arXiv:2609.36942v1 [cs.LG] 29 Sep 2026

AI FOR INDUSTRY
The Italian Institute of Artificial Intelligence

Safe-by-Design LEARNING VIA ENERGY-BASED NEURAL NETWORKS

Simone Betteti
RIAS Lab
The Italian Institute of Artificial Intelligence for Industry
Torino, 10129, ITA
{simone.betteti}@ai4i.it

Morteza Lahijanian
RECUV
University of Colorado Boulder
Boulder, CO 80303, USA
{morteza.lahijanian}@colorado.edu

Luca Laurenti*
RIAS Lab
The Italian Institute of Artificial Intelligence for Industry
Torino, 10129, ITA
{luca.laurenti}@ai4i.it

September 30, 2026

ABSTRACT

Learning neural-network models of dynamical systems with safety guarantees is a fundamental requirement for their deployment in safety-critical settings. Safety is commonly established by proving the invariance of a desired subset in state-space, ensuring that every trajectory initialized in this subset remains confined to it for all time under admissible inputs. Existing frameworks, however, either rely on computationally expensive post-hoc verification or employ safety-enforcing mechanisms without formal correctness guarantees. In this paper, we introduce a novel neural architecture grounded in energy-based modern Hopfield networks to guarantee *safety-by-design* while retaining sufficient expressiveness to model complex nonlinear dynamics. Specifically, we integrate modern Hopfield networks with a port-Hamiltonian neural ODE, enabling by design the construction of barrier functions yielding explicit admissible-input sets and quantitative robustness radii. Across several benchmarks, including an 12-dimensional nanodrone model, our framework achieves state-of-the-art performance while producing certified invariant sets that are more robust to external solicitations than comparable existing approaches.

1 Introduction

Learning dynamical systems from data is increasingly central to robotics, autonomous systems, and scientific modeling, where neural state-space models can reconstruct complex nonlinear trajectories with remarkable accuracy [Karniadakis et al., 2021, Legaard et al., 2023]. In safety-critical applications, however, predictive accuracy over observed trajectories is not sufficient. A model that performs well over the finite horizons represented during training may develop unstable or physically undesirable behavior when propagated under new inputs or for longer time intervals. *Safety* is particularly important when learned dynamics are intended for applications such as flight, autonomous operation, or human-robot interaction, where the behavior of the model outside densely sampled training regions is itself operationally relevant [Ashmore et al., 2021, Brunke et al., 2022, Gyevnár and Kasirzadeh, 2025]. The central question is therefore not only whether a neural model can accurately identify nonlinear dynamics, but whether useful properties of its future evolution can be certified *before* deployment.

*L. Laurenti is also with the Delft Center for Systems and Control, TU Delft, Delft, 2600 AA, The Netherlands.

Safety in learning-enabled dynamical systems is commonly enforced through external mechanisms on top of an unconstrained predictor, including safety filters, runtime monitors, constrained controllers, and post-hoc verification [Brunke et al., 2022, Hewing et al., 2020]. Certifiability mechanisms are indispensable components of safety-critical systems, but they leave open a complementary question: can certifiability be made an intrinsic property of the learned dynamics themselves? Neural ordinary differential equations (NODEs) [Chen et al., 2018, Greydanus et al., 2019] have motivated several approaches towards safety-certified dynamics, using Lyapunov functions, dissipativity constraints, or algebraic structure to control behavior beyond the training trajectories [Kolter and Manek, 2019, Lawrence et al., 2020, Kojima and Okamoto, 2022, Kang et al., 2021, Yang et al., 2022]. Port-Hamiltonian neural networks are particularly attractive because they decompose the vector field into energy-preserving interconnection, energy dissipation, and external actuation through explicit input ports [Van der Schaft, 2007, Massaroli et al., 2020]. Their energy balance provides a natural structural primitive for stability and control without requiring the learned vector field to remain unconstrained.

A central tension nevertheless remains between *certifiability* and *expressivity*. Strong stability guarantees in recent port-Hamiltonian architectures are commonly obtained through restrictive Hamiltonian parameterizations, including convex energies designed around a single equilibrium [Roth et al., 2025]. Yet nonconvex energy landscapes are precisely what allow energy-based models to represent multiple stable regimes and complex nonlinear organization. Modern Hopfield networks provide a particularly expressive realization of this principle: advances in energy-based architectures have shown that rich, nonconvex landscapes can coexist with an explicit energy structure [Krotov and Hopfield, 2020, Hoover et al., 2022, 2023]. For controlled physical dynamics, this suggests a different route to safe learning: rather than simplifying the Hamiltonian until certification becomes tractable, preserve an expressive energy landscape and exploit its geometry directly to determine which external inputs are compatible with safe evolution.

We pursue this approach through a *port-Hamiltonian energy-based model* (pH-EBM), in which a radially unbounded modern Hopfield energy (Fig. 1b) parametrizes the Hamiltonian of a controlled neural ODE (Fig. 1a). Radial unboundedness confines bounded-energy trajectories to compact sublevel sets, while the interior Hamiltonian remains free to develop nonconvex and multi-well structure. The port-Hamiltonian decomposition provides an explicit energy balance between dissipation and external actuation. We then use the same learned Hamiltonian to define energy barrier functions: for a prescribed sublevel component, the barrier condition identifies the external inputs that cannot inject sufficient energy to drive the learned trajectory across its boundary (Fig. 1c). Safety is therefore not introduced as an additional network or post-processing stage, but derived from the energy geometry already governing the neural dynamics. The resulting guarantees concern invariance of prescribed regions of the *learned model state space* under certified classes of inputs; when bounded disturbances or model mismatch are explicitly characterized, the same construction extends to robust certificates.

Our contributions are threefold. **First**, we introduce a pH-EBM that combines the dissipative structure of port-Hamiltonian NODEs with a coercive, globally nonconvex modern-Hopfield Hamiltonian, separating boundedness at large state norm from the geometry required to represent complex operating regimes. **Second**, we derive explicit set-valued safety certificates from the learned energy, such as exact state-dependent admissible-input sets, state-independent robustness radii for entire energy regions, and robust extensions under bounded perturbations. We further show how the certified margin is governed by the geometry of the energy shell, dissipation normal to its boundary, and exposure of this direction to the input port, with local curvature providing tractable closed-form bounds. **Third**, we evaluate the framework across nonlinear system-identification and controlled-mechanics benchmarks spanning classical physical systems, nonconvex multi-attractor dynamics, adversarial out-of-distribution forcing, and high-dimensional flight identification. The experiments show that pH-EBM can offer competitive predictive accuracy, often outperforming state-of-the-art methods, while providing non-trivial safety guarantees by design.

2 Problem formulation

We consider the learning of an unknown controlled continuous-time dynamical system described by

$$\dot{x}^\star(t) = f^\star\big(x^\star(t), u(t)\big), \quad y(t) = h^\star\big(x^\star(t), u(t)\big) + v(t), \tag{1}$$

where $x^\star(t) \in \mathbb{R}^{n_\star}$ is the generally unobserved physical state, $u(t) \in \mathcal{U} \subseteq \mathbb{R}^k$ is a known bounded external input, $y(t) \in \mathbb{R}^p$ is the measured output, and v denote measurement perturbations. Throughout the paper, we assume that $f^\star$ satisfies the standard Carathéodory conditions, i.e., locally Lipschitz continuous in x and locally essentially bounded over the admissible values of u. These assumptions guarantee the existence and uniqueness of forward-complete Carathéodory solutions [Khalil, 2002]. Both the physical state $x^\star$ (including its dimensionality) and the vector field $f^\star$ are unknown, and the system can be observed only through sampled inputs u and outputs y. The available dataset therefore consists of M experiments, possibly collected from different unknown initial conditions and under different

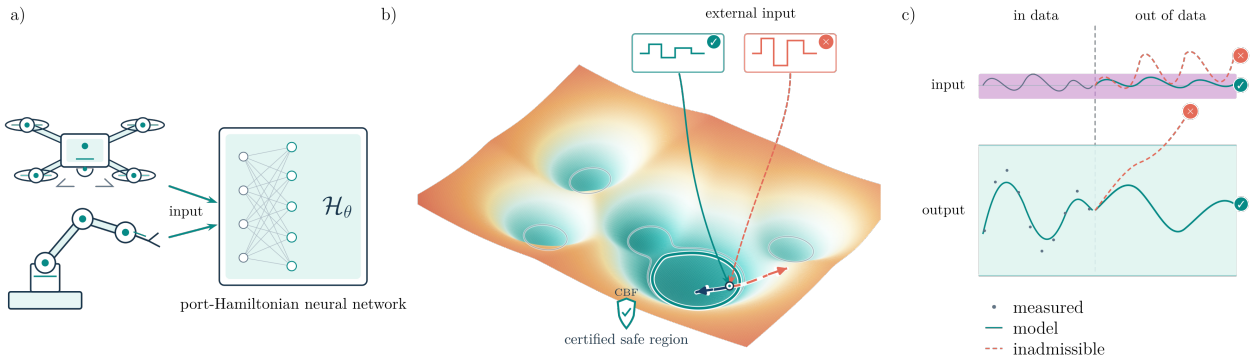


Figure 1: ***Safe-by-design* Energy-Based Model.** a) Input-output trajectories from physical or simulated systems are used to train the pH-EBM: the external input u drives the neural dynamics, while the measured output y supervises reconstruction. b) The learned Hamiltonian, parameterized by a modern Hopfield energy, organizes the neural ODE into a continuous energy landscape whose low-energy regions encode stable operating regimes. Beyond interpolating between sparsely observed trajectories, this geometry identifies energy sublevel sets that can be certified as invariant for prescribed classes of inputs. c) Inputs lying within the certified admissible set therefore generate trajectories that remain confined to the desired operating region, turning the learned energy landscape into an explicit quantitative safety certificate.

input signals, i.e.,

$$\mathcal{D} := \left\{ \left(u^{(j)}(t_i), y^{(j)}(t_i) \right)_{i=0}^{N_j} \right\}_{j=1}^{M}. \tag{2}$$

Since the physical state $x^\star$ is unobserved, our objective is not necessarily to recover its original realization, but to learn a latent state-space model that reproduces the observed input-output behaviour and generalizes within a neighbourhood sufficiently covered by the data. Specifically, we seek a continuous-time neural state-space model of the form

$$\dot{z}(t) = f_\theta\big(z(t), u(t)\big), \quad \hat{y}(t) = h_\phi\big(z(t), u(t)\big), \tag{3}$$

where $z(t) \in \mathbb{R}^d$ is the latent model state, θ and ϕ are trainable parameters, h_ϕ is a continuous latent-to-output decoder function, and $\hat{y}(t) \in \mathbb{R}^p$ is the measurement of the latent state $z(t)$. Note that $\hat{y}$ is in the same space as y of System (1). The initial latent state associated with each experiment is inferred jointly with the model parameters, or produced by an appropriate encoder. Our objective is not merely to fit the trajectories in $\mathcal{D}$. Instead, we seek a *safe-by-design* neural architecture. To give this notion a physical interpretation, let $\mathcal{Y}_{\text{safe}} \subseteq \mathbb{R}^p$ be a prescribed closed set of outputs that satisfy the relevant operational safety requirements. Safety of the learned model is established by jointly constructing a robustly invariant latent set $\mathcal{Z}_{\text{safe}}$ whose corresponding outputs remain in $\mathcal{Y}_{\text{safe}}$.

Definition 1 (ρ-Robust Invariant Sets)**.** *Let $\rho > 0$ be an input-intensity threshold and define the associated input set $u \in \mathcal{U}_\rho := \{u \in \mathcal{U} : \|u\| \le \rho\}$. We say that the compact set of latent states $\mathcal{C} \subset \mathbb{R}^d$ is ρ-Robust invariant if:*

- *for all $z \in \mathcal{C}$ and $u \in \mathcal{U}_\rho$ it holds that $h_\phi(z, u) \in \mathcal{Y}_{\text{safe}}$ and*
- *for all $z(0) \in \mathcal{C}$ and input signals u such that that $u(t) \in \mathcal{U}_\rho$ for all $t > 0$, it holds that $z(t) \in \mathcal{C}$ for all $t > 0$.*

In this paper, we will establish results for the latent representations z, and then exploit the continuity of the decoder map h_ϕ to extend safety to compacts in measurement space. Importantly, in our setting the geometry of $\mathcal{Z}_{\text{safe}}$ is generally unknown before training. Since the latent representation z, its dynamics f_θ, and the output map h_Φ are learned jointly from data, an invariant set in latent space $\mathcal{Z}_{\text{safe}}$ cannot generally be prescribed a priori. It must instead be learned jointly with the neural state-space realization. This leads to the following problem:

Problem 1. ***Learning and certifying neural ODEs.*** *Given a dataset $\mathcal{D}$, a set $\mathcal{Z}_{\text{safe}}$, and an input-intensity parameter $\rho > 0$, learn a neural ODE $\dot{z}(t) = f_\theta(z(t), u(t))$ with measured output $\hat{y}(t) = h_\phi(z(t), u(t))$ such that for all $\{u(t) \in \mathcal{U}_\rho\}_{t \ge 0}$ returns a maximal ρ-Robust Invariant Set $\mathcal{C} \subseteq \mathcal{Z}_{\text{safe}}$.*

Many autonomous systems admit nonempty positively invariant regions, while systems operating in closed loop with stabilising feedback often admit nonempty robust positively invariant regions under bounded disturbances [Blanchini, 1999, Blanchini and Miani, 2015]. The objective of Problem 1 is therefore to endow an input-output model learned from data with this structural property: we seek to identify a latent state-space realization while simultaneously guaranteeing

that a nontrivial latent safe set is robustly positively invariant. Solving Problem 1 is, however, particularly challenging because generic post-hoc methods for computing or certifying invariant sets for neural dynamics, based for instance on gridding, mixed-integer programming, or SMT solving, are computationally demanding and scale poorly with the state dimension and network size [Abate et al., 2021, Dawson et al., 2023]. We therefore pursue a complementary route and consider neural architectures that admit a guaranteed ρ-robust invariant set by construction. In particular, in Section 3 we introduce a port-Hamiltonian neural architecture whose Hamiltonian sublevel components provide barrier certificates under explicit dissipation and input conditions. The barrier certificates ultimately allow us to characterize safe ρ-robustly forward-invariant sets. In Section 4 we instead present the performances of our pH-EBM model in complex dynamics learning tasks, displaying how the model achieves state-of-the-art accuracy and provides explicit safety certificates, using drones' bounded long time rollouts as a paradigmatic proof of concept.

3 Safe-by-design port-Hamiltonian neural ODEs

The solution of Problem 1 with a neural ODE prompts the choice of a vector field structure that simultaneously combines two key ingredients: (i) a stability-inducing mechanism to confine trajectories and (ii) an input port. Thereby, following [Roth et al., 2025], we specialize the neural state-space model in (3) to the port-Hamiltonian vector field

$$\dot{z} = f_\theta(z,u) = [J_{\theta_J}(z) - R_{\theta_R}(z)]\,\nabla\mathcal{H}_{\theta_H}(z) + G_{\theta_G}(z)u, \tag{4}$$

where $z \in \mathbb{R}^d$, $u \in \mathcal{U} \subseteq \mathbb{R}^k$, and $\mathcal{H}_{\theta_H} : \mathbb{R}^d \to \mathbb{R}$ is a continuously differentiable Hamiltonian. Moreover, $J_{\theta_J}(z) = -J_{\theta_J}(z)^\top \in \mathbb{R}^{d\times d}$ is skew-symmetric, $0 \preceq R_{\theta_R}(z) = R_{\theta_R}(z)^\top \in \mathbb{R}^{d\times d}$ is symmetric positive semidefinite, and $G_{\theta_G}(z) \in \mathbb{R}^{d\times k}$ is the input map. The symmetry constraints on J_{θ_J} and R_{θ_R} are enforced directly through their parameterizations. We assume that J_{θ_J}, R_{θ_R}, G_{θ_G}, and $\nabla\mathcal{H}_{\theta_H}$ are locally Lipschitz in z and that $u : [0,\infty) \to \mathcal{U}$ is locally essentially bounded. Consequently, (4) admits a unique Carathéodory solution on its maximal interval of existence. Equation 4 is a standard port-Hamiltonian ODE and satisfies the energy balance [Van der Schaft, 2007, Willems, 2007]

$$\frac{\mathrm{d}}{\mathrm{d}t}\mathcal{H}_{\theta_H}(z) = -\nabla\mathcal{H}_{\theta_H}(z)^\top R_{\theta_R}(z)\nabla\mathcal{H}_{\theta_H}(z) + \nabla\mathcal{H}_{\theta_H}(z)^\top G_{\theta_G}(z)u. \tag{5}$$

Thus, skew-symmetry of J_{θ_J} makes the interconnection term energy preserving, positive semidefiniteness of R_{θ_R} makes the internal dynamics dissipative, and $G_{\theta_G}u$ accounts for energy exchanged with the environment.

Importantly, this balance does not require convexity of the Hamiltonian. Consequently, we now define a non-convex Hamiltonian, which provides the key architectural ingredient for constructing compact confinement regions in the latent space. In particular, we employ the hybrid modern Hopfield Hamiltonian introduced in [Betteti and Laurenti, 2026]:

$$\mathcal{H}_{\theta_H}(z) = \frac{1}{2}\|z\|_2^2 - \sum_{h=1}^{L} g_h(z)^\top b_h - \sum_{h=2}^{L}\mathcal{F}_h(g_h(z))\,, \tag{6}$$

where $L \geq 2$ is the number of layers, with widths $N_1 = d, N_2, \ldots, N_L$. We set $g_1(z) = z$, $\Psi_1(z) = z$ and recursively define the hidden feature vectors as

$$g_h(z) = W_{h(h-1)}\Psi_{h-1}(g_{h-1}(z))\,, \qquad h = 2,\ldots,L. \tag{7}$$

Here, $\mathcal{F}_h : \mathbb{R}^{N_h} \to \mathbb{R}$ are convex differentiable potentials and $\Psi_h = \nabla\mathcal{F}_h$ their continuous gradients, $W_{h(h+1)} \in \mathbb{R}^{N_h\times N_{h+1}}$ are interaction matrices between consecutive layers, with tied weights $W_{(h+1)h} = W_{h(h+1)}^\top$, and $b_h \in \mathbb{R}^{N_h}$ are bias vectors entering the Hamiltonian directly. The parameters θ_H collect the interaction matrices, biases, and any trainable parameters of the potentials. Notice that each $g_h(z)$ is obtained by composing all preceding layers, allowing increasingly complex features to be encoded throughout the architecture. In addition, each hidden layer g_H is a standard MLP, to which the universal approximation theorem applies [**?**]: consequently, the model can in principle approximate any vector field given sufficient data and parameters. The following lemma guarantees that a bounded first hidden activation is sufficient to preserve the quadratic growth of the Hamiltonian at infinity.

Lemma 2. *Assume that the first hidden activation Ψ_2 has bounded image. Then, $\mathcal{H}_{\theta_H}$ is coercive. That is, there exist constants $c_0, c_1 \geq 0$ such that $\mathcal{H}_{\theta_H}(z) \geq \frac{1}{2}\|z\|_2^2 - c_1\|z\|_2 - c_0$.*

Lemma 2 follows by the boundedness of $\Psi_2 = \nabla\mathcal{F}_2$, which implies that $\mathcal{F}_2$ grows at most linearly, while all feature vectors $g_h(z)$ with $h \geq 3$ remain bounded. Hence, the negative terms in (6) grow at most linearly in $\|z\|_2$ and are dominated by its quadratic leading term. Coercivity and continuity of $\mathcal{H}_{\theta_H}$ guarantee that its sublevel sets are compact despite the potentially nonconvex hidden-layer contributions. This is a key property we rely on in the following subsection to construct ρ-Robust Invariant Sets.

3.1 ρ-Robust Invariant Set Computation

To characterize ρ-Robust Invariant Sets for (4), we rely on barrier certificates [Ames et al., 2017]. In particular, we show that an isolated local minimum of $\mathcal{H}_{\theta_H}$ induces a compact energy sublevel component whose robustness to external inputs can be explicitly quantified. To show that, we let $z_\star$ be an isolated local minimizer of $\mathcal{H}_{\theta_H}$ and define the shifted energy $V_\star(z) := \mathcal{H}_{\theta_H}(z) - \mathcal{H}_{\theta_H}(z_\star)$. For $\epsilon > 0$, consider

$$\mathcal{C}_{\epsilon,\star} := \mathrm{Comp}_{z_\star}\left\{z : h_{\epsilon,\star}(z) \geq 0\right\}, \qquad h_{\epsilon,\star}(z) := \epsilon - V_\star(z), \qquad \Gamma_{\epsilon,\star} := \partial\mathcal{C}_{\epsilon,\star}, \tag{8}$$

where $\mathrm{Comp}_{z_\star}$ denotes the connected component containing $z_\star$. Since $\mathcal{H}_{\theta_H}$ is coercive, $\mathcal{C}_{\epsilon,\star}$ is compact. We assume that $\mathcal{H}_{\theta_H}(z_\star) = \epsilon$ is a regular value of $\mathcal{H}_{\theta_H}$, so that $\nabla\mathcal{H}_{\theta_H}(z) \neq 0$ for every $z \in \Gamma_{\epsilon,\star}$. Given a norm $\|\cdot\|$ on the input space, we denote the corresponding dual norm by $\|\cdot\|_*$ and define

$$d_H(z) := \nabla\mathcal{H}_{\theta_H}(z)^\top R_{\theta_R}(z)\nabla\mathcal{H}_{\theta_H}(z) \geq 0, \qquad a_H(z) := G_{\theta_G}(z)^\top\nabla\mathcal{H}_{\theta_H}(z) \in \mathbb{R}^k.$$

Then, in Theorem 3 below, we rely on the fact that along the dynamics (4) it holds that

$$\dot{h}_{\epsilon,\star}(z,u) = d_H(z) - a_H(z)^\top u \geq d_H(z) - \|a_H(z)\|_*\|u\|, \tag{9}$$

to prove that $h_{\epsilon,\star}(z,u)$ is a valid barrier function for System 4.

Theorem 3 (**ρ-robust invariant energy sets**). *Let $\mathcal{H}_{\theta_H}$ be defined in (6), and consider the port-Hamiltonian dynamics in (4). For every $z \in \Gamma_{\epsilon,\star}$, define*

$$\rho_{\mathrm{BF}}(z) := \begin{cases} \frac{d_H(z)}{\|a_H(z)\|_*}, & a_H(z) \neq 0, \\ +\infty, & a_H(z) = 0. \end{cases} \tag{10}$$

Then the vector field is inward-pointing or tangent to $\mathcal{C}_{\epsilon,\star}$ at z for every input satisfying $\|u\| \leq \rho_{\mathrm{BF}}(z)$. Consequently, $\mathcal{C}_{\epsilon,\star}$ in (8) is an $\rho_{\epsilon,\star}$-robust invariant set, where

$$\rho_{\epsilon,\star} := \inf_{z\in\Gamma_{\epsilon,\star}} \rho_{\mathrm{BF}}(z). \tag{11}$$

Proof. For every $z \in \Gamma_{\epsilon,\star}$, the barrier function condition implies $\dot{h}_{\epsilon,\star}(z,u) \geq 0$ whenever $\|u\| \leq \rho_{\mathrm{BF}}(z)$. The uniform radius in (11) therefore makes the vector field inward-pointing or tangent at every boundary point. Invariance then follows from the standard barrier-certificate conditions introduced in Appendix B.3. □

Theorem 3 provides an explicit robustness radius quantifying the tolerance of each energy component $\mathcal{C}_{\epsilon,\star}$ to external inputs.

Remark 1. *The construction extends directly to bounded additive perturbations. In particular, consider $\dot{z} = f_\theta(z,u) + w$, $w \in \mathcal{W}(z)$ where $\mathcal{W}(z)$ is nonempty and compact, and define its support function as $\sigma_{\mathcal{W}(z)}(q) := \sup_{w\in\mathcal{W}(z)} q^\top w$. In this case, $\dot{h}_{\epsilon,\star} \geq d_H(z) - \|a_H(z)\|_*\|u\| - \sigma_{\mathcal{W}(z)}\big(\nabla\mathcal{H}_{\theta_H}(z)\big)$. Hence, provided that $d_H(z) \geq \sigma_{\mathcal{W}(z)}\big(\nabla\mathcal{H}_{\theta_H}(z)\big)$ for all $z \in \Gamma_{\epsilon,\star}$, the disturbance-robust radius is*

$$\rho^{\mathcal{W}}_{\epsilon,\star} := \inf_{\substack{z\in\Gamma_{\epsilon,\star} \\ a_H(z)\neq 0}} \frac{d_H(z) - \sigma_{\mathcal{W}(z)}\big(\nabla\mathcal{H}_{\theta_H}(z)\big)}{\|a_H(z)\|_*}, \tag{12}$$

with the quotient interpreted as $+\infty$ when $a_H(z) = 0$.

Remark 2. *The norm-ball certificate can also be generalized to arbitrary compact convex input sets. In particular, a compact convex set $\mathcal{U}$ is robustly admissible if $\sigma_{\mathcal{U}}(a_H(z)) + \sigma_{\mathcal{W}(z)}(\nabla\mathcal{H}_{\theta_H}(z)) \leq d_H(z)$, for all $z \in \Gamma_{\epsilon,\star}$. Equivalently, the largest input envelope certified by this condition is*

$$\overline{\mathcal{U}}^{\mathrm{rob}}_{\epsilon,\star} := \mathcal{U} \cap_{z\in\Gamma_{\epsilon,\star}} \left\{u \in \mathbb{R}^k : a_H(z)^\top u \leq d_H(z) - \sigma_{\mathcal{W}(z)}(\nabla\mathcal{H}_{\theta_H}(z))\right\} \tag{13}$$

This set is convex because it is the intersection of $\mathcal{U}$ with a family of half-spaces. Thus, the construction accommodates norm balls, boxes, polytopes, ellipsoids, and asymmetric actuator constraints.

Remark 3. *A consequence of Theorem 15 is that if $\mathcal{H}_{\theta_H}$ has finitely many isolated local minima $\{z_i^\star\}_{i=1}^M$, repeating the construction for suitable regular energy levels ϵ_i gives the robustness atlas*

$$\mathfrak{A} := \{(\mathcal{C}_{\epsilon_i,i}, \overline{\mathcal{U}}^{\mathrm{rob}}_{\epsilon_i,i}, \rho_{\epsilon_i,i})\}_{i=1}^M, \tag{14}$$

which assigns a certified operating envelope to each learned dynamical regime.

3.2 Computable robustness bounds from local geometry

The exact robust radius in (11) requires minimizing a state-dependent ratio over the generally nonconvex boundary $\Gamma_{\epsilon,\star}$. Although exact, this characterization may be computationally demanding and does not explicitly reveal how the robustness margin depends on the local geometry of the learned model. We therefore derive a conservative lower bound in terms of the steepness of the energy shell, dissipation in its normal direction, and exposure of the normal direction to the input port. We then bound the shell-steepness term using a local Polyak-Łojasiewicz condition and derive an input-to-energy estimate under persistent forcing.

For readability, in what follows, we suppress the parameter subscripts in $\mathcal{H}_{\theta_H}$, R_{θ_R}, and G_{θ_G}. We restrict attention to energy levels for which $\mathcal{H}(z_\star)+\epsilon$ is a regular value of $\mathcal{H}$, so that $\nabla\mathcal{H}(z)\neq 0$ on $\Gamma_{\epsilon,\star}$ and the outward energy-normal direction $\eta(z) := \frac{\nabla\mathcal{H}(z)}{\|\nabla\mathcal{H}(z)\|_2}$ is well defined. Since $\Gamma_{\epsilon,\star}$ is compact, regularity also implies that the gradient norm is uniformly bounded away from zero on the boundary. We now define the following quantities

$$r_{\epsilon,\star} := \inf_{z\in\Gamma_{\epsilon,\star}} \eta(z)^\top R(z)\eta(z), \quad \bar{g}^{\perp}_{\epsilon,\star} := \sup_{z\in\Gamma_{\epsilon,\star}} \|G(z)^\top\eta(z)\|_*, \quad \kappa_{\epsilon,\star} := \inf_{z\in\Gamma_{\epsilon,\star}} \|\nabla\mathcal{H}(z)\|_2. \tag{15}$$

Here, $r_{\epsilon,\star}$ measures normal dissipation, $\bar{g}^{\perp}_{\epsilon,\star}$ measures exposure of the energy shell to the input port, and $\kappa_{\epsilon,\star}$ measures the minimum steepness of the energy shell. Regularity and compactness imply $\kappa_{\epsilon,\star}>0$, whereas $r_{\epsilon,\star}>0$ is an additional dissipativity assumption: positive semidefiniteness of R alone allows the dissipation to vanish in the normal direction. We are now ready to state Theorem 4.

Theorem 4 (Geometric lower bound on the robustness radius)**.** *Suppose that $r_{\epsilon,\star}>0$ and $\bar{g}^{\perp}_{\epsilon,\star}\neq 0$. Then*

$$\rho_{\epsilon,\star} \geq \underline{\rho}^{\mathrm{geo}}_{\epsilon,\star} := \frac{r_{\epsilon,\star}\kappa_{\epsilon,\star}}{\bar{g}^{\perp}_{\epsilon,\star}}. \tag{16}$$

Consequently, $\mathcal{C}_{\epsilon,\star}$ is a $\underline{\rho}^{\mathrm{geo}}_{\epsilon,\star}$-robust invariant set.

The proof is provided in Appendix B.6. The theorem guarantees that a large overall input gain is not necessarily detrimental: only its component normal to the energy shell consumes the available robustness margin.

We next lower-bound $\kappa_{\epsilon,\star}$ using a local Polyak–Łojasiewicz (PL) condition [Karimi et al., 2016, Pengyun et al., 2023]. In our framework, the Hamiltonian is a known scalar, while its gradient is implicitely defined: the local PL condition bounds the growth $\nabla\mathcal{H}$ on a shell $\Gamma_{\epsilon,\star}$with the specified energy budget $\epsilon>0$, and therefore readily provides a local estimate of the $\rho_{\epsilon,\star}$-robustness radius. Since $\mathcal{H}_{\theta_H}$ is C^1 with locally Lipschitz gradient, the appropriate local curvature object is its Clarke generalized Hessian, $\partial^2_C\mathcal{H}_{\theta_H} := \partial_C(\nabla\mathcal{H}_{\theta_H})$.

Proposition 5 (Local Clarke-Hessian certificate for PL)**.** *Let $\nabla\mathcal{H}_{\theta_H}(z_\star)=0$, and suppose that $\mathcal{H}_{\theta_H}$ is C^1 with locally Lipschitz gradient in a neighbourhood of $z_\star$. Define $m_\star := \min_{Q\in\partial^2_C\mathcal{H}_{\theta_H}(z_\star)}\lambda_{\min}(Q)$. If $m_\star>0$, then, for every $0<\mu_\star<m_\star$, there exists a convex neighbourhood $\mathcal{N}_\delta$ of $z_\star$ on which*

$$\frac{1}{2}\|\nabla\mathcal{H}_{\theta_H}(z)\|_2^2 \geq \mu_\star\left(\mathcal{H}_{\theta_H}(z)-\mathcal{H}_{\theta_H}(z_\star)\right). \tag{17}$$

The proposition is strictly local and leaves the global Hamiltonian free to be nonconvex and multi-well. When $\mathcal{H}$ is C^2 near $z_\star$, it reduces to the classical positive-definite Hessian test. The proof and activation-specific formulas are provided in Appendix B.9. To translate the PL condition into a robustness bound, the selected component must satisfy $\mathcal{C}_{\epsilon,\star}\subseteq\mathcal{N}_\delta$. We also require uniform dissipation along the energy gradient, i.e., the existence of $r_\star>0$ such that $d_H(z)\geq r_\star\|\nabla\mathcal{H}(z)\|_2^2$ on $\mathcal{N}_\delta$, which is satisfied for neighbourhood of isolated local minima with no flat directions.

Corollary 6 (PL-based robustness bound)**.** *Suppose (17) holds for some $0<\mu_\star<m_\star$. Then $\kappa_{\epsilon,\star}\geq\sqrt{2\mu_\star\epsilon}$ and $r_{\epsilon,\star}\geq r_\star$. Consequently, if $\bar{g}^{\perp}_{\epsilon,\star}>0$, then*

$$\rho_{\epsilon,\star} \geq \rho^{\mathrm{PL}}_{\epsilon,\star} := \frac{r_\star\sqrt{2\mu_\star\epsilon}}{\bar{g}^{\perp}_{\epsilon,\star}}. \tag{18}$$

The same local PL induced geometry gives more than invariance of one prescribed shell, and extends the idea of energy-driven convergence to energy-tube confiment of sollicited trajectories.

Proposition 7 (Input-to-energy tube)**.** *Let the uniform bounds $\|G(z)\|_2\leq\bar{g}$, and $\|u(t)\|_2\leq\bar{u}$ hold in the component $\mathcal{C}_{\epsilon,\star}$ and set $c:=\bar{g}\bar{u}$. If $c\leq r_\star\sqrt{2\mu_\star\epsilon}$ then $\mathcal{C}_{\epsilon,\star}$ is a $\bar{u}$-Robust Invariant set. In particular,*

$$\liminf_{t\to\infty} h_{\epsilon,\star}(z(t)) \geq \epsilon - \frac{c^2}{2r_\star^2\mu_\star}. \tag{19}$$

The condition $c\leq r_\star\sqrt{2\mu_\star\epsilon}$ guarantees that the tube centered at $z_\star$ with radius $c/(2r_\star^2\mu_\star)$ fits inside $\mathcal{C}_{\epsilon,\star}$, recovering for $c=r_\star\sqrt{2\mu_\star\epsilon}$ the robust boundary margin from a dynamic energy estimate.

Table 1: **Prediction accuracy and certified robustness across nonlinear dynamical systems.** *Comparative performance of the pH-EBM across benchmarks addressing nonlinear system identification, nonconvex dynamics, and higher-dimensional controlled systems.*

Benchmark Metrics	Published RMSE ↓ $\rho_{\epsilon,\star}$ ↑	portHNN-u RMSE ↓ $\rho_{\epsilon,\star}$ ↑	pH-EBM RMSE ↓ $\rho_{\epsilon,\star}$ ↑
Silverbox	[**0.293**,-]	[47.821, 2e-2]	[0.422, **2e-1**]
CED	[**0.054**, -]	[0.217, 3e-2]	[0.064, **6e-1**]
Duffing double-well	[-, -]	[0.254, 5e-4]	[**0.048**, **4e-1**]
3-link pendulum	[0.051, -]	[0.276, 5e-3]	[**0.014**, **2e-1**]
NanoDrone 0.5 s	[13.712, -]	[-, -]	[**12.521**, **6e-2**]
NanoDrone 5 s	[1363.701, -]	[-, -]	[**369.025**, **6e-2**]

4 Experimental validation

Benchmark suite and evaluation protocol. The experimental suite spans established nonlinear system identification (Silverbox [Wigren and Schoukens, 2013] and CED [T. and M., 2017])[2], nonconvex controlled mechanics (Duffing [Molero et al., 2012] and the n-link pendulum [Okamoto and Kojima, 2025]), and high-dimensional long-horizon flight dynamics (NanoDrone [Busetto et al., 2026]); the tasks and main results are summarized in Table 1. We compare against benchmark-specific literature references, portHNN-u, and dissipative or black-box NODE baselines where available. The reported pH-EBM models are trained with a Huber observation loss and a mixed AdamW-to-SGD optimization schedule. Prediction is evaluated using each benchmark's native metric, while robustness is assessed using the common energy-shell and normalized-radius protocol detailed in Appendix C.7. The pH-EBM framework is available at `https://github.com/sim1bet/energy-safe-dynamics`.

Accuracy and robustness across benchmarks. Across the suite, the pH-EBM combines competitive or improved predictive accuracy with substantially larger certified input margins (Table 1). On Silverbox and CED, it remains close to the strongest published errors (0.422 vs. 0.293 and 0.064 vs. 0.054, respectively) while reducing the portHNN-u error and increasing the corresponding robustness radius by approximately $10\times$ and $20\times$. The advantage is strongest on the nonconvex mechanical systems: on Duffing, RMSE decreases from 0.254 to 0.048 while the radius increases from 5×10^{-4} to 4×10^{-1}; on the 3-link pendulum, RMSE decreases from 0.276 to 0.014 while the radius increases from 5×10^{-3} to 2×10^{-1}. The latter also improves on the published RMSE of 0.051, and retains lower error under the Brownian and step out-of-distribution forcing tests reported in Appendix C.1, while requiring approximately $1/20$ of the training time of the dissipative NODE reference. Finally, on NanoDrone the pH-EBM improves the reported RMSE at both $0.5\,\mathrm{s}$ (12.521 vs. 13.712) and, more markedly, at $5\,\mathrm{s}$ (369.025 vs. 1363.701), while retaining a normalized certificate of 6×10^{-2}. These results show that the structural constraints do not impose a systematic accuracy–robustness trade-off: across the tested systems, improved certification is compatible with competitive—and in the nonconvex and long-horizon regimes, substantially improved—prediction.

Robust long time-horizon rollouts in high-dimensional Nanodrone application. The NanoDrone benchmark considers identification of flight dynamics from collections of short, disjoint $0.5,\mathrm{s}$ trajectories acquired under markedly different excitation distributions (Random, Chirp, Square, and Melon). We evaluate the pH-EBM against the reference black-box neural model of Busetto et al. [2026] under two increasingly demanding settings: **(i)** long-horizon prediction over the training-supported S3 distributions (Random, Chirp, and Square), and **(ii)** long-horizon prediction under the unseen Melon distribution, corresponding to the original extrapolation task. In both cases, models are rolled out for $5,\mathrm{s}$, i.e., ten times beyond the duration of the trajectories available during training. This setting therefore tests not only whether the learned dynamics fit short flight segments, but whether they remain predictive and operationally well behaved when recursively propagated far beyond the training horizon. On the S3 distributions, the pH-EBM achieves substantially lower prediction error than the reference black-box model, including state-of-the-art short-horizon validation accuracy together with markedly improved long-horizon reconstruction. Figure 2.a shows that its mean absolute error (MAE) remains consistently smaller throughout the $5,\mathrm{s}$ rollout, demonstrating that the structured Hamiltonian dynamics retain the expressivity required for high-dimensional flight identification without sacrificing long-term stability. More importantly, this predictive advantage is accompanied by an explicit safety certificate. Figures 2.(b-c) show the learned trajectory evolving inside the certified energy component $\mathcal{C}_{\epsilon,\star}$: as it approaches the boundary $\Gamma_{\epsilon,\star}$, the pH-EBM dynamics remain inward-pointing and the trajectory does not leave the certified operating region. In contrast, the black-box NODE accurately tracks the Random trajectory over approximately the $0.5,\mathrm{s}$ training

[2]Dataset specifications and benchmark protocols are available at `https://www.nonlinearbenchmark.org`.

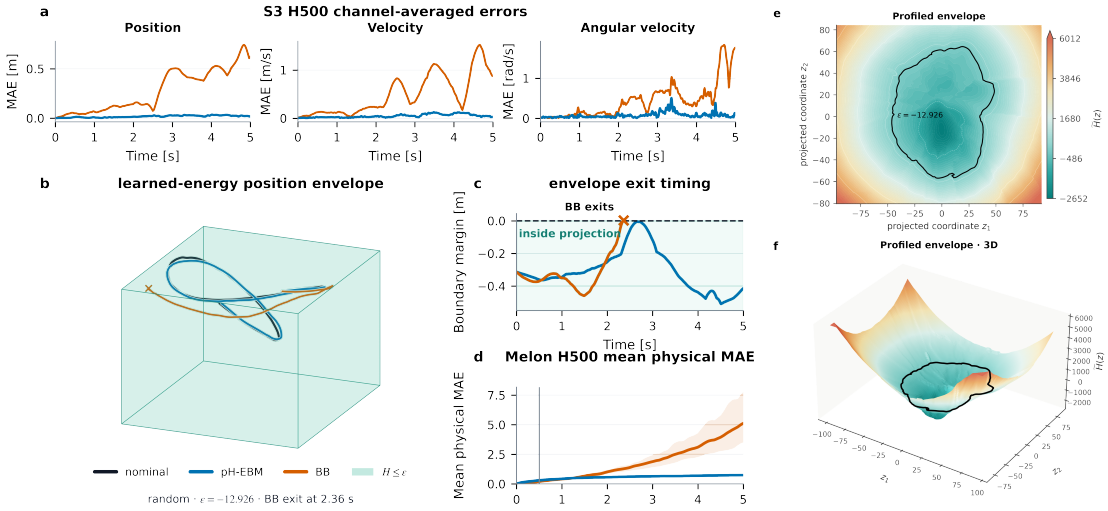


Figure 2: ***Safe* long-horizon reconstruction of NanoDrone flight dynamics with pH-EBM.** a) Mean absolute error (MAE) of pH-EBM and the black-box baseline [Busetto et al., 2026] over a $5,\mathrm{s}$ Random rollout, ten times the $0.5,\mathrm{s}$ training horizon. The pH-EBM maintains substantially lower error, while the BB error rapidly grows beyond the training window. b-c) Random-trajectory reconstruction relative to the projected certified region $\mathcal{C}_{\epsilon,\star}$ at $\epsilon = -12.926$. The pH-EBM remains inside the certified set and turns inward near $\Gamma_{\epsilon,\star}$, whereas the BB trajectory exits the region after approximately $2,\mathrm{s}$. d) Extrapolation to the unseen Melon excitation. Although BB is more accurate over the original $0.5,\mathrm{s}$ horizon, its long-rollout error quickly increases, while pH-EBM remains bounded, highlighting the distinction between short-horizon accuracy and safe recursive deployment. e-f) Two- and three-dimensional projections of the learned Hamiltonian and shell $\Gamma_{-12.926,\star}$, revealing a strongly nonconvex yet radially unbounded energy landscape that supports certified long-horizon evolution.

horizon, but subsequently drifts away and eventually exits the same projected safe region. Figures 2.(e-f) visualize the corresponding two- and three-dimensional projections of the learned Hamiltonian and its certified boundary shell, making the relation between the learned energy geometry and the long-horizon trajectory explicit. The unseen Melon distribution exposes a complementary and particularly relevant trade-off. Over the original short-horizon extrapolation task, the pH-EBM is less accurate than the reference black-box model. Yet this ordering reverses qualitatively as the rollout is extended: while the black-box prediction error grows rapidly and the trajectory develops an unstable mode, the pH-EBM remains confined to the certified region and its average MAE approaches a bounded long-horizon regime (Fig. 2.d). The experiment therefore separates two notions that short-horizon benchmark scores alone cannot distinguish: *local predictive fitness* and *safe recursive deployment*. A model may provide the smaller error immediately outside the training distribution while still generating unstable trajectories when propagated for longer horizons. The pH-EBM instead sacrifices some short-horizon Melon accuracy while retaining an *a priori* forward-invariance certificate and preventing the unstable long-term behavior observed in the unconstrained NODE. For safety-critical flight dynamics, where prediction is repeatedly propagated and instability can be more consequential than a moderate increase in instantaneous error, this distinction is central.

Nonconvex geometry and attractor-dependent certification. The Duffing oscillator [Molero et al., 2012] provides a canonical controlled system with a nonconvex double-well potential, where external forcing drives transitions between distinct operating regimes. We compare the pH-EBM against the port-Hamiltonian portHNN-u neural architecture of Desai et al. [2021], evaluating both predictive reconstruction and the geometry of the resulting ρ-safety certificate. In the non-chaotic regime, both models recover the characteristic double-well energy landscape from input-output trajectories (Fig. 3.a), showing that accurate identification can preserve the underlying nonconvex structure. The pH-EBM further reveals how certification depends on the learnt geometry and NODE specifications. The pointwise radius $\rho_{\mathrm{BF}}(z)$ decreases near the saddle separating the two wells, and the state-independent radius $\rho_{\epsilon,\star}$ correspondingly weakens as the certified shell approaches the critical energy before increasing again beyond it (Fig. 3.b). Saddle-proximity weakenanig of the certificate directly illustrates the theoretical link between critical energy levels and reduced input tolerance. Importantly, similar predictive reconstructions do not imply comparable safety margins: under the adopted configuration, the pH-EBM attains a state-independent certificate approximately two orders of magnitude larger than portHNN-u (Fig. 3.c). Robustness is therefore governed not by prediction error alone, but by the energy geometry and dissipation structure induced by the learned architecture.

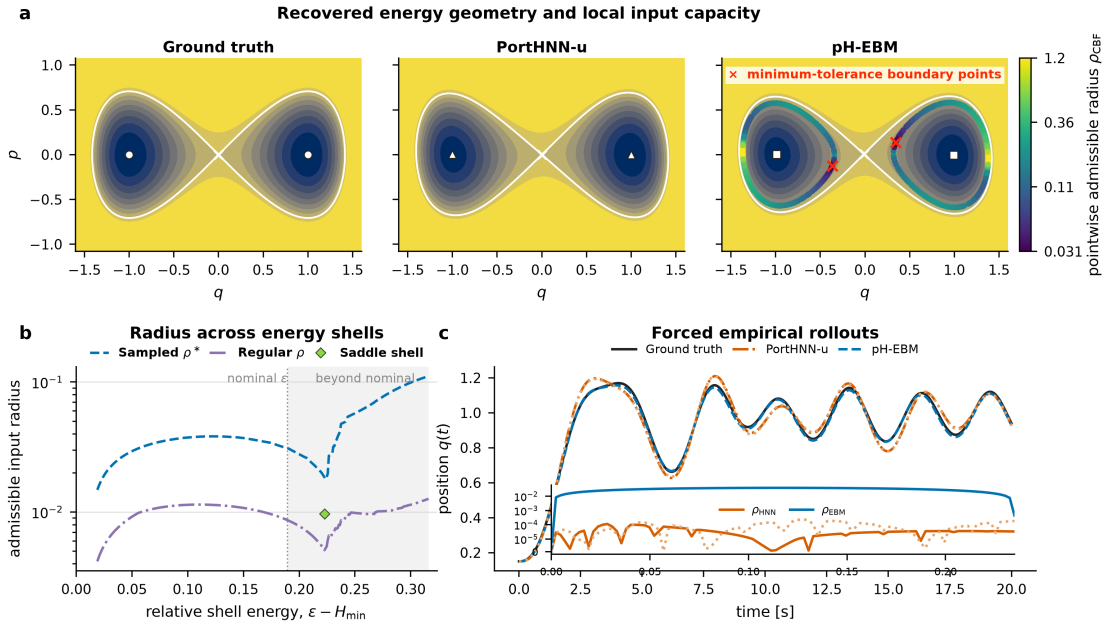


Figure 3: **Geometry reconstruction, accuracy, and certified robustness in the 2D Duffing oscillator.** a) Ground-truth Duffing Hamiltonian (left) and energies reconstructed by portHNN-u (center) and pH-EBM (right). Both recover the characteristic nonconvex double-well geometry. The pH-EBM panel additionally reports the pointwise input tolerance, whose minimum occurs near the saddle separating the two wells. b) Theoretical (purple) and empirical (blue) state-independent radii $\rho_{\epsilon,\star}$ across an energy-level sweep. The certificate decreases as the shell approaches the saddle energy, then increases again after the certified component expands beyond it, revealing reduced input tolerance near transitions between operating regimes. c) Reconstruction error and empirical robustness radius for portHNN-u and pH-EBM. While both models achieve accurate trajectory reconstruction, the pH-EBM yields certified radii approximately 2-3 orders of magnitude larger than portHNN-u.

Out-of-distribution robustness and mechanical scaling. The n-link pendulum provides a challenging nonlinear mechanical benchmark whose complexity increases rapidly with the number of coupled joints. For $n = 2$ and $n = 3$, we compare the pH-EBM against the naive and dissipative neural ODEs studied in Okamoto and Kojima [2025] and the portHNN-u model. Across these systems, the pH-EBM improves on the reference models in three key respects: predictive accuracy, robustness to adversarial inputs, and training efficiency (Appendix C.1). In particular, under both Brownian and step out-of-distribution forcing, the pH-EBM combines lower prediction error with a certified ρ-robust invariant region. Moreover, training requires approximately $1/20$ of the time for the dissipative reference, illustrating that stability and safety need not come at the computational cost of externally imposed dissipativity mechanisms.

5 Conclusions

Limitations of the pH-EBM framework. The proposed port-Hamiltonian energy-based model combines expressive nonlinear dynamics with explicit bounded-rollout and safety certificates, but introduces two practical costs: training complexity and hyperparameter sensitivity. In our experiments, pH-EBMs generally require longer training and more targeted model selection than unconstrained NODEs, although their computational cost remains comparable to portHNN-u and can be substantially lower than that of alternative stable architectures such as dissipative NODEs. A more fundamental limitation concerns the scope of the safety guarantee: the certificates derived here apply directly to the *learned model dynamics*, not automatically to the unknown physical plant. Extending invariance guarantees to the true system requires a validated enclosure of model mismatch and external disturbances, which must then be incorporated explicitly into the robust barrier condition. Reducing this model-to-plant gap, together with the training and model-selection burden, is therefore central to future deployment in safety-critical settings.

Future directions. The pH-EBM provides a general substrate for learning nonlinear dynamics while retaining stability and quantitative safety guarantees. A natural next step is to extend the framework to physical interactions not explicitly modeled here, including contact and hybrid dynamics, and to integrate the learned model with feedback controllers

rather than treating external inputs as prescribed signals. Developing scalable certificate-aware training and closed-loop safety guarantees would move the framework from certified system identification toward end-to-end deployment in safety-critical applications such as autonomous flight, robotic manipulation, and human–robot interaction.

AI usage disclosure. Large language models, including ChatGPT and Claude, were used to assist with language polishing, codebase organization and optimization, and the implementation of dataset interfaces for the experimental framework. They were not used for research ideation, formulation of the scientific contributions, development of the theoretical results, or derivation and verification of mathematical proofs. All scientific content, experimental design, implementation choices, and reported results were reviewed and validated by the authors.

Reproducibility statement. We provide extensive material to facilitate reproduction of both the theoretical and experimental results. The Appendix states the assumptions underlying the theoretical guarantees and contains complete proofs of the main results. It also documents the datasets, preprocessing, model architectures, training protocols, hyperparameters, evaluation procedures, and certificate computations used throughout the experiments. The accompanying supplementary material includes the implementation of the pH-EBM framework together with the experiment configurations and scripts required to reproduce the reported benchmarks and robustness analyses.

References

G. E. Karniadakis, I. G. Kevrekidis, L. Lu, P. Perdikaris, S. Wang, and L. Yang. Physics-informed machine learning. *Nature Reviews Physics*, 3(6):422–440, May 2021. ISSN 2522-5820. doi:10.1038/s42254-021-00314-5.

C. Legaard, T. Schranz, G. Schweiger, J. Drgoňa, B. Falay, C. Gomes, A. Iosifidis, M. Abkar, and P. Larsen. Constructing neural network based models for simulating dynamical systems. *ACM Computing Surveys*, 55(11):1–34, February 2023. ISSN 1557-7341. doi:10.1145/3567591.

R. Ashmore, R. Calinescu, and C. Paterson. Assuring the machine learning lifecycle: Desiderata, methods, and challenges. *ACM Computing Surveys*, 54(5):1–39, May 2021. ISSN 1557-7341. doi:10.1145/3453444.

L. Brunke, M. Greeff, A. W. Hall, Z. Yuan, S. Zhou, J. Panerati, and A. P. Schoellig. Safe learning in robotics: From learning-based control to safe reinforcement learning. *Annual Review of Control, Robotics, and Autonomous Systems*, 5(1):411–444, May 2022. ISSN 2573-5144. doi:10.1146/annurev-control-042920-020211.

B. Gyevnár and A. Kasirzadeh. Ai safety for everyone. *Nature Machine Intelligence*, 7(4):531–542, April 2025. ISSN 2522-5839. doi:10.1038/s42256-025-01020-y.

L. Hewing, K. P. Wabersich, M. Menner, and M. N. Zeilinger. Learning-based model predictive control: Toward safe learning in control. *Annual Review of Control, Robotics, and Autonomous Systems*, 3(1):269–296, May 2020. ISSN 2573-5144. doi:10.1146/annurev-control-090419-075625.

R. T. Q. Chen, Y. Rubanova, J. Bettencourt, and D. K. Duvenaud. Neural ordinary differential equations. volume 31. Curran Associates, Inc., 2018.

S. Greydanus, M. Dzamba, and J. Yosinski. Hamiltonian neural networks. volume 32. Curran Associates, Inc., 2019.

J. Z. Kolter and G. Manek. Learning stable deep dynamics models. volume 32. Curran Associates, Inc., 2019.

N. Lawrence, P. Loewen, M. Forbes, J. Backstrom, and B. Gopaluni. Almost surely stable deep dynamics. volume 33, pages 18942–18953. Curran Associates, Inc., 2020.

R. Kojima and Y. Okamoto. Learning deep input-output stable dynamics. volume 35, pages 8187–8198. Curran Associates, Inc., 2022.

Q. Kang, Y. Song, Q. Ding, and W. P. Tay. Stable neural ode with lyapunov-stable equilibrium points for defending against adversarial attacks. In *Advances in Neural Information Processing Systems*, volume 34, pages 14925–14937. Curran Associates, Inc., 2021.

A. Yang, J. Xiong, M. Raginsky, and E. Rosenbaum. Input-to-state stable neural ordinary differential equations with applications to transient modeling of circuits. In *Proceedings of The 4th Annual Learning for Dynamics and Control Conference*, volume 168 of *Proceedings of Machine Learning Research*, pages 663–675. PMLR, 23–24 Jun 2022.

A. Van der Schaft. *Port-Hamiltonian systems: an introductory survey*. EMS Press, May 2007. ISBN 9783985475384. doi:10.4171/022-3/65.

S. Massaroli, M. Poli, M. Bin, J. Park, A. Yamashita, and H. Asama. Stable neural flows, 2020.

F. J. Roth, D. K. Klein, M. Kannapinn, J. Peters, and O. Weeger. Stable port-hamiltonian neural networks. *arXiv*, 2025. doi:10.48550/ARXIV.2502.02480.

D. Krotov and J. J. Hopfield. Large associative memory problem in neurobiology and machine learning. In *International Conference on Learning Representations*, 2020. doi:10.48550/arXiv.2008.06996.

B. Hoover, D. H. Chau, H. Strobelt, and D. Krotov. A universal abstraction for hierarchical hopfield networks. In *The Symbiosis of Deep Learning and Differential Equations II*, 2022.

B. Hoover, Y. Liang, B. Pham, R. Panda, H. Strobelt, D. H. Chau, M. Zaki, and D. Krotov. Energy transformer. *Advances in neural information processing systems*, 36:27532–27559, 2023.

H. K. Khalil. *Nonlinear systems*. Prentice-Hall, Upper Saddle River, NJ, 2002. The book can be consulted by contacting: PH-AID: Wallet, Lionel.

F. Blanchini. Set invariance in control. *Automatica*, 35(11):1747–1767, November 1999. ISSN 0005-1098. doi:10.1016/s0005-1098(99)00113-2.

F. Blanchini and S. Miani. *Set-Theoretic Methods in Control*. Springer International Publishing, 2015. ISBN 9783319179339. doi:10.1007/978-3-319-17933-9.

A. Abate, D. Ahmed, A. Edwards, M. Giacobbe, and A. Peruffo. Fossil: a software tool for the formal synthesis of lyapunov functions and barrier certificates using neural networks. In *Proceedings of the 24th International Conference on Hybrid Systems: Computation and Control*, HSCC '21, page 1–11. ACM, May 2021. doi:10.1145/3447928.3456646.

C. Dawson, S. Gao, and C. Fan. Safe control with learned certificates: A survey of neural lyapunov, barrier, and contraction methods for robotics and control. *IEEE Transactions on Robotics*, 39(3):1749–1767, 2023. ISSN 1941-0468. doi:10.1109/tro.2022.3232542.

J. C. Willems. Dissipative dynamical systems. *European Journal of Control*, 13(2-3):134–151, January 2007. ISSN 0947-3580. doi:10.3166/ejc.13.134-151.

S. Betteti and L. Laurenti. Hybrid energy-based models for physical ai: Provably stable identification of port-hamiltonian dynamics, 2026.

A. D. Ames, X. Xu, J. W. Grizzle, and P. Tabuada. Control barrier function based quadratic programs for safety critical systems. *IEEE Transactions on Automatic Control*, 62(8):3861–3876, August 2017. ISSN 1558-2523. doi:10.1109/tac.2016.2638961.

H. Karimi, J. Nutini, and M. Schmidt. *Linear Convergence of Gradient and Proximal-Gradient Methods Under the Polyak-Łojasiewicz Condition*, page 795–811. Springer International Publishing, 2016. ISBN 9783319461281. doi:10.1007/978-3-319-46128-1_50.

Y. Pengyun, F. Cong, and L. Zhouchen. On the lower bound of minimizing polyak-Łojasiewicz functions. In Gergely Neu and Lorenzo Rosasco, editors, *Proceedings of Thirty Sixth Conference on Learning Theory*, volume 195 of *Proceedings of Machine Learning Research*, pages 2948–2968. PMLR, 12–15 Jul 2023.

T. Wigren and J. Schoukens. Three free data sets for development and benchmarking in nonlinear system identification. In *2013 European Control Conference (ECC)*, page 2933–2938. IEEE, 2013. doi:10.23919/ecc.2013.6669201.

Wigren T. and Schoukens M. *Coupled electric drives data set and reference models*. Number 024 in Technical Report Uppsala Universitet. Uppsala University Sweden, November 2017.

F. J. Molero, M. Lara, S. Ferrer, and F. Céspedes. 2-d duffing oscillator: Elliptic functions from a dynamical systems point of view. *Qualitative Theory of Dynamical Systems*, 12(1):115–139, May 2012. ISSN 1662-3592. doi:10.1007/s12346-012-0081-1.

Y. Okamoto and R. Kojima. Learning deep dissipative dynamics. *Proceedings of the AAAI Conference on Artificial Intelligence*, 39(18):19749–19757, April 2025. ISSN 2159-5399. doi:10.1609/aaai.v39i18.34175.

R. Busetto, E. Cereda, M. Forgione, G. Maroni, D. Piga, and D. Palossi. Nonlinear system identification for a nano-drone benchmark. *Control Engineering Practice*, 172:106871, 2026. ISSN 0967-0661. doi:10.1016/j.conengprac.2026.106871.

S. A. Desai, M. Mattheakis, D. Sondak, P. Protopapas, and S. J. Roberts. Port-hamiltonian neural networks for learning explicit time-dependent dynamical systems. *Physical Review E*, 104(3), 2021. ISSN 2470-0053. doi:10.1103/physreve.104.034312.

J. M. Lee. *Introduction to Smooth Manifolds*. Springer New York, 2012. ISBN 9781441999825. doi:10.1007/978-1-4419-9982-5.

F. H. Clarke. Generalized gradients and applications. *Transactions of the American Mathematical Society*, 205:247–247, 1975. ISSN 0002-9947. doi:10.1090/s0002-9947-1975-0367131-6.

A. D. Ames, S. Coogan, M. Egerstedt, G. Notomista, K. Sreenath, and P. Tabuada. Control barrier functions: Theory and applications. In *2019 18th European Control Conference (ECC)*, page 3420–3431. IEEE, 2019. doi:10.23919/ecc.2019.8796030.

A Preliminaries

Notation. $\mathbb{1}_d$ denotes the d-dimensional vector with all ones, $\mathbb{0}_d$ the d-dimensional vector with all zeros, while $\mathcal{I}_d$ denotes the identity matrix in $\mathbb{R}^{d\times d}$. We denote compact subsets of $\mathbb{R}^d$ as sets $\mathcal{B} \subset\subset \mathbb{R}^d$, and their boundary as $\partial\mathcal{B}$. For two real vectors x, y of the same dimension, $x^\top y$ denotes the standard inner product. Let $f : \mathbb{R}^d \to \mathbb{R}$; for $f \in C^k(\mathbb{R}^d)$, the function is k-times continuously differentiable. For $f \in C^2(\mathbb{R}^d)$, the gradient of f is denoted as ∇f, and the Hessian as $D(\nabla f) = D^2 f$. The partial derivative of f with respect to the variable x_i is denoted as $\partial f/\partial x_i$. Θ bounds the growth of $f \sim \Theta(g)$ through the existence of constants $a, b > 0$ such and a function g such that $a\ g(x) < f(x) < b\ g(x)$. Given functions $g : \mathbb{R}^n \to \mathbb{R}^d$ and $f : \mathbb{R}^d \to \mathbb{R}$, we denote the composition of the two functions as $f \circ g(y)$, for $y \in \mathbb{R}^n$. The abbreviation a.e. stands for almost everywhere, that is for all $\mathcal{B} \subset \mathbb{R}^d$ except zero measure sets. Given a matrix $A \in \mathbb{R}^{d\times d}$, we denote with $A^\top$ its transpose. In case A is symmetric, $\lambda_{\min}(A)$, $\lambda_{\max}(A) \in \mathbb{R}$ denote its minimum and its maximum eigenvalue. We denote with $B_r(x) \subset \mathbb{R}^d$ the ball of radius $r > 0$ and center $x \in \mathbb{R}^d$.

A.1 Lie and Clarke's derivatives

Definition 8 (Lie derivative)**.** *Let $X : \mathbb{R}^d \to \mathbb{R}^d$ be a locally Lipschitz continuous vector field and let $\Phi_X : \mathbb{R}_{\geq 0} \to \mathbb{R}^d$ denote the associated local flow. For a smooth function $g : \mathbb{R}^d \to \mathbb{R}$, the Lie derivative of g along X is defined as*

$$\mathcal{L}_X g(x) = \frac{d}{dt}|_{t=0}(g\ \circ \Phi_X^t)(x), \tag{20}$$

and coincides with the directional derivative

$$\mathcal{L}_X g(x) = \nabla_X g(x) = \nabla_x g(x)^\top X(x). \tag{21}$$

Lie derivatives [Lee, 2012, Chapters 3, 9] will be central for the characterization of the negative definiteness of the *energy* function along the trajectories generated by the EBM dynamics. We extend the notion of derivatives to nonsmooth functions by recalling Clarke's generalized gradient and directional derivative [Clarke, 1975], which apply to any locally Lipschitz function.

Definition 9 (Generalized gradient)**.** *Let $f : \mathbb{R}^d \to \mathbb{R}$ be a locally Lipschitz function. The generalized gradient of f at $x \in \mathbb{R}^d$ is denoted as $\partial f(x)$ and is the convex hull of the set of limits*

$$\lim_{k\to+\infty} \nabla f(x + h_k), \tag{22}$$

with $h_k \xrightarrow{k\to+\infty} \mathbb{0}_d$ and f differentiable at $x + h_k \in \mathbb{R}^d$ for all $k \in$.

Definition 10 (Generalized directional derivative)**.** *Let $f : \mathbb{R}^d \to \mathbb{R}$ be a locally Lipschitz function and take $v \in \mathbb{R}^d$. Then the Clarke's generalized directional derivative at $x \in \mathbb{R}^d$ is defined as*

$$f^\circ(x; v) = \lim_{h\to\mathbb{0}_d} \sup_{\delta\to 0} \frac{f(x + h + \delta v) - f(x + h)}{\delta} \tag{23}$$

$$= \max_{\xi\in\partial f(x)} \xi^\top v. \tag{24}$$

In particular, combining (21) with (24), the Clarke Lie derivative of a locally Lipschitz function f along a vector v is given by

$$\mathcal{L}_v f(x) := f^\circ(x; v) = \max_{\xi\in\partial f(x)} \xi^\top v, \tag{25}$$

extending the classical smooth definition to the nonsmooth setting.

B Technical derivations

B.1 Port-Hamiltonian energy balance

A port-Hamiltonian model [Van der Schaft, 2007] is a dynamical system characterized by dynamics

$$\dot{x} = [J(x) - R(x)]\nabla\mathcal{H} + G(x)u + w \tag{26}$$

$$y_p = G(x)^\top \nabla\mathcal{H} \tag{27}$$

where $x \in \mathbb{R}^d$ is the state of the model, $J(x) = -J(x)^\top \in \mathbb{R}^{d\times d}$ is the skew-symmetric interconnection matrix, $0 \preceq R(x) = R(x)^\top \in \mathbb{R}^{d\times d}$ is the symmetric positive-semidefinite dissipation matrix, $G(x) \in \mathbb{R}^{d\times k}$ is the input port, $u \in \mathbb{R}^k$ the external time varying input, $w \in \mathbb{R}^d$ is an unknown but bounded disturbance, and $y_p \in \mathbb{R}^k$ the power-conjugate output port. Along (4),

$$\mathcal{L}_f\mathcal{H} = \nabla\mathcal{H}^\top(J-R)\nabla\mathcal{H} + \nabla\mathcal{H}^\top(Gu+w) \tag{28}$$

$$= \nabla\mathcal{H}^\top J\nabla\mathcal{H} - \nabla\mathcal{H}^\top R\nabla\mathcal{H} + y_{\mathrm{p}}^\top u + \nabla\mathcal{H}^\top w. \tag{29}$$

For every real vector q, skew-symmetry gives

$$q^\top Jq = (q^\top Jq)^\top = q^\top J^\top q = -q^\top Jq, \tag{30}$$

so $q^\top J_{\theta_J} q = 0$. Therefore, dissipativity reduces to

$$\mathcal{L}_f\mathcal{H} = -\nabla\mathcal{H}^\top R\nabla\mathcal{H} + y_p^\top u + \nabla\mathcal{H}^\top w. \tag{31}$$

No convexity property of $\mathcal{H}$ is required for dissipativity. Coercivity is used instead to make energy sublevel sets compact and thereby preclude escape to infinity along bounded-energy trajectories. Convergence to a particular minimum requires additional information on the invariant set where the dissipation vanishes and is not implied by coercivity alone.

B.2 Derivation and coercivity of the hybrid Hamiltonian

The original energy function for a L-layered modern Hopfield network has been introduced in [Krotov and Hopfield, 2020, Hoover et al., 2022]. We identify the visible state with $g_1 = z \in \mathbb{R}^d$ and denote the hidden variables by $g_h \in \mathbb{R}^{N_h}$, $h = 2, \dots, L$. For each layer let $\mathcal{F}_h : \mathbb{R}^{N_h} \to \mathbb{R}$ be a proper convex differentiable potential and define the activation $\Psi_h(g_h) = \nabla\mathcal{F}_h(g_h)$. Adjacent layers are coupled by matrices $W_{h(h+1)} \in \mathbb{R}^{N_h\times N_{h+1}}$ with $W_{(h+1)h} = W_{h(h+1)}^\top$ and each layer has bias term $b_h \in \mathbb{R}^{N_h}$. The modern Hopfield energy is then

$$\begin{aligned}\mathcal{E}(g_1,\dots,g_L) = & -\sum_{h=1}^{L-1} \Psi_h(g_h)^\top W_{h(h+1)}\Psi_{h+1}(g_{h+1}) \\ & + \sum_{h=1}^{L}\left[g_h^\top\left(\Psi_h(g_h) - b_h\right) - \mathcal{F}_h(g_h)\right] \qquad (32)\\ = \; & g_1^\top\Psi_1(g_1) - \mathcal{F}_1(g_1) - g_1^\top b_1 \\ & + \sum_{h=2}^{L}\left[-\Psi_{h-1}(g_{h-1})^\top W_{(h-1)h}\Psi_h(g_h) + g_h^\top\left(\Psi_h(g_h) - b_h\right) - \mathcal{F}_h(g_h)\right]. \qquad (33)\end{aligned}$$

The recurrence of the dynamics, with all hidden layers taking contributions from both the layers above and below, makes training the full dynamical architecture a computational and temporal challenge. For this reason, in Betteti and Laurenti [2026] a hybrid modern Hopfield energy has been introduced, where only the visible layer maintains its dynamical characterization and all the hidden layers are expressed as static maps ϕ_h of the layer below. Specifically, by expressing $x_{(h)} = \phi_h(x_{(h-1)})$ for $h = 2, \dots, L$, both the energy and the autonomous energy-gradient contribution $\dot{x} \propto -\nabla\mathcal{E}(x)$ can be expressed in terms of the visible layer state.

With $g_1 = z$, the visible layer internal contribution to the Hamiltonian is

$$z^\top\Psi_1(z) - \mathcal{F}_1(z) = z^\top z - \frac{1}{2}\|z\|_2^2 \tag{34}$$

$$= \frac{1}{2}\|z\|_2^2. \tag{35}$$

For every hidden layer $h \geq 2$, the feedforward constraint (7) yields

$$g_h^\top\Psi_h(g_h) = \Psi_{h-1}(g_{h-1})^\top W_{h(h-1)}^\top\Psi_h(g_h) \tag{36}$$

$$= \Psi_{h-1}(g_{h-1})^\top W_{(h-1)h}\Psi_h(g_h), \tag{37}$$

which cancels exactly the coupling term between layers $h-1$ and h in (33). Applying this cancellation for $h = 2, \dots, L$ leaves precisely

$$\begin{aligned}\mathcal{E}(z, g_2, \dots, g_L) &\equiv \mathcal{H}_{\theta_H}(z) \\ &= \frac{1}{2}\|z\|_2^2 - z^\top b_1 - \sum_{h=2}^{L}\left(\mathcal{F}_h(g_h(z)) + g_h(z)^\top b_h\right) \qquad (38)\end{aligned}$$

and the associated neural network action has autonomous contribution

$$\dot{z} \propto -\nabla \mathcal{H}_{\theta_H}(z) \tag{39}$$

We next prove (43). Suppose the first hidden activation has bounded image: there exists $M_2 < +\infty$ such that $\|\Psi_2(x)\|_2 \leq M_2$ for all $x \in \mathbb{R}^{N_2}$. Since $\Psi_2 = \nabla \mathcal{F}_2$, the mean-value inequality gives

$$|\mathcal{F}_2(z) - \mathcal{F}_2(0)| \leq M_2 \|z\|_2, \tag{40}$$

so

$$\mathcal{F}_2(W_{21} z) \leq |\mathcal{F}_2(0)| + M_2 \|W_{21}\|_2 \|z\|_2. \tag{41}$$

Moreover, $g_3 = W_{32}\Psi_2(g_2)$ belongs to a compact set independent of z. By continuity of Ψ_3, its image over this compact set is bounded; recursively, every $g_h(z)$ for $h \geq 3$ lies in a compact set. Continuity of $\mathcal{F}_h$ and compactness of the layer variables g_h then yields constants $C_h < +\infty$ such that

$$\mathcal{F}_h(g_h(z)) + g_h(z)^\top b_h \leq C_h, \qquad h = 3, \ldots, L. \tag{42}$$

Substituting these estimates into (6) yields

$$\mathcal{H}_{\theta_H}(z) \geq \frac{1}{2}\|z\|_2^2 - c_1 \|z\|_2 - c_0 \tag{43}$$

with

$$c_1 = \|W_{21}\|_2 (M_2 + b_2) + b_1, \qquad c_0 = |\mathcal{F}_2(0)| + \sum_{h=3}^{L} C_h. \tag{44}$$

The quadratic term dominates the affine lower bound as $\|z\|_2 \to +\infty$, proving coercivity.

B.3 Barrier functions and invariance

Consider a controlled system $\dot{z} = f(z, u, w)$ and a continuously differentiable function $h : \mathbb{R}^d \to \mathbb{R}$. The superlevel set

$$\mathcal{C} := \{z : h(z) \geq 0\} \tag{45}$$

is *invariant* if every solution initialized in $\mathcal{C}$ remains in $\mathcal{C}$ for all future times for which the solution exists. A standard way to certify this property is through a zeroing barrier function. Let α be an extended class-$\mathcal{K}$ function, i.e., a continuous strictly increasing function satisfying $\alpha(0) = 0$. For a fixed admissible input/disturbance pair, the inequality

$$\mathcal{L}_f h(z, u, w) + \alpha(h(z)) \geq 0 \tag{46}$$

prevents h from decreasing through zero. Under the usual solution-existence assumptions, if (46) holds on a neighborhood of $\mathcal{C}$, then $\mathcal{C}$ is invariant. If it holds for every u in a prescribed input set and every w in a prescribed disturbance set, the same set is robustly invariant for those exogenous signals [Ames et al., 2017, 2019].

On a regular boundary point $z \in \partial\mathcal{C}$, where $h(z) = 0$ and $\nabla h(z) \neq 0$, the condition reduces to the geometric inward-pointing requirement

$$\mathcal{L}_f h(z, u, w) \geq 0. \tag{47}$$

Equivalently, the vector field must lie in the tangent cone of $\mathcal{C}$; this is the local form of Nagumo's invariance condition.

B.4 BF invariance, robust input sets, and boundary geometry

We first distinguish three levels of certification that are used in the main text: a pointwise zeroing-BF condition, a boundary-only invariance condition, and a state-independent set of exogenous inputs that satisfies the boundary condition everywhere. In the following treatment, we will refer to a specific condition being *robustly* satisfied when it holds over the entire support $\mathcal{W}$ of the unknown disturbances.

Let $h_\epsilon(z) = \epsilon - \mathcal{H}(z)$ and consider the perturbed port-Hamiltonian dynamics

$$\dot{z} = [J(z) - R(z)]\nabla \mathcal{H}(z) + G(z)u + w, \qquad w \in \mathcal{W}(z), \tag{48}$$

where $\mathcal{W}(z) \subset \mathbb{R}^d$ is nonempty and compact. Write $d_H := \nabla \mathcal{H}^\top R \nabla \mathcal{H}$, and $a_H := G^\top \nabla \mathcal{H}$. Then

$$\mathcal{L}_f h_\epsilon = d_H - a_H^\top u - \nabla \mathcal{H}^\top w. \tag{49}$$

Therefore the zeroing-BF inequality $\mathcal{L}_f h_\epsilon + \alpha(h_\epsilon) \geq 0$ for every $w \in \mathcal{W}(z)$ is equivalent to

$$a_H(z)^\top u + \sigma_{\mathcal{W}(z)}(\nabla\mathcal{H}(z)) \leq d_H(z) + \alpha(h_\epsilon(z)). \tag{50}$$

Setting $\mathcal{W}(z) = \{0\}$ recovers the result without bounded perturbayions. If $a_H(z) = 0$, no division by $\|a_H(z)\|_*$ is performed: the energy effect of the input is identically zero to first order and feasibility depends only on dissipation, the barrier slack, and the disturbance term.

For any norm $\|\cdot\|$ on input space with dual norm $\|\cdot\|_*$, Hölder's inequality gives

$$a_H^\top u \leq \|a_H\|_* \|u\|. \tag{51}$$

Thus (10) is sufficient in the nominal case. It is the largest origin-centered norm ball contained in the unconstrained half-space because

$$\sup_{\|u\| \leq \rho} a_H^\top u = \rho \|a_H\|_*. \tag{52}$$

B.5 Exact robust boundary certificate

Let $z_\star$ be an isolated local minimum and define

$$\mathcal{C}_{\epsilon,\star} := \mathrm{Comp}_{z_\star}\{z : h_{\epsilon,\star}(z) \geq 0\}, \qquad h_{\epsilon,\star}(z) := \epsilon - V_\star(z), \qquad \Gamma_{\epsilon,\star} := \partial\mathcal{C}_{\epsilon,\star}. \tag{53}$$

The energy component is both connected (by definition) and compact by coercivity of $\mathcal{H}$, and the solutions of (48) exist uniquely for the admissible measurable inputs under consideration.

On $\Gamma_{\epsilon,\star}$ one has by definition that $h_{\epsilon,\star} = 0$, therefore the robust inward-pointing condition is

$$a_H(z)^\top u + \sigma_{\mathcal{W}(z)}(\nabla\mathcal{H}(z)) \leq d_H(z). \tag{54}$$

Before proceeding with the first theorem, we provide a general notion from differential geometry, that is, that of regular hypersurface, which will be used in reference to the boundaries of our candidate safe sets $\mathcal{C}_{\epsilon,\star}$.

Proposition 11 (Regular $(d-1)$-dim hypersurface)**.** *Let $f : \mathbb{R}^d \to \mathbb{R}$ be a C^1 function, and for $s \in \mathbb{R}$ define the level set*

$$\Gamma_s := \{z \in \mathbb{R}^d : \ f(z) = s\} \tag{55}$$

Then we say that Γ_s is a regular $(d-1)$-dim C^1-hypersurface if

$$\nabla f(z) \neq 0 \qquad \forall z \in \Gamma_s. \tag{56}$$

When the Hamiltonian $\mathcal{H}$ has isolated critical points, the shells $\Gamma_{\epsilon,\star}$ are regular almost everywhere $\epsilon > 0$, as there will be finitely many values of ϵ such that a saddle point or local maxima lie in $\Gamma_{\epsilon,\star}$. The existence of entire connected open subsets $\mathcal{V} \subset \mathbb{R}^d$ where the regularity assumption is not satisfied, and hence such that $\mathcal{H}(z) = c$ for all $z \in \mathcal{V}$, is possible but unlikely. The existence of such regions would require the following two conditions to hold simultaneously

$$\nabla\mathcal{H}(z) = 0, \quad \det\nabla^2\mathcal{H}(z) = 0 \qquad \forall z \in \mathcal{V}. \tag{57}$$

More specifically, if the Hamiltonian $\mathcal{H}$ is constituted only by primitives $\{\mathcal{F}_h\}_{h=1}^L$ of real analytic activation functions - such as the $\mathrm{softmax}$, $\tanh$, $\mathrm{logistic}$, and $\mathrm{softplus}$ - then the existence of open subsets $\mathcal{V}$ with constant energy is impossible due to analytic continuation. If such $\mathcal{V}$ existed, it would imply that $\mathcal{H} \equiv c$ on all $\mathbb{R}^d$, which contraddicts the coercivity of $\mathcal{H}$. Therefore, the condition in (57) can only be satisfied by neural networks characterized by less regular activation functions with broad plateaus - such as ReLU - and even in such cases remains unlikely. In the following theorem, we specialize the definition of safe sets as defined by the energy BF to our port-Hamiltonian framework.

Theorem 12 (Maximal uniform robust input set)**.** *Let $\Gamma_{\epsilon,\star}$ be the regular boundary of the zero level set $h_{\epsilon,\star} = 0$ in the sense of Proposition 11, and define*

$$\overline{\mathcal{U}}_{\epsilon,\star}^{\mathrm{rob}} := \mathcal{U} \cap \bigcap_{z \in \Gamma_{\epsilon,\star}} \left\{u : a_H(z)^\top u + \sigma_{\mathcal{W}(z)}(\nabla\mathcal{H}(z)) \leq d_H(z)\right\}. \tag{58}$$

Then

(i) for any locally essentially bounded input satisfying $u(t) \in \overline{\mathcal{U}}_{\epsilon,\star}^{\mathrm{rob}}$ almost everywhere, $\mathcal{C}_{\epsilon,\star}$ is robustly invariant for every measurable disturbance $w(t) \in \mathcal{W}(z(t))$;

(ii) $\overline{\mathcal{U}}^{\mathrm{rob}}_{\epsilon,\star}$ *is maximal with respect to set inclusion among state-independent subsets of* $\mathcal{U}$ *whose every element satisfies* (54) *at every boundary state.*

Proof. (i) let $z \in \Gamma_{\epsilon,\star}$, and observe that $a_H(z)^\top u + \sigma_{\mathcal{W}(z)}(\nabla\mathcal{H}(z)) \le d_H(z) \iff \mathcal{L}_f h_{\epsilon,\star}(z,u,w) \ge 0$ for every $w \in \mathcal{W}(z)$. Intersecting these pointwise admissible half-spaces over the full boundary yields exactly $\overline{\mathcal{U}}^{\mathrm{rob}}_{\epsilon,\star}$. $\mathcal{L}_f h \ge 0$ on $\Gamma_{\epsilon,\star}$ implies that the neural network action is in the tangent cone of the sublevel component, so Nagumo's tangent condition yields robust invariance.

(ii) Let $\mathcal{U}^-_{\epsilon,\star}$ be such that $\mathcal{U}^-_{\epsilon,\star}/\overline{\mathcal{U}}^{\mathrm{rob}}_{\epsilon,\star} \neq \emptyset$. Then there exists $u \in \mathcal{U}^-_{\epsilon,\star}$ such that

$$a_H(z)^\top u + \sigma_{\mathcal{W}(z)}(\nabla\mathcal{H}(z)) > d_H(z) \tag{59}$$

and $\mathcal{U}^-_{\epsilon,\star}$ does not make $\mathcal{C}_{\epsilon,\star}$ robustly invariant.

□

It is now easy to generalize the previous definition of safety ensured input set to arbitrary geometries instead of normed balls.

Corollary 13 (Certification of arbitrary compact input sets)**.** *Let* $\mathcal{U}_0 \subseteq \mathcal{U}$ *be compact. Then* $\mathcal{U}_0 \subseteq \overline{\mathcal{U}}^{\mathrm{rob}}_{\epsilon,\star}$ *if and only if*

$$\sigma_{\mathcal{U}_0}(a_H(z)) + \sigma_{\mathcal{W}(z)}(\nabla\mathcal{H}(z)) \le d_H(z) \qquad \forall z \in \Gamma_{\epsilon,\star} \tag{60}$$

Proof. Let $\zeta(z,u) = a_H^\top u + \sigma_{\mathcal{W}(z)}(\nabla\mathcal{H}(z)) - d_H(z)$. The function ζ is continuous in both of its arguments. Since both $\Gamma_{\epsilon,\star}$ and $\mathcal{U}_0$ are compact, then by Weierstrass theorem ζ constrained to this sets admits maximum and minimum with respect to both of its arguments.

The certification of arbitrary compact input sets then requires

$$\begin{aligned} 0 &< \min_{z\in\Gamma_{\epsilon,\star}} \max_{u\in\mathcal{U}_0} \zeta(z,u) \\ &= \min_{z\in\Gamma_{\epsilon,\star}} \left[\max_{u\in\mathcal{U}_0} a_H^\top u\right] + \sigma_{\mathcal{W}(z)}(\nabla\mathcal{H}(z)) - d_H(z) \\ &= \min_{z\in\Gamma_{\epsilon,\star}} \sigma_{\mathcal{U}_0}(a_H(z)) + \sigma_{\mathcal{W}(z)}(\nabla\mathcal{H}(z)) - d_H(z) \end{aligned} \tag{61}$$

□

Define the residual boundary dissipation after worst-case disturbance injection,

$$\delta(z) := d_H(z) - \sigma_{\mathcal{W}(z)}(\nabla\mathcal{H}(z)). \tag{62}$$

Assume $\delta(z) \ge 0$ on the boundary, so that zero input is itself robustly admissible. The exact radius of the largest origin-centered norm ball contained in (58) is

$$\rho_{\epsilon,\star} := \inf_{\substack{z\in\Gamma_{\epsilon,\star} \\ a_H(z)\neq 0}} \frac{\delta(z)}{\|a_H(z)\|_*}, \tag{63}$$

Boundary points with $a_H(z) = 0$ impose no input-radius restriction as long as $\delta(z) \ge 0$. Consequently, every $\rho \le \rho_{\epsilon,\star}$ satisfies $\{u : \|u\| \le \rho\} \cap \mathcal{U} \subseteq \overline{\mathcal{U}}^{\mathrm{rob}}_{\epsilon,\star}$.

A frequently useful but more conservative separable bound follows from

$$\frac{\min_{\Gamma_{\epsilon,\star}} \delta(z)}{\max_{\Gamma_{\epsilon,\star}} \|a_H(z)\|_*} \le \inf_{\substack{z\in\Gamma_{\epsilon,\star} \\ a_H(z)\neq 0}} \frac{\delta(z)}{\|a_H(z)\|_*}. \tag{64}$$

This inequality corrects the tempting but generally false replacement of the infimum of a ratio by the ratio of independent extrema.

B.6 Regular energy shells and normal robustness decomposition

The regularity condition for the shell has deeper geometric implications, since it allows to clearly outline the vectors resulting in inward and outward motion with respect to the set $\mathcal{C}_{\epsilon,\star}$.

Proposition 14 (Regular energy boundary)**.** *Assume that*

$$\nabla\mathcal{H}(z) \neq 0, \qquad z \in \Gamma_{\epsilon,\star}. \tag{65}$$

Then every point of $\Gamma_{\epsilon,\star}$ belongs locally to the regular level set $\{z : h_{\epsilon,\star} = 0\}$, which is a C^1 embedded hypersurface of dimension $(d-1)$, and we can characterize the tangent space and the tangent cone at $\Gamma_{\epsilon,\star}$ as

$$T_z\Gamma_{\epsilon,\star} = \{v : \nabla\mathcal{H}(z)^\top v = 0\}, \tag{66}$$

$$T_{\mathcal{C}_{\epsilon,\star}}(z) = \{v : \nabla\mathcal{H}(z)^\top v \leq 0\}. \tag{67}$$

In addition, if the boundary is compact, then $\kappa_{\epsilon,\star} := \inf_{\Gamma_{\epsilon,\star}} \|\nabla\mathcal{H}\|_2 > 0$.

Proof. Consider the BF $h_{\epsilon,\star}(z)$, and its directional derivative $Dh_{\epsilon,\star}(z)[v] = -\nabla\mathcal{H}(z)^\top v$. Since $h_{\epsilon,\star}$ maps $\mathbb{R}^d$ to $\mathbb{R}$, the derivative is surjective at all points $z \in \Gamma_{\epsilon,\star}$ such that $\nabla\mathcal{H}(z) \neq 0$. Since by assumption $\nabla\mathcal{H}(z) \neq 0$, we have that $h_{\epsilon,\star} = 0$ is a submersion and $\Gamma_{\epsilon,\star}$ is a C^1 $(d-1)$-hypersurface.

Because $\mathcal{C}_{\epsilon,\star}$ is the superlevel set $h_{\epsilon,\star} \geq 0$, the feasible first-order directions are defined by $Dh_{\epsilon,\star}(z)[v] \geq 0$,or equivalently $\nabla\mathcal{H}(z)^\top v \leq 0$, giving (67). Finally, by Weierstrass theorem the continuous function $\|\nabla\mathcal{H}\|_2$ attains a positive maximum and minimum on the compact boundary, which must be strictly positive given the assumption $\nabla\mathcal{H}(z) \neq 0$ on $\Gamma_{\epsilon,\star}$. □

The characterization of tangent space and tangent cone of $\Gamma_{\epsilon,\star}$, and in particular the nonzero-gradient condition is essential for a first-order barrier test. If $\nabla\mathcal{H} = 0$ at a boundary point, $\mathcal{L}_f h_{\epsilon,\star}$ vanishes and $Dh_{\epsilon,\star}(z)[v] \equiv 0$ for any vector $v \in \mathbb{R}^d$: therefore, the BF condition fails to distinguish inward from outward motion. Thereby, regularity makes the energy gradient a valid boundary normal for the BF argument.

For $z \in \Gamma_{\epsilon,\star}$ define the outward energy normal

$$\eta(z) := \frac{\nabla\mathcal{H}(z)}{\|\nabla\mathcal{H}(z)\|_2}. \tag{68}$$

Positive homogeneity of support functions and the identities

$$d_H(z) = \|\nabla\mathcal{H}(z)\|_2^2 \eta(z)^\top R(z)\eta(z), \tag{69}$$

$$\|a_H(z)\|_* = \|\nabla\mathcal{H}(z)\|_2 \|G(z)^\top\eta(z)\|_*, \tag{70}$$

$$\sigma_{\mathcal{W}(z)}(\nabla\mathcal{H}(z)) = \|\nabla\mathcal{H}(z)\|_2 \sigma_{\mathcal{W}(z)}(\eta(z)) \tag{71}$$

show that, whenever $G(z)^\top\eta(z) \neq 0$,

$$\frac{\delta(z)}{\|a_H(z)\|_*} = \frac{\|\nabla\mathcal{H}(z)\|_2 \eta(z)^\top R(z)\eta(z) - \sigma_{\mathcal{W}(z)}(\eta(z))}{\|G(z)^\top\eta(z)\|_*}. \tag{72}$$

Interestingly, (72) reveals that, with respect to the input port, only the directions that are normal to the level set consumes the energy margin, while tangent directions leave it unaffected. In particular, if $\|G(z)^\top\eta(z)\|_* = 0$ for all $z \in \Gamma_{\epsilon,\star}$, the input is tangent to the Hamiltonian shells over the entire boundary and the energy certificate imposes no magnitude restriction on u; only the disturbance margin must remain nonnegative.

Let

$$r_{\epsilon,\star} := \inf_{\Gamma_{\epsilon,\star}} \eta^\top R\eta, \tag{73}$$

$$\bar{g}^\perp_{\epsilon,\star} := \sup_{\Gamma_{\epsilon,\star}} \|G^\top\eta\|_*, \tag{74}$$

$$\bar{w}^\perp_{\epsilon,\star} := \sup_{\Gamma_{\epsilon,\star}} \sigma_{\mathcal{W}(z)}(\eta). \tag{75}$$

If $\bar{g}^\perp_{\epsilon,\star} > 0$, and $r_{\epsilon,\star}\kappa_{\epsilon,\star} \geq \bar{w}^\perp_{\epsilon,\star}$, then (72) yields

$$\rho_{\epsilon,\star} \geq \frac{r_{\epsilon,\star}\kappa_{\epsilon,\star} - \bar{w}^\perp_{\epsilon,\star}}{\bar{g}^\perp_{\epsilon,\star}}, \tag{76}$$

which proves (72).

For each attractor component one may define the maximal regular-shell energy

$$\epsilon_\star^{\mathrm{reg}} := \sup\{\bar{\epsilon} > 0 : \nabla\mathcal{H}(z) \neq 0 \quad \forall z \in \Gamma_{e,\star},\ \forall \epsilon \in (0, \bar{\epsilon})\}. \tag{77}$$

Every compact shell below this threshold is eligible for the regular-boundary certificate. When a shell approaches a critical energy containing a saddle or another stationary point, $\kappa_{\epsilon,\star}$ can collapse toward zero, quantitatively revealing the loss of first-order robustness. After the saddle point, the ρ-safety certificate exit resumes, but the new certified region will contain the sets on "both" sides of the saddle.

B.7 Bounds under local PL and persistent forcing

Assume the component $\mathcal{C}_{\epsilon,\star}$ lies inside a neighborhood $\mathcal{N}_\star$ of a local minimum $z_\star \in \mathbb{R}^d$ on which the local Polyak-Łojasiewicz (PL) condition [Karimi et al., 2016] holds for the Hamiltonian $\mathcal{H}$.

$$\frac{1}{2}\|\nabla\mathcal{H}(z)\|_2^2 \geq \mu_\star\left(\mathcal{H}(z) - \mathcal{H}(z_\star)\right) = \mu_\star V_\star(z), \qquad d_H(z) \geq r_\star\|\nabla\mathcal{H}(z)\|_2^2, \tag{78}$$

for constants $\mu_\star, r_\star > 0$ where $r_\star \leq \min_{z\in\mathcal{C}_{\epsilon,\star}} \lambda_{\min}(R(z))$. The PL condition is associated to a region of convex growth around a minima, and since we are requiring it to be satisfied for all points in $\mathcal{N}_\star$, then necessarily we will have that $\mathcal{N}_\star \subset \mathcal{C}_{\epsilon_\star^{\mathrm{reg}},\star}$, where $\epsilon_\star^{\mathrm{reg}}$ has been defined in (77). Indeed, it holds that

$$\lim_{\epsilon\to\epsilon_\star^{\mathrm{reg}}} \kappa_{\epsilon,\star} = 0 \tag{79}$$

and $\mathcal{N}_\star \supset \mathcal{C}_{\epsilon_\star^{\mathrm{reg}},\star}$ would lead, for a point $z \in \Gamma_{\epsilon_\star^{\mathrm{reg}},\star}$ to the contraddiction

$$0 \geq \underbrace{\mu_\star V_\star(z)}_{>0} \tag{80}$$

Let $R(x) \succ 0$ over the entire component $\mathcal{C}_{\epsilon,\star}$. Then the PL condition implies

$$d_H(z) \geq 2r_\star\mu_\star V_\star(z). \tag{81}$$

For the zeroing-BF condition inside the component, (81) show that

$$a_H(z)^\top u \leq 2r_\star\mu_\star V_\star(z) + \alpha\big(\epsilon - V_\star(z)\big) \tag{82}$$

is a sufficient conservative approximation of the exact admissible half-space. If $\alpha(s) = \gamma s$, the right-hand side is affine in $q = V_\star(z) \in [0, \epsilon]$. Considering the first derivative of $l(q) = 2r_\star\mu_\star q + \gamma(\epsilon - q)$, we have $\dot{l}(q) = 2r_\star\mu_\star - \gamma$. Consequently, if $2r_\star\mu_\star > \gamma$, then l is strictly increasing over $[0, \epsilon]$ and its minimum on the interval is $\epsilon\gamma$. Conversely, if $\gamma > 2r_\star\mu_\star$, the l is strictly decreasing over $[0, \epsilon]$ and its minimum is $2r_\star\mu_\star\epsilon$. Therefore, the minimum of the r.h.s. can be expressed as

$$\beta_\epsilon = \epsilon\min\{2r_\star\mu_\star, \gamma\}. \tag{83}$$

Consequently, with $A_\epsilon := \sup_{z\in\mathcal{C}_{\epsilon,\star}} \|a_H(z)\|_* < \infty$, the bound

$$\|u\| \leq \frac{\beta_\epsilon}{A_\epsilon} \tag{84}$$

ensures the chosen zeroing-BF inequality throughout the component whenever $A_\epsilon > 0$. This is stronger than boundary-only invariance and is correspondingly more conservative.

On $\Gamma_{\epsilon,\star}$, PL yields

$$\|\nabla\mathcal{H}(z)\|_2 \geq \sqrt{2\mu_\star\epsilon} > 0. \tag{85}$$

Hence the boundary is regular by Proposition 14 and $\kappa_{\epsilon,\star} \geq \sqrt{2\mu_\star\epsilon}$. Substitution in the normal robustness bound gives the robust PL corollary

$$\rho_{\epsilon,\star} \geq \frac{r_{\epsilon,\star}\sqrt{2\mu_\star\epsilon} - \bar{w}_{\epsilon,\star}^\perp}{\bar{g}_{\epsilon,\star}^\perp}, \tag{86}$$

whenever the numerator is nonnegative. This is an invariance-only certificate; it need not satisfy the selected zeroing-BF inequality at every interior point.

B.8 Input-to-energy robustness under persistent forcing

In this subsection we are going to present an additional bound on the confinement of the port-Hamiltonian trajectories $z(t)$ whenever they are in a component $\mathcal{C}_{\epsilon,\star}$ and we have uniform bounds on the input u, the input-port G and the disturbances w.

Proposition 15 (Input-to-energy tube)**.** *Let the uniform bounds*

$$\|G(z)\|_2 \leq \bar{g}, \qquad \|u(t)\|_2 \leq \bar{u}, \qquad \|w(t)\|_2 \leq \bar{w}. \tag{87}$$

hold in the component $\mathcal{C}_{\epsilon,\star}$ and set $c := \bar{g}\bar{u} + \bar{w}$. If

$$c \leq r_\star \sqrt{2\mu_\star \epsilon}, \tag{88}$$

then $\mathcal{C}_{\epsilon,\star}$ is robustly invariant and every trajectory initialized in the component satisfies (**??**)*. In particular,*

$$\limsup_{t\to\infty} V_\star(z(t)) \leq \frac{c^2}{2r_\star^2 \mu_\star}. \tag{89}$$

Proof. Along the trajectories of the latent port-Hamiltonian EBM model, defined by dynamics (48) with vector field $f(z) = [J(z) - R(z)]\nabla\mathcal{H}(z) + G(z)u + w$, we have the following Lie derivative on the difference function $V_\star$.

$$\begin{aligned}
\mathcal{L}_f V_\star(z,u) &= \mathcal{L}_f \mathcal{H}\,(z,u) \\
&= -\nabla\mathcal{H}(z)R(z)\nabla\mathcal{H}(z) + (G(z)u + w)^\top \nabla\mathcal{H}(z) \\
&\leq -r_\star \|\nabla\mathcal{H}(z)\|_2^2 + c\|\nabla\mathcal{H}(z)\|_2 \\
&\leq -\frac{r_\star}{2}\|\nabla\mathcal{H}(z)\|_2^2 + \frac{c^2}{2r_\star} \\
&\leq -r_\star\mu_\star V_\star(z) + \frac{c^2}{2r_\star}
\end{aligned} \tag{90}$$

where in the fourth passage we have used Young's inequality $ab \leq \varepsilon a^2/2 + b^2/(\varepsilon 2)$ with $a = \|\nabla\mathcal{H}\|$, $b = c$, and $\varepsilon = r_\star$ and in the last passage the PL condition.

Under Assumption 88 we actually have on the level set $h_{\epsilon,\star} = 0$

$$\begin{aligned}
-\mathcal{L}_f h_{\epsilon,\star}\,(z,u) &= \mathcal{L}_f V_\star\,(z,u) \\
&\leq -r_\star\mu_\star\epsilon + \frac{(r_\star\sqrt{2\mu_\star\epsilon})^2}{2r_\star} = 0
\end{aligned} \tag{91}$$

and consequently we also have that $\mathcal{C}_{\epsilon,\star}$ is robustly invariant for all (u,w) satisfying Assumption 87. Finally, applying the comparison lemma on (90) we obtain

$$V_\star(t) \leq e^{-r_\star\mu_\star t}V_\star(0) + \frac{c^2}{2r_\star^2\mu_\star}\left(1 - e^{-r_\star\mu_\star t}\right). \tag{92}$$

Taking the limit $t \to +\infty$ we conclude. □

Notice finally that when $c < r_\star\sqrt{2\mu_\star\epsilon}$, we are actually stating that asymptotically the trajectories of the system will be bounded to level sets $h_{\epsilon,\star} = s$ with $s \geq \epsilon - c^2/(2r_\star\mu_\star)$.

B.9 Activation-dependent guidance for local PL

We begin the current subsection by proving the general result of Proposition 5, and the proceed to specialize the result to an example architecture and different classes of activation functions.

Proof. Since $\mathcal{H} \in C^{1,1}_{\text{loc}}(\mathbb{R}^d)$, then the Hessian $\nabla^2\mathcal{H}(z)$ is defined a.e. $z \in \mathbb{R}^d$, and at the points of non-differentiability we relay on Clarke's generalized Jacobian $z \mapsto \partial(\nabla\mathcal{H})(z)$ as defined in Definition 9. The set $\partial(\nabla\mathcal{H})(z)$ is nonempty, compact, convex, and upper semi-continuous, and consequently

$$Q(z) = Q(z)^\top \qquad \forall Q(z) \in \partial(\nabla\mathcal{H})(z). \tag{93}$$

Let $m_\star = \min_{Q(z)\in\partial(\nabla\mathcal{H})(z)} \lambda_{\min}(Q(z))$, and since $z_\star$ is a local minimum, by compactness we have that $Q(z) \succ 0$ for all $Q(z) \in \partial(\nabla\mathcal{H})(z)$ and consequently $m_\star > 0$.

By upper semicontinuity of the generalized Jacobian there exists $\delta > 0$ and a neighbourhood $\mathcal{N}_\delta$ of the local minimum $z_\star$ such that

$$\|Q(z) - Q(z_\star)\|_2 \le \delta \qquad \forall z \in \mathcal{N}_\delta \tag{94}$$

Fix $\mu_\star \in (0, m_\star)$ such that $\delta < m_\star - \mu_\star$. Then is holds that

$$Q(z) \succ \mu_\star I_d \qquad \forall Q(z) \in \partial(\nabla\mathcal{H})(z),\ \forall z \in \mathcal{N}_\delta \tag{95}$$

By Taylor expansion of $\mathcal{H}$ at $z_\star$ in $\mathcal{N}_\delta$ we obtain

$$\begin{aligned}\mathcal{H}(z_\star) &\ge \mathcal{H}(z) + \nabla\mathcal{H}(z)^\top(z_\star - z) + \tfrac{1}{2}(z_\star - z)^\top Q(z)(z_\star - z) \quad Q(z) \in \partial(\nabla\mathcal{H})(z)\\ &\ge \mathcal{H}(z) + \nabla\mathcal{H}(z)^\top(z_\star - z) + \tfrac{\mu_\star}{2}\|z_\star - z\|_2^2\end{aligned} \tag{96}$$

where in the last passage we have applied (95). Re-ordering the terms, it holds that

$$\begin{aligned}\mathcal{H}(z) - \mathcal{H}(z_\star) &\le -\nabla\mathcal{H}(z)^\top(z_\star - z) - \tfrac{\mu_\star}{2}\|z_\star - z\|_2^2\\ &\le \nabla\mathcal{H}(z)^\top(z - z_\star) - \tfrac{\mu_\star}{2}\|z - z_\star\|_2^2\end{aligned} \tag{97}$$

The r.h.s. is a function $ar - \mu_\star r^2/2$ that reaches a maximum in $r > 0$ for $r = a/\mu_\star$. Taking now $a = \nabla\mathcal{H}$ and $r = z - z_\star$ we obtain

$$\mathcal{H}(z) - \mathcal{H}(z_\star) \le \tfrac{1}{2\mu_\star}\|\nabla\mathcal{H}(z)\|_2^2. \tag{98}$$

□

For a one-hidden-layer hybrid Hamiltonian,

$$\mathcal{H}(z) = \frac{1}{2}\|z\|_2^2 - \mathcal{F}_2(W_{21}z), \qquad \nabla\mathcal{H}(z) = z - W_{21}^\top\Psi_2(W_{21}z). \tag{99}$$

If Ψ_2 is locally Lipschitz, the Clarke sum and chain rules give the sound inclusion

$$\partial(\nabla\mathcal{H})(z_\star) \subseteq \left\{I - W_{21}^\top Q W_{21} : Q \in \partial(\Psi_2)(W_{21}z_\star)\right\}. \tag{100}$$

Thus the tractable sufficient condition

$$\inf_{Q\in\partial(\Psi_2)(W_{21}z_\star)} \lambda_{\min}\left(I - W_{21}^\top Q W_{21}\right) > 0 \tag{101}$$

implies Proposition 5. The inclusion may be strict for nonsmooth compositions, but positivity over the displayed outer set remains a sound certificate.

For the one-hidden-layer Hamiltonian (99), the gradient is the affine term z minus the composition of the locally Lipschitz activation Ψ_2 with two linear maps. The Clarke generalized-Jacobian sum and chain rules yield (100). Therefore, requiring every matrix $I - W_{21}^\top Q W_{21}$ with $Q \in \partial(\Psi_2)(W_{21}z_\star)$ to be uniformly positive definite is a directly checkable sufficient condition for Proposition 5. The following specializations make this condition explicit.

Softmax. Let

$$\mathcal{F}_2(s) = \frac{1}{\beta}\log\left(\sum_{i=1}^{N_2} e^{\beta s_i}\right), \qquad p(s) = \Psi_2(s) = \mathrm{softmax}(\beta s), \qquad \beta > 0. \tag{102}$$

Then $\partial(\Psi_2)(s) = \{D\Psi_2(s)\}$ with

$$D\Psi_2(s) = \beta\left[\mathrm{diag}(p(s)) - p(s)p(s)^\top\right], \tag{103}$$

and a sufficient local PL condition at $z_\star$ is

$$I - \beta W_{21}^\top\left[\mathrm{diag}(p_\star) - p_\star p_\star^\top\right] W_{21} \succ 0, \qquad p_\star = p(W_{21}z_\star). \tag{104}$$

Softmax is smooth and bounded, so it is also compatible with the bounded-first-hidden-layer condition used to establish coercivity.

Hyperbolic tangent. For the componentwise activation $\Psi_{2,i}(s_i) = \tanh(\beta s_i)$, one may take

$$\mathcal{F}_2(s) = \sum_i \frac{1}{\beta} \log\cosh(\beta s_i), \tag{105}$$

with

$$\partial(\Psi_2)(s) = \left\{\beta \operatorname{diag}\left(\operatorname{sech}^2(\beta s_i)\right)\right\}. \tag{106}$$

The generalized-Hessian certificate reduces to

$$I - \beta W_{21}^\top \operatorname{diag}\left(\operatorname{sech}^2(\beta(W_{21} z_\star)_i)\right) W_{21} \succ 0. \tag{107}$$

The activation is smooth and bounded and therefore satisfies both the local regularity requirement and the first-hidden-layer boundedness requirement.

Logistic sigmoid. For $\Psi_{2,i}(s_i) = \sigma(\beta s_i)$, with $\sigma(r) = (1 + e^{-r})^{-1}$,

$$\partial(\Psi_2)(s) = \{\beta \operatorname{diag}(\sigma(\beta s_i)(1 - \sigma(\beta s_i)))\}, \tag{108}$$

and local PL follows from

$$I - \beta W_{21}^\top \operatorname{diag}(\sigma_i^\star(1 - \sigma_i^\star))\, W_{21} \succ 0, \qquad \sigma_i^\star = \sigma(\beta(W_{21} z_\star)_i). \tag{109}$$

Sigmoid is again smooth and bounded.

ReLU and activation kinks. For the componentwise ReLU activation $\Psi_{2,i}(s_i) = [s_i]_+$, a convex $C^{1,1}$ primitive is $\mathcal{F}_{2,i}(s_i) = \frac{1}{2}[s_i]_+^2$. Its Clarke generalized Jacobian is

$$\partial(\Psi_2)(s) = \left\{ \operatorname{diag}(q) : \begin{array}{ll} q_i = 1, & s_i > 0, \\ q_i = 0, & s_i < 0, \\ q_i \in [0,1], & s_i = 0 \end{array} \right\}. \tag{110}$$

Consequently, local PL is certified even at a kink whenever

$$I - W_{21}^\top D W_{21} \succ 0 \qquad \text{for every } D \in \partial(\Psi_2)(W_{21} z_\star). \tag{111}$$

Let $D_{\max}$ set the slope to one for every active or zero preactivation and to zero for every strictly inactive preactivation. Since $0 \preceq D \preceq D_{\max}$ for every admissible D, condition (111) is equivalent to the single worst-case check

$$I - W_{21}^\top D_{\max} W_{21} \succ 0. \tag{112}$$

This removes the need to assume a fixed activation pattern around $z_\star$. ReLU remains unbounded, however, so it is not compatible with the bounded-first-hidden-layer hypothesis used above to establish global coercivity when it is chosen as Ψ_2; the local Clarke certificate is still applicable when ReLU appears only in deeper layers or when coercivity is guaranteed by another architectural mechanism.

Bounded piecewise-linear activations. The generalized certificate also covers nonsmooth activations that are compatible with the hybrid coercivity argument. For example, let $\Psi_{2,i}(s_i) = \operatorname{clip}(s_i, -1, 1)$ (hard-$\tanh$). A convex $C^{1,1}$ primitive is

$$\mathcal{F}_{2,i}(s_i) = \begin{cases} \frac{1}{2} s_i^2, & |s_i| \le 1, \\ |s_i| - \frac{1}{2}, & |s_i| > 1. \end{cases} \tag{113}$$

The generalized slopes are one for $|s_i| < 1$, zero for $|s_i| > 1$, and any value in $[0, 1]$ at $s_i = \pm 1$. Thus the same worst-case diagonal test as (112), with $D_{\max}$ defined from the unsaturated and kink coordinates, certifies local PL. Unlike ReLU, hard-$\tanh$ is bounded and therefore can simultaneously satisfy the first-hidden-layer boundedness condition used for coercivity.

Deeper hybrid networks. For deeper architectures, the activation-independent statement remains Proposition 5: it is sufficient that every element of the full projected Clarke generalized Hessian at the learned equilibrium be positive definite. Computing that set exactly may be conservative or expensive because generalized-Jacobian chain rules through the feedforward composition can yield set-valued inclusions. Any tractable outer approximation $\mathcal{Q}_\star$ satisfying

$$\partial(\nabla \mathcal{H})(z_\star) \subseteq \mathcal{Q}_\star \tag{114}$$

therefore provides a sound sufficient certificate whenever every $M \in \mathcal{Q}_\star$ has a common positive spectral margin. The one-hidden-layer formulas above are the simplest exact or directly computable instances of this principle.

C Experimental details and reproducibility

This section records the experimental choices used in the released code and complements the compact discussion in Section 4. The benchmark-specific architectures are intentionally allowed to vary: the common object across tasks is the pH-EBM construction, i.e., a learned Hopfield storage function coupled to free port-Hamiltonian interconnection, dissipation, and input maps. Unless stated otherwise, continuous-time rollouts are integrated with fourth-order Runge–Kutta (RK4), and all reported certificates are evaluated after training without modifying the learned parameters.

Table 2: **Prediction accuracy and certified robustness across nonlinear dynamical systems.** *Comparative performance of the pH-EBM across benchmarks addressing nonlinear system identification, nonconvex dynamics, and higher-dimensional controlled systems. Results reported within square brackets,* $[..]$*, correspond to evaluations on distinct test sets of the same benchmark, while entries separated by semicolons refer to variants of the reference architecture.*

Benchmark	Metric ↓	Published	portHNN-u	pH-EBM
Silverbox	RMSE	$[0.289, 0.334, 0.257]$	$[51.343, 50.981, 41.138]$	$[0.395, 0.545, 0.327]$
CED	RMSE	$[0.062, 0.047]$	$[0.241, 0.366]$	$[0.073, 0.055]$
Duffing double-well	RMSE	-	0.254	0.048
3-link pendulum	RMSE	$0.162; 0.162$	$0.051; 0.051$	0.014
NanoDrone	RMSE	-	$[1.532, 5.260, 2.150, 15.541]$	$[1.368, 3.550, 1.343, 11.648]$
NanoDrone (Melon)	RMSE	$[4.051, 12.581, 6.059, 35.5735]$	$[3.939, 12.445, 5.574, 30.802]$	$[11.860, 28.014, 7.169, 36.597]$

C.1 Datasets, splits, and preprocessing

Silverbox. We use the canonical SISO Silverbox benchmark through the `nonlinear_benchmarks` interface. Input and output are standardized using the mean and standard deviation of the training record only; the same affine transformation is then applied unchanged to every test record. The final model uses a four-dimensional latent state. Its initial state is inferred from a burn-in window before the free rollout begins. The released evaluation scores all three official test records—multisine, full arrow, and arrow without extrapolation - rather than selecting a single favorable trajectory. The RMSE score reported in the main is the average RMSE over the three test datasets: check Table 2 for the individual scores. The champion is trained on windows of 192 prediction steps with stride 16 and batch size 32. The benchmark-native output RMSE is reported in mV, together with the dimensionless NRMSE used internally for diagnostics.

Coupled Electric Drive (CED). CED contains two SISO realizations corresponding to different actuation amplitudes. Both training realizations are concatenated and standardized with statistics computed jointly from the combined training data, while the two test records remain separate and are rolled out independently, with no state carried between them. The benchmark-provided state-initialization window has length 10. The selected configuration uses a four-dimensional latent state, 64-step training windows, unit stride, and batch size 16. We report each test RMSE in ticks/s and retain the mean NRMSE only as an aggregate diagnostic. This separation is important because the two records probe distinct operating amplitudes rather than repeated measurements of the same trajectory.

Duffing double well. The Duffing experiment is state observed and therefore differs deliberately from the output-only identification benchmarks. Data are generated from the physical dynamics $\dot{q} = p$, $\dot{p} = q - q^3 - rp + u$ with $r = 0.4$, sampling interval $\Delta t = 0.02$ s, and 20 s trajectories. The frozen dataset contains 200 training, 40 validation, and 100 nominal-test trajectories. Initial conditions alternate between the two wells, with 20% of the trajectories drawn from a higher-energy regime. Training inputs combine piecewise-constant, pseudo-random binary, and multisine forcing. Four held-out forcing families - step, chirp, unseen-frequency sinusoid, and long constant forcing - are generated only for OOD evaluation. Each OOD family contains 25 trajectories in the supplied artifact. No state or input normalization is used: the learned state is directly $z = (q, p)$, the physical initial condition is provided to the model, and the state itself is the prediction target. This makes the learned Hamiltonian directly comparable with the analytic mechanical energy.

n-link pendulum. The pendulum experiments import the arrays produced by the reference Deep Dissipative Dynamics data generator; the pH-EBM code does not regenerate an "equivalent" dataset. Inputs, partial observations, and full states are loaded from the corresponding `.npy` files. The primary mode is output-only: the observation consists of the first-joint angle/velocity pair, whereas the complete physical state is retained only for diagnostics. All supplied trajectories use $\Delta t = 0.05$ s and duration 10 s. The pH-EBM uses an eight-dimensional latent state, an input-aware 10-sample encoder, and the same output-only convention for both the two- and three-link experiments. The final released configurations use 50-step windows, stride 50, batch size 8, and 500 epochs. The supplementary figures report

Table 3: **Per-benchmark pH-EBM configurations.** Architectures and active parameter counts reconstructed from the supplied champion configurations and frozen provenance artifacts. Entries marked "not in release" correspond to rows already present in the draft for which no experiment code or champion configuration was supplied.

Benchmark	d_z	Hamiltonian architecture	Activation	J	R	G	Parameters
Silverbox	4	$128 \to 64$	softmax(2)→poly(4)	free skew	$LL^\top$	free $G \in \mathbb{R}^{4\times 1}$	≈ 9.1k
CED	4	$96 \to 48$	softmax(2)→poly(4)	free skew	$LL^\top$	free $G \in \mathbb{R}^{4\times 1}$	≈ 5.25k
Duffing double-well	2	$64 \to 32$	$\tanh(2)$ →poly(4)	free skew	$LL^\top$	free $G \in \mathbb{R}^{2\times 1}$	≈ 2.28k
n-link pendulum	8	$64 \to 32$	$\tanh(2)$ →poly(4)	free skew	$LL^\top$	free $G \in \mathbb{R}^{8\times 1}$	≈ 2.99k
NanoDrone†	12	$128 \to 64$	$\tanh$ →poly(4)	free pH field	learned PSD	unrestricted 4-port	43,738

both nominal reconstruction/OOD diagnostics and the learned projected energy geometry. The supplied release contains complete champion artifacts for $n = 2$ and $n = 3$.

NanoDrone. The NanoDrone loader follows a frozen external dataset revision and validates all file names, timestamps, and quaternion norms before training. The training families are Square, Random, and Chirp, with four runs per family; runs 1-3 form the development training split and run 4 is used for development validation. The three Melon trajectories form the held-out extrapolation family and are explicitly excluded from scaler fitting and from the S3 model-selection protocol. Samples are spaced by $\Delta t = 0.01$ s. The physical state is represented by 12 coordinates: position, linear velocity, the $SO(3)$ logarithm of attitude, and angular velocity. The four inputs are rotor angular speeds. State and input scalers are fitted only on the training trajectories, and the primary pH-EBM is trained in state-observed mode with an identity readout. The final S3 confirmation campaign evaluates horizons of 50, 100, 250, 500, and 1000 steps; the primary long-horizon comparison is 500 steps (5 s), ten times the 50-step (0.5 s) training/evaluation horizon emphasized by the benchmark. Four frozen seeds (11, 23, 47, 89) are used in the paired pH/black-box confirmation protocol, with model selection performed on the training-supported families only.

C.2 Per-benchmark model configurations

The selected models share the same pH-EBM construction but not a forced common width. This is intentional: state dimension, observation structure, and input dimension differ substantially across the tasks. In all code-supported configurations below, J is free and skew-symmetric by construction, R is learned through a Cholesky factor and is positive semidefinite (strictly positive definite when the learned factor has nonzero diagonal), and G is learned rather than tied to the scalar Hamiltonian parameterization. "Active parameters" count only terms used by the selected computational graph.

C.3 Training, validation, and model selection

Across the standard pH-EBM interfaces, optimization is performed on finite rollout windows with a Huber observation loss, Adam/AdamW updates, global gradient clipping, and RK4 integration. The first training epochs use a stop-gradient rollout warm-up to limit long-horizon gradient pathologies; subsequent epochs differentiate through the complete rollout. This warm-up changes the optimization path only and does not alter the model evaluated at test time.

Benchmark-specific schedules. The Silverbox champion uses 800 epochs, learning rate 4.59×10^{-4}, 192-step windows, stride 16, and batch size 32. CED uses a longer 8000-epoch schedule, learning rate 10^{-4}, 64-step windows, stride 1, and batch size 16. Duffing uses 1000 epochs, learning rate 1.5×10^{-3}, 100-step windows, stride 25, and batch size 64 in the frozen champion. The two supplied n-link champions use 500 epochs, learning rate 3×10^{-4}, 50-step windows, stride 50, and batch size 8. For NanoDrone, the final S3 confirmation protocol fine-tunes a frozen parent model for 120 epochs with fresh Adam, batch size 64, a five-epoch warm-up, and a learning-rate schedule peaking at 3×10^{-5} and ending at 10^{-7}. Validation is evaluated at a fixed list of epochs, and the S3 campaign is repeated for four seeds. The Melon family is inaccessible to this selection stage by protocol.

Initial conditions and readouts. Silverbox, CED, and the primary n-link experiments are output-only systems. Their latent initial condition is inferred from a short observation window; CED and n-link use an input-aware encoder so the burn-in actuation is available when estimating $z(0)$. Their measured output is distinct from the power-conjugate port output. Duffing and NanoDrone instead use observed-state training: the physical state initializes the rollout directly and no learned state decoder is required. This distinction is preserved deliberately because the corresponding datasets expose different information to the learner.

C.4 Baselines and comparison protocol

The baseline is chosen to answer the scientific question posed by each experiment rather than forcing one architecture across unrelated datasets. Silverbox and CED are compared with benchmark-native literature references under the same published test records and physical error units. Duffing uses the released portHNN-u implementation, including a version with free full-matrix J, R, G, so that the comparison isolates the effect of the Hamiltonian parameterization rather than fixing the pH operators to a scalar ansatz. The n-link comparison imports the exact upstream Deep Dissipative Dynamics trajectories and uses its reference model on those same arrays. NanoDrone uses the accompanying black-box neural model and a paired confirmation protocol: pH and black-box models see the same S3 families, horizons, validation starts, and seed set. The frozen protocol records 43,738 active parameters for the pH model and 18,508 for the reference black-box model; the comparison is therefore a benchmark-reference comparison rather than a parameter-matched capacity ablation. Missing certification for an unconstrained baseline is denoted as unavailable, not as a zero robustness radius.

C.5 Prediction metrics and statistical reporting

For scalar-output identification benchmarks we report the benchmark-native root-mean-square error

$$\mathrm{RMSE} = \sqrt{\frac{1}{N}\sum_{i=1}^{N}(\hat{y}_i - y_i)^2}, \tag{115}$$

and, for internal cross-dataset diagnostics, $\mathrm{NRMSE} = \mathrm{RMSE}/\,\mathrm{std}(y)$. Silverbox RMSE is converted back to mV and CED is reported in ticks/s. Duffing and the state-observed diagnostics use the same definitions directly in physical state coordinates. NanoDrone is evaluated in physical units after undoing the training scalers. Its primary paired S3 estimand is the difference between pH-EBM and black-box physical trajectory MAE at 500 steps, with additional reporting at 50, 100, 250, and 1000 steps and channel-family diagnostics for position, velocity, orientation, and angular velocity. Quaternion observations are converted to $SO(3)$ logarithmic coordinates for the 12-state model; orientation error is additionally tracked geometrically in the evaluation code. For consistency with official validation script, Nanodrone errors are also computed as per-output channel RMSE.

C.6 Operational energy-shell selection

The certificate calculations are performed after fitting. We first search the learned Hamiltonian for low-energy wells by multi-start gradient descent. Half of the starts are drawn from states visited by the fitted model when such states are available, and the remaining starts are randomized around their empirical center and scale. Descents are restricted to the operational state ball used by the numerical model, and terminal points are clustered into candidate wells. Candidates with a large energy-gradient norm are rejected before they are used as stationary anchors. This avoids treating a low-energy but nonstationary optimizer endpoint as a physical attractor.

For a chosen threshold ϵ, the full-state shell $\mathcal{H}(z) = \epsilon$ is traced from every discovered well whose energy lies below the threshold. Random unit directions are launched from the well; along each ray the algorithm expands an outer bracket until the first crossing of the target energy is found and then refines the crossing by bisection. A short damped correction in the local gradient-normal direction finally reduces the residual $|\mathcal{H}(z) - \epsilon|$. Rays that do not reach the shell within the configured search radius are recorded as failures rather than silently projected to an unrelated point. Consequently, all extrema reported from this procedure are sampled shell estimates; a finite ray set by itself is not presented as an exhaustive proof of a continuous high-dimensional boundary.

C.7 Numerical computation and verification of certificates

Hamiltonian projection. A two-dimensional picture of a high-dimensional Hamiltonian is necessarily a reduction, so we use two complementary views and label them separately. Let $c \in \mathbb{R}^d$ be the selected well, let $U \in \mathbb{R}^{d\times 2}$ contain two orthonormal display directions, and let V span the orthogonal complement. The *affine slice*

$$H_{\mathrm{slice}}(\zeta) = \mathcal{H}(c + U\zeta) \tag{116}$$

is an exact cross-section: all unshown coordinates are frozen. The *profiled envelope*

$$\widetilde{H}(\zeta) = \min_{\omega} \mathcal{H}(c + U\zeta + V\omega) \tag{117}$$

instead minimizes over the unshown coordinates. Up to numerical optimization error, the set $\{\zeta : \widetilde{H}(\zeta) \le \epsilon\}$ is the planar projection of the full sublevel set $\{z : \mathcal{H}(z) \le \epsilon\}$. The minimization is solved independently over the display

grid with Adam and multiple warm starts, including the orthogonal coordinates of discovered wells. We retain the orthogonal-gradient norm as a stationarity diagnostic for the profile.

The display plane is chosen from physically or dynamically informative directions. With multiple discovered wells, the first direction follows the inter-well geometry and the second is taken from the dominant orthogonal variation of the sampled boundary/trajectory cloud. With a single well, the default is PCA of the sampled safety boundary; ordinary trajectory PCA is used only when boundary information is unavailable. Duffing is already two dimensional and therefore requires no projection. The plane limits are derived from robust quantiles of the projected reference states, boundary samples, and discovered wells, with padding added only for readability. The supplementary slice/profile pairs in Figs. 4, 5, 7–8, and 9, 10 use this procedure.

Input-radius computation. At every sampled shell state we evaluate the raw Hamiltonian gradient and the learned R and G matrices, then compute the pointwise boundary margin described in Section 3.1. For a chosen input norm, the sampled "exact" shell radius is the minimum of the pointwise ratio

$$\frac{d_H(z) - \sigma_{\mathcal{W}(z)}(\nabla \mathcal{H}(z))}{\|G(z)^\top \nabla \mathcal{H}(z)\|_*} \tag{118}$$

over the traced states for which the input couples to the energy gradient. In parallel, we compute the regular-boundary lower estimate from the sampled extrema of $\|\nabla H\|$, normal dissipation, normal input gain, and disturbance support. The implementation records both values because they answer different questions: the first measures the least favorable sampled direction directly, while the second exposes the geometric factors entering the analytic lower bound. Both remain numerical shell estimates unless the continuous extrema are independently bounded.

Epsilon sweeps. The epsilon sweeps in the supplementary figures do not rescale a single boundary. For every energy level, the full-state shell is traced again from scratch and both radii are recomputed. By default, 300 levels are sampled in relative energy coordinates $\epsilon - H_{\min}$, from 0.10 to $5/3$ times the nominal energy gap $\epsilon_{\text{nom}} - H_{\min}$. Thus the sweep covers the certified operating shell and extends by two thirds of its nominal gap to reveal what happens near and beyond critical levels. The boundary search radius is enlarged by a factor 1.5 during this sweep so that outer shells are not artificially truncated. For Duffing, the range crosses the saddle energy; after components merge, the shell is treated globally rather than assigning points to their original well. Failed levels are stored as missing values and are not interpolated. Every sweep artifact stores the absolute ϵ, $H_{\min}$, relative energy, sampled radius, regular-boundary estimate, number of successful boundary points, failed-ray fraction, and component summaries so the plotted curves can be audited independently.

Hamiltonian-normalized empirical certificate. The robustness radius $\rho_{\epsilon,\star}$ introduced above is an exact certificate defined at a specific energy level ϵ and expressed in the native coordinates of the model input. As a consequence, its numerical value may vary under simple rescalings of either the energy function or the input variables, making direct comparisons across architectures difficult. To obtain a dimensionless quantity suitable for cross-model evaluation, we introduce an empirical normalization procedure based exclusively on validation data.

Common energy level. Given a fixed validation rollout

$$\{(z_i, u_i)\}_{i=1}^{N_{\text{val}}},$$

we define the reference energy level as the q-quantile of the validation energy distribution,

$$\epsilon_q := Q_q\Big(\{\mathcal{H}(z_i)\}_{i=1}^{N_{\text{val}}}\Big), \qquad q = 0.99. \tag{119}$$

This choice ensures that all models are evaluated on a shell corresponding to the same validation-state occupancy rather than at an arbitrary absolute energy value.

Input normalization. To remove the dependence on the physical units and scaling of the inputs, we normalize the input space using the empirical second moment

$$M_u := \frac{1}{N_{\text{val}}} \sum_{i=1}^{N_{\text{val}}} u_i u_i^\top = L_u L_u^\top, \tag{120}$$

and introduce normalized coordinates

$$u = L_u v.$$

When M_u is rank-deficient, L_u is restricted to the subspace effectively excited by the validation data. In these coordinates, distances are measured relative to the variability observed during validation rather than in the original input units.

Normalized robustness margin. Using the runtime vector field f_θ^{run}, define the directional energy dissipation rate

$$\ell_H(z,u) := \nabla\mathcal{H}(z)^\top f_\theta^{\mathrm{run}}(z,u). \tag{121}$$

A negative value of ℓ_H corresponds to local energy dissipation, whereas $\ell_H \geq 0$ identifies inputs capable of halting or reversing this decrease. We therefore define the normalized robustness margin at state z as

$$r_H(z) := \inf\left\{\|v\|_2 : \ell_H(z, L_u v) \geq 0\right\}. \tag{122}$$

Equivalently, $r_H(z)$ is the smallest perturbation, measured in normalized input coordinates, capable of eliminating the local decrease of the energy function. By convention, $r_H(z) = 0$ whenever $\ell_H(z,0) \geq 0$. For input-affine dynamics,

$$f_\theta^{\mathrm{run}}(z,u) = f_\theta^{\mathrm{run}}(z,0) + B_\theta(z)u,$$

the above optimization admits the closed-form expression

$$r_H(z) = \frac{-\nabla\mathcal{H}(z)^\top f_\theta^{\mathrm{run}}(z,0)}{\left\|L_u^\top B_\theta(z)^\top \nabla\mathcal{H}(z)\right\|_2}, \tag{123}$$

whenever the denominator is nonzero. For nonlinear forcing structures, $r_H(z)$ is obtained numerically by locating the first zero crossing of ℓ_H along the normalized input direction.

Empirical normalized radius. Let $\widehat{\Gamma}_q$ denote a numerical reconstruction of the energy shell

$$\Gamma_q := \{z : \mathcal{H}(z) = \epsilon_q\}$$

obtained from validation data. We then define

$$\widehat{\rho}_{q,\mathrm{norm}} := \min_{z\in\widehat{\Gamma}_q} r_H(z), \qquad q = 0.99. \tag{124}$$

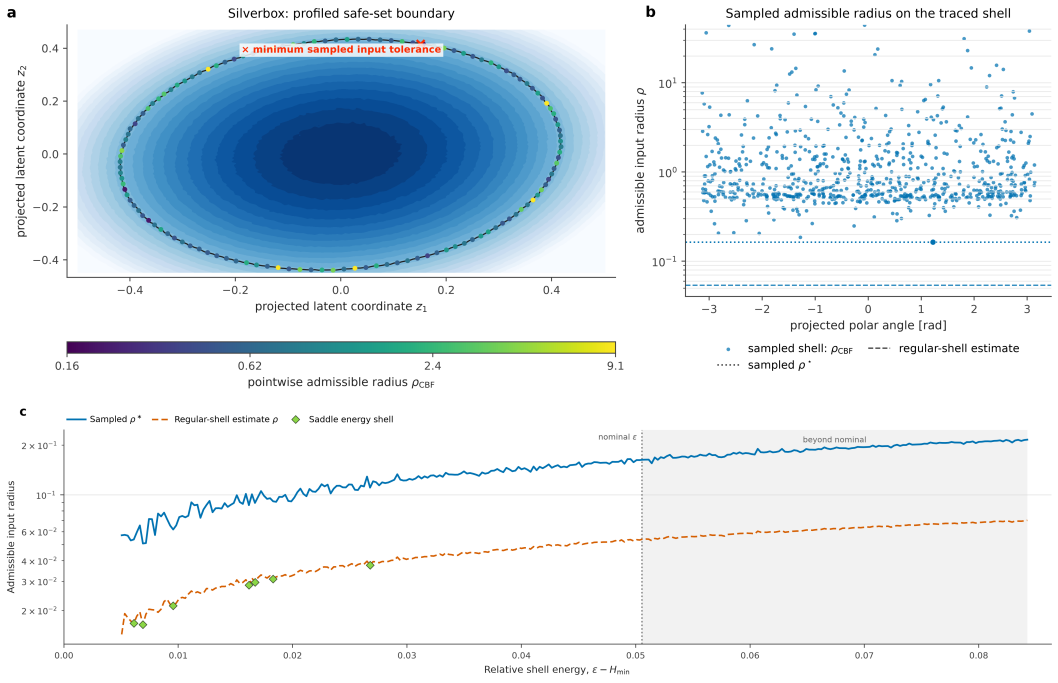
The quantity $\widehat{\rho}_{q,\mathrm{norm}}$ measures the smallest normalized input perturbation required to destroy local energy dissipation on the validation shell. By construction, it is invariant under any strictly increasing reparameterization of $\mathcal{H}$ and uses a common, data-derived input geometry for all models. Since the shell itself is reconstructed numerically from finite validation trajectories, $\widehat{\rho}_{0.99,\mathrm{norm}}$ should be interpreted as an empirical benchmarking metric for cross-model comparison. In contrast, the radius $\rho_{\epsilon,\star}$ remains the exact continuous-state robustness certificate associated with a prescribed energy level.

C.8 Additional benchmark results

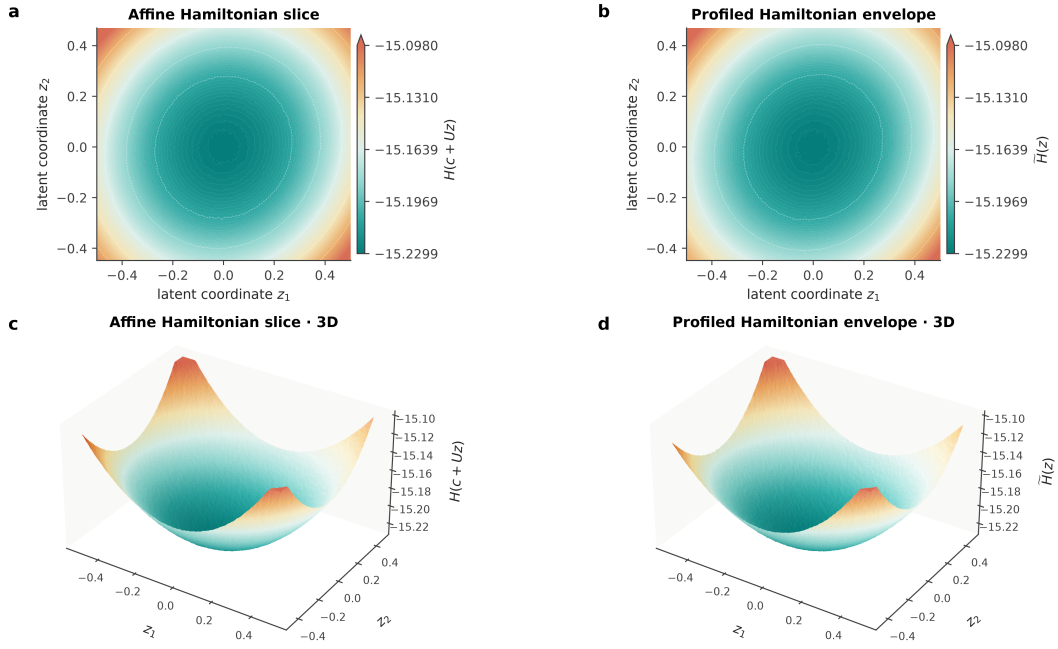
Figures 4-11 collect the supplementary outputs generated by the released experiment pipeline. In particular, the Silverbox and CED plates show that the same certificate machinery can be applied when the learned energy is effectively single-well (or almost); the Duffing figure validates recovery of a genuinely nonconvex physical energy; the n-link figures compare prediction/OOD behavior with high-dimensional projected geometry; and the NanoDrone figures expose the distinction between short-horizon fit and long-horizon safe recursive deployment. These visualizations are diagnostic companions to the quantitative table in the main paper and do not replace the full-state certificate computations from which the shell radii are obtained.

C.9 Ablations and sensitivity analyses

The released experiments prioritize ablations that alter the scientific certificate rather than exhaustive optimizer sweeps. Duffing compares the modern-Hopfield Hamiltonian with the portHNN-u reference while keeping the port-Hamiltonian interpretation explicit. The n-link family changes mechanical dimension under a common eight-dimensional latent pH-EBM template. NanoDrone pairs the structured model with the benchmark black-box model over identical S3 starts and horizons. Finally, the epsilon sweeps provide a post-training sensitivity analysis of the certificate itself: they reveal how the minimum shell gradient, normal dissipation, input exposure, and admissible-input radius change as the certified operating region expands toward critical energy levels.

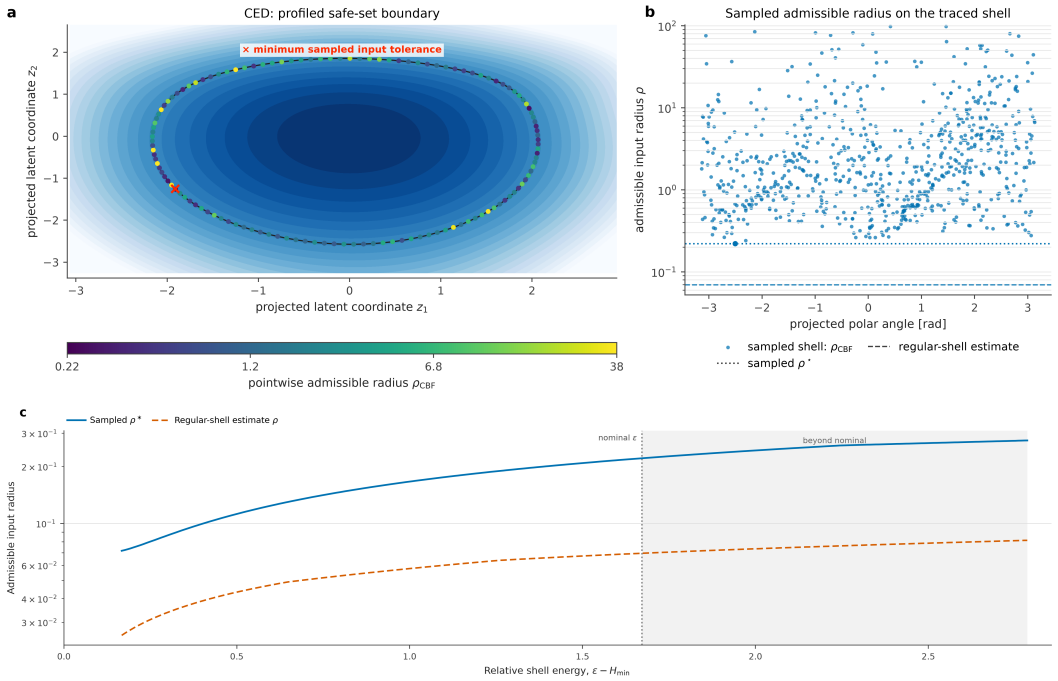


(a) Projected safe-set geometry and input-radius sweep.

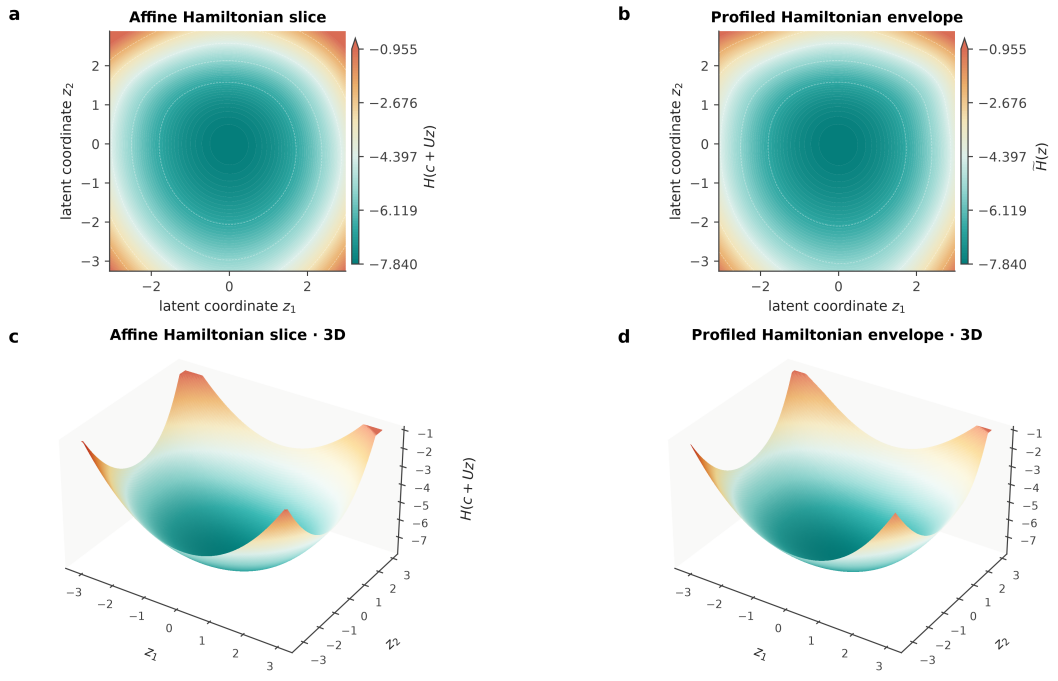


(b) Affine Hamiltonian slice and profiled energy envelope.

Figure 4: **Silverbox: learned energy geometry and certificate diagnostics.** The upper composite shows the profiled two-dimensional energy view at the nominal shell: traced boundary states are colored by their pointwise admissible-input radius, the red cross marks the least tolerant sampled state, and the adjacent panel resolves the same radii along the shell. The lower sweep compares the sampled shell minimum with the regular-shell estimate as the relative energy budget increases; green diamonds mark detected saddle-energy shells. The second composite contrasts an affine Hamiltonian slice with the profiled envelope in both 2D and 3D. Their close agreement indicates that the selected plane captures the dominant single-well geometry over the explored region. These plots visualize the *learned* Hamiltonian and should not be interpreted as recovery of a unique physical Silverbox potential.



(a) Projected safe-set boundary and epsilon sweep.



(b) Affine slice and profiled Hamiltonian view.

Figure 5: **CED: learned energy geometry and certificate diagnostics.** The upper composite shows the profiled safe-set boundary, colored by pointwise admissible-input radius, together with the shell-wise radius distribution and the least tolerant sampled state. The energy sweep compares the sampled minimum radius with the regular-shell estimate from the nominal shell into higher-energy regions. The second composite compares the affine slice and profiled envelope in 2D and 3D. The smooth, nearly monotone radius trend and close slice/profile agreement are consistent with an effectively single-well learned geometry over the explored region; they do not establish convexity of an underlying physical energy.

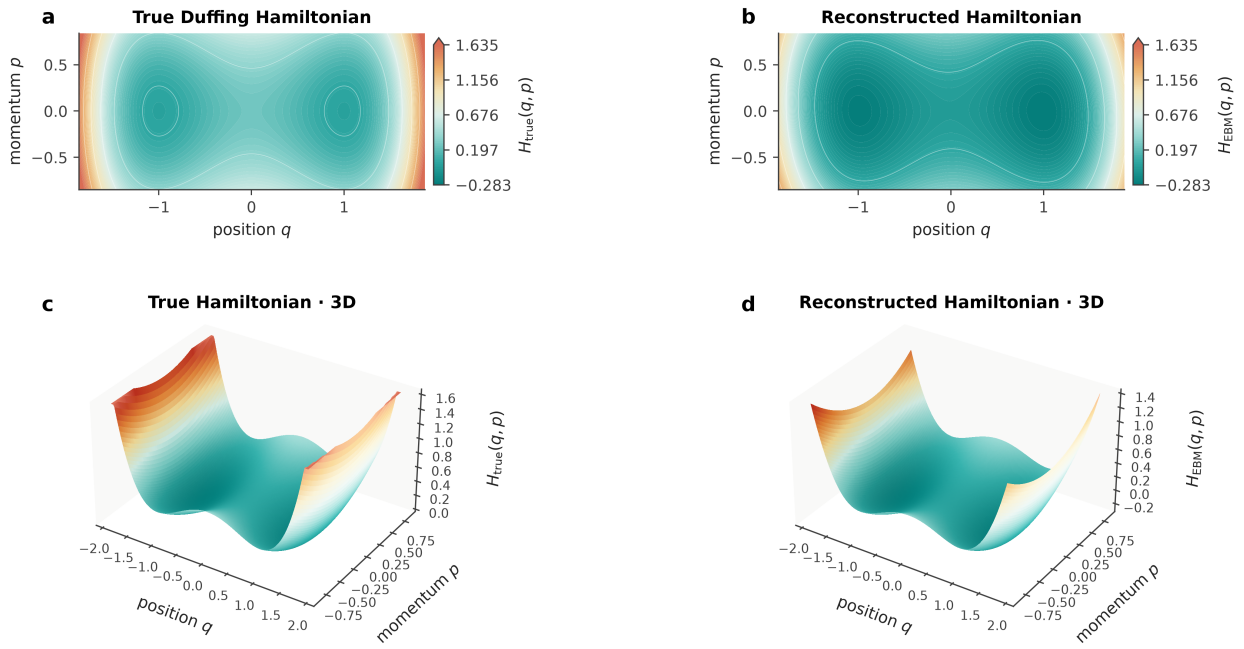
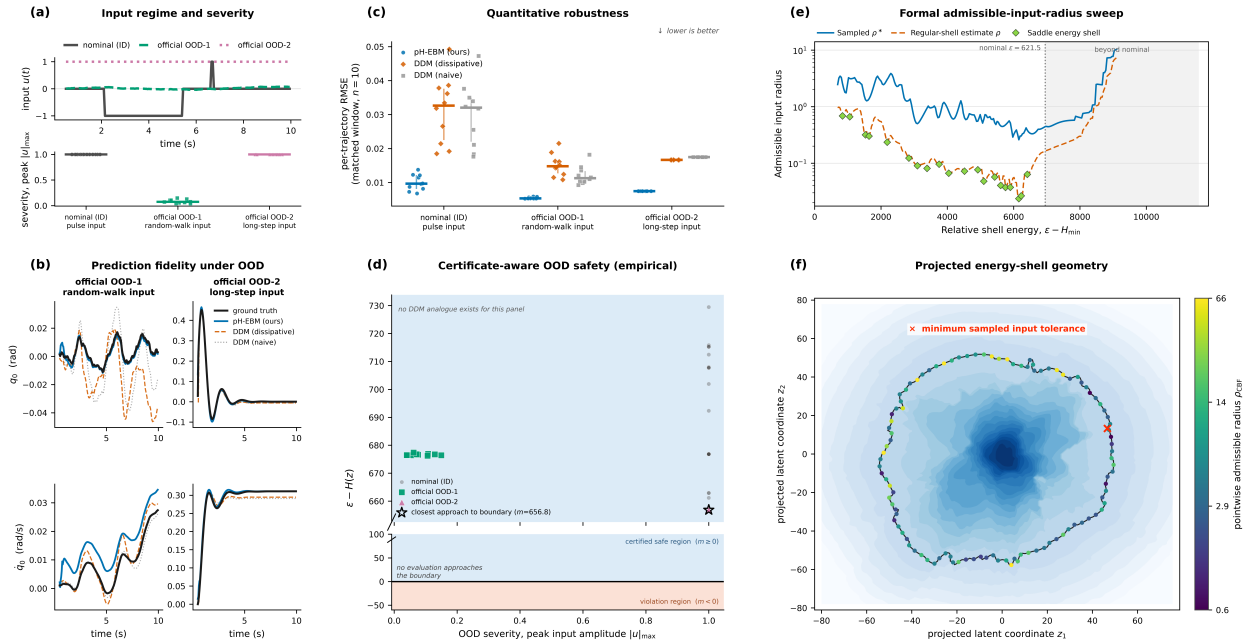
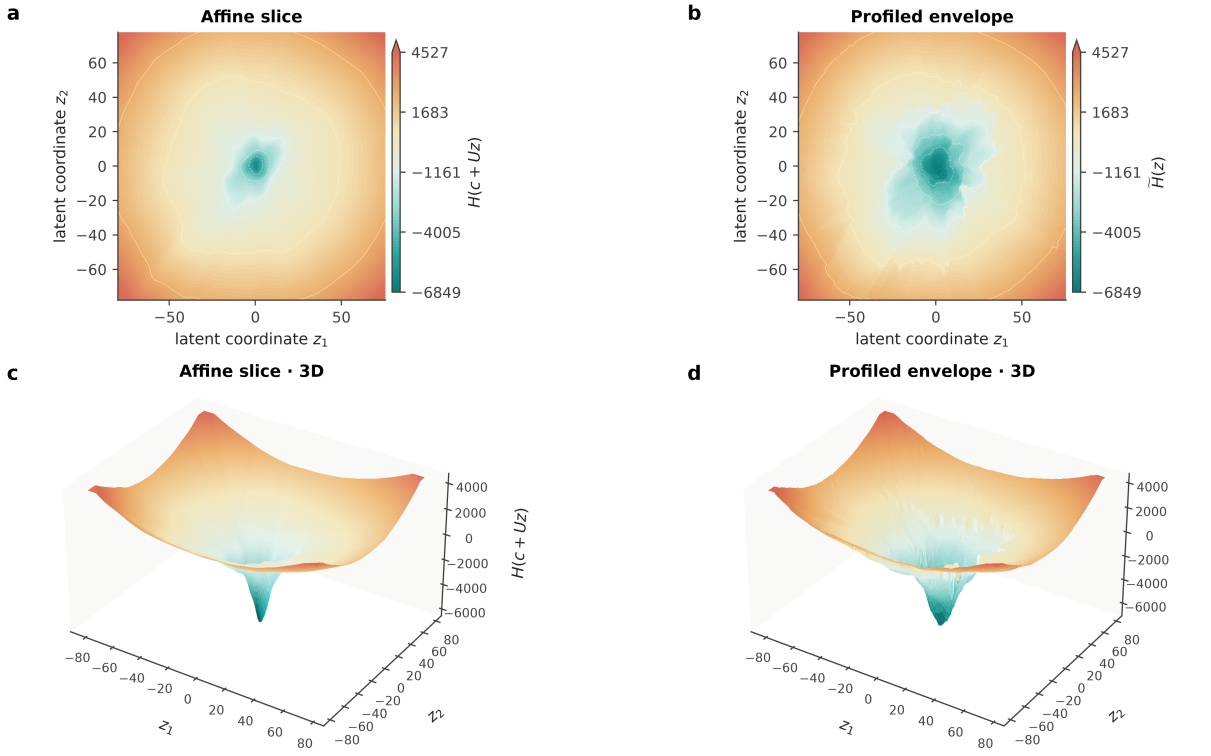


Figure 6: **Duffing Hamiltonian reconstruction in physical coordinates.** Analytic (left) and learned pH-EBM (right) energies are shown directly in (q, p), so no latent projection is involved. Both the contour and 3D views recover the characteristic double-well topology and the central saddle separating the two basins. The comparison supports qualitative recovery of the physical energy geometry; it does not imply unique or exact identification of the underlying Hamiltonian.

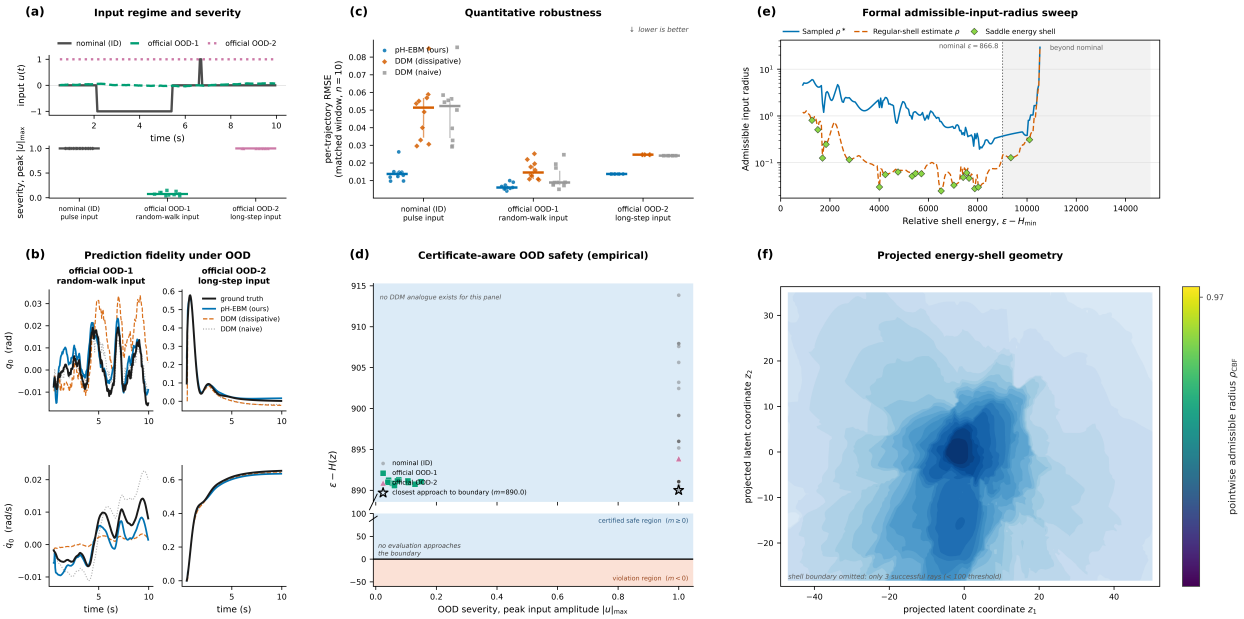


(a) $n = 2$: prediction, OOD tests, and certificate diagnostics.

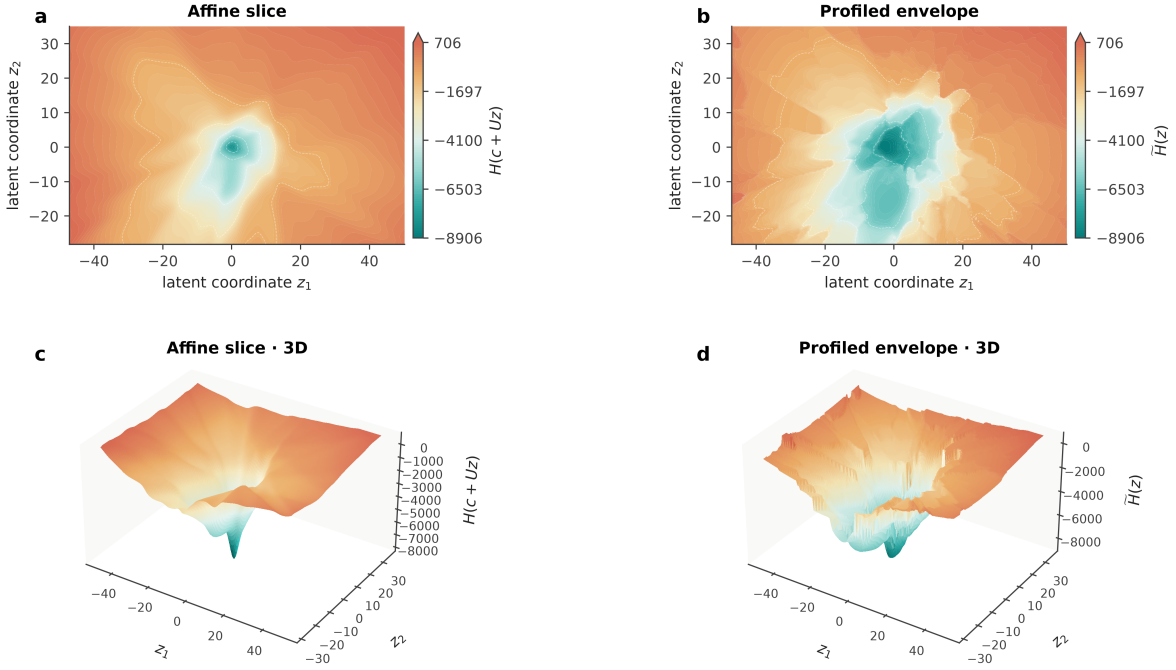


(b) $n = 2$: affine slice versus profiled envelope.

Figure 7: 2**-link pendulum: prediction, OOD forcing, and certificate geometry.** The first composite reports the nominal pulse input and the two official OOD inputs, representative trajectory reconstructions, per-trajectory RMSE across pH-EBM and dissipative/naive DDM baselines, the empirical safety margin of the evaluated OOD trajectories, the admissible-input-radius sweep across energy shells, and the projected shell colored by pointwise radius. The radius decreases near several detected saddle-energy shells before increasing beyond the nominal operating energy. The second composite compares an affine Hamiltonian slice with the profiled envelope, showing the nonconvex structure that is hidden by any single trajectory projection.

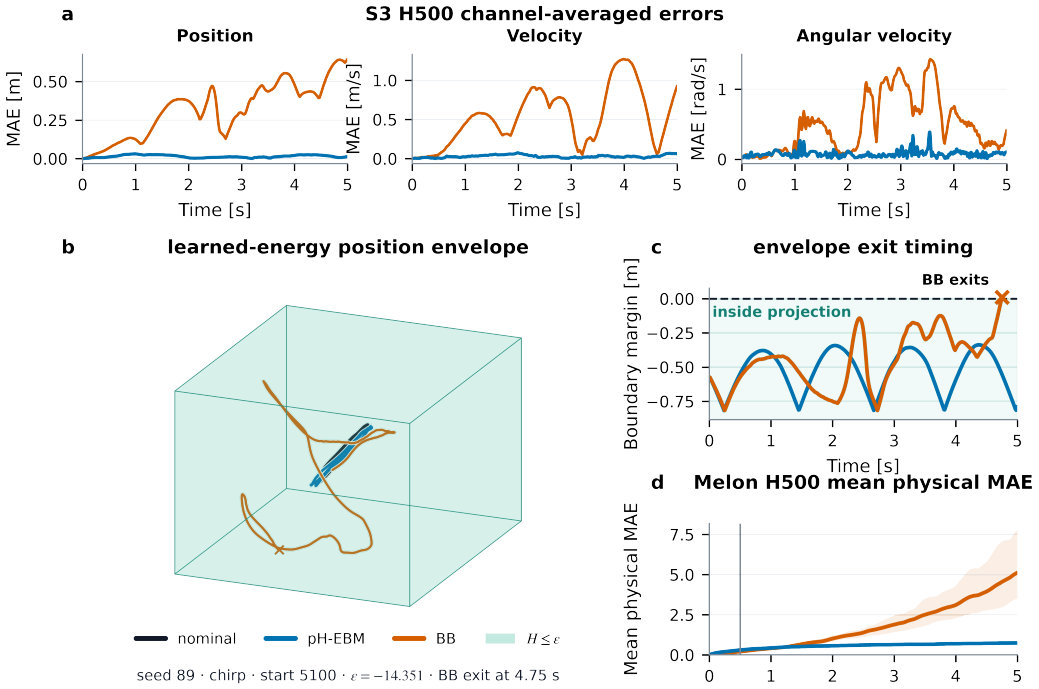


(a) $n = 3$: prediction, OOD tests, and certificate diagnostics.

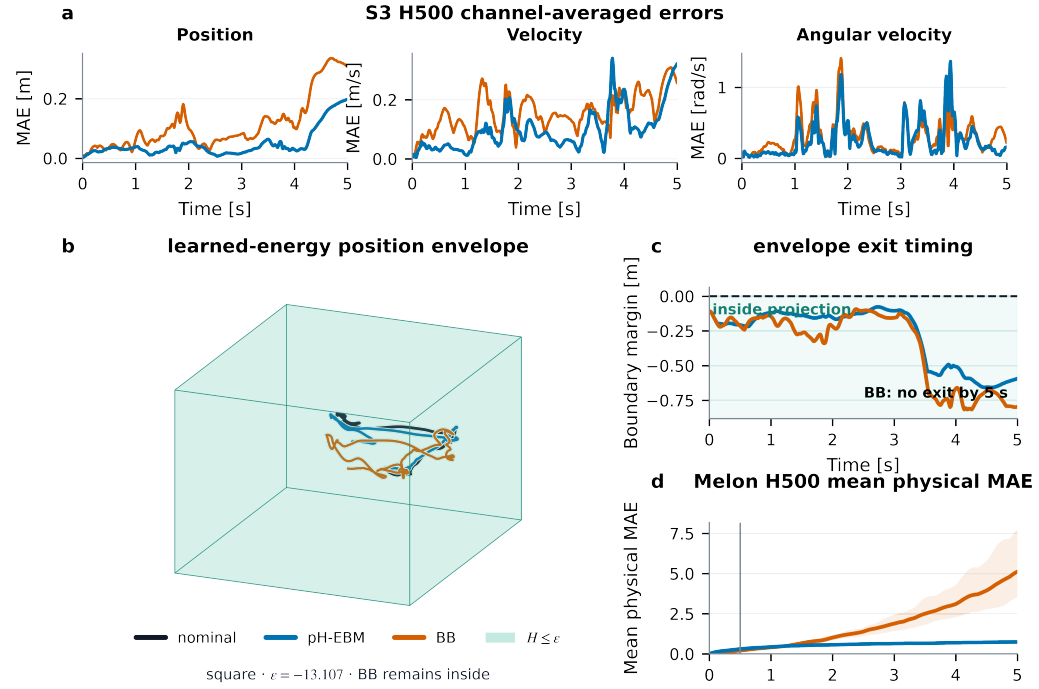


(b) $n = 3$: affine slice versus profiled envelope.

Figure 8: 3**-link pendulum: prediction, OOD forcing, and certificate geometry.** The first composite mirrors the 2-link protocol: nominal and official OOD inputs, representative reconstructions, per-trajectory RMSE, empirical safety margins, and an energy-shell sweep of the sampled and regular-shell admissible radii. Detected saddle-energy shells coincide with pronounced changes in the radius curve. The projected energy field is shown for geometric context, but the shell boundary is intentionally omitted because too few rays met the reconstruction threshold; it should therefore not be read as a complete projected boundary certificate. The second composite compares affine-slice and profiled-envelope views of the learned nonconvex Hamiltonian.

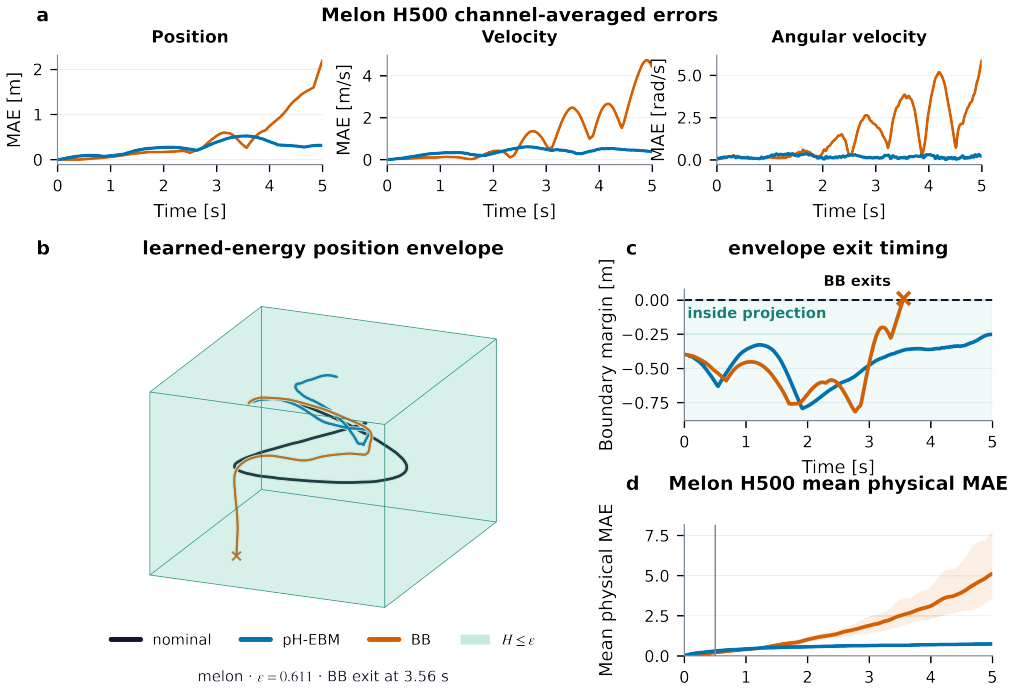


(a) Chirp long-horizon prediction and learned safe envelope.

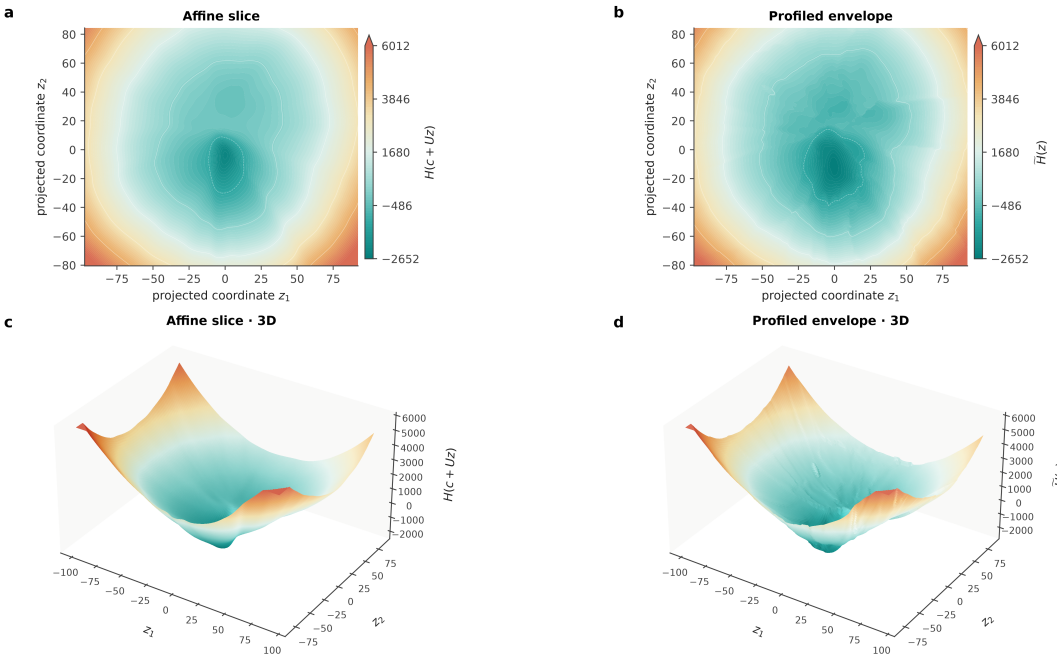


(b) Square long-horizon prediction and learned safe envelope.

Figure 9: **NanoDrone: training-supported long-horizon Chirp and Square rollouts.** For each excitation, the top panels report channel-averaged position, velocity, and angular-velocity errors; the 3D panel compares the nominal trajectory with pH-EBM and black-box rollouts inside the projected learned-energy envelope; the exit-timing panel tracks the most critical projected coordinate; and the final panel reports physical MAE over the extended horizon. The pH-EBM remains inside the displayed envelope for the shown 5 s rollouts, whereas the black-box trajectory exits at the indicated time. The envelope is a visualization of the high-dimensional certificate, not a substitute for its full-state evaluation.



(a) Held-out Melon long-horizon prediction.



(b) Projected Hamiltonian and profiled envelope.

Figure 10: **NanoDrone: held-out Melon extrapolation and learned Hamiltonian geometry.** The first composite evaluates the unseen Melon family over a 5 s rollout, reporting channel-wise error, projected trajectory evolution relative to the learned-energy envelope, exit timing, and physical MAE. The black-box prediction is initially competitive but develops rapidly growing long-horizon error and leaves the displayed envelope, whereas the pH-EBM remains bounded in the shown rollout. The second composite compares an affine two-dimensional Hamiltonian slice with the profiled envelope obtained by optimizing over unshown coordinates, in both contour and 3D views, exposing the nonconvex geometry of the learned high-dimensional energy.

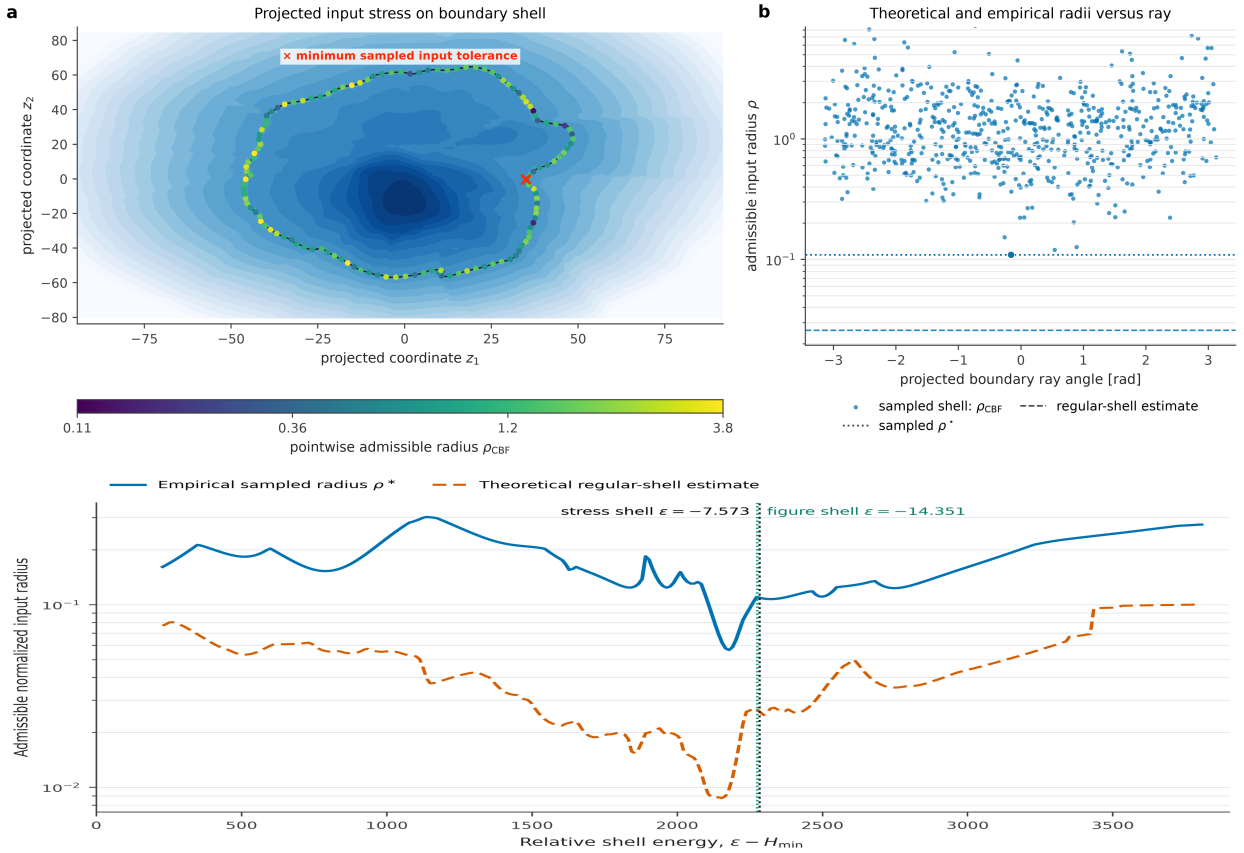


Figure 11: **NanoDrone projected certificate.** Left: projection of the traced full-state energy shell, with boundary samples colored by their pointwise admissible-input radius and the least tolerant sampled state marked explicitly. Right: the same pointwise radii ordered along the projected shell, together with the sampled shell minimum and regular-shell estimate. The projection is diagnostic only: certification is evaluated in the full model state space, and the displayed contour is not asserted to contain every high-dimensional boundary extremum.