. <https://github.com/google/jax>, 2018.
- [7] Detlef Dürr and Alexander Bach. The Onsager–Machlup function as Lagrangian for the most probable path of a diffusion process. *Communications in Mathematical Physics*, 60(2):153–170, 1978. doi: 10.1007/BF01609446.
- [8] Herbert Goldstein, Charles Poole, and John Saffo. *Classical Mechanics*. Addison Wesley, third edition, 2002.
- [9] Neil J. Gordon, David J. Salmond, and Adrian F. M. Smith. Novel approach to nonlinear/non-Gaussian Bayesian state estimation. *IEEE Proceedings F (Radar and Signal Processing)*, 140(2): 107–113, 1993. doi: 10.1049/ip-f-2.1993.0015.
- [10] Kairui Hao and Ilias Bilonis. An information field theory approach to Bayesian state and parameter estimation in dynamical systems. *Journal of Computational Physics*, 512:113139, 2024. doi: 10.1016/j.jcp.2024.113139.
- [11] Kairui Hao and Ilias Bilonis. Neural information field filter. *Mechanical Systems and Signal Processing*, 226:112253, 2025. doi: 10.1016/j.ymssp.2024.112253.
- [12] Matthew D. Hoffman and Andrew Gelman. The No-U-Turn sampler: Adaptively setting path lengths in Hamiltonian Monte Carlo. *Journal of Machine Learning Research*, 15(47):1593–1623, 2014. URL <http://jmlr.org/papers/v15/hoffman14a.html>.
- [13] Daphne Koller and Nir Friedman. *Probabilistic Graphical Models: Principles and Techniques*. MIT Press, Cambridge, MA, 2009.
- [14] Kevin Linka, Amelie Schäfer, Xuhui Meng, Zongren Zou, George Em Karniadakis, and Ellen Kuhl. Bayesian physics-informed neural networks for real-world nonlinear dynamical systems. *Computer Methods in Applied Mechanics and Engineering*, 402:115346, 2022. doi: 10.1016/j.cma.2022.115346.
- [15] Jun S. Liu and Rong Chen. Sequential Monte Carlo methods for dynamic systems. *Journal of the American Statistical Association*, 93(443): 1032–1044, 1998. URL <http://www.jstor.org/stable/2669847>.
- [16] Hyeonbin Moon, Hanbin Cho, Wabi Demeke, Byungki Ryu, and Seunghwa Ryu. Thermal conductivity estimation of thermoelectric materials with uncertainty quantification using Bayesian physics-informed neural networks, 2025. arXiv:2510.16723.
- [17] Iain Murray, Zoubin Ghahramani, and David J. C. MacKay. MCMC for doubly-intractable distributions. In *Proceedings of the Twenty-Second Conference on Uncertainty in Artificial Intelligence*, pages 359–366. AUAI Press, 2006.
- [18] Richard M. Murray, Zexiang Li, and S. Shankar Sastry. *A Mathematical Introduction to Robotic Manipulation*. CRC Press, 1994.
- [19] Lars Onsager and Stefan Machlup. Fluctuations and irreversible processes. *Physical Review*, 91(6): 1505–1512, 1953. doi: 10.1103/PhysRev.91.1505.
- [20] Andrew Pensoneault and Xueyu Zhu. Efficient Bayesian physics informed neural networks for inverse problems via ensemble Kalman inversion. *Journal of Computational Physics*, 508:113006, 2024. doi: 10.1016/j.jcp.2024.113006.

- [21] Du Phan, Neeraj Pradhan, and Martin Jankowiak. Composable effects for flexible and accelerated probabilistic programming in NumPyro, 2019. arXiv:1912.11554.
- [22] Maziar Raissi, Paris Perdikaris, and George Em Karniadakis. Physics-informed neural networks: A deep learning framework for solving forward and inverse problems involving nonlinear partial differential equations. *Journal of Computational Physics*, 378:686–707, 2019. doi: 10.1016/j.jcp.2018.10.045.
- [23] Simo Särkkä and Arno Solin. *Applied Stochastic Differential Equations*, volume 10 of *Institute of Mathematical Statistics Textbooks*. Cambridge University Press, 2019. doi: 10.1017/9781108186735.
- [24] Khemraj Shukla, Zongren Zou, Theo Kaeufer, Michael Triantafyllou, and George Em Karniadakis. Uncertainty quantification in PINNs for turbulent flows: Bayesian inference and repulsive ensembles, 2026. arXiv:2604.17156.
- [25] Simon Stock, Jochen Stiasny, Davood Babazadeh, Christian Becker, and Spyros Chatzivasileiadis. Bayesian physics-informed neural networks for robust system identification of power systems. In *2023 IEEE Belgrade PowerTech*, pages 1–6. IEEE, 2023. doi: 10.1109/PowerTech55446.2023.10202692.
- [26] Luning Sun and Jian-Xun Wang. Physics-constrained Bayesian neural network for fluid flow reconstruction with sparse and noisy data. *Theoretical and Applied Mechanics Letters*, 10(3):161–169, 2020. doi: 10.1016/j.taml.2020.01.031.
- [27] Sifan Wang, Hanwen Wang, and Paris Perdikaris. On the eigenvector bias of Fourier feature networks: From regression to solving multi-scale PDEs with physics-informed neural networks. *Computer Methods in Applied Mechanics and Engineering*, 384:113938, 2021. doi: 10.1016/j.cma.2021.113938.
- [28] Eugene Wong and Moshe Zakai. On the convergence of ordinary integrals to stochastic integrals. *The Annals of Mathematical Statistics*, 36(5):1560–1564, 1965. doi: 10.1214/aoms/1177699916.
- [29] 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.

A SUPPLEMENTARY MATERIAL

A.1 Laplace Approximation of the Collider Bias for Global Surrogates

This appendix derives the two-term decomposition (16) for a global surrogate such as a neural network. Step 1 approximates Z_λ by Laplace’s method, Step 2 identifies the two factors through which λ enters the resulting curvature, Steps 3 and 4 differentiate with respect to these factors, and Step 5 collects the terms; a worked example for the oscillator follows. Along drift-invariant directions, the metric bias admits exact statements, derived in Appendix A.2.

Setup. We write the residual (7) in continuous time,

$$r_\theta(t, \lambda) := A_\lambda(z_\theta(t)) (\dot{v}_\theta(t) - a_\lambda(z_\theta(t), t)), \quad v_\theta := \dot{x}_\theta, \quad (23)$$

so that $r_n(\theta, \lambda) = r_\theta(t_n, \lambda)$; since v_θ is the derivative of x_θ , only the acceleration equation can be violated. The surrogate carries the weight prior $p(\theta) = \mathcal{N}(0, P^{-1})$ of Section 3, where $P \rightarrow 0$ corresponds to a flat reference measure. Chain and collider share this prior and the energy (8) and differ only in their normalizers, Z_λ in (12) and Z_r in (10), so that $\Delta = \log Z_\lambda - \log Z_r + \text{const}$ by (14). Only Z_λ depends on λ ; since Σ_r is fixed, $\log Z_r$ is a constant.

Step 1: Laplace approximation. Let $\theta^*(\lambda)$ minimize $E[\theta, \lambda] - \log p(\theta)$, the negative logarithm of the data-free integrand of Z_λ . A second-order expansion around θ^* gives

$$\log Z_\lambda \approx -E[\theta^*, \lambda] - \frac{1}{2} \theta^{*\top} P \theta^* - \frac{1}{2} \log \det H_\lambda + \text{const}, \quad H_\lambda := \nabla_\theta^2 E|_{\theta^*} + P, \quad (24)$$

where the constant ($\frac{1}{2} \log \det P$) does not depend on λ . We drop the first two terms, assuming that the surrogate (nearly) solves the ODE, $E[\theta^*, \lambda] \approx 0$, and that θ^* follows an ODE solution rather than the prior mode, so that $\frac{1}{2} \theta^{*\top} P \theta^*$ depends only weakly on λ . This is exact for linear, unforced dynamics and a surrogate with $x_\theta \equiv 0$ at $\theta = 0$, where $\theta^* = 0$ for every λ . Since the residuals vanish at θ^* , so does the part of $\nabla_\theta^2 E$ that is proportional to them, which leaves $H_\lambda \approx G_\lambda + P$ with the Gauss–Newton curvature

$$G_\lambda := h \sum_{n=1}^N \left(\frac{\partial r_\theta(t_n, \lambda)}{\partial \theta} \right)^\top \Sigma_r^{-1} \frac{\partial r_\theta(t_n, \lambda)}{\partial \theta}, \quad \log Z_\lambda \approx -\frac{1}{2} \log \det(G_\lambda + P) + \text{const}. \quad (25)$$

Since $\log Z_r$ is constant, this gives Eq. (15); the remaining steps evaluate its first term.

Step 2: Two factors of the curvature. Let $B(t) := \partial z_\theta(t) / \partial \theta$ be the tangent map of the surrogate at θ^* , and $Df_\lambda(t) := \partial f_\lambda / \partial z$ the drift Jacobian along the fitted trajectory $z(t) := z_{\theta^*}(t)$. By the chain rule,

$$\frac{\partial r_\theta(t, \lambda)}{\partial \theta} \approx A_\lambda(z(t)) (L_\lambda B)_v(t), \quad L_\lambda := \frac{d}{dt} - Df_\lambda(t), \quad (26)$$

where $(\cdot)_v$ denotes the velocity block. The position block of $L_\lambda B$ vanishes because $v_\theta = \dot{x}_\theta$, and the term containing the derivative of A_λ multiplies the residual, which vanishes at θ^* . Stacking over the collocation points gives

$$G_\lambda = \mathbf{B}^\top \mathbf{L}_\lambda^\top \mathbf{A}_\lambda^\top \mathbf{S}^{-1} \mathbf{A}_\lambda \mathbf{L}_\lambda \mathbf{B}, \quad \mathbf{A}_\lambda = \text{diag}(A_\lambda(z(t_1)), \dots, A_\lambda(z(t_N))), \quad \mathbf{S} = h^{-1} \text{diag}(\Sigma_r, \dots, \Sigma_r). \quad (27)$$

The stacked tangent map $\mathbf{B}$ depends on λ only through $\theta^*(\lambda)$, which we hold fixed. Since $\mathbf{S}$ is fixed, λ enters G_λ through two factors: the residual metric $\mathbf{A}_\lambda$ and the drift operator $\mathbf{L}_\lambda$. Since these factors are rectangular, $\log \det(G_\lambda + P)$ does not split into a sum over them; we therefore differentiate it with respect to each factor in turn, holding the other fixed.

Step 3: Residual metric (Mechanism 1). Let $A_\lambda(z) = \kappa_\lambda \bar{A}(z)$ with a scalar $\kappa_\lambda > 0$ and a λ -independent $\bar{A}$, e.g., $A_\lambda = m$ for the oscillator. Then the metric scale factors out of (27),

$$G_\lambda = \kappa_\lambda^2 \tilde{G}, \quad \tilde{G} := \mathbf{B}^\top \mathbf{L}_\lambda^\top \bar{\mathbf{A}}^\top \mathbf{S}^{-1} \bar{\mathbf{A}} \mathbf{L}_\lambda \mathbf{B}, \quad (28)$$

where $\bar{\mathbf{A}} := \text{diag}(\bar{A}(z(t_1)), \dots, \bar{A}(z(t_N)))$, and $\tilde{G}$ does not depend on κ_λ . Since $\partial G_\lambda / \partial \log \kappa_\lambda = 2G_\lambda$, Jacobi's formula gives

$$\frac{\partial}{\partial \log \kappa_\lambda} \left(-\frac{1}{2} \log \det(G_\lambda + P) \right) = -\text{tr}[(G_\lambda + P)^{-1} G_\lambda] = -n_{\text{eff}}. \quad (29)$$

When is n_{eff} constant? For $P \succ 0$, let $\{g_k\}$ be the eigenvalues of $P^{-1/2} \tilde{G} P^{-1/2} \succeq 0$. Then $\det(G_\lambda + P) = \det(P) \prod_k (1 + g_k \kappa_\lambda^2)$, so that

$$-\frac{1}{2} \log \det(G_\lambda + P) = -\frac{1}{2} \sum_k \log(1 + g_k \kappa_\lambda^2) + \text{const}, \quad n_{\text{eff}} = \sum_k \frac{g_k \kappa_\lambda^2}{1 + g_k \kappa_\lambda^2}. \quad (30)$$

Each summand lies in $[0, 1)$ and is nonzero only for $g_k > 0$, so $n_{\text{eff}} < \text{rank } G_\lambda \leq \min(Nd, \dim \theta)$ whenever $G_\lambda \neq 0$. Once the physics dominates the prior on every direction it constrains ($g_k \kappa_\lambda^2 \gg 1$), n_{eff} saturates at $\text{rank } G_\lambda$ and no longer depends on κ_λ ; we call this the saturated regime. For $P = 0$ and full-rank G_λ (which requires $\dim \theta \leq Nd$), $n_{\text{eff}} = \dim \theta$ for every κ_λ .

Mechanism 1 (residual metric). Since Z_r does not depend on κ_λ , Eq. (29) gives $\partial \Delta / \partial \log \kappa_\lambda = -n_{\text{eff}}$. With $\sum_n \log |\det A_\lambda(z(t_n))| = Nd \log \kappa_\lambda + \text{const}$, this integrates, for constant n_{eff} , to the metric term $-\frac{n_{\text{eff}}}{Nd} \sum_n \log |\det A_\lambda(z(t_n))|$ of (16). For a state-dependent $A_\lambda(z(t))$, the term holds locally: on time windows that are short compared with the variation of $A_\lambda(z(t))$ but long compared with the resolution of the surrogate, A_λ is nearly constant, and if the constrained directions are spread uniformly over time and components, each window contributes its share n_{eff}/Nd of the local sum of $\log |\det A_\lambda|$. This neglects the coupling between windows and the anisotropy of A_λ relative to Σ_r , both absent in the scalar case.

State-space latent. On the Markovian latent of Section 4, θ consists of the Td velocities under a flat prior ($P = 0$), and $\partial r / \partial v$ is block lower-triangular with invertible diagonal blocks $A_\lambda(z_{i-1})/h$. Hence $n_{\text{eff}} = Td = Nd$ exactly, and the metric term becomes the exact collider bias (22).

Step 4: Drift operator (Mechanism 2). We set $A_\lambda = I_d$, so that λ enters (27) only through $\mathbf{L}_\lambda$. We assume that the span of the surrogate's tangent map is approximately invariant under L_λ , with a λ -independent Gram determinant, and that the physics dominates P on the resolved modes. Then $\frac{1}{2} \log \det(G_\lambda + P)$ equals $\log |\det L_\lambda|$ restricted to the resolved modes, up to λ -independent terms. On the position component, L_λ acts as $\frac{d^2}{dt^2} - \frac{\partial a_\lambda}{\partial v} \frac{d}{dt} - \frac{\partial a_\lambda}{\partial x}$, whose matrix symbol at frequency ω has, by the Schur complement, the same determinant as $\hat{L}_\lambda(\omega) := i\omega I_{2d} - Df_\lambda$. For constant Df_λ (relaxed below), a surrogate with effective bandwidth Ω on $[0, T]$ resolves real Fourier modes at density $\mathcal{T}/\pi$, so that

$$\frac{1}{2} \log \det(G_\lambda + P) \approx \frac{\mathcal{T}}{\pi} \int_0^\Omega \log |\det \hat{L}_\lambda(\omega)| d\omega. \quad (31)$$

The eigenvalue integral. Let $\{s_k\}_{k=1}^{2d}$ be the eigenvalues of Df_λ , so that $\log |\det \hat{L}_\lambda(\omega)| = \sum_k \log |i\omega - s_k|$. The integral (31) diverges as $\Omega \rightarrow \infty$, but at a rate independent of λ ; differentiating with respect to the eigenvalues isolates the λ -dependent part. For a complex pair $s = -\rho \pm i\beta$ with $\beta > 0$ ($\rho > 0$ damped, $\rho < 0$ growing), $I(\rho, \beta) := \int_0^\Omega \log |(i\omega - s)(i\omega - \bar{s})| d\omega$ satisfies

$$\frac{\partial I}{\partial \rho} = \int_0^\Omega \left(\frac{\rho}{(\omega - \beta)^2 + \rho^2} + \frac{\rho}{(\omega + \beta)^2 + \rho^2} \right) d\omega = \arctan \frac{\Omega - \beta}{\rho} + \arctan \frac{\Omega + \beta}{\rho} = \pi \text{sign}(\rho) + \mathcal{O}(1/\Omega), \quad (32)$$

$$\frac{\partial I}{\partial \beta} = \frac{1}{2} \log \frac{(\Omega + \beta)^2 + \rho^2}{(\Omega - \beta)^2 + \rho^2} = \mathcal{O}(\beta/\Omega), \quad (33)$$

where the boundary terms at $\omega = 0$ cancel. Up to $\mathcal{O}(1/\Omega)$, the damping rate thus enters with weight π and the frequency β not at all, so each pair contributes $\pi|\rho|$. A real eigenvalue $s = -\rho$ gives $\partial_\rho \int_0^\Omega \log |i\omega + \rho| d\omega = \arctan(\Omega/\rho) = \frac{\pi}{2} \text{sign}(\rho) + \mathcal{O}(1/\Omega)$, i.e., a contribution $\frac{\pi}{2}|\rho|$. Every eigenvalue therefore contributes $\frac{\pi}{2}|\text{Re } s_k|$, and multiplying by the mode density $\mathcal{T}/\pi$ gives

$$\frac{1}{2} \log \det(G_\lambda + P) = \frac{\mathcal{T}}{2} \sum_{k=1}^{2d} |\text{Re } s_k| + \text{const} + \mathcal{R}, \quad \mathcal{R} = \mathcal{O}(\mathcal{T}/\Omega). \quad (34)$$

Without growing modes ($\text{Re } s_k \leq 0$ for all k), $\sum_k |\text{Re } s_k| = -\text{tr}(Df_\lambda) = -\nabla \cdot f_\lambda$, and inserting (34) into (15) gives

$$\Delta_{\text{diss}}(\lambda) = \frac{\mathcal{T}}{2} \nabla \cdot f_\lambda + \text{const} + \mathcal{R}. \quad (35)$$

Otherwise, $\Delta(\lambda)$ is not determined by the divergence alone: growing modes are penalized like damped ones, and a divergence that changes sign along the orbit enters through its absolute value. Appendix A.3.6 discusses both cases for conservative systems.

Nonlinear drift. For nonlinear dynamics, we treat $Df_\lambda(z(t))$ as constant on windows that are short compared with $\mathcal{T}$ but long compared with the correlation time Ω^{-1} of the surrogate and the relaxation times $|\text{Re } s_k|^{-1}$. Summing over windows replaces $\mathcal{T} \nabla \cdot f_\lambda$ by $\int_0^\mathcal{T} \nabla \cdot f_\lambda(z(t), t) dt$ and gives Eq. (17). Its leading term is the Onsager–Machlup correction to the path action [7, 19], obtained here as the fluctuation determinant of a band-limited surrogate. The remainder $\mathcal{R}$ collects the $\mathcal{O}(1/\Omega)$ tails of (32) and (33), the residual floor $E[\theta^*, \lambda]$ of a finite-capacity fit, and the windowing. Since all of these grow with $\mathcal{T}$ at the same rate as the leading term, their relative size is set by Ω compared with the time scales of the system, not by the length of the record.

State-space latent. Under the Euler–Maruyama transition (18), v_i depends only on the noise increments $\Delta W_j := \sqrt{h} \xi_j$ up to step i , so the Jacobian of $(\Delta W_1, \dots, \Delta W_T) \mapsto (v_1, \dots, v_T)$ is block lower-triangular with diagonal blocks $A_\lambda(z_{i-1})^{-1} \Sigma_r^{1/2}$, which do not involve the drift. This is why the exact collider bias (22) has no dissipation term: the Markovian latent normalizes the path density step by step (the Itô pre-point form), whereas a global surrogate, which imposes no ordering, yields the midpoint (Onsager–Machlup) form, consistent with the Wong–Zakai limit of smooth approximations to Brownian paths [28].

Step 5: Collecting the terms. Adding the metric term of Step 3 and the dissipation term of Step 4 gives Eq. (16); without growing modes, its dissipation term is the divergence integral (17). Of the four simplifications listed in Section 3, three need a caveat in practice. Simplification (i) fails when the data-free integrand of Z_λ is dominated by the prior rather than the physics; the terms dropped from (24) then depend on λ . For (ii), when $x \equiv 0$ solves the ODE, $\theta^* = 0$, where the tangent map of a neural surrogate degenerates; we therefore evaluate the curvature at a representative state on the solution manifold (Appendix A.3.7). Since the factors in (27) are rectangular, the additivity (iv) holds only to first order. Eq. (16) thus captures the dominant dependence of Δ on λ and shows where the collider bias is large, but adding $-\Delta$ to the log-posterior of a B-PINN need not recover the chain posterior exactly. Table 2 summarizes both mechanisms.

Table 2: Taxonomy of structural biases in collider B-PINNs. Scaling denotes the asymptotic rate of the collider bias $\Delta = \log(p_{\text{collider}}/p_{\text{chain}})$. Null conditions indicate when Δ is invariant to inferred parameters; $\{s_k\}$ are eigenvalues of Df_λ .

Mechanism	Physical Origin	Scaling	Null Condition
Residual Metric	State-space map A_λ depends on λ	$\Theta(n_{\text{eff}})$ or $\Theta(\mathcal{T})$	A_λ invariant to λ
Dissipation	Phase-space volume contraction	$\Theta(\mathcal{T})$	Markovian latent, or $\sum_k \text{Re } s_k $ invariant to λ

Worked example: the damped linear oscillator. For $m\ddot{x} + c\dot{x} + kx = \sigma_r \xi$ with force residuals ($A_\lambda = m$, $d = 1$), the drift $f_\lambda(x, v) = (v, -\frac{k}{m}x - \frac{c}{m}v)^\top$ has the constant Jacobian $Df_\lambda = \begin{bmatrix} 0 & 1 \\ -k/m & -c/m \end{bmatrix}$ with divergence $\nabla \cdot f_\lambda = -c/m$. Its eigenvalues are $-\rho \pm i\beta$ with $\rho = c/(2m)$ and $\beta = \sqrt{k/m - c^2/(4m^2)}$ if underdamped, and real and negative if overdamped; in both cases $\sum_k |\text{Re } s_k| = c/m$. No windowing is needed, and (16) reads

$$\Delta(\lambda) \approx \underbrace{-n_{\text{eff}} \log m}_{\text{metric}} - \underbrace{\frac{c\mathcal{T}}{2m}}_{\text{dissipation}} + \text{const}. \quad (36)$$

The dissipation term penalizes large c/m in proportion to $\mathcal{T}$, so the collider underestimates damping. The stiffness enters only through β , which by (33) contributes no leading-order term; the $\mathcal{O}(\beta/\Omega)$ remainder, however, still grows with $\mathcal{T}$ and rewards smaller k/m , so restoring only the leading-order terms leaves part of the stiffness

bias (Appendix A.3.7). Along the drift-invariant rescaling $(m, c, k) \rightarrow \alpha(m, c, k)$, only the metric term varies, $\partial\Delta/\partial\log\alpha = -n_{\text{eff}}$: the collider is rewarded without bound for shrinking the mass scale, opposed only by the prior and by the process-noise level, which identifies σ_r/m . On the state-space latent, the same holds exactly with $n_{\text{eff}} = N = T$, since (22) gives $\Delta_{\text{Markov}} = -T\log m + \text{const}$; Appendix A.2 discusses its effect on the collider posterior.

A.2 Exact Collider Bias along Drift-Invariant Directions

While the decomposition derived in Appendix A.1 is generally approximate, the metric bias (Mechanism 1) admits exact statements along drift-invariant directions.

Consider a drift-invariant direction parameterized by jointly scaling masses and forces by a factor α . Suppose the scaled components follow independent log-normal priors with log-scale standard deviations s_j . Measuring $\log\alpha$ relative to the prior medians $\tilde{\lambda}_j$ via the precision-weighted mean

$$\log\alpha = \frac{\sum_j s_j^{-2} \log(\lambda_j/\tilde{\lambda}_j)}{\sum_j s_j^{-2}}, \quad (37)$$

the prior factorizes into independent components along $\log\alpha$ and the identified directions, with a Gaussian marginal of standard deviation $(\sum_j s_j^{-2})^{-1/2}$ along $\log\alpha$.

Under acceleration residuals, the residuals themselves remain invariant under this scaling. Consequently, neither the residual energy (8) nor the observational likelihood depends on α , and Z_λ is independent of α . Both the chain and the collider joint then depend on α solely through the prior, so both posteriors leave the distribution of $\log\alpha$ exactly at its prior, for any latent representation.

Under force residuals, $A_\lambda = M_\lambda$ scales with α while Σ_r is fixed, so the process noise on the velocities scales with α^{-1} : the data constrain α through the noise level, and the chain posterior of $\log\alpha$ differs from its prior. On top of this, the collider joint acquires an additional explicit α -dependence: on the state-space latent, $\Delta_{\text{Markov}} = -Td\log\alpha + \text{const}$ exactly by Eq. (22), and on a global surrogate, $\Delta \approx -n_{\text{eff}}\log\alpha + \text{const}$ by the metric term of Eq. (16). The collider posterior of $\log\alpha$ is therefore the chain posterior reweighted by α^{-Td} (respectively $\alpha^{-n_{\text{eff}}}$). This reward grows with the number of time steps (respectively constrained directions), whereas the constraint from the noise level is set by the finitely many observations; on the oscillator, where the exact Kalman likelihood is available, it places the state-space collider at $\log\alpha = -42.4$ (Table 1).

A.3 Experimental Details and Full Results

This appendix supports Section 5 and follows its order. Appendix A.3.1 specifies the setup, and Appendix A.3.2 verifies that our method samples the chain posterior correctly, since every bias in this paper is measured against it. Appendix A.3.3 reports all parameters of the runs summarized in Table 1, and Appendix A.3.4 shows that the bias on the oscillator is reproducible across datasets, and Appendix A.3.5 how it depends on the residual scale. The next two subsections support the discussion of individual mechanisms in Section 5.2: the dissipation term on the two conservative systems (Appendix A.3.6) and the closed-form correction of the neural surrogate (Appendix A.3.7). Appendix A.3.8 details the scaling-law experiments of Figs. 3 and 4.

A.3.1 Setup

Evaluation grid. Each of the three systems is combined with two residual metrics (acceleration and force/torque residuals) and with the inference variants listed below. In every configuration, the process noise is isotropic, $\Sigma_r = \sigma_r^2 I_d$, with σ_r fixed at its true value. Within a system, all runs share the dataset, ground truth and prior, the time discretization, number of particles, sampler budget and random seed, and the network and collocation grid of the surrogate. Two runs that differ in a single factor of the grid therefore isolate the effect of that factor, for instance of the residual metric.

Systems and data. Table 3 lists the settings that differ between systems. The data are simulated with the same semi-implicit Euler–Maruyama scheme (18) that the inference uses, but on a finer time grid: $2\times$ finer for the oscillator and about $50\times$ finer for the pendulum and the relativistic particle (Appendix A.3.2 shows that the results do not rely on this match). The number of Euler–Maruyama steps T used for inference and the number

of particles were chosen per system by a convergence study, and are identical across all state-space runs of that system, so that chain and collider remain comparable.

Table 3: **Per-system settings.** The MLP surrogate maps a Fourier-feature embedding of dimension 33 through two hidden layers of width 24 to the d output coordinates. Nd is the number of residual components, i.e., collocation points times dimension.

	Oscillator	Double pendulum	Relativistic particle
ground truth λ	$(m, c, k) = (1, 1.2, 4)$	$(m_1, m_2, \ell_1, \ell_2) = (1, 0.5, 1, 0.8)$	$(m_0, q, c) = (1, 1, 1)$
residual scale σ_r (fixed)	0.30	0.02	0.02
horizon $\mathcal{T}$	24	12	12
observations K	40	20	20
observation noise σ_{obs}	0.02	0.02 rad	0.005
Euler–Maruyama steps T	1248	1235	12160
particles	1000	1000	2000
MLP layer widths	33-24-24-1	33-24-24-2	33-24-24-1
$\dim \theta$	1441	1466	1441
residual components Nd	128	720	300

Inference variants.

- **Chain (state space)** is our method (Section 4). Its latent is the Euler–Maruyama path $z_{0:T}$ with the normalized transition kernel (19); a bootstrap particle filter estimates the marginal likelihood $\hat{p}(y | \lambda)$, and PMMH samples λ .
- **Collider (state space)** uses the same latent, particle filter and sampler, but omits Z_λ . It is a control that differs from our method only in the normalizer.
- **Collider B-PINN** is the standard formulation of Section 2.2: a Fourier-feature MLP surrogate x_θ with a Gaussian weight prior, sampled with NUTS jointly over (θ, λ) .
- **Chain (Fourier MLP)** is the Collider B-PINN with the closed-form correction $-\Delta$ for the missing normalizer added to its log-density (Appendix A.3.7). It is run on the oscillator only (Section 5.1).

Because the oscillator is linear, the marginal likelihood of its state-space latent is also available exactly from a Kalman filter. We use this in one place. Under force residuals, PMMH cannot reach the posterior of the state-space collider (see below), so on the oscillator we sample it with an adaptive random-walk Metropolis sampler instead, targeting the exact chain likelihood multiplied by $e^{\Delta(\lambda)} = m^{-T}$ from Eq. (22) (4 chains, $\hat{R} \leq 1.001$). These entries are marked with ‡. Applied to the chain model, the same sampler reproduces the PMMH posterior (e.g., $\log m$: 0.014 ± 0.139 against 0.006 ± 0.138).

Priors. All physical parameters have log-normal priors with median at the ground truth. The standard deviations of the log-parameters are 0.4 for masses, charges and the damping coefficient c of the oscillator, and 0.25 for lengths, stiffnesses and the light speed c of the relativistic particle. The Gaussian weight prior of the B-PINN follows LeCun scaling at unit gain: each weight of a layer with n_{in} inputs has variance $1/n_{\text{in}}$, and each bias has variance 1.

Samplers. PMMH runs for 24,000 iterations. The first 4,000 adapt the proposal, and the first 30% of the remaining 20,000 are discarded as burn-in, which leaves 14,000 draws. NUTS runs 4 chains, each with 3,000 warm-up iterations and 6,000 draws at a maximum tree depth of 11, which gives 24,000 draws in total. All samplers use random seed 0. Each run uses a single GPU (NVIDIA Quadro RTX 8000 or NVIDIA L40) on an internal compute cluster. On the oscillator, for instance, PMMH takes 442s and the B-PINN 5,834s (NUTS) for these budgets.

Software and licenses. All methods are implemented in JAX [6]; NUTS and the autocorrelation-based effective sample sizes use NumPyro [4, 21]. The nested baselines of Appendix A.4 use the public reimplementations of Bolt [5], which additionally depends on Optax. JAX, NumPyro and Optax are released under the Apache License 2.0; the reimplementations do not state a license. We release our code under the MIT license.

Reporting conventions. Unless stated otherwise, tables report the posterior mean divided by the ground truth, so that 1.00 means no error. There are two exceptions. Table 1 reports the drift-invariant direction as the precision-weighted log scale factor $\log \alpha$ of Eq. (37). Tables 7 and 9 report *displacements* in log space, i.e., the posterior mean of the log parameter minus the log of its true value. Since the mean of a logarithm is not the logarithm of the mean, these displacements are not exactly the logarithms of the ratios reported elsewhere; for instance, the damping ratio of the Collider B-PINN under acceleration residuals has a displacement of -0.526 (Table 9), and $\exp(-0.526) = 0.59$, whereas Table 1 reports the ratio 0.61.

Runs without valid posterior samples (†). Two kinds of runs do not yield valid posterior samples. The first is the state-space collider under force residuals. Along the drift-invariant direction, its log-density gains $-Td \log \alpha$, which rewards small masses. Together with the log-normal prior, this still defines a proper posterior, but its mode lies far outside the range the sampler can represent; for the oscillator, where the exact Kalman likelihood locates it, at $\log \alpha \approx -42$. PMMH moves toward it until the Euler–Maruyama recursion becomes numerically unstable; from then on, the particle filter returns non-finite likelihoods for a large fraction of the proposals, and only a handful of the 20,000 post-adaptation proposals are accepted. The second kind are B-PINN runs whose chains have not mixed ($\hat{R} > 1.05$). For both kinds, we report the median of the draws, which describes where the sampler stopped rather than the posterior. Marginal plots show such runs as an arrow at the edge of the axis, labelled with the value reached.

A.3.2 Validation of the Reference Posterior

All biases in this paper are measured against the chain posterior that our method samples. Three checks confirm that this reference is correct: the first tests the likelihood estimator, the second the simulation of the data, and the third the sampler.

Particle filter against the exact likelihood. For the linear oscillator, the Kalman filter gives the exact log-likelihood $\log p(y | \lambda)$. At three points along the drift-invariant direction, the particle-filter estimate deviates from it by $+0.14$, -0.10 and $+0.27$ nats, against Monte Carlo standard deviations of 0.27, 0.38 and 0.38 across five independent filter runs. The likelihood estimator underlying PMMH thus shows no bias at the resolution relevant for our comparisons.

Simulation step. Because the data are simulated with the same scheme that the inference uses, the chain model could benefit from a matched discretization. To rule this out, we regenerated ten datasets of the oscillator configuration at a $32\times$ finer step and computed the exact (Kalman) chain posterior on each. Its displacement from the truth is -0.01 ± 0.07 in $\log(c/m)$ and $+0.04 \pm 0.02$ in $\log(k/m)$ (mean $\pm$ standard error over datasets), and its 90% credible intervals contain the truth in 8 or 9 of the 10 datasets for every identified parameter.

A different sampler. On the double pendulum and the relativistic particle, we also sampled both models on the state-space latent with NUTS instead of PMMH. NUTS runs jointly over (z_0, ξ, λ) , where $\xi = (\xi_1, \dots, \xi_T)$ are the standardized noise increments of Eq. (18). The path $z_{0:T}$ is a deterministic function of these variables, and the standard normal prior on ξ does not depend on λ . In this non-centred parameterization, the chain model therefore needs no explicit normalizer, while the collider differs from it by $\Delta(\lambda)$ in (22). The resulting posteriors agree with those obtained with PMMH, and the collapse of the collider under torque residuals is reproduced; neither result is therefore an artefact of the particle filter. NUTS is thus a valid alternative to PMMH for the chain model, but a more expensive one on fine time grids: per effective sample of the drift-invariant scale, it takes 4.9 s against 2.7 s for PMMH on the pendulum (1,235 steps), and about 1,800 s against 6 s on the relativistic particle (12,160 steps; 51 h of NUTS in total).

Table 4: **Oscillator.** The three inference legs under both residual metrics, on one dataset and one prior, with σ_r held at its true value. Posterior mean $\pm$ standard deviation over 14000 draws (state space) and 24000 draws (B-PINN). †: a walk at the Euler map’s stability boundary, not a posterior; its median is quoted; ‡: chains that have not merged ($\hat{R} > 1.05$), likewise a median rather than a posterior.

residual	leg	m (1.00)	c (1.20)	k (4.00)	c/m (1.20)	k/m (4.00)
force residual	Chain (state space)	1.015 ± 0.14	1.209 ± 0.20	3.981 ± 0.51	1.209 ± 0.24	3.948 ± 0.43
	Collider (state space)	$9.3 \cdot 10^{-6} \ddagger$	$3.7 \cdot 10^{-4} \ddagger$	$0.03387 \ddagger$	$39.74 \ddagger$	$3656 \ddagger$
	Collider (B-PINN)	0.2391 ± 0.050	0.5663 ± 0.13	1.832 ± 0.25	2.462 ± 0.73	7.935 ± 1.7
acceleration residual	Chain (state space)	1.035 ± 0.22	1.219 ± 0.28	4.030 ± 0.79	1.191 ± 0.20	3.927 ± 0.41
	Collider (state space)	1.035 ± 0.22	1.219 ± 0.28	4.030 ± 0.79	1.191 ± 0.20	3.927 ± 0.41
	Collider (B-PINN)	1.314 ± 0.28	0.9419 ± 0.24	4.115 ± 0.81	0.7311 ± 0.18	3.162 ± 0.35

Table 5: **Double pendulum.** The three inference legs under both residual metrics, on one dataset and one prior, with σ_r held at its true value. Posterior mean $\pm$ standard deviation over 14000 draws (state space) and 24000 draws (B-PINN). †: a walk at the Euler map’s stability boundary, not a posterior; its median is quoted; ‡: chains that have not merged ($\hat{R} > 1.05$), likewise a median rather than a posterior.

residual	leg	m_1 (1.00)	m_2 (0.500)	ℓ_1 (1.00)	ℓ_2 (0.800)	m_1+m_2 (1.50)	m_2/m_1 (0.500)
force residual	Chain (state space)	1.078 ± 0.26	0.5367 ± 0.13	1.011 ± 0.011	0.7890 ± 0.010	1.615 ± 0.39	0.4976 ± 0.012
	Collider (state space)	$0.1713 \ddagger$	$0.02538 \ddagger$	$1.386 \ddagger$	$0.07589 \ddagger$	$0.1967 \ddagger$	$0.1481 \ddagger$
	Collider (B-PINN)	0.08769 ± 0.020	0.02269 ± 0.0054	0.1441 ± 0.019	0.2137 ± 0.034	0.1104 ± 0.022	0.2709 ± 0.085
acceleration residual	Chain (state space)	1.021 ± 0.30	0.5062 ± 0.15	1.011 ± 0.012	0.7926 ± 0.011	1.527 ± 0.45	0.4958 ± 0.011
	Collider (state space)	1.021 ± 0.30	0.5062 ± 0.15	1.011 ± 0.012	0.7926 ± 0.011	1.527 ± 0.45	0.4958 ± 0.011
	Collider (B-PINN)	$1.004 \ddagger$	$0.4715 \ddagger$	$1.023 \ddagger$	$0.7808 \ddagger$	$1.476 \ddagger$	$0.4699 \ddagger$

A.3.3 Full Posterior Results

Tables 4–6 report the posterior of every parameter for every inference variant, including the parameters and parameter combinations that Table 1 omits. They show the same pattern as Table 1. The chain model keeps the drift-invariant scale near the truth and recovers every identified quantity to within 1.3 posterior standard deviations. The errors of the collider, in contrast, concentrate where Eq. (16) predicts: under force residuals on the parameters that enter the residual metric, and on the parameters that enter the dissipation term (the damping of the oscillator, the light speed of the relativistic particle).

A.3.4 Replication over Datasets

All runs above use a single dataset per system. On one dataset, a posterior can lie away from the truth by chance; a bias means that it does so systematically, on average over datasets. To distinguish the two, we repeat the oscillator configuration of Fig. 2 (acceleration residuals) on independent datasets, each with a new SDE path and new observation noise. We do this in two complementary ways.

Sampled posteriors. We rerun the B-PINN (NUTS) and our chain model (state space, PMMH) with the settings of Appendix A.3.1 on ten datasets, including the one of Fig. 2. All B-PINN runs reach $\hat{R} \leq 1.001$.

Exact posteriors. To remove sampling error and convergence issues altogether, we use ten further datasets and replace the neural network by a surrogate on which all posteriors can be computed exactly. The collider is a B-PINN on a linear cosine basis (65 coefficients with prior standard deviation 0.3, $Nd = 128$ collocation residuals). Because the residual is linear in the coefficients, they can be integrated out in closed form, and Z_λ can be restored exactly on the same basis, which gives a chain model on the surrogate. The exact Kalman likelihood of the Euler–Maruyama latent gives the state-space chain. We evaluate all three posteriors on a grid over the identified coordinates ($\log(c/m), \log(k/m)$), with the log-normal prior of Appendix A.3.1; less than 10^{-3} of each posterior’s mass lies at the boundary of the grid.

Table 6: **Relativistic particle.** The three inference legs under both residual metrics, on one dataset and one prior, with σ_r held at its true value. Posterior mean $\pm$ standard deviation over 14000 draws (state space) and 24000 draws (B-PINN). †: a walk at the Euler map’s stability boundary, not a posterior; its median is quoted; ‡: chains that have not merged ($\hat{R} > 1.05$), likewise a median rather than a posterior.

residual	leg	m_0 (1.00)	q (1.00)	c (1.00)	q/m_0 (1.00)
force residual	Chain (state space)	0.8233 ± 0.15	0.8681 ± 0.15	0.8939 ± 0.12	1.057 ± 0.055
	Collider (state space)	$2.1 \cdot 10^{-5} \ddagger$	$0.00383 \ddagger$	$4133 \ddagger$	$182.8 \ddagger$
	Collider (B-PINN)	0.1630 ± 0.036	0.1485 ± 0.032	1.608 ± 0.27	0.9127 ± 0.036
acceleration residual	Chain (state space)	0.9941 ± 0.29	1.049 ± 0.31	0.8881 ± 0.087	1.056 ± 0.046
	Collider (state space)	0.9941 ± 0.29	1.049 ± 0.31	0.8881 ± 0.087	1.056 ± 0.046
	Collider (B-PINN)	1.101 ± 0.33	0.9872 ± 0.29	1.783 ± 0.25	0.8969 ± 0.015

Table 7: **Replication of the oscillator configuration over datasets** (acceleration residuals). Top: the neural B-PINN and our chain model, sampled as in the main text, on ten datasets. Bottom: exactly computed posteriors on a linear surrogate, on ten further datasets. *displ.*: mean over datasets of the posterior mean of the log parameter minus the log truth, $\pm$ its standard error. *sd*: mean posterior standard deviation. *90% cov.*: number of datasets whose central 90% credible interval contains the truth.

Model (latent)	log(c/m)			log(k/m)		
	displ.	sd	90% cov.	displ.	sd	90% cov.
<i>Sampled posteriors, ten datasets</i>						
Chain (state space)	$+0.02 \pm 0.05$	0.17	10/10	$+0.02 \pm 0.02$	0.11	10/10
Collider B-PINN (MLP)	-0.42 ± 0.03	0.24	4/10	-0.12 ± 0.03	0.11	7/10
<i>Exact posteriors on a linear surrogate, ten datasets</i>						
Chain (state space)	-0.03 ± 0.05	0.17	8/10	$+0.02 \pm 0.04$	0.11	9/10
Chain (linear surrogate)	-0.05 ± 0.05	0.18	9/10	$+0.02 \pm 0.03$	0.09	9/10
Collider (linear surrogate)	-0.46 ± 0.04	0.23	1/10	-0.08 ± 0.03	0.10	7/10

Results. Table 7 confirms that the displacements reported in the main text are a systematic bias. Over ten datasets, the B-PINN underestimates c/m by 0.42 ± 0.03 log units and k/m by 0.12 ± 0.03 , and it lies below the truth on every dataset in both coordinates. Measured relative to the chain model on the same dataset, which removes the variation between datasets, the displacements are -0.45 ± 0.02 and -0.14 ± 0.02 . The B-PINN posterior is not narrower than the chain’s (mean standard deviation 0.24 against 0.17 in $\log(c/m)$), but it is displaced by almost two of its own standard deviations, so its 90% credible intervals contain the truth in only 4 and 7 of the ten datasets, whereas those of the chain model contain it in all ten. The dataset of Fig. 2 lies at the upper end: its displacement in $\log(c/m)$ (-0.53) is the third largest of the ten, against a median of -0.39 .

The exact posteriors on the linear surrogate show the same picture, with displacements of -0.46 ± 0.04 and -0.08 ± 0.03 and coverage in 1 and 7 of ten datasets. Restoring Z_λ exactly on the same surrogate removes the bias in both coordinates and restores nominal coverage, as does the state-space chain. The bias is therefore reproducible across datasets, and it is caused by the missing normalizer rather than by the particular dataset or by the limited expressiveness of the surrogate.

A.3.5 Sensitivity to the Residual Scale

All experiments above fix σ_r at the value that generated the data. To test how the bias depends on this value, we repeat the exact-posterior comparison of Appendix A.3.4 on the same ten oscillator datasets (acceleration residuals), varying σ_r in the data and in the models. For $\sigma_r \leq 0.1$, we evaluate the posteriors on a finer grid, since they are narrower than its original spacing.

Table 8: **Sensitivity to the residual scale** (oscillator, acceleration residuals, exact posteriors, ten datasets). Displacement of $\log(c/m)$ from the truth (mean over datasets $\pm$ standard error) and number of datasets whose central 90% credible interval contains the truth. Top: σ_r matched to the data. Bottom: σ_r misspecified.

σ_r		Chain (state space)		Chain (linear surrogate)		Collider (linear surrogate)	
data	model	displ.	90% cov.	displ.	90% cov.	displ.	90% cov.
0.05	0.05	-0.01 ± 0.02	9/10	-0.03 ± 0.02	9/10	-0.07 ± 0.02	7/10
0.1	0.1	-0.01 ± 0.03	9/10	-0.02 ± 0.03	9/10	-0.12 ± 0.03	7/10
0.3	0.3	-0.03 ± 0.05	8/10	-0.05 ± 0.05	9/10	-0.46 ± 0.04	1/10
1.0	1.0	-0.04 ± 0.07	8/10	-0.16 ± 0.07	9/10	-0.74 ± 0.03	0/10
0	0.3	$+0.40 \pm 0.01$	0/10	$+0.37 \pm 0.01$	1/10	-0.15 ± 0.01	10/10
0.3	0.9	$+0.97 \pm 0.05$	0/10	$+0.96 \pm 0.05$	0/10	-0.30 ± 0.02	10/10

With σ_r matched to the data, the collider bias is present at every scale and grows with it, from -0.07 to -0.74 in $\log(c/m)$, while both chain models remain close to the truth (Table 8, top). At small σ_r , the displacement is small in absolute terms but still about one posterior standard deviation, and coverage remains below nominal.

A misspecified residual scale introduces a different bias, which affects the chain model as well. For noise-free data, or for a residual scale three times too large, the chain overestimates the damping, and its intervals miss the truth (Table 8, bottom). The downward collider bias then partly offsets this error, so that the collider’s wider intervals cover the truth; this reflects two biases of opposite sign cancelling, not a correct model.

Ideally, the residual scale is therefore matched to the data. When it is unknown, a natural remedy is to infer σ_r jointly with λ , as noted in Section 6. For the chain model, this is a direct extension of PMMH, since the per-step normalizers (21) remain analytic in σ_r . For the collider, however, the bias Δ then also depends on σ_r , which gives rise to a further bias mechanism; we leave its analysis to future work.

A.3.6 The Dissipation Term on the Conservative Systems

The double pendulum and the relativistic particle are both conservative, yet the dissipation term of a global surrogate need not vanish for them. We first explain why, and then evaluate the term for each system.

Why conservative systems can retain a dissipation term. Suppose that the forces derive from a (possibly time-dependent) potential and that $M_\lambda = \partial^2 \mathcal{L} / \partial v \partial v^\top$ is the velocity Hessian of the Lagrangian $\mathcal{L}$. Then the flow is Hamiltonian in the canonical coordinates (x, p) with $p = \partial \mathcal{L} / \partial v$, and divergence-free in these coordinates (Liouville’s theorem). Since the change of variables $(x, v) \mapsto (x, p)$ has Jacobian determinant $\det M_\lambda$,

$$\nabla \cdot f_\lambda(z(t), t) = -\frac{d}{dt} \log \det M_\lambda(z(t)), \quad \int_0^{\mathcal{T}} \nabla \cdot f_\lambda dt = \log \frac{\det M_\lambda(z(0))}{\det M_\lambda(z(\mathcal{T}))}, \quad (38)$$

a boundary term that does not grow with $\mathcal{T}$. On a global surrogate, however, $\Delta(\lambda)$ depends on $\sum_k |\operatorname{Re} s_k|$ (Eq. (34)) rather than on the divergence, and the two can differ in two ways. First, the eigenvalues of a Hamiltonian linearization come in pairs $\pm s$: oscillatory pairs ($\operatorname{Re} s = 0$) contribute nothing, but hyperbolic pairs (real $\pm s$) contribute $\mathcal{T}|s|$ although their divergence vanishes, so a chaotic system can retain a λ -dependent term that grows with $\mathcal{T}$. Second, if Df_λ has a single eigenvalue with nonzero real part, as for a one-dimensional system with velocity-dependent inertia, then $\sum_k |\operatorname{Re} s_k| = |\nabla \cdot f_\lambda|$ pointwise, and the windowed dissipation term becomes $-\frac{1}{2} \int_0^{\mathcal{T}} |\nabla \cdot f_\lambda| dt$. This grows linearly in $\mathcal{T}$ whenever the divergence changes sign along the orbit, although $\int_0^{\mathcal{T}} \nabla \cdot f_\lambda dt$ stays bounded. The pendulum is an instance of the first case and the relativistic particle of the second. We evaluate both quantities along the noise-free trajectory at the ground truth; this explains why the dissipation term leaves the pendulum results in Table 1 unaffected but biases the relativistic particle.

Double pendulum. The time average of $\sum_k |\operatorname{Re} s_k|$ over $[0, \mathcal{T}]$ is 0.508, so the surrogate retains a term of $\frac{\mathcal{T}}{2} \langle \sum_k |\operatorname{Re} s_k| \rangle = 3.05$ nats at $\mathcal{T} = 12$, although the mean divergence is only $1.2 \cdot 10^{-2}$. This term does not affect the shift $\log \alpha$ along the drift-invariant direction reported in Table 1: scaling $(m_1, m_2) \rightarrow \alpha(m_1, m_2)$ leaves the acceleration $a_\lambda = M_\lambda^{-1} F_\lambda$, and hence Df_λ , unchanged, and indeed the time average agrees to six digits at

$\alpha = 0.5, 1$ and 2 . Off the direction, it is negligible as well: it varies by less than $2 \cdot 10^{-3}$ over ± 0.05 in $\log \ell_1$, i.e., $|\partial \Delta / \partial \log \ell_1| \approx 0.04$ nats per unit of $\log \ell_1$, which is small against a prior with log-standard deviation 0.25 .

Relativistic particle. The drive returns the particle to rest at both ends of the record ($\gamma(0) = \gamma(\mathcal{T}) = 1$), so $\int_0^{\mathcal{T}} \nabla \cdot f_\lambda dt = 0$ to machine precision: the divergence term is not merely bounded but absent. The drift Jacobian $Df_\lambda = \begin{bmatrix} 0 & 1 \\ 0 & \nabla \cdot f_\lambda \end{bmatrix}$, however, has the eigenvalues 0 and $\nabla \cdot f_\lambda = -3qE(t)v/(m_0 c^2 \gamma)$, which changes sign twice per drive period. Hence $\sum_k |\text{Re } s_k| = |\nabla \cdot f_\lambda|$, and $\int_0^{\mathcal{T}} |\nabla \cdot f_\lambda| dt = 10.36$, so that $\Delta_{\text{diss}} = -5.18$ nats at the truth, a term that grows linearly in $\mathcal{T}$. Differentiating with respect to the light speed at a fixed trajectory gives $\partial \Delta_{\text{diss}} / \partial \log c = +9.5$ nats per unit of $\log c$. The collider is thus rewarded for a larger c , because a larger c keeps γ closer to one and thereby flattens the divergence. This reward has the sign and the order of magnitude of the observed error, and it is too strong for a prior with log-standard deviation 0.25 to counteract. The B-PINN reports $c = 1.78$ ($\log c$ displaced by $+0.569$), against 0.89 for the chain. We do not derive a quantitative prediction from these numbers, because Δ_{diss} is strongly nonlinear in c (roughly $\propto -c^{-2}$) over the range that the posterior covers.

A.3.7 Restoring the Normalizer on the Neural Surrogate

In principle, the bias could stem from the probabilistic model (collider or chain), the latent (state space or neural surrogate), or the estimator (PMMH or NUTS). On the state-space latent, the model alone decides: chain and collider share latent and estimator and coincide exactly whenever $\Delta(\lambda)$ is constant (Section 5.2), and NUTS reproduces the PMMH results (Appendix A.3.2). Here, we hold the neural surrogate fixed and change only the model.

Construction. Chain (Fourier MLP) adds the term $-\Delta = \log Z_r - \log Z_\lambda$ to the log-density of the oscillator B-PINN. By Eq. (13), this turns the collider into the chain, up to the approximations of Appendix A.1. For the oscillator, Eq. (16) at $d = 1$, with $\det A_\lambda = m$ under force residuals and $\det A_\lambda = 1$ under acceleration residuals, gives

$$\log Z_r - \log Z_\lambda = \underbrace{\frac{c\mathcal{T}}{2m}}_{\text{dissipation}} + \underbrace{n_{\text{eff}} \log m}_{\text{metric, force residuals only}} + \text{const.} \quad (39)$$

Everything else (architecture, weight prior, data, collocation grid, sampler and seed) is identical to the B-PINN.

Effective number of constraints. The correction requires $n_{\text{eff}} = \text{tr}[G_\lambda(G_\lambda + P)^{-1}]$. We compute it once, from the Gauss–Newton curvature at $\lambda = \lambda_{\text{true}}$, and keep it fixed during sampling. This is the idealization under which the closed forms are derived (assumption (iii) in Appendix A.1): by Eq. (30), n_{eff} depends on λ only through κ_λ and is constant in the saturated regime. We evaluate the curvature at the final network weights of the corresponding B-PINN run rather than at the data-free mode $\theta^* = 0$, where $x_\theta \equiv 0$ solves the ODE but the network Jacobian is degenerate (assumption (ii)). With $Nd = 128$ and $\dim \theta = 1441$, this gives $n_{\text{eff}} = 110.2$ for force residuals. For acceleration residuals, only the dissipation term is active and no n_{eff} is needed.

This value depends only moderately on λ but strongly on the network weights. Evaluating the curvature at the chain’s posterior mean instead of λ_{true} changes it by less than 0.5% , and over draws from the chain posterior it varies by about $\pm 10\%$ (98–116 around 110.2). At the final state of the acceleration-residual B-PINN, by contrast, it is 99.9, and at a MAP fit of the network it drops to 44. Moreover, n_{eff} is not constant along the B-PINN’s own posterior (55–78, where the B-PINN has collapsed the mass scale), so the saturated regime of assumption (iii) holds only approximately. This sensitivity to the expansion point is the main source of the overcorrection discussed below.

Results. Table 9 reports two configurations, each of which activates a different subset of the terms in (39). All runs converge ($\hat{R} = 1.000$ over 24,000 draws), so these are valid posteriors. How much of the bias the correction removes follows how accurate the closed forms are in each configuration:

- With acceleration residuals, only the dissipation term is active and no n_{eff} enters. The correction removes 97.5% of the displacement of $\log(c/m)$.

Table 9: **Restoring the normalizer on the neural surrogate (oscillator)**. Displacements from the truth in log space (posterior mean of the log parameter minus the log truth) for the Collider B-PINN and for Chain (Fourier MLP), and the fraction of the displacement removed by the correction, $1 - |\text{chain}|/|\text{collider}|$. *Active terms* are the terms of Eq. (39) that depend on the inferred parameters in that configuration. The last configuration is the $\mathcal{T} = 24$ run of the horizon study in Fig. 3.

Configuration	Active terms	Direction	Collider	Chain (MLP)	Removed
force residuals	metric, dissipation	$\log m$	-1.453	+0.497	65.8%
acceleration residuals	dissipation	$\log(c/m)$	-0.526	-0.013	97.5%

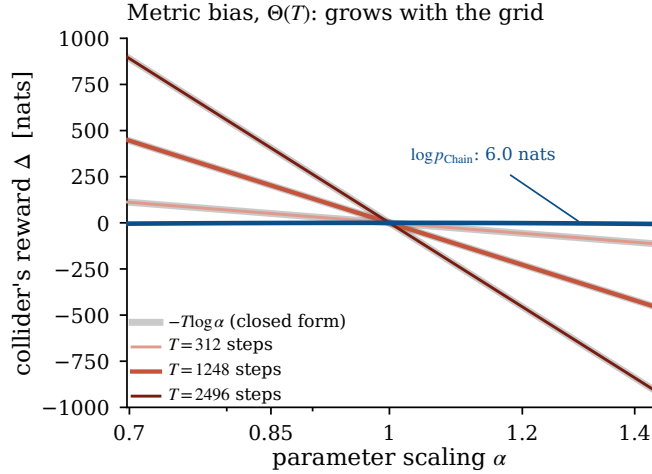


Figure 4: **Exact closed form of the residual metric bias against measurements on the oscillator** (Eq. (22)). Closed form in grey, the collider on the Markovian state-space latent in dark red, the chain in blue. On the state-space latent, $\Delta(\lambda)$ along the drift-invariant direction matches $-T \log \alpha$ to within $2.8 \cdot 10^{-12}$ nats for each number of time steps shown ($T = 312, 1,248, 2,496$; $T = 624$, omitted for clarity, agrees equally), whereas the chain’s profile stays flat.

- With force residuals, the approximate n_{eff} matters, and the correction can overshoot. The collider’s $\log m = -1.45$ becomes $+0.50$ ($m \approx 1.64$ against a truth of 1.00): the correction crosses the truth and leaves a displacement of about one third of the original on the other side. A term linear in $\log m$ with a fixed coefficient n_{eff} replaces a function of κ_λ and can therefore overcorrect.

As cautioned at the end of Appendix A.1, the closed forms thus locate the bias and remove most of it, but they do not calibrate it exactly.

Remaining stiffness bias. The correction also leaves part of the stiffness bias: with acceleration residuals, Chain (Fourier MLP) reports $k/m = 3.48$, against 3.16 for the B-PINN and 3.93 for the chain on the state-space latent (truth 4.00). This remainder is still collider bias rather than a representation error of the surrogate. On the linear cosine surrogate of Appendix A.3.4, where Z_λ can be restored exactly, the collider displaces $\log(k/m)$ from the truth by -0.08 ± 0.03 (mean $\pm$ standard error over ten datasets), whereas the exactly restored chain ($+0.02 \pm 0.03$) and the state-space chain ($+0.02 \pm 0.04$) are unbiased. The leading-order correction misses this part because k enters $\log Z_\lambda$ only through the $\mathcal{O}(\beta/\Omega)$ remainder of Eq. (33).

A.3.8 Details on the Scaling-Law Experiments

Figs. 3 and 4 use the oscillator configuration of Appendix A.3.1, in each case with the mechanism that is not under test switched off. Fig. 4 checks the metric term of $\Delta(\lambda)$ in a setting where its closed form is exact; Fig. 3 tests the approximate dissipation term on the neural surrogate.

Residual metric bias (Fig. 4). The figure shows $\Delta(\lambda)$ along the drift-invariant direction $(m, c, k) \rightarrow \alpha(m, c, k)$ on the state-space latent, for force residuals; the Markovian latent switches off the dissipation term. We use four discretizations, $T \in \{312, 624, 1248, 2496\}$. The measured $\Delta(\lambda)$ matches $-T \log \alpha$ to within $2.8 \cdot 10^{-12}$ nats, i.e., to machine precision, as expected since Eq. (22) is exact on this latent. Over the swept range of α , $\Delta(\lambda)$ varies by 225, 449, 899 and 1797 nats, doubling with the number of steps, whereas the chain’s log-posterior varies by only 6.0 nats over the same range.

Dissipation bias (Fig. 3). Unlike the metric term, the dissipation term is not a property of the latent alone, so we measure its effect on the posterior rather than $\Delta(\lambda)$ itself. We use acceleration residuals, which switch off the metric bias, and vary the horizon over $\mathcal{T} \in \{12, 24, 48\}$; the run at $\mathcal{T} = 24$ is the standard oscillator configuration. Across horizons, we keep the surrogate’s bandwidth per unit time and the Euler step fixed and scale K , N and T with $\mathcal{T}$, so that only the amount of data changes. The panel shows the posterior mean of $\log(c/m)$ minus its true value for Chain (state space), Chain (Fourier MLP) and the Collider B-PINN, each with a bar of ± 1 posterior standard deviation. The state-space collider is not plotted separately, since under acceleration residuals $\Delta(\lambda)$ is constant and its draws are identical to the chain’s.

The B-PINN lies at -0.377 , -0.526 and -0.669 , i.e., at 1.55, 2.08 and 2.67 of its own posterior standard deviations, which barely change (0.243, 0.253, 0.251). The chain stays near the truth, both on the state-space latent (-0.086 , -0.022 , $+0.041$, with standard deviations 0.198, 0.176, 0.143) and on the surrogate with the correction (-0.095 , -0.013 , $+0.141$). The chain’s posterior thus contracts as the record grows, whereas the B-PINN’s does not, so the B-PINN’s error grows relative to its own stated uncertainty.

To compare with the closed form (36), we reweight the chain posterior of $\log(c/m)$, approximated as Gaussian, by $e^{\Delta_{\text{diss}}} = e^{-(c/m)\mathcal{T}/2}$. This predicts shifts relative to the chain of -0.21 , -0.32 and -0.41 , whereas the B-PINN is shifted by -0.29 , -0.50 and -0.71 . The leading-order term thus predicts the direction of the bias and its growth with the record, but only between three quarters and three fifths of its size; it also predicts a posterior that contracts with $\mathcal{T}$, which the B-PINN’s does not.

A.4 Comparison with Nested Inference

To our knowledge, the only prior work that adopts the chain model, by Alberts, Hao, and Bilonis [1, 10, 11], handles the unknown normalizer Z_λ by nested inference: an inner loop estimates the gradient $\partial_\lambda \log Z_\lambda$ anew at every step of an outer loop over the joint state. We compare our method, the particle filter with PMMH (Section 4.3), against the two nested schemes that target the same chain posterior.

- **NSVI** [10] (nested stochastic variational inference) fits an outer Gaussian guide over path coefficients, physical parameters and initial state by maximizing an ELBO. The normalizer gradient in this ELBO is supplied by a second, persistent inner guide, which is re-fitted at every outer step.
- **NPSGLD** [11] (nested preconditioned stochastic-gradient Langevin dynamics) replaces both guides by preconditioned Langevin chains: a persistent auxiliary chain over the path supplies the normalizer gradient to a main chain over the joint state.

We use the public JAX reimplementation of both methods by Bolt [5] (commit `dab4403`), which is not by the original authors; we run it unmodified and supply only the problem definition and the evaluation. To isolate the inference algorithm, all methods use the Euler–Maruyama path as latent, on which Z_λ is in fact available in closed form. The comparison therefore measures how expensive and how accurate nested estimation of Z_λ is where it could be avoided. The baselines were designed for latents without a closed-form normalizer, a setting we do not test.

Setup. We use two systems in the configurations of Section 5.1: the damped oscillator under both residual metrics, and the double pendulum under torque residuals. All methods share the model, the latent ($T = 1248$ Euler–Maruyama steps for the oscillator and $T = 1235$ for the pendulum), the observations, and the log-normal priors with median at the truth. As in all experiments, σ_r is fixed at its true value. Instead of tuning a single run per baseline, we sweep their hyperparameters. For NSVI, we vary the outer learning rate in $[5 \cdot 10^{-5}, 10^{-3}]$ and the number of inner iterations from 8 to 256, with budgets of up to $1.2 \cdot 10^6$ iterations. For NPSGLD, we vary the step size in $[10^{-8}, 10^{-3}]$, use 4 or 16 chains and both of its preconditioners (`rmsprop` and `dense_fisher`),

and switch step-size annealing on or off, with budgets of up to 10^7 iterations. This amounts to 23 runs on the oscillator and 21 on the pendulum. All runs, including those of our method, use the same Quadro RTX 8000 GPU, so that run times are directly comparable.

Evaluation metrics. Wall-clock time alone favours samplers that have not mixed, since they produce many draws that carry little information. We therefore measure cost in *seconds per effective sample*, i.e., wall-clock time divided by the effective sample size (ESS), and report the split- $\hat{R}$ alongside it. As in Table 1, runs with $\hat{R} > 1.05$ do not yield a valid posterior; we report their median and mark them with a dagger. For our method, the ESS is the smallest over the components of λ ; for NPSGLD, it is the ESS of the drift-invariant scale. This choice favours the baseline: on the drift-invariant scale alone, the ESS of our method on the pendulum is about 870 rather than 343, which would lower its cost from 6.71 to 2.6s per effective sample. The cost and $\hat{R}$ of our method refer to a single PMMH chain. To check its mixing, we additionally ran four independent chains (sampler seeds 0–3) per configuration; they give split- $\hat{R} \leq 1.001$ for all parameters, and their posterior means of the drift-invariant scale agree to within 0.02 (oscillator) and 0.003 (pendulum). NSVI draws independent samples from a fitted guide, so $\hat{R}$ and ESS do not apply to it; we assess it by its bias instead. Finally, the *width* of a run is the ratio of its posterior standard deviation to that of our method. It measures what matters for uncertainty quantification: a method can locate the posterior correctly and still misrepresent its uncertainty.

Table 10: **Nested inference compared with our particle filter with PMMH.** Ranges are over the hyperparameter sweep of each method. *Location* is the drift-invariant scale (m for the oscillator, m_1+m_2 for the pendulum), reported as posterior mean $\pm$ standard deviation where $\hat{R} \leq 1.05$ and as a daggered median otherwise. *Width* is the posterior standard deviation divided by that of our method; 1.00 is correct, and 0.03 means thirty-fold overconfidence. *Cost* is in seconds per effective sample. A run counts as *diverged* if its location is off by more than a factor of 100 or its spread exceeds 100 reference standard deviations.

System	Method	Runs	Location	$\hat{R}$	ESS	Width	Cost
Oscillator force residuals	PF + PMMH (ours)	1	1.015 ± 0.140	1.00	604	1.00	0.81
	NSVI	8	0.78–1.43	—	—	0.15–0.28	—
	NPSGLD	10	1.00–1.12 †	1.3–3.0	3–32	0.03–1.25	56–2385
Oscillator accel. residuals	PF + PMMH (ours)	1	1.035 ± 0.220	1.00	875	1.00	0.51
	NSVI	2	1.28	—	—	1.36–1.38	—
	NPSGLD	3	1.00–1.12 †	1.1–1.5	6–86	1.09–1.90	21–190
Double pendulum torque residuals	PF + PMMH (ours)	1	1.615 ± 0.394	1.00	343	1.00	6.71
	NSVI	16	0.17–1.58	—	—	0.03–0.10	—
	NPSGLD	5	1.50–1.92 †, 2 diverged	2.5–277	3–4	0.07–0.94	141–495

Cost. Our method produces one effective sample every 0.5–6.7s. NPSGLD, the only baseline that produces MCMC samples, needs 21 to 2,385s per effective sample, i.e., $21\times$ to $2,900\times$ more, depending on the configuration. Its cheapest pendulum run finishes about five times sooner than our method, but it is $21\times$ more expensive per effective sample, since its 423s yield only three effective samples. A larger budget does not improve its mixing either: the longest NPSGLD run, with 10^7 iterations over 11,444s, yields five effective samples.

NSVI converges to a biased posterior. On the oscillator, NSVI is fast, and most of its settings agree on the drift-invariant scale; but the posterior they agree on is the wrong one. In six of its eight force-residual runs, the damping ratio lies at $c/m = 0.78$ – 0.81 , whereas the truth is 1.20 and our method gives 1.209 ± 0.236 , an underestimate of about 1.8 reference standard deviations. The other two runs, both with 10^5 outer iterations, end at $c/m = 0.69$ and 0.94 . More computation does not reduce the bias: raising the outer iterations from $3 \cdot 10^4$ to $3 \cdot 10^5$ and the inner iterations from 32 to 256 increases the run time 113-fold, from 128s to 14,474s, but moves c/m only from 0.790 to 0.800. The bias is therefore a property of the converged approximation, not of insufficient computation. On the chaotic pendulum, NSVI does not converge at all: across 16 settings, the mass scale ranges from 0.17 to 1.58, so its final estimate reflects where the budget ran out rather than the posterior. It is also strongly overconfident there: its posterior standard deviations are only 0.03–0.10 times that of our method (Table 10), i.e., ten to thirty times too small, so that its scattered estimates come with misleadingly narrow uncertainty.

NPSGLD misrepresents the uncertainty. None of the 18 NPSGLD runs, on either system and in any configuration, reaches $\hat{R} \leq 1.05$. Because its chains do not mix, its narrow posteriors reflect insufficient exploration rather than precision. On the pendulum, NPSGLD reports $m_1+m_2 = 1.495$ with a standard deviation fourteen times too small; on the oscillator with force residuals, it reports $m = 1.000$ with a standard deviation thirty times too small. At the same time, its posteriors over the identified ratios are up to $2.5\times$ too wide. The NPSGLD posterior is therefore not only mislocated but also misshapen. Our method, in contrast, converges in every configuration ($\hat{R} \leq 1.001$ over four independent chains) and is asymptotically exact.

Why the two systems differ. Both baselines perform considerably better on the oscillator than on the pendulum, for structural reasons. The oscillator is linear, so its path posterior is close to Gaussian, and the Gaussian outer guide of NSVI is nearly well specified. Moreover, with force residuals, the comparatively large process noise ($\sigma_r = 0.30$) determines the drift-invariant scale, which is the direction that nested schemes find most difficult. The pendulum, in contrast, is chaotic, its mass matrix depends on the state, and its path posterior over 1,235 steps and two degrees of freedom is strongly non-Gaussian. Nested inference therefore does not fail everywhere, but it succeeds only where the problem is close to what its approximations assume, and even there NPSGLD costs more than $40\times$ as much per effective sample as our method.