IMPROVING HAMILTONIAN LEARNING THROUGH LOCAL ENERGY DECOMPOSITION

Anonymous authors
Paper under double-blind review

ABSTRACT

Hamiltonian neural networks provide a principled framework for learning conservative dynamics by parameterizing a scalar Hamiltonian and deriving the vector field through Hamilton's equations. However, preserving Hamiltonian structure does not guarantee that the energy function itself is represented accurately. When the Hamiltonian is modeled by a conventional multilayer perceptron, it inherits the spectral bias of coordinate-based neural representations, learning smooth components more readily than fine-scale variations that can strongly affect the dynamics after differentiation. We introduce the Local Multiscale Hamiltonian Neural Network (LM-HNN), which restructures the Hamiltonian according to its physical interaction structure. LM-HNN decomposes the energy into kinetic and potential components and represents them as sums of shared local functions over multiple interaction neighborhoods. This replaces a single high-dimensional approximation problem with lower-dimensional, reusable interaction laws, making fine-scale energy structure easier to resolve while preserving Hamiltonian dynamics. Across low-dimensional, interacting, and fast–slow systems, LM-HNN consistently achieves more accurate Hamiltonian derivatives and long-horizon rollouts than competing structure-preserving models, while substantially extending the effective spectral range of the learned energy. The same structured representation also generalizes more reliably to energy levels not observed during training. These results show that accuracy depends on the Hamiltonian representation as well as its structure, and that incorporating locality, decomposition, and shared interaction laws provides a simple and effective route to more accurate and generalizable Hamiltonian learning.

1 INTRODUCTION

Learning Hamiltonian dynamics can be challenging when the underlying energy function contains fine-scale variations. Hamiltonian neural networks (HNNs) obtain a vector field by differentiating a learned scalar energy $H_\theta(\boldsymbol{q}, \boldsymbol{p})$, making the quality of this representation central to prediction (Greydanus et al., 2019). A standard HNN uses a global Hamiltonian representation, where a single network maps the full phase-space state to the total energy. Standard HNNs therefore inherit the spectral bias (Rahaman et al., 2019; Tancik et al., 2020) of coordinate networks, learning low-frequency energy variations more readily than fine-scale structure. Since the energy is differentiated, unresolved high-frequency components can cause substantial vector-field errors even when their energy amplitude is small, and these errors can accumulate during rollout (Sitzmann et al., 2020; Gruver et al., 2022).

Efforts to address spectral bias have improved the recovery of fine-scale signals by changing coordinate encodings and activation functions (Tancik et al., 2020; Sitzmann et al., 2020). These modifications change which frequencies a network learns efficiently, making them natural candidates for improving Hamiltonian approximation. Hamiltonian learning has also benefited from decomposing the energy across temporal scales (Li et al., 2026). However, improving frequency resolution or separating timescales does not directly encode how the energy is assembled from physical interactions. A global Hamiltonian model must still infer which state variables interact, how their energy contributions combine, and which interaction laws recur across the system.

Encoding these interactions explicitly allows fine-scale energy variation to be learned through local functions, each depending on only a small subset of the state variables. Many mechanical systems

Reviewers: please read the Reviewer Guidelines (iclr.cc/Conferences/2027/ReviewerGuidelines) and the AI Policy for Reviewers (iclr.cc/Conferences/2027/AIPolicyForReviewers). If you used AI to expand, edit, or polish your review, please provide the input text to the LLM. Better still, consider skipping the LLM and submitting your original text: we, and the authors, are much more interested in your unedited thoughts than in what an LLM has to say. AI-assisted or not, you are putting your name and reputation behind your review: LLM-generated falsehoods, hallucinations or misrepresentations are subject to disciplinary action, which may include desk-rejecting all papers you have authored.

admit a kinetic–potential decomposition (Goldstein et al., 2002; Marion, 2013), and finite-range potentials can be represented as sums of local energy contributions, a principle established in neural potential-energy models (Behler & Parrinello, 2007). Repeated instances of the same interaction provide supervision for a shared local function. This suggests a mechanism for mitigating spectral bias by reducing the dimensionality of the functions being learned and concentrating shared capacity on the underlying interaction laws. The same factorization can support generalization when globally unseen states contain local configurations similar to those encountered during training.

Building on these insights, we introduce the Local Multiscale Hamiltonian Neural Network (LM-HNN). The model learns shared local functions for kinetic and potential energies, then sums their outputs to form the total Hamiltonian. Evaluating these functions over several neighborhood radii accommodates different interaction ranges without committing to a single radius in advance. This local factorization targets spectral bias by concentrating each network's capacity on a lower-dimensional interaction law and its fine-scale energy variations. Aggregating the learned contributions also supports generalization to unseen energy levels when new total energies arise from combinations of local configurations represented during training. The assembled scalar Hamiltonian generates the dynamics through Hamilton's equations, preserving Hamiltonian structure by construction. For systems with repeated interactions and fixed neighborhood sizes, parameter sharing also makes the parameter count independent of system size. We evaluate these properties through spectral diagnostics, tests at unseen energy levels, and prediction accuracy across six conservative systems. Comparisons with structure-preserving baselines and spectral interventions show that LM-HNN achieves lower derivative and rollout errors, while ablations identify locality and kinetic–potential decomposition as the main sources of these gains.

Concretely, our contributions are as follows:

- We analyze the spectral bias of conventional HNNs, and show that it is measurable directly on the learned Hamiltonian rather than only from prediction error.
- We propose LM-HNN, which combines kinetic and potential decomposition with shared local energy functions evaluated across multiple neighborhood scales. The representation extends the spectral range over which the Hamiltonian is resolved and improves derivative and rollout accuracy against structure-preserving baselines.
- We show that LM-HNN generalizes to energy levels outside the training range, where the baselines lose accuracy, supporting more accurate dynamics beyond the training-energy range.

2 Related Work

Structure-preserving learning of dynamical systems. Data-driven modeling of dynamical systems spans sparse system identification (Brunton et al., 2016), continuous-time neural models (Chen et al., 2018), and neural operators for spatially distributed systems (Li et al., 2020; Lu et al., 2021). For conservative dynamics, a related line of work incorporates mechanics directly into the model class. Hamiltonian neural networks learn a scalar energy and derive the vector field through Hamilton's equations (Greydanus et al., 2019; Bertalan et al., 2019), while Lagrangian networks parameterize the Lagrangian and recover the corresponding equations of motion (Lutter et al., 2019; Cranmer et al., 2020). Other approaches preserve geometric structure through symplectic integration or directly parameterized symplectic maps (Chen et al., 2019; Zhong et al., 2020; Jin et al., 2020; Hairer et al., 2006), with related extensions to non-canonical and PDE settings (Eidnes & Lye, 2024). These methods primarily constrain the geometry of the learned dynamics. Our focus is complementary: even with a valid Hamiltonian structure, prediction quality still depends on how accurately the underlying scalar Hamiltonian can be represented.

Implicit neural representations and spectral bias. Coordinate-based MLPs, or implicit neural representations (INRs), learn continuous signals directly from their coordinates. A key limitation is spectral bias: standard MLPs tend to fit low-frequency components much faster than high-frequency ones (Rahaman et al., 2019; Tancik et al., 2020). Here, frequency is defined over the coordinate domain rather than time. To improve high-frequency representation, prior work modifies either the representation or the optimization process. Fourier features enrich the input spectrum (Tancik et al., 2020), SIREN uses periodic activations (Sitzmann et al., 2020), and methods such as WIRE, FINER, and BACON provide more explicit control over spatial-frequency content and bandwidth (Saragadam et al., 2023; Liu et al., 2024; Lindell et al., 2022). Other approaches instead modify training dynamics,

for example through neural-tangent-kernel-based gradient reweighting (Shi et al., 2025). A standard HNN can also be viewed as an INR over phase space. So spectral error is particularly important here for a better approximation of the dynamic system. High-frequency approximation errors can therefore be amplified through differentiation and directly distort the learned vector field. Rather than introducing a generic frequency encoding or activation, we create a new method to approximate the Hamiltonian using its kinetic–potential structure, locality, and repeated interactions, making the target function easier for the underlying coordinate networks to represent.

Local and multiscale representations. Multiscale representations have a long history in numerical analysis, including multigrid and multiresolution methods that organize information across spatial scales (Mandel et al., 1988; Mallat, 1989; Hughes, 1995). Similar ideas appear in modern neural representations. NeRF-style models represent continuous spatial fields with shared coordinate networks (Mildenhall et al., 2020), while mip-NeRF and multiresolution hash encodings combine information across multiple spatial scales (Barron et al., 2021; Müller et al., 2022). LM-HNN adopts this representation principle rather than the rendering objective. It expresses the Hamiltonian through shared local energy functions evaluated over multiple neighborhood radii, reducing the complexity seen by each network while directly encoding finite-range interactions. Sharing the same local energy law across sites further allows the representation to be reused across spatial locations and system sizes. In this sense, local multiscale construction serves as a physics-guided representations for making the Hamiltonian easier to approximate, rather than explicit Fourier decomposition of the dynamics.

3 BACKGROUND

Hamiltonian Dynamics. Hamiltonian mechanics describes a conservative system using generalized coordinates $\boldsymbol{q}$, conjugate momenta $\boldsymbol{p}$, and a scalar Hamiltonian $H(\boldsymbol{q}, \boldsymbol{p})$ representing the total energy (Goldstein et al., 2002; Marion, 2013). With $\boldsymbol{z} = (\boldsymbol{q}, \boldsymbol{p})$, the dynamics are

$$\dot{\boldsymbol{z}} = J\nabla_{\boldsymbol{z}} H(\boldsymbol{z}), \quad \text{where} \quad J = \begin{bmatrix} 0 & I \\ -I & 0 \end{bmatrix}, \tag{1}$$

and the exact flow conserves H and preserves the symplectic structure (Hairer et al., 2006). Hamiltonian neural networks (HNNs) (Greydanus et al., 2019; Toth et al., 2019) approximate H with a differentiable model H_θ, yielding vector field $J\nabla H_\theta$. Given M observed state–derivative pairs, the model is trained by minimizing

$$\mathcal{L}_{\text{deriv}} = \frac{1}{M} \sum\nolimits_{i=1}^{M} \| J\nabla H_\theta(\boldsymbol{z}_i) - \dot{\boldsymbol{z}}_i \|_2^2 . \tag{2}$$

Implicit Neural Representation. A standard HNN can be viewed as an implicit neural representation (INR) of the Hamiltonian over phase space, where a coordinate network maps the continuous state $\boldsymbol{z}$ to the scalar energy $H_\theta(\boldsymbol{z})$, whose gradient determines the dynamics. Hamiltonian mechanics constrains the structure of the vector field, while the neural representation determines how accurately the energy and its derivatives are approximated. Preserving Hamiltonian structure alone therefore does not guarantee accurate dynamics. The energy residual $\delta H = H_\theta - H$ affects the learned vector field through its gradient, giving

$$\| J\nabla H_\theta(\boldsymbol{z}) - \dot{\boldsymbol{z}} \|_2 = \| \nabla \delta H(\boldsymbol{z}) \|_2 . \tag{3}$$

where the equality follows from $J^\top J = I$. A constant offset has no effect, whereas a small but rapidly varying residual can produce a large vector-field error. Conventional HNNs represent the total energy using a single network over the full state, leaving kinetic–potential decomposition, locality, and repeated interactions to be learned implicitly from data.

Spectral Bias of Coordinate-Based Networks. Coordinate-based MLPs exhibit spectral bias, typically learning low-frequency target components more rapidly than high-frequency ones under gradient-based optimization (Rahaman et al., 2019; Xu et al., 2020). Here, frequency refers to variation over phase-space coordinates rather than temporal frequency along a trajectory. The neural tangent kernel (NTK) provides an explanation through the learning rates of different residual modes (Jacot et al., 2018). Under linearized gradient flow for squared-error function fitting, the component of the residual $r = f_\theta - f$ along eigenfunction v_j of the fixed NTK operator Θ evolves as

$$\langle v_j, r(t) \rangle = e^{-\lambda_j t} \langle v_j, r(0) \rangle, \tag{4}$$

where λ_j is the corresponding eigenvalue and t denotes optimization time. Higher-frequency components are often associated with smaller eigenvalues in conventional fully connected networks, explaining their slower convergence (Basri et al., 2019; Cao et al., 2021; Xu et al., 2020). Fourier features (Tancik et al., 2020) and periodic activations (Sitzmann et al., 2020) modify this spectral behavior and improve the recovery of fine-scale variation.

For HNNs, unresolved high-frequency structure is particularly consequential because the learned energy is differentiated to obtain the dynamics. For the energy residual δH, Parseval's identity gives

$$\int \|\nabla \delta H(\mathbf{z})\|_2^2 \, \mathrm{d}\mathbf{z} = \int \|\mathbf{k}\|_2^2 \left|\widehat{\delta H}(\mathbf{k})\right|^2 \mathrm{d}\mathbf{k}. \tag{5}$$

Higher-wavenumber components therefore receive greater weight in the vector-field error, even when their energy amplitude is small. Derivative supervision emphasizes these errors during training, but does not by itself guarantee their accurate recovery (Wang et al., 2022). This motivates improving how the Hamiltonian is represented. LM-HNN uses shared local energy functions with lower-dimensional inputs, allowing repeated interactions to be learned through a common representation. The next section develops this construction while retaining Hamiltonian structure.

4 METHOD

We design the Local Multiscale Hamiltonian Neural Network (LM-HNN) to improve the approximation of fine-scale energy variations by representing the Hamiltonian through lower-dimensional energy functions. We first decompose the Hamiltonian into kinetic and potential components (Sec. 4.1), then represent each through shared local functions over multiple neighborhood scales (Sec. 4.2). We discuss how this lower-dimensional representation supports the recovery of fine-scale energy structure and generalization to unseen energy levels (Sec. 4.3). The dynamics follow from the resulting scalar Hamiltonian through Hamilton's equations, preserving Hamiltonian structure.

4.1 ENERGY DECOMPOSITION AND LOCAL NEIGHBORHOODS

Following the decomposition of the Hamiltonian into kinetic and potential energy (Goldstein et al., 2002), we parameterize the two components using separate functions,

$$H_\theta(\boldsymbol{q}, \boldsymbol{p}) = T_\varphi(\boldsymbol{q}, \boldsymbol{p}) + V_\phi(\boldsymbol{q}), \tag{6}$$

where T_φ and V_ϕ denote the kinetic and potential energies respectively. The potential energy depends only on the configuration, while the kinetic energy may depend on both coordinates and momenta, accommodating configuration-dependent inertia and momentum coupling. This separation allows each component to use a parameterization suited to its physical structure. For kinetic energies that are quadratic and diagonal in momentum, we encode the quadratic dependence explicitly and learn the inverse-mass coefficients.

To describe the interaction structure within these components, let $\mathcal{N}_r(i)$ denote the interaction neighborhood of degree of freedom i at radius r. The corresponding local variables are

$$\boldsymbol{q}_i^r = \boldsymbol{q}_{\mathcal{N}_r(i)}, \qquad \boldsymbol{p}_i^r = \boldsymbol{p}_{\mathcal{N}_r(i)}. \tag{7}$$

These neighborhoods specify which variables enter each local energy function. They are determined by the system's interaction structure, with increasing r providing broader local context. On the chains, $\mathcal{N}_r(i)$ is a window of bond strains for the potential and a window of sites for the kinetic energy; the inputs on every system are listed in Appendix A.1. The next subsection uses these variables to construct shared energy functions and combine their contributions across neighborhood scales.

4.2 LOCAL MULTISCALE ENERGY PARAMETERIZATION

Potential Energy. We represent the potential energy as a weighted sum of shared local contributions,

$$V_\theta(\boldsymbol{q}) = \sum\nolimits_{r\in\mathcal{R}} \pi_r^V \sum\nolimits_{i\in\mathcal{I}_V} v_{\theta,r}\left(\boldsymbol{q}_i^r\right), \tag{8}$$

where $\mathcal{R}$ is the set of neighborhood radii and $\mathcal{I}_V$ is the index set of local contributions to the potential energy. The scale weights are then parameterized as

$$\pi_r^V = \frac{\exp(\eta_r^V)}{\sum_{r'\in\mathcal{R}} \exp(\eta_{r'}^V)}, \tag{9}$$

where η_r^V are learnable scale logits. At each radius, the shared network $v_{\theta,r}$ learns a common interaction law across neighborhoods with the same local structure. Combining contributions across scales incorporates both short- and longer-range interaction context.

Kinetic Energy. For kinetic energies that are quadratic and diagonal in momentum, we use

$$T_\theta(\boldsymbol{q}, \boldsymbol{p}) = \frac{1}{2} \sum\nolimits_{i=1}^{N} m_{\theta,i}^{-1}(\boldsymbol{q}) p_i^2, \tag{10}$$

with locally parameterized inverse-mass coefficients

$$m_{\theta,i}^{-1}(\boldsymbol{q}) = \epsilon + \mathrm{softplus}\left(\sum\nolimits_{r \in \mathcal{R}} \pi_r^T a_{\theta,r}\left(\boldsymbol{q}_i^r\right) \right). \tag{11}$$

At each radius r, the same function $a_{\theta,r}$ is applied to the local coordinates around every site, while π_r^T weights their contributions across radii. This parameterization explicitly encodes the quadratic dependence on momentum and learns the configuration-dependent inverse masses.

When momentum coupling is present, we instead use

$$T_\theta(\boldsymbol{q}, \boldsymbol{p}) = \sum\nolimits_{r \in \mathcal{R}} \pi_r^T \sum\nolimits_{i \in \mathcal{I}_T} t_{\theta,r}\left(\boldsymbol{q}_i^r, \boldsymbol{p}_i^r\right), \tag{12}$$

where $\mathcal{I}_T$ is the index set of local terms used to construct the learned kinetic energy. The shared functions $t_{\theta,r}$ depend jointly on the coordinates and momenta within each neighborhood, allowing them to represent local momentum coupling. These constructions assemble the Hamiltonian from shared, lower-dimensional energy functions matched to the system's interaction structure.

4.3 Spectral Resolution and Energy Generalization

For systems with repeated local interactions, local factorization reduces the dimensionality of Hamiltonian approximation by expressing the total energy as a sum of shared local functions,

$$H_\theta(\boldsymbol{z}) = \sum\nolimits_i h_\theta\left(\boldsymbol{z}_{\mathcal{N}(i)}\right), \tag{13}$$

where each local function receives only the variables involved in an interaction. Fine-scale variation is therefore represented directly in the coordinates on which it depends. Parameter sharing aggregates the contributions of repeated interactions, such that

$$\nabla_\theta H_\theta(\boldsymbol{z}) = \sum\nolimits_i \nabla_\theta h_\theta\left(\boldsymbol{z}_{\mathcal{N}(i)}\right). \tag{14}$$

These interactions are therefore captured by a common set of parameters, allowing the same local energy law to be learned across the system. Lower-dimensional inputs and repeated supervision can thereby support the recovery of fine-scale Hamiltonian structure.

The same factorization provides a mechanism for generalization to unseen energy levels. Let $\mathcal{S}_{\mathrm{train}}$ denote the training region in phase space and let $P_i \boldsymbol{z} = \boldsymbol{z}_{\mathcal{N}(i)}$ be the local projection. An unseen state may lie outside $\mathcal{S}_{\mathrm{train}}$ while its local configurations remain within or near the projected domains $P_i \mathcal{S}_{\mathrm{train}}$. Under this condition, the representation

$$H_\theta(\boldsymbol{z}) = \sum\nolimits_i h_\theta\left(P_i \boldsymbol{z}\right) \tag{15}$$

evaluates new combinations of familiar local configurations, reducing the need for extrapolation within each local function. The kinetic–potential decomposition further reduces the approximation burden. When the diagonal quadratic kinetic form is used, momentum dependence is encoded explicitly, leaving the networks to learn configuration-dependent inverse-mass coefficients and interaction potentials. These functions can then be reused across different states and energy levels.

5 Experiments

We evaluate the proposed representation through the learned Hamiltonian, its vector field, and the resulting rollouts. We first examine spectral bias and compare with existing mitigation techniques, then assess prediction accuracy within and beyond the training energy range across six conservative systems against structure-preserving baselines. Ablations examine which components contribute to performance and how their benefits vary with interaction structure, system size, and nonlinearity.

5.1 Dynamical Systems and Baselines

Dynamical systems. We evaluate LM-HNN on pendulum and double pendulum, Fermi–Pasta–Ulam-Tsingou (FPUT) chain (Fermi et al., 1955; Benettin et al., 2013) under nearest-neighbor, Toda (Toda, 1967), and Morse (Morse, 1929) potential, and the fast-slow Hamiltonian oscillator described in Duruisseaux et al. (2023). These six conservative systems span low-dimensional and many-body dynamics, separable and non-separable kinetic energies, different interaction laws, and separated timescales, providing a controlled setting for studying representation accuracy, spectral bias mitigation, and energy generalization. We include detailed definitions of these systems in Appendix A.2.

To study spectral bias in a controlled setting, we introduce a FPUT variant with perturbed potential

$$\phi_k(s) = \phi_{\mathrm{FPUT}}(s) + \frac{a}{k}(1 - \cos ks), \tag{16}$$

by adding force term $a \sin(ks)$. We fix $a = 0.5$ and vary k, thereby increasing the frequency of the Hamiltonian variation while keeping the force amplitude fixed.

Baselines. HNN (Greydanus et al., 2019) models the Hamiltonian using a single network, providing the direct comparison with our local decomposition under the same Hamiltonian formulation. SympNet (Jin et al., 2020) learns a symplectic discrete-time map, testing whether comparable accuracy can be achieved by preserving the geometry of evolution without explicitly learning the energy. FS-HNN (Li et al., 2026) learns separate Hamiltonian branches for temporal frequency bands, allowing us to compare decomposition by timescale and decomposition by physical interaction. The HNN and the variants of Table 1 receive the same inputs as LM-HNN (Appendix A.1).

5.2 Spectral Bias of the Learned Hamiltonian

Fine-scale Hamiltonian components can contribute substantially to the vector field even when their energy amplitudes are small, making their recovery important for accurate dynamics. We therefore test whether the proposed local decomposition makes these components easier to learn by examining Hamiltonian spectral errors during training on the unperturbed FPUT chain. Figure 1 (a) and (b) tracks the spectral power of the difference between the learned and true Hamiltonians during training. HNN and LM-HNN both learn low-wavenumber components first, but LM-HNN reduces higher-wavenumber residuals more rapidly at matched parameter counts and within the same training budget. Lower Hamiltonian residuals and vector-field errors show that the improvement is present in the learned energy and its derivatives before numerical integration (Table 10 and Figure 3, Appendix C.1).

To test whether the improved recovery on the unperturbed chain extends to finer energy variations, we use the perturbed FPUT potential in Equation 16. We increase the perturbation wavenumber k, which controls how rapidly the added force $a \sin(ks)$ varies with bond displacement s, while keeping its amplitude a fixed. Figure 1 (c) and (d) shows that HNN's recovery of this added force component deteriorates from approximately $k = 3$, whereas LM-HNN reaches nearly complete recovery through $k = 6$. The improved force recovery is accompanied by lower rollout MSE (Table 12), reflecting more accurate state predictions over time, and lower peak error (Table 13), indicating better agreement with the reference trajectory amplitudes (Appendix C.2).

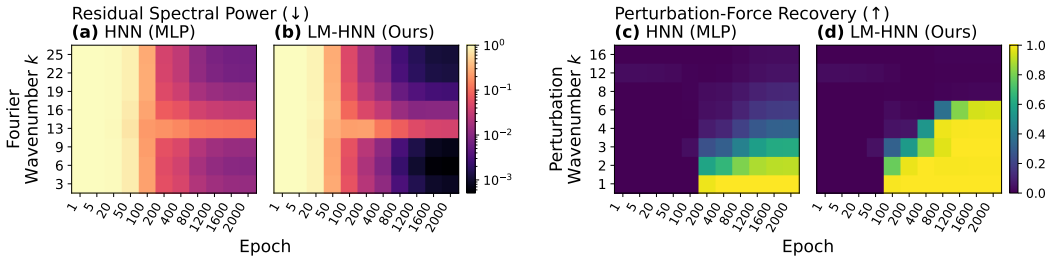

Figure 1: **Spectral bias in the learned Hamiltonian on the FPUT chain (N = 8). (a, b)** Residual Hamiltonian spectral power during training on the unperturbed FPUT chain, normalized by the target power at each Fourier wavenumber k. **(c, d)** Recovery of the added force component during training on the perturbed chain, where k now denotes the prescribed perturbation wavenumber, varied while keeping the force amplitude fixed. At matched parameter counts, LM-HNN reduces higher-wavenumber residuals more rapidly and recovers finer imposed interactions.

Table 1: **Baselines for mitigating spectral bias.** Derivative and rollout MSE for FPUT and pendulum for baseline mitigation techniques applied to HNN at matched parameter count. The structural approach of LM-HNN outperforms both Fourier features and sinusoidal activations.

	FPUT		Pendulum	
Model	Deriv. MSE	Rollout MSE	Deriv. MSE	Rollout MSE
HNN	$(3.6\pm0.3)\times10^{-2}$	$(2.0\pm0.1)\times10^{-1}$	$(1.9\pm0.7)\times10^{-6}$	$(2.7\pm0.6)\times10^{-5}$
Fourier (σ=0.5)	$(1.8\pm0.0)\times10^{-1}$	$(4.1\pm0.4)\times10^{-1}$	$(1.2\pm0.2)\times10^{-5}$	$(3.0\pm3.2)\times10^{-2}$
Fourier (σ=2)	$(1.8\pm0.1)\times10^{0}$	$(3.7\pm0.6)\times10^{0}$	$(3.7\pm0.2)\times10^{-4}$	$(9.3\pm3.7)\times10^{-2}$
SIREN (ω_0=10)	$(8.1\pm0.4)\times10^{-1}$	$(3.3\pm0.3)\times10^{0}$	$(1.9\pm0.2)\times10^{-7}$	$(3.3\pm2.8)\times10^{-4}$
SIREN (ω_0=30)	$(3.0\pm0.1)\times10^{0}$	$(5.4\pm0.5)\times10^{0}$	$(1.6\pm1.1)\times10^{-5}$	$(1.4\pm0.0)\times10^{0}$
LM-HNN	$\mathbf{(7.7\pm1.4)\times10^{-4}}$	$\mathbf{(4.9\pm0.6)\times10^{-3}}$	$\mathbf{(7.4\pm5.6)\times10^{-9}}$	$\mathbf{(4.1\pm2.5)\times10^{-8}}$

To test whether these gains can be achieved without decomposing the Hamiltonian, we compare LM-HNN with HNNs using Fourier features (Tancik et al., 2020) or sinusoidal activations (Sitzmann et al., 2020). Table 1 reports derivative and rollout errors at matched parameter counts. In the tested configurations, Fourier features increase both errors. Sinusoidal activations can improve derivative accuracy, but these gains do not extend to rollouts from unseen initial conditions. LM-HNN improves both metrics over HNN on FPUT and pendulum and achieves the lowest rollout errors, indicating that local energy decomposition provides benefits not reproduced by these spectral interventions.

5.3 Generalization to Unseen Energy Levels

Having established improved recovery of fine-scale Hamiltonian structure, we next examine whether LM-HNN's learned energy components support accurate dynamics beyond the training energy range. We evaluate the trained models from higher-energy initial conditions for the pendulum and chain systems and use a scaled-initial-condition test for the fast–slow oscillator, multiplying its test states by 1.5, which places 9 of the 30 above the largest training energy (Appendix A.2); the settings are summarized in Table 2.

Figure 2 compares errors in the predicted state derivatives for the pendulum and FPUT chain, with blue curves marking the maximum training energy. HNN's errors grow substantially beyond this boundary, whereas LM-HNN maintains lower errors and exhibits a smaller increase.

Table 2 shows that this advantage extends to rollout prediction at unseen energies. The baselines generally reach normalized errors of order one, comparable to the variance of the reference trajectories, whereas LM-HNN's mean error remains below 0.2 in every row, including the scaled-initial-condition test on the fast–slow oscillator. The separation is especially large for the pendulum systems, where LM-HNN retains very small prediction errors. LM-HNN's rollout errors increase on the chains, where higher energies produce larger bond strains, but remain substantially below the baselines across FPUT, FPUT-Toda, and FPUT-Morse.

Table 2: **Normalized rollout MSE at unseen energy levels.** The FPUT chains are trained on $E/N \in [0.05, 2.0]$ and tested at 2.5 and 3.0 per site; the pendulum is trained on $E \in [0.05, 4]$ and tested at 6 and 8; and the double pendulum is trained on $E \in [0.5, 5]$ and tested at 7 and 9. The fast–slow row is a scaled-initial-condition test: the test states are multiplied by 1.5, which places 9 of the 30 above the largest training energy of 2.48 (Appendix A.2). Baseline methods exhibit increased rollout error at unseen energy levels whilst LM-HNN maintains low error.

System	HNN	SympNet	FS-HNN	LM-HNN
Pendulum	$(1.3\pm0.2)\times10^{0}$	$(1.0\pm0.6)\times10^{0}$	$(6.1\pm1.2)\times10^{-1}$	$\mathbf{(2.7\pm2.4)\times10^{-8}}$
Double Pendulum	$(1.7\pm0.6)\times10^{0}$	$(3.5\pm0.6)\times10^{0}$	$(2.2\pm0.2)\times10^{0}$	$\mathbf{(8.1\pm5.3)\times10^{-6}}$
FPUT	$(1.8\pm0.1)\times10^{0}$	$(1.8\pm0.1)\times10^{0}$	$(2.2\pm0.1)\times10^{0}$	$\mathbf{(6.5\pm1.1)\times10^{-2}}$
FPUT-Toda	$(1.7\pm0.2)\times10^{0}$	$(1.9\pm0.2)\times10^{0}$	$(2.3\pm0.4)\times10^{0}$	$\mathbf{(1.3\pm0.4)\times10^{-2}}$
FPUT-Morse	$(1.2\pm0.1)\times10^{0}$	$(1.5\pm0.2)\times10^{0}$	$(1.9\pm0.1)\times10^{0}$	$\mathbf{(4.8\pm1.5)\times10^{-3}}$
Fast-Slow	$(4.8\pm0.8)\times10^{-1}$	$(1.9\pm0.4)\times10^{0}$	$(2.7\pm0.1)\times10^{0}$	$\mathbf{(1.4\pm0.4)\times10^{-1}}$

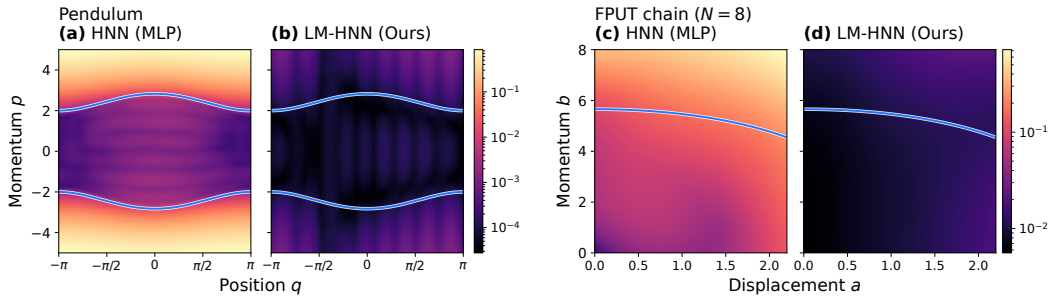

Figure 2: **Vector-field accuracy beyond the training-energy range.** Vector-field error $\|J\nabla H_\theta(\boldsymbol{z}) - \dot{\boldsymbol{z}}\|_2$, shown as a local root-mean-square for **(a, b)** the pendulum phase plane and **(c, d)** the FPUT chain ($N = 8$) on the slice $\boldsymbol{q} = a\mathbf{u}$, $\boldsymbol{p} = b\mathbf{v}$ along fixed random unit directions. Blue curves mark the maximum training energy, and each pair shares a logarithmic color scale. LM-HNN maintains lower errors beyond this boundary, supporting generalization to unseen energy levels.

On the pendulum the angular encoding is shared with the HNN, so it does not explain the gap; energies above the range are faster rotations with larger $|p|$, which LM-HNN evaluates through its fixed quadratic form $\frac{1}{2}m_\theta^{-1}(q)\,p^2$ while the HNN must extrapolate an unconstrained dependence on p. The chain results support the usefulness of shared local potentials over a wider range of configurations. Table 15 further shows that smaller errors between the learned and true Hamiltonians on validation states accompany lower rollout errors at unseen energies (Appendix C.4), connecting Hamiltonian approximation quality to generalization.

5.4 Accuracy of the Learned Hamiltonian Across Systems

The spectral and unseen-energy results motivate testing whether local energy decomposition improves prediction across different kinetic couplings, interaction potentials, and timescales. We evaluate all six systems from unseen initial conditions within the training distribution, using comparable parameter budgets (Table 14, Appendix C.3).

Table 3 shows that LM-HNN achieves the lowest mean rollout error across all six systems, demonstrating consistent benefits across different kinetic couplings, interaction laws, and timescales. The accompanying reduction in validation derivative errors (Table 3) shows that these gains extend to the underlying vector fields, connecting more accurate learned dynamics to better trajectory predictions.

Table 3: **Rollout and derivative MSE across six conservative systems.** SympNet learns a discrete symplectic map and therefore does not provide a continuous vector field for the derivative comparison. LM-HNN achieves the lowest rollout and derivative error on every system, showing consistent improvement across low-dimensional, lattice, and fast–slow dynamics.

System	HNN	SympNet	FS-HNN	LM-HNN (ours)
Rollout MSE				
Pendulum	$(2.7\pm0.6)\times10^{-5}$	$(2.9\pm1.2)\times10^{-2}$	$(6.2\pm2.8)\times10^{-6}$	$\mathbf{(4.1\pm2.5)\times10^{-8}}$
Double Pendulum	$(4.3\pm5.0)\times10^{-2}$	$(1.7\pm0.3)\times10^{0}$	$(1.9\pm0.9)\times10^{-2}$	$\mathbf{(3.2\pm3.2)\times10^{-6}}$
FPUT	$(2.0\pm0.1)\times10^{-1}$	$(6.2\pm0.6)\times10^{-2}$	$(1.2\pm0.4)\times10^{-2}$	$\mathbf{(4.9\pm0.6)\times10^{-3}}$
FPUT-Toda	$(7.0\pm0.9)\times10^{-1}$	$(1.7\pm0.7)\times10^{-1}$	$(2.6\pm0.8)\times10^{-1}$	$\mathbf{(6.7\pm4.7)\times10^{-3}}$
FPUT-Morse	$(5.1\pm0.5)\times10^{-1}$	$(1.3\pm0.4)\times10^{-1}$	$(2.1\pm0.3)\times10^{-1}$	$\mathbf{(1.5\pm0.4)\times10^{-3}}$
Fast-Slow	$(8.7\pm3.7)\times10^{-3}$	$(1.3\pm0.6)\times10^{0}$	$(2.1\pm0.9)\times10^{-3}$	$\mathbf{(1.8\pm0.5)\times10^{-3}}$
Derivative MSE				
Pendulum	$(1.9\pm0.7)\times10^{-6}$	–	$(5.0\pm1.7)\times10^{-6}$	$\mathbf{(7.4\pm5.6)\times10^{-9}}$
Double Pendulum	$(1.4\pm0.2)\times10^{-3}$	–	$(2.2\pm0.2)\times10^{-4}$	$\mathbf{(3.3\pm2.2)\times10^{-6}}$
FPUT	$(6.0\pm0.1)\times10^{-3}$	–	$(3.4\pm3.3)\times10^{-3}$	$\mathbf{(5.6\pm0.9)\times10^{-4}}$
FPUT-Toda	$(1.5\pm0.1)\times10^{-2}$	–	$(2.7\pm0.1)\times10^{-3}$	$\mathbf{(4.5\pm0.4)\times10^{-4}}$
FPUT-Morse	$(6.2\pm0.4)\times10^{-3}$	–	$(7.1\pm1.2)\times10^{-3}$	$\mathbf{(2.3\pm0.6)\times10^{-4}}$
Fast-Slow	$(9.8\pm0.2)\times10^{-7}$	–	$(6.1\pm1.4)\times10^{-7}$	$\mathbf{(3.1\pm0.6)\times10^{-7}}$

Table 4: **Structural ablations of LM-HNN.** Normalized rollout MSE on the FPUT and Toda chains ($N = 8$). Locality and kinetic–potential decomposition improve accuracy, while weight sharing reduces model size without a consistent accuracy penalty. Full metrics are reported in Appendix B.

Ablation Variant	Param.	FPUT	FPUT-Toda
LM-HNN	26,892	$(4.9_{\pm 0.6})\times 10^{-3}$	$(\mathbf{6.7_{\pm 4.7}})\times \mathbf{10^{-3}}$
w/o kinetic–potential split	26,736	$(3.8_{\pm 0.6})\times 10^{-2}$	$(7.1_{\pm 1.2})\times 10^{-2}$
w/o locality	26,932	$(5.5_{\pm 0.5})\times 10^{-2}$	$(3.8_{\pm 1.0})\times 10^{-1}$
w/o weight sharing	228,537	$(\mathbf{3.5_{\pm 0.8}})\times \mathbf{10^{-3}}$	$(1.9_{\pm 0.2})\times 10^{-2}$

Figure 4 illustrates the resulting trajectory behavior through reduced phase and amplitude errors on the chains and more accurate tracking of the slow drift and fast oscillation phase in the fast–slow system (Appendix C.3).

5.5 Ablations

To identify which structural choices drive LM-HNN's accuracy gains, we ablate its components on the FPUT and FPUT-Toda chains. Removing locality produces the largest increase in rollout error, followed by removing the kinetic–potential split (Table 4). Both variants retain Hamiltonian structure and similar parameter counts, supporting the importance of organizing the energy approximation through local functions and separate energy components. Removing weight sharing increases the parameter count by approximately $8.5\times$ without a consistent accuracy gain. Locality and energy decomposition therefore contribute most directly to accuracy, while sharing represents repeated interactions efficiently.

The receptive-field study clarifies the role of multiple scales (Table 6, Appendix B). A radius-zero neighborhood contains a single bond strain and performs best on the nearest-neighbor chain, but cannot represent the added second-neighbor coupling, for which the best single radius is $\mathcal{R} = \{1\}$. The multiscale set $\mathcal{R} = \{0, 1, 2\}$ accommodates both interaction ranges and achieves the lowest derivative error on the two-scale chain, while matched single radii yield lower rollout errors. The kinetic-head comparison similarly favors diagonal quadratic forms for uncoupled momenta, local sums for finite-range momentum coupling, and a dense representation for global coupling (Table 7).

The chain-length sweep further supports the shared local representation (Table 8). From $N = 4$ to $N = 16$, LM-HNN retains the same parameter count and keeps mean rollout error below 10^{-2}, while its advantage over parameter-matched HNNs widens. This is consistent with learning local interaction laws whose complexity does not grow with the number of repeated interactions. The accuracy advantage also persists across the tested interaction nonlinearities (Table 9).

6 Conclusion

Hamiltonian structure determines how a learned energy function generates dynamics, but accurate long-term prediction ultimately depends on how well that energy itself is represented. We show that a monolithic neural representation of the Hamiltonian inherits the spectral bias of coordinate-based neural networks, and because the dynamics depend on gradients of the energy, its unresolved fine-scale components translate directly into substantial vector-field and rollout errors. LM-HNN addresses this by decomposing the Hamiltonian into kinetic and potential components represented by shared local energy functions over multiple interaction scales. As a result, LM-HNN extends the effective spectral range of the learned Hamiltonian, consistently improves derivative and long-horizon rollout accuracy across low-dimensional, interacting, and fast–slow systems, and remains accurate at unseen energy levels.

Future Work. LM-HNN's shared local energy decomposition could be extended to more general interaction structures and systems described by spatially varying fields.

Limitations. The current formulation assumes conservative dynamics and does not explicitly model dissipation or external forcing. Additionally, the current formulation assumes access to the full system state and does not explicitly address partially observed dynamics.

AI USE STATEMENT

In this work, we used generative AI tools for generating synthetic datasets, implementing methods, supporting qualitative and thematic data analysis, and designing and running experiments. We have not used generative AI tools for developing theoretical models or conceptual frameworks, formulating mathematical claims, providing critical ingredients for proving mathematical claims, or interpreting results, and assisting in writing proofs, proposing or refining hypotheses, assisting with translation, cleaning and reformatting datasets are not applicable to this work. Additionally, we used generative AI tools for creating or editing software code, drafting parts of the manuscript, and editing the manuscript to improve readability. We have reviewed all AI-assisted work. The authors reviewed the implementation and experimental procedures, verified the reported results, and checked the manuscript for accuracy. We take responsibility for the final content of this work, including text, claims or artifacts produced with the aid of generative AI.

ETHICS STATEMENT

This work is methodological. All data are generated by numerically integrating analytic Hamiltonians of textbook mechanical systems, namely pendulums, nonlinear lattices, and a fast–slow oscillator, so the study involves no human subjects, personal data, or content that could identify individuals. The intended use is the data-driven modeling and simulation of conservative physical systems, and we are not aware of a direct path from the method to harmful applications. As with any learned dynamics model, the accuracy of LM-HNN is established only within the regimes evaluated here, and deployment in safety-relevant settings, such as the control of physical systems, would require validation on the target system. Our use of generative AI tools is disclosed in the AI use statement.

REPRODUCIBILITY STATEMENT

We describe the experimental settings and implementation details to support reproducibility, and will release the source code alongside the paper upon acceptance. Section 4 specifies the model formulation, and Section 5.1 introduces the benchmark systems and baselines. Appendix A.2 provides the Hamiltonian definitions, data-generation procedures, and training and test energy ranges. Appendix A.1 describes the data splits, optimization and model-selection procedures, numerical integration, rollout horizons, evaluation metrics, and reporting across independent training seeds. Appendix A.3 defines the spectral diagnostics and their computation, while Appendix B describes the ablation configurations and controlled comparisons. Model parameter counts are reported in Table 14.

REFERENCES

Jonathan T. Barron, Ben Mildenhall, Matthew Tancik, Peter Hedman, Ricardo Martin-Brualla, and Pratul P. Srinivasan. Mip-NeRF: A multiscale representation for anti-aliasing neural radiance fields. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 5855–5864, 2021.

Ronen Basri, David Jacobs, Yoni Kasten, and Shira Kritchman. The convergence rate of neural networks for learned functions of different frequencies. In *Advances in Neural Information Processing Systems*, volume 32, 2019.

Jörg Behler and Michele Parrinello. Generalized neural-network representation of high-dimensional potential-energy surfaces. *Physical review letters*, 98(14):146401, 2007.

Giancarlo Benettin, Helen Christodoulidi, and Antonio Ponno. The fermi-pasta-ulam problem and its underlying integrable dynamics. *Journal of Statistical Physics*, 152(2):195–212, 2013.

Tom Bertalan, Felix Dietrich, Igor Mezić, and Ioannis G Kevrekidis. On learning hamiltonian systems from data. *Chaos: An Interdisciplinary Journal of Nonlinear Science*, 29(12), 2019.

Steven L Brunton, Joshua L Proctor, and J Nathan Kutz. Discovering governing equations from data by sparse identification of nonlinear dynamical systems. *Proceedings of the national academy of sciences*, 113(15):3932–3937, 2016.

Yuan Cao, Zhiying Fang, Yue Wu, Ding-Xuan Zhou, and Quanquan Gu. Towards understanding the spectral bias of deep learning. In *Proceedings of the Thirtieth International Joint Conference on Artificial Intelligence (IJCAI)*, pp. 2205–2211, 2021.

Ricky TQ Chen, Yulia Rubanova, Jesse Bettencourt, and David K Duvenaud. Neural ordinary differential equations. *Advances in neural information processing systems*, 31, 2018.

Zhengdao Chen, Jianyu Zhang, Martin Arjovsky, and Léon Bottou. Symplectic recurrent neural networks. *arXiv preprint arXiv:1909.13334*, 2019.

Miles Cranmer, Sam Greydanus, Stephan Hoyer, Peter Battaglia, David Spergel, and Shirley Ho. Lagrangian neural networks. In *ICLR 2020 Workshop on Integration of Deep Neural Models and Differential Equations*, 2020.

Valentin Duruisseaux, Joshua W. Burby, and Qi Tang. Approximation of nearly-periodic symplectic maps via structure-preserving neural networks. *Scientific Reports*, 13:8351, 2023.

Sølve Eidnes and Kjetil Olsen Lye. Pseudo-hamiltonian neural networks for learning partial differential equations. *Journal of Computational Physics*, 500:112738, 2024.

Enrico Fermi, John Pasta, Stanislaw Ulam, and Mary Tsingou. Studies of nonlinear problems. Technical Report LA-1940, Los Alamos Scientific Laboratory, 1955.

Herbert Goldstein, Charles P. Poole, and John L. Safko. *Classical Mechanics*. Addison-Wesley, 3 edition, 2002. ISBN 9780201657029.

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

Nate Gruver, Marc Finzi, Samuel Stanton, and Andrew Gordon Wilson. Deconstructing the inductive biases of hamiltonian neural networks. *arXiv preprint arXiv:2202.04836*, 2022.

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

Thomas JR Hughes. Multiscale phenomena: Green's functions, the dirichlet-to-neumann formulation, subgrid scale models, bubbles and the origins of stabilized methods. *Computer methods in applied mechanics and engineering*, 127(1-4):387–401, 1995.

Arthur Jacot, Franck Gabriel, and Clément Hongler. Neural tangent kernel: Convergence and generalization in neural networks. In *Advances in Neural Information Processing Systems*, volume 31, 2018.

Pengzhan Jin, Zhen Zhang, Aiqing Zhu, Yifa Tang, and George Em Karniadakis. Sympnets: Intrinsic structure-preserving symplectic networks for identifying hamiltonian systems. *Neural Networks*, 132:166–179, 2020.

Yaojun Li, Yulong Yang, and Christine Allen-Blanchette. Frequency-separable Hamiltonian neural network for multi-timescale dynamics. *arXiv preprint arXiv:2603.06354*, 2026.

Zongyi Li, Nikola Kovachki, Kamyar Azizzadenesheli, Burigede Liu, Kaushik Bhattacharya, Andrew Stuart, and Anima Anandkumar. Fourier neural operator for parametric partial differential equations. *arXiv preprint arXiv:2010.08895*, 2020.

David B. Lindell, Dave Van Veen, Jeong Joon Park, and Gordon Wetzstein. BACON: Band-limited coordinate networks for multiscale scene representation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 16252–16262, 2022.

Zhen Liu, Hao Zhu, Qi Zhang, Jingde Fu, Weibing Deng, Zhan Ma, Yanwen Guo, and Xun Cao. FINER: Flexible spectral-bias tuning in implicit NEural representation by variable-periodic activation functions. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 2713–2722, 2024.

Lu Lu, Pengzhan Jin, Guofei Pang, Zhongqiang Zhang, and George Em Karniadakis. Learning nonlinear operators via deeponet based on the universal approximation theorem of operators. *Nature machine intelligence*, 3(3):218–229, 2021.

Michael Lutter, Christian Ritter, and Jan Peters. Deep lagrangian networks: Using physics as model prior for deep learning. *arXiv preprint arXiv:1907.04490*, 2019.

Stephane G Mallat. Multiresolution approximations and wavelet orthonormal bases of l^2 (r). *Transactions of the American mathematical society*, 315(1):69–87, 1989.

Jan Mandel, Steve McCormick, and John Ruge. An algebraic theory for multigrid methods for variational problems. *SIAM Journal on Numerical Analysis*, 25(1):91–110, 1988.

Jerry B Marion. *Classical dynamics of particles and systems*. Academic Press, 2013.

Ben Mildenhall, Pratul P. Srinivasan, Matthew Tancik, Jonathan T. Barron, Ravi Ramamoorthi, and Ren Ng. NeRF: Representing scenes as neural radiance fields for view synthesis. In *European Conference on Computer Vision*, pp. 405–421, 2020. doi: 10.1007/978-3-030-58452-8_24.

Philip M. Morse. Diatomic molecules according to the wave mechanics. II. Vibrational levels. *Physical Review*, 34(1):57–64, 1929.

Thomas Müller, Alex Evans, Christoph Schied, and Alexander Keller. Instant neural graphics primitives with a multiresolution hash encoding. *ACM Transactions on Graphics*, 41(4):1–15, 2022. doi: 10.1145/3528223.3530127.

Nasim Rahaman, Aristide Baratin, Devansh Arpit, Felix Draxler, Min Lin, Fred Hamprecht, Yoshua Bengio, and Aaron Courville. On the spectral bias of neural networks. In *Proceedings of the 36th International Conference on Machine Learning*, volume 97 of *Proceedings of Machine Learning Research*, pp. 5301–5310. PMLR, 2019.

Vishwanath Saragadam, Daniel LeJeune, Jasper Tan, Guha Balakrishnan, Ashok Veeraraghavan, and Richard G. Baraniuk. WIRE: Wavelet implicit neural representations. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 18507–18516, 2023.

Kexuan Shi, Hai Chen, Leheng Zhang, and Shuhang Gu. Inductive gradient adjustment for spectral bias in implicit neural representations. In *Proceedings of the 42nd International Conference on Machine Learning*, volume 267 of *Proceedings of Machine Learning Research*, pp. 54864–54891. PMLR, 2025.

Vincent Sitzmann, Julien N. P. Martel, Alexander W. Bergman, David B. Lindell, and Gordon Wetzstein. Implicit neural representations with periodic activation functions. In *Advances in Neural Information Processing Systems*, volume 33, 2020.

Matthew Tancik, Pratul P. Srinivasan, Ben Mildenhall, Sara Fridovich-Keil, Nithin Raghavan, Utkarsh Singhal, Ravi Ramamoorthi, Jonathan T. Barron, and Ren Ng. Fourier features let networks learn high frequency functions in low dimensional domains. In *Advances in Neural Information Processing Systems*, volume 33, 2020.

Morikazu Toda. Vibration of a chain with nonlinear interaction. *Journal of the Physical Society of Japan*, 22(2):431–436, 1967.

Peter Toth, Danilo Jimenez Rezende, Andrew Jaegle, Sébastien Racanière, Aleksandar Botev, and Irina Higgins. Hamiltonian generative networks. *arXiv preprint arXiv:1909.13789*, 2019.

Sifan Wang, Xinling Yu, and Paris Perdikaris. When and why PINNs fail to train: A neural tangent kernel perspective. *Journal of Computational Physics*, 449:110768, 2022.

Zhi-Qin John Xu, Yaoyu Zhang, Tao Luo, Yanyang Xiao, and Zheng Ma. Frequency principle: Fourier analysis sheds light on deep neural networks. *Communications in Computational Physics*, 28(5):1746–1767, 2020.

Yaofeng Desmond Zhong, Biswadip Dey, and Amit Chakraborty. Symplectic ODE-net: Learning hamiltonian dynamics with control. In *International Conference on Learning Representations*, 2020.

A SYSTEMS, DATA GENERATION, AND EVALUATION PROTOCOL

This appendix describes the experimental setup underlying the results in the main text, including how trajectories from the benchmark systems are generated and used for model training and evaluation. The spectral diagnostics complement these evaluations by examining errors in the learned Hamiltonians and their derivatives.

A.1 TRAINING AND EVALUATION PROTOCOL

Data generation. Unless stated otherwise, trajectories are generated with the classical fourth-order Runge–Kutta (RK4) method from initial conditions sampled on prescribed energy shells,

$$H(\boldsymbol{z}_0) = E.$$

States from the generated training trajectories are pooled and randomly split into 80% training and 20% validation sets. For Hamiltonian models, analytic derivatives

$$\dot{\boldsymbol{z}} = J\nabla H(\boldsymbol{z})$$

are used as supervision, and the models are trained with the derivative loss in Eq. (2). SympNet is trained on the same underlying trajectories using its one-step prediction objective, while FS-HNN follows its original multiscale training procedure on the same datasets.

Model inputs. On a chain with fixed ends $q_0 = q_{N+1} = 0$, the potential sums over bonds $b = 0, \dots, N$ and its local variables are the strain window $\mathcal{S}_b^r = (s_{b-r}, \dots, s_{b+r})$ with $s_b = q_{b+1} - q_b$, zero-padded at the ends, so the neighborhood of Eq. (7) is made of differences of neighboring positions; the kinetic energy sums over sites and uses the position window $(q_{i-r}, \dots, q_{i+r})$, joined by the momentum window in Eq. (12). At $r = 0$ the potential network sees one strain, so $v_{\theta,0}(s_b)$ is a nearest-neighbor bond potential, and the inverse mass of site i depends on q_i alone; the same $\mathcal{R}$ is used for both components. On the angle systems each angle enters every local function as $(\sin q, \cos q)$ and the momenta only through the quadratic form: $T_\theta = \frac{1}{2} m_\theta^{-1}(q)\, p^2$ for the pendulum, and $T_\theta = \frac{1}{2} \boldsymbol{p}^\top M_\theta^{-1}(q_1 - q_2)\, \boldsymbol{p}$ with a Cholesky-factorized M_θ^{-1} and $V_\theta = v_1(q_1) + v_2(q_2) + v_{12}(q_1 - q_2)$ for the double pendulum. The HNN and its Fourier-feature and SIREN variants receive the same inputs as LM-HNN; SympNet and FS-HNN act on the raw state, with angle trajectories unwrapped so that they remain trajectories of the same system.

Optimization and model selection. Hamiltonian models are optimized with AdamW and a cosine learning-rate schedule. We train for 2000 epochs on the chain systems and for 600–1000 epochs on the remaining systems. Model selection is based on the validation derivative error. The derivative MSE reported in the tables is the squared error averaged over the state components as well as over the states, that is, the training loss of Eq. (2) divided by the state dimension $2N$; the RMS vector-field error of the spectral diagnostics is the square root of $2N$ times this quantity. Unless stated otherwise, all reported results are the mean $\pm$ standard deviation over three independently trained model seeds.

Rollout evaluation. Generalization is evaluated from initial conditions that are not used in any training trajectory. In-distribution test initial conditions are sampled independently from the same energy ranges used for training, while the unseen-energy experiments use energy levels above the training range as specified in Appendix A.2.

For Hamiltonian models, rollouts are generated by integrating the learned vector field with RK4. SympNet instead predicts the discrete evolution directly and is iterated autoregressively. We use 120 rollout steps for the chain systems, 200 steps for the pendulum systems, and 1000 steps for the fast–slow oscillator. For the fast–slow system, evaluation uses the 30 held-out reference test trajectories.

We report the final-step normalized rollout error,

$$\mathcal{E}_{\mathrm{roll}} = \frac{\mathrm{MSE}(\hat{\boldsymbol{z}}(T), \boldsymbol{z}(T))}{\mathrm{Var}(\boldsymbol{z}(t))_{t\in[0,T]}}, \tag{17}$$

where the denominator is the variance of the corresponding reference trajectory over the rollout horizon.

For the chain systems, we additionally measure errors in the predicted trajectory amplitudes using

$$\mathcal{E}_{\text{peak}} = \frac{1}{2}\frac{|\max_t \hat{x}(t) - \max_t x(t)| + |\min_t \hat{x}(t) - \min_t x(t)|}{\max_t x(t) - \min_t x(t)}, \tag{18}$$

averaged over all state variables. We refer to this quantity as the *peak error*.

The two measures answer different questions, and we report both because a model can do well on one while failing on the other. The normalized rollout error compares states at a single instant, so it is dominated by phase: once a predicted trajectory drifts in phase, its pointwise difference from the reference saturates at the scale of the motion itself, whether or not the predicted motion still has the right extent. The peak error instead compares the turning points of the motion and does not depend on when they are reached. This distinction is physically meaningful here, because in a conservative system the amplitude of a trajectory is fixed by the level set of the Hamiltonian passing through its initial state. A model whose predicted extrema are wrong has therefore learned an energy surface of the wrong shape, rather than a correct trajectory that has merely fallen out of step. On the chains the turning points are also where the anharmonic part of the bond potential is probed most strongly, so the peak error reports directly on the part of the potential that a smooth global approximation resolves least well. We report the peak error for the chain systems only: the angular coordinates of the pendulum systems are wrapped to $(-\pi, \pi]$, so their extrema carry no amplitude information.

A.2 Dynamical Systems and Data Generation

Pendulum. We consider the unit-mass, unit-length pendulum with unit gravitational acceleration,

$$H(q, p) = \frac{1}{2}p^2 + (1 - \cos q). \tag{19}$$

The training set contains sixteen energy shells evenly spaced over $E \in [0.05, 4.0]$, spanning both librational and rotational motion; the separatrix occurs at $E = 2$.

For each energy shell, we generate eight initial conditions. The angle is sampled uniformly from $(-\pi, \pi]$ and rejected whenever $V(q) > E$. The momentum is then determined by the remaining kinetic energy,

$$p = \pm\sqrt{2(E - V(q))}, \tag{20}$$

with the sign chosen randomly. Each initial condition is integrated for 300 steps with $\Delta t = 0.05$, and the angle is wrapped to $(-\pi, \pi]$, yielding 38,400 states in total.

In-distribution rollouts start from two unseen initial conditions on each of the shells

$$E \in \{0.5, 1.5, 2.5, 3.5\},$$

and run for 200 steps. Angular errors are computed modulo 2π. For unseen-energy evaluation, we use

$$E \in \{6, 8\},$$

both above the maximum training energy of 4.

Double pendulum. We use a double pendulum with $m_1 = m_2 = l_1 = l_2 = g = 1$ and canonical coordinates (q_1, q_2, p_1, p_2). Its Hamiltonian is

$$H(\boldsymbol{q}, \boldsymbol{p}) = \frac{1}{2}\boldsymbol{p}^\top M(\boldsymbol{q})^{-1}\boldsymbol{p} + (m_1 + m_2)gl_1(1 - \cos q_1) + m_2 gl_2(1 - \cos q_2), \tag{21}$$

$$M(\boldsymbol{q}) = \begin{pmatrix} (m_1 + m_2)l_1^2 & m_2 l_1 l_2 \cos(q_1 - q_2) \\ m_2 l_1 l_2 \cos(q_1 - q_2) & m_2 l_2^2 \end{pmatrix}. \tag{22}$$

Training uses sixteen energy shells evenly spaced over $E \in [0.5, 5.0]$, with eight initial conditions on each shell. The angles are sampled uniformly and rejected if $V(\boldsymbol{q}) > E$. A random direction $\mathbf{u}$ in velocity space is then scaled so that

$$\frac{1}{2}s^2\mathbf{u}^\top M(\boldsymbol{q})\mathbf{u} = E - V(\boldsymbol{q}), \tag{23}$$

after which the canonical momentum is set to

$$\boldsymbol{p} = M(\boldsymbol{q})\, s\mathbf{u}. \tag{24}$$

Each trajectory contains 300 steps with $\Delta t = 0.02$.

In-distribution rollouts use two unseen initial conditions from each of

$$E \in \{1, 2, 3, 4\}$$

and run for 200 steps. The unseen-energy shells are

$$E \in \{7, 9\},$$

compared with a maximum training energy of 5.

FPUT, Toda, and Morse chains. We consider fixed-boundary chains of N unit masses with $q_0 = q_{N+1} = 0$. Defining the bond strain

$$s_b = q_{b+1} - q_b, \qquad b = 0, \dots, N,$$

the Hamiltonian takes the form

$$H(\boldsymbol{q}, \boldsymbol{p}) = \frac{1}{2}\boldsymbol{p}^\top A \boldsymbol{p} + \sum_{b=0}^{N} \phi(s_b). \tag{25}$$

We use $A = I$ except in the kinetic-coupling ablation of Table 7.

For the FPUT chain,

$$\phi_{\mathrm{FPUT}}(s) = \frac{1}{2}s^2 + \frac{\alpha}{3}s^3 + \frac{\beta}{4}s^4, \tag{26}$$

with $(\alpha, \beta) = (0.25, 0.10)$ unless stated otherwise. We additionally consider the Toda potential

$$\phi_{\mathrm{Toda}}(s) = e^s - s - 1, \tag{27}$$

and the Morse potential

$$\phi_{\mathrm{Morse}}(s) = D(1 - e^{-as})^2, \qquad D = 8, \qquad a = \frac{1}{\sqrt{2D}}, \tag{28}$$

where the parameters are chosen so that all three interaction potentials have unit curvature at $s = 0$.

For the controlled spectral-bias experiment of Section 5.2, we augment the FPUT interaction with an oscillatory component,

$$\phi_k(s) = \phi_{\mathrm{FPUT}}(s) + \frac{a}{k}\big(1 - \cos ks\big), \qquad a = 0.5. \tag{29}$$

The corresponding bond force therefore contains the additional term $a \sin(ks)$. Its amplitude is fixed as k varies, while the energy amplitude of the perturbation decreases as $1/k$. This construction allows us to vary the wavenumber of the Hamiltonian perturbation without simultaneously increasing the magnitude of the associated force.

For the receptive-field study of Appendix B we also use a chain carrying two interaction ranges at once, obtained by adding a harmonic second-neighbor coupling to the FPUT bonds,

$$H(\boldsymbol{q}, \boldsymbol{p}) = \frac{1}{2}\|\boldsymbol{p}\|^2 + \sum_{b=0}^{N} \phi_{\mathrm{FPUT}}(s_b) + \frac{g}{2}\sum_{i=0}^{N-1} \big(q_{i+2} - q_i\big)^2, \qquad g = 0.3, \tag{30}$$

again with $q_0 = q_{N+1} = 0$. At the training amplitudes the second-neighbor term carries about 21% of the potential energy, so it is a genuine but secondary contribution. The point of this system is that the two ranges are separated in the strain coordinates that the local potential heads read: since $q_{i+2} - q_i = s_i + s_{i+1}$, a window of radius 0 contains a single strain and cannot represent the second-neighbor term at all, whereas radius 1 and above can. Data generation and evaluation follow the protocol used for the other chains.

For each chain system, the training data contain eight energy shells evenly spaced over the energy-per-site range

$$E/N \in [0.05, 2.0].$$

At each energy, we generate four trajectories of 100 steps with $\Delta t = 0.05$.

Two trajectories begin from momentum-dominated states. Positions are sampled uniformly from

$$\boldsymbol{q} \in [-0.35, 0.35]^N,$$

and the remaining energy is assigned to the momentum along a random direction $\mathbf{u}$:

$$\boldsymbol{p} = s\mathbf{u}, \qquad \frac{1}{2}s^2\mathbf{u}^\top A\mathbf{u} = E - V(\boldsymbol{q}). \tag{31}$$

The other two trajectories begin from potential-dominated states. A random direction $\mathbf{u}$ in configuration space is scaled by bisection until

$$V(s\mathbf{u}) = fE, \qquad f \sim \mathcal{U}(0.55, 0.95), \tag{32}$$

and the remaining energy is assigned to the momentum. These near-turning-point states increase coverage of the larger-strain configurations that determine the extrema of the resulting trajectories.

The data-generation seed is $100 + N$. For $N = 8$, the resulting training dataset contains 3,200 states. In-distribution rollouts start from two unseen momentum-dominated initial conditions at each of

$$E/N \in \{0.25, 0.75, 1.25, 1.75\}$$

and run for 120 steps. Unseen-energy evaluation uses four initial conditions at each of

$$E/N \in \{2.5, 3.0\},$$

both above the maximum training value of 2.0.

Fast–slow oscillator. We use the nearly periodic Hamiltonian system of Duruisseaux et al. (2023). With $\boldsymbol{z} = (q_1, q_2, p_1, p_2)$ and $\varepsilon = 0.01$,

$$H_\varepsilon = \frac{1}{2}\big(q_1^2 + p_1^2\big) + \frac{\varepsilon}{2}\big(q_2^2 + p_2^2\big) + \varepsilon q_1 q_2 \sin(2q_1 + 2q_2). \tag{33}$$

The pair (q_1, p_1) oscillates at unit frequency, while (q_2, p_2) evolves on the slower scale ε, with nonlinear coupling of order ε.

We use the reference dataset of Duruisseaux et al. (2023), consisting of 300 trajectories of 3000 steps with $\Delta t = 0.1$. Initial conditions are drawn from three regions of phase space: one general window and two windows containing small slow-oscillator orbits centered to the left and right. The trajectories are divided into 240 training, 30 validation, and 30 test trajectories.

For derivative-trained Hamiltonian models, we subsample every twentieth state from the training and validation trajectories, yielding 40,770 states with analytic derivative targets. SympNet and FS-HNN use the full trajectories according to their respective training procedures.

Rollouts start from the initial states of the 30 held-out test trajectories and run for 1000 steps. In addition to the full-state error, we report the normalized rollout error of the slow pair (q_2, p_2). The scaled-initial-condition test multiplies the same 30 initial states by 1.5, which raises their energy by a factor between 2.1 and 2.5. The training trajectories have energies in $[0.04, 2.48]$ and the test states in $[0.02, 2.02]$; after scaling the energies lie in $[0.05, 4.56]$, and 9 of the 30 states exceed the largest training energy. The other states move to larger amplitudes of both oscillators at energies that training still covers, so this row of Table 2 measures extrapolation in amplitude rather than in energy alone.

A.3 Spectral Diagnostics

To characterize errors in the learned Hamiltonian directly, we evaluate several spectral diagnostics on 2048 validation states from each system using double precision.

Let

$$\delta H(\boldsymbol{z}) = H_\theta(\boldsymbol{z}) - H(\boldsymbol{z}) \tag{34}$$

denote the Hamiltonian residual. Because the Hamiltonian is identifiable only up to an additive constant from derivative supervision, we subtract the mean of δH before computing energy-error statistics. We report its RMS magnitude,

$$\sigma_H = \sqrt{\mathbb{E}\Big[\big(\delta H - \mathbb{E}[\delta H]\big)^2\Big]}, \tag{35}$$

together with the RMS gradient error

$$\|\nabla \delta H\|_{\mathrm{rms}}. \tag{36}$$

Since the canonical symplectic matrix J is orthogonal,

$$\|J\nabla H_\theta - \dot{\boldsymbol{z}}\| = \|J\nabla\delta H\| = \|\nabla\delta H\|, \tag{37}$$

so this quantity is exactly the RMS vector-field error of the Hamiltonian model.

We additionally define the effective residual wavenumber

$$k_{\mathrm{eff}} = \frac{\|\nabla\delta H\|_{\mathrm{rms}}}{\sigma_H}. \tag{38}$$

This quantity characterizes the characteristic spatial frequency of the remaining Hamiltonian residual: for fixed residual magnitude, larger k_{eff} corresponds to more rapidly varying error. It is therefore interpreted jointly with σ_H and the vector-field error rather than as an accuracy metric on its own.

To inspect the residual spectrum more directly, we construct 96 one-dimensional probe lines through validation states. For chain systems, each line varies one site while holding the remaining coordinates fixed; for the pendulum systems, one angular coordinate is varied. Along each line, we compute the power spectra of the target Hamiltonian H, the residual δH, and the corresponding directional derivative of the residual using a Blackman–Harris window. We define k_{99} as the wavenumber below which 99% of the target Hamiltonian power is contained.

The heat maps in Figure 1 are constructed as follows. For the perturbed FPUT chain, we quantify recovery of the added wavenumber-k force component using

$$\lambda = \frac{\mathrm{MSE}_0 - \mathrm{MSE}_k + \mathrm{MSE}_{\mathrm{ignored}}}{2\,\mathrm{MSE}_{\mathrm{ignored}}}, \tag{39}$$

where MSE_k is the derivative error with respect to the full perturbed vector field, MSE_0 is the error with respect to the perturbation-free vector field, and $\mathrm{MSE}_{\mathrm{ignored}}$ is the error incurred by omitting the oscillatory perturbation entirely. When the error in the smooth component is orthogonal to the perturbation, λ equals the fraction of the perturbation force recovered by the learned model.

B Ablations

We perform a sequence of ablations to identify which aspects of LM-HNN are responsible for its improvement and under what conditions the proposed representation is most effective. We first isolate the contributions of kinetic–potential decomposition, locality, and weight sharing. We then study how the receptive field should be chosen, how the kinetic representation should adapt to different coupling structures, and how the advantage of LM-HNN changes with system size and interaction complexity. Finally, we compare the proposed physical decomposition with generic architectural remedies for spectral bias. Throughout this appendix the tables report the derivative MSE on validation states, the normalized final-step rollout error of Eq. (17), and the peak error of Eq. (18), which measures how accurately the amplitude of the predicted motion is preserved.

Contributions of decomposition, locality, and weight sharing. We first ablate the three main structural components of LM-HNN on the FPUT and Toda chains. Table 5 shows that the gain does not come from a single architectural modification, but from combining physically meaningful decomposition with local and shared representations.

Removing the kinetic–potential split and replacing the separate energy branches with a single local network over joint position–momentum neighborhoods substantially increases both derivative and rollout errors. The degradation is larger when locality is removed entirely and the kinetic and potential energies are represented by global networks over the full state. These results indicate that

Table 5: **Ablation of the LM-HNN structure.** We perform ablation on the FPUT and FPUT-Toda chains ($N = 8$). Removing locality causes the largest degradation, followed by removing the kinetic–potential decomposition. Removing weight sharing increases the parameter count substantially without providing a consistent accuracy gain. Single-radius variants are included here for completeness and studied in more detail in Table 6.

FPUT				
Variant	Params	Deriv. MSE	Rollout MSE	Peak err.
LM-HNN, $\mathcal{R} = \{0, 1, 2\}$ (full)	26,892	$(5.6 \pm 0.9) \times 10^{-4}$	$(4.9 \pm 0.6) \times 10^{-3}$	$0.83 \pm 0.06\%$
single radius $\mathcal{R} = \{0\}$	8,708	$\mathbf{(2.6 \pm 0.4) \times 10^{-4}}$	$\mathbf{(2.5 \pm 0.8) \times 10^{-4}}$	$\mathbf{0.17 \pm 0.04\%}$
single radius $\mathcal{R} = \{2\}$	9,220	$(1.5 \pm 0.2) \times 10^{-3}$	$(2.0 \pm 0.8) \times 10^{-2}$	$1.65 \pm 0.30\%$
no kinetic–potential split	26,736	$(2.1 \pm 0.2) \times 10^{-3}$	$(3.8 \pm 0.6) \times 10^{-2}$	$1.58 \pm 0.14\%$
no locality (global T, V networks)	26,932	$(4.0 \pm 0.1) \times 10^{-3}$	$(5.5 \pm 0.5) \times 10^{-2}$	$3.00 \pm 0.14\%$
no weight sharing	228,537	$(4.0 \pm 0.2) \times 10^{-4}$	$(3.5 \pm 0.8) \times 10^{-3}$	$0.59 \pm 0.05\%$
FPUT-Toda				
Variant	Params	Deriv. MSE	Rollout MSE	Peak err.
LM-HNN, $\mathcal{R} = \{0, 1, 2\}$ (full)	26,892	$(4.5 \pm 0.4) \times 10^{-4}$	$\mathbf{(6.7 \pm 4.7) \times 10^{-3}}$	$\mathbf{0.85 \pm 0.17\%}$
no kinetic–potential split	26,736	$(3.2 \pm 0.4) \times 10^{-3}$	$(7.1 \pm 1.2) \times 10^{-2}$	$2.67 \pm 0.19\%$
no locality (global T, V networks)	26,932	$(1.6 \pm 0.1) \times 10^{-2}$	$(3.8 \pm 1.0) \times 10^{-1}$	$7.19 \pm 0.67\%$
no weight sharing	228,537	$\mathbf{(4.3 \pm 0.8) \times 10^{-4}}$	$(1.9 \pm 0.2) \times 10^{-2}$	$0.99 \pm 0.07\%$

kinetic–potential decomposition simplifies the learning problem, while locality provides the larger gain by reducing the dimensionality of the functions that each network must approximate.

Weight sharing plays a complementary role. Assigning independent networks to different sites and bonds increases the parameter count by nearly an order of magnitude without yielding a consistent improvement. Sharing the local interaction functions therefore acts not merely as a parameter-saving device, but as an appropriate inductive bias for systems built from repeated interactions.

Choice of receptive field. The previous ablation establishes the importance of locality, but a local model still needs a choice of interaction neighborhood. Table 6 compares single radii against the multiscale set on two chains whose interaction ranges differ: the nearest-neighbor FPUT chain, and the two-scale chain of Eq. (30), which adds a second-neighbor coupling to the same bonds.

On the nearest-neighbor chain, the radius matched to the interaction, $\mathcal{R} = \{0\}$, is the most accurate choice, and a single radius that is too large loses accuracy: $\mathcal{R} = \{2\}$ raises the rollout error by roughly two orders of magnitude. On the two-scale chain the ranking reverses. A radius-0 window

Table 6: **Ablation of the receptive-field.** We perform ablation on two chains with different interaction ranges ($N = 8$, $\alpha = 0.25$, $\beta = 0.10$): the nearest-neighbor FPUT chain, and the two-scale chain of Eq. (30). The best single radius changes between the two systems, and the radius that wins on the nearest-neighbor chain, $\mathcal{R} = \{0\}$, is the one that cannot represent the second-neighbor interaction at all. The multiscale set $\mathcal{R} = \{0, 1, 2\}$ is close to the best choice on both chains and attains the lowest derivative error of any model on the two-scale chain.

System	Radii $\mathcal{R}$	Params	Deriv. MSE	Rollout MSE	Peak err.
Nearest-neighbor chain	$\{0\}$	8,708	$\mathbf{(2.6 \pm 0.4) \times 10^{-4}}$	$\mathbf{(2.5 \pm 0.7) \times 10^{-4}}$	$\mathbf{0.17 \pm 0.03\%}$
	$\{1\}$	8,964	$(1.0 \pm 0.2) \times 10^{-3}$	$(4.1 \pm 1.8) \times 10^{-3}$	$0.72 \pm 0.20\%$
	$\{2\}$	9,220	$(1.5 \pm 0.1) \times 10^{-3}$	$(2.0 \pm 0.6) \times 10^{-2}$	$1.65 \pm 0.24\%$
	$\{0, 1, 2\}$	26,892	$(5.6 \pm 0.7) \times 10^{-4}$	$(4.9 \pm 0.5) \times 10^{-3}$	$0.83 \pm 0.05\%$
	HNN	26,971	$(6.0 \pm 0.1) \times 10^{-3}$	$(2.0 \pm 0.1) \times 10^{-1}$	$4.86 \pm 0.11\%$
Two-scale chain	$\{0\}$	8,708	$(5.0 \pm 0.0) \times 10^{-2}$	$(8.7 \pm 0.0) \times 10^{-1}$	$9.56 \pm 0.01\%$
	$\{1\}$	8,964	$(6.3 \pm 0.6) \times 10^{-4}$	$\mathbf{(9.1 \pm 2.2) \times 10^{-4}}$	$\mathbf{0.43 \pm 0.04\%}$
	$\{2\}$	9,220	$(1.1 \pm 0.2) \times 10^{-3}$	$(5.4 \pm 2.0) \times 10^{-3}$	$0.93 \pm 0.21\%$
	$\{0, 1, 2\}$	26,892	$\mathbf{(4.4 \pm 0.3) \times 10^{-4}}$	$(2.1 \pm 0.3) \times 10^{-3}$	$0.54 \pm 0.03\%$
	HNN	26,971	$(4.9 \pm 0.1) \times 10^{-3}$	$(1.3 \pm 0.1) \times 10^{-1}$	$4.20 \pm 0.10\%$

Table 7: **Kinetic coupling and kinetic-head design for $N = 8$.** Systems include $A = I$ for the standard FPUT chain; tridiagonal coupling $A = I + 0.3\,\mathrm{tridiag}(1)$, which introduces $p_i p_{i+1}$ terms; and global coupling $A = I + \frac{1}{N}\mathbf{1}\mathbf{1}^\top$. We compare the diagonal quadratic form of Eq. (10), the local-summation form of Eq. (12), and a dense Cholesky form $\frac{1}{2}\boldsymbol{p}^\top L_\theta(\boldsymbol{q})L_\theta(\boldsymbol{q})^\top\boldsymbol{p}$. The results show that the kinetic representation should follow the coupling structure: diagonal for uncoupled momenta, local for finite-range coupling, and dense for global coupling.

Kinetic coupling	Kinetic head	Deriv. MSE	Rollout MSE	Peak err.
Standard FPUT ($A = I$)	quadratic, diagonal	$\mathbf{(5.6\pm0.9)\times10^{-4}}$	$\mathbf{(4.9\pm0.6)\times10^{-3}}$	$\mathbf{0.83\pm0.06\%}$
	local sum, $\mathcal{R} = \{0\}$	$(6.7\pm0.8)\times10^{-4}$	$(1.0\pm0.2)\times10^{-2}$	$1.04\pm0.12\%$
	local sum, $\mathcal{R} = \{0,1,2\}$	$(6.5\pm0.9)\times10^{-4}$	$(1.1\pm0.3)\times10^{-2}$	$0.95\pm0.10\%$
Tridiagonal coupling	quadratic, diagonal	$(9.0\pm0.0)\times10^{-2}$	$(1.6\pm0.0)\times10^{0}$	$16.11\pm0.01\%$
	local sum, $\mathcal{R} = \{0\}$	$(9.0\pm0.0)\times10^{-2}$	$(1.6\pm0.0)\times10^{0}$	$16.23\pm0.04\%$
	local sum, $\mathcal{R} = \{0,1,2\}$	$\mathbf{(6.6\pm0.8)\times10^{-4}}$	$\mathbf{(1.5\pm0.5)\times10^{-2}}$	$\mathbf{1.42\pm0.26\%}$
Global coupling	local sum, $\mathcal{R} = \{0,1,2\}$	$(8.1\pm0.3)\times10^{-3}$	$(1.7\pm0.1)\times10^{-1}$	$5.50\pm0.08\%$
	dense Cholesky	$\mathbf{(9.7\pm1.3)\times10^{-4}}$	$\mathbf{(2.9\pm0.3)\times10^{-2}}$	$\mathbf{2.04\pm0.07\%}$

holds one bond strain and therefore cannot represent the second-neighbor term at all; its derivative error stalls near 5×10^{-2} across seeds and its rollout error exceeds that of the conventional HNN. The best single radius there is $\mathcal{R} = \{1\}$, the smallest window containing both interaction ranges.

The multiscale set is close to the best choice on both chains without being told which one it is. On the two-scale chain it reaches the lowest derivative error of any model in the table, below every single radius: the sharp nearest-neighbor potential is represented by the one-dimensional radius-0 branch while the smooth second-neighbor term is carried by the wider branches, so each component is learned in the coordinates it actually depends on. Its rollout error there stays within a factor of about two of the matched radius, and on the nearest-neighbor chain it remains an order of magnitude below the conventional HNN. Since the interaction range of a physical system is generally unknown in advance, and since the radius that wins on one chain is the one that fails on the other, the full model uses $\mathcal{R} = \{0, 1, 2\}$ throughout.

Kinetic representation and coupling structure Locality is beneficial only when it reflects the structure of the underlying Hamiltonian. To test this explicitly, Table 7 varies the momentum coupling while changing the kinetic head accordingly. For the standard FPUT system with diagonal kinetic energy, the diagonal quadratic form in Eq. (10) is both sufficient and most accurate. Once nearest-neighbor momentum coupling is introduced, a diagonal head can no longer represent the correct kinetic energy, while the local-summation form of Eq. (12) with $\mathcal{R} = \{0, 1, 2\}$ recovers the coupled dynamics accurately.

For genuinely global momentum coupling, however, the local-summation representation is no longer sufficient and a dense Cholesky parameterization performs substantially better. Thus, LM-HNN does not assume that every Hamiltonian is local. Rather, its advantage arises when the representation is matched to the interaction structure present in the system.

Scaling with system size. A central benefit of the shared local representation is that its parameter count does not grow with the number of repeated interactions. We test whether this structural advantage also translates into better accuracy as the chain becomes longer. Table 8 compares LM-HNN with parameter-matched conventional HNNs from $N = 4$ to $N = 16$.

The conventional HNN becomes progressively harder to train as the dimension of the global Hamiltonian increases: its derivative error rises by roughly an order of magnitude and its rollout error grows from approximately 3×10^{-2} to 5×10^{-1}. In contrast, LM-HNN uses the same 26,892 parameters for every chain length and its rollout error remains in the 10^{-3} range throughout the sweep. The widening gap supports the interpretation that LM-HNN learns a reusable interaction law whose complexity is largely independent of how many times that interaction is repeated.

Sensitivity to interaction nonlinearity. We next study whether the advantage of the local representation changes as the interaction law becomes more difficult to approximate. Table 9 varies the

Table 8: **Scaling with chain length N at matched parameter count.** We scale the chain length on the FPUT system with $\alpha = 0.25$ and $\beta = 0.10$. From $N = 4$ to $N = 16$, the rollout error of LM-HNN remains between approximately 9×10^{-4} and 6×10^{-3}, while that of the conventional HNN grows from 3×10^{-2} to 5×10^{-1}.

	Params		Derivative MSE		Rollout MSE	
N	HNN	LM-HNN	HNN	LM-HNN	HNN	LM-HNN
4	27,031	26,892	$(9.7\pm0.8)\times10^{-4}$	$\mathbf{(1.5\pm0.2)\times10^{-4}}$	$(3.2\pm0.5)\times10^{-2}$	$\mathbf{(9.4\pm0.6)\times10^{-4}}$
6	27,005	26,892	$(3.3\pm0.1)\times10^{-3}$	$\mathbf{(5.0\pm0.4)\times10^{-4}}$	$(9.6\pm2.6)\times10^{-2}$	$\mathbf{(2.6\pm0.4)\times10^{-3}}$
8	26,971	26,892	$(6.0\pm0.1)\times10^{-3}$	$\mathbf{(5.6\pm0.9)\times10^{-4}}$	$(2.0\pm0.1)\times10^{-1}$	$\mathbf{(4.9\pm0.6)\times10^{-3}}$
10	26,929	26,892	$(8.8\pm1.1)\times10^{-3}$	$\mathbf{(4.3\pm0.8)\times10^{-4}}$	$(2.3\pm0.2)\times10^{-1}$	$\mathbf{(5.1\pm1.6)\times10^{-3}}$
12	26,879	26,892	$(1.1\pm0.0)\times10^{-2}$	$\mathbf{(7.0\pm1.7)\times10^{-4}}$	$(5.2\pm0.4)\times10^{-1}$	$\mathbf{(4.6\pm1.1)\times10^{-3}}$
16	26,755	26,892	$(1.7\pm0.1)\times10^{-2}$	$\mathbf{(7.6\pm1.6)\times10^{-4}}$	$(5.2\pm0.4)\times10^{-1}$	$\mathbf{(6.0\pm1.4)\times10^{-3}}$

nonlinear coefficients (α, β) of the FPUT potential while keeping the model comparison otherwise fixed.

As the interaction becomes more nonlinear, the conventional HNN degrades substantially: its peak error increases from 4.9% to as high as 9.7%, and its derivative and rollout errors increase correspondingly. LM-HNN remains much more stable, with peak errors between approximately 0.8% and 1.2% across the sweep. This result is consistent with the proposed representation mechanism. Stronger nonlinearities make the full global energy landscape increasingly difficult to approximate, whereas LM-HNN continues to learn the same low-dimensional bond-level interaction law.

Comparison with generic spectral-bias interventions. Because one motivation for LM-HNN is to improve the representation of fine-scale Hamiltonian structure, we finally compare it with generic methods designed to alleviate spectral bias in coordinate networks. Table 1 applies random Fourier features and sinusoidal activations to the conventional HNN at matched parameter count.

For this comparison the derivative error is measured on held-out trajectories, generated by the same energy-shell sampler with a data seed never used in training, so that it reports the accuracy of the learned vector field on new trajectories rather than along the training ones; the rollout error starts from unseen initial conditions as everywhere else.

Random Fourier features raise both errors on both systems. Sinusoidal activations fit the training trajectories closely, but the field they learn between those trajectories is inaccurate, so their derivative error on held-out trajectories is large and their rollouts from unseen initial conditions are more than an order of magnitude worse than those of the original HNN. LM-HNN has the lowest derivative and rollout error in every column of Table 1, on the chain by a factor of about fifty over the next model.

This comparison indicates that increasing access to high-frequency functions is not sufficient by itself. For Hamiltonian learning, the fine-scale structure must also be organized in a way that is consistent with the physical decomposition and interaction structure of the energy so that improvements in local fitting translate into accurate dynamics away from the observed trajectories.

Table 9: **Parameter matched comparison of LM-HNN and HNN.** FPUT chain ($N = 8$) across elastic parameters (α, β), comparing the parameter-matched HNN and LM-HNN. As the interaction becomes more nonlinear, the conventional HNN degrades substantially, while LM-HNN remains comparatively stable; its peak error stays between approximately 0.8% and 1.2% across the sweep.

	Deriv. MSE		Rollout MSE		Peak err. (%)	
(α, β)	HNN	LM-HNN	HNN	LM-HNN	HNN	LM-HNN
(0.25, 0.10)	$(6.0\pm0.1)\times10^{-3}$	$\mathbf{(5.6\pm0.9)\times10^{-4}}$	$(2.0\pm0.1)\times10^{-1}$	$\mathbf{(4.9\pm0.6)\times10^{-3}}$	4.86 ± 0.13	$\mathbf{0.83\pm0.06}$
(0.50, 0.10)	$(6.6\pm0.6)\times10^{-3}$	$\mathbf{(8.7\pm2.5)\times10^{-4}}$	$(3.7\pm0.7)\times10^{-1}$	$\mathbf{(1.8\pm1.0)\times10^{-2}}$	6.03 ± 0.54	$\mathbf{0.96\pm0.10}$
(1.00, 0.30)	$(1.8\pm0.0)\times10^{-2}$	$\mathbf{(1.2\pm0.2)\times10^{-3}}$	$(8.2\pm0.5)\times10^{-1}$	$\mathbf{(2.2\pm0.8)\times10^{-2}}$	9.68 ± 0.88	$\mathbf{1.24\pm0.12}$
(0.25, 0.30)	$(2.3\pm0.2)\times10^{-2}$	$\mathbf{(9.0\pm1.6)\times10^{-4}}$	$(2.9\pm0.5)\times10^{-1}$	$\mathbf{(6.1\pm1.2)\times10^{-3}}$	5.88 ± 0.35	$\mathbf{0.93\pm0.06}$
(0.00, 0.50)	$(4.6\pm0.4)\times10^{-2}$	$\mathbf{(7.5\pm0.8)\times10^{-4}}$	$(4.6\pm0.4)\times10^{-1}$	$\mathbf{(3.9\pm1.2)\times10^{-3}}$	8.29 ± 0.38	$\mathbf{0.76\pm0.11}$

Ablation summary. Taken together, the ablations support four conclusions. First, locality provides the largest accuracy gain by reducing the dimensionality of the Hamiltonian representation, while kinetic–potential decomposition further simplifies the functions that must be learned. Second, weight sharing makes this local representation scalable by allowing the same interaction law to be reused across repeated structures without increasing the parameter count. Third, the local representation is most effective when its receptive field and kinetic head match the coupling structure of the underlying Hamiltonian. Fourth, when the interaction range is not known in advance, combining multiple neighborhood scales provides a robust alternative to selecting a single radius, remaining competitive with the best matched single-scale choice across different interaction structures.

These experiments also distinguish LM-HNN from generic approaches that increase the spectral capacity of a coordinate network. Its advantage grows with both system size and interaction nonlinearity, and cannot be reproduced simply by adding Fourier features or sinusoidal activations. Together, the results support the central claim of this work: accurate Hamiltonian learning requires not only sufficient capacity to represent fine-scale variation, but also a representation that organizes that variation according to the physical structure of the underlying energy.

C Additional Results

This section provides additional evidence supporting the main experimental results. We first examine the Hamiltonian representation and its spectral error, then study the controlled perturbed-FPUT test, report accuracy across all systems, and finally relate Hamiltonian approximation error to rollout performance and generalization beyond the training energy range. We conclude with detailed results on the fast–slow oscillator and a preliminary extension to field-valued dynamics.

Table 10: **Spectral diagnostics on validation states.** RMS Hamiltonian residual σ_H, effective residual wavenumber k_{eff}, and RMS vector-field error $\|\nabla\delta H\|_{\text{rms}}$ for HNN, FS-HNN, and LM-HNN. LM-HNN has the smallest Hamiltonian residual and vector-field error on every system. Its remaining residual can have a larger effective wavenumber because the lower-frequency error has already been substantially reduced. The diagnostics follow the Fourier-domain analysis of spectral bias in Rahaman et al. (2019) and the frequency principle of Xu et al. (2020), which examine the residual of a coordinate network by wavenumber over its input domain and across training; k_{eff} is defined in Equation 38 from Parseval's identity, and the link between the wavenumber content of the residual and the vector-field error follows the kernel analysis of Wang et al. (2022).

System	Model	σ_H	k_{eff}	$\|\nabla\delta H\|_{\text{rms}}$
Pendulum	HNN	$(4.5\pm0.8)\times10^{-4}$	4.1	$(1.9\pm0.4)\times10^{-3}$
	FS-HNN	$(1.3\pm0.9)\times10^{-3}$	2.9	$(3.0\pm0.6)\times10^{-3}$
	LM-HNN	$\mathbf{(2.7\pm0.9)\times10^{-5}}$	**4.2**	$\mathbf{(1.1\pm0.5)\times10^{-4}}$
Double Pendulum	HNN	$(1.5\pm0.2)\times10^{-2}$	4.8	$(7.1\pm0.7)\times10^{-2}$
	FS-HNN	$(8.4\pm1.5)\times10^{-2}$	3.6	$(2.9\pm0.1)\times10^{-1}$
	LM-HNN	$\mathbf{(3.8\pm1.5)\times10^{-4}}$	**9.1**	$\mathbf{(3.4\pm1.2)\times10^{-3}}$
FPUT	HNN	$(1.7\pm0.1)\times10^{-1}$	1.8	$(3.1\pm0.0)\times10^{-1}$
	FS-HNN	$(1.7\pm0.7)\times10^{-1}$	1.3	$(2.2\pm1.1)\times10^{-1}$
	LM-HNN	$\mathbf{(4.4\pm0.0)\times10^{-2}}$	**2.1**	$\mathbf{(9.4\pm0.7)\times10^{-2}}$
FPUT-Toda	HNN	$(2.7\pm0.1)\times10^{-1}$	1.8	$(4.9\pm0.1)\times10^{-1}$
	FS-HNN	$(1.5\pm0.1)\times10^{-1}$	1.4	$(2.1\pm0.0)\times10^{-1}$
	LM-HNN	$\mathbf{(3.0\pm0.4)\times10^{-2}}$	**2.9**	$\mathbf{(8.4\pm0.4)\times10^{-2}}$
FPUT-Morse	HNN	$(2.7\pm0.3)\times10^{-1}$	1.2	$(3.2\pm0.1)\times10^{-1}$
	FS-HNN	$(2.1\pm0.3)\times10^{-1}$	1.6	$(3.4\pm0.3)\times10^{-1}$
	LM-HNN	$\mathbf{(3.2\pm0.5)\times10^{-2}}$	**1.9**	$\mathbf{(6.1\pm0.8)\times10^{-2}}$
Fast–slow	HNN	$(2.7\pm0.1)\times10^{-4}$	5.9	$(1.6\pm0.0)\times10^{-3}$
	FS-HNN	$(1.2\pm0.1)\times10^{-3}$	5.2	$(6.4\pm0.4)\times10^{-3}$
	LM-HNN	$\mathbf{(1.7\pm0.2)\times10^{-4}}$	**6.9**	$\mathbf{(1.1\pm0.1)\times10^{-3}}$

C.1 Spectral Diagnostics of the Learned Hamiltonian

Section 5.2 tracks the residual spectrum of the learned Hamiltonian over training on the FPUT chain. This subsection applies the same view to all six systems at the end of training and connects the residual to the vector-field error that drives the rollouts. Table 10 reports the RMS Hamiltonian residual σ_H, the effective residual wavenumber k_{eff} of Eq. (38), and the RMS vector-field error $\|\nabla\delta H\|_{\text{rms}}$ on the validation states of Appendix A.3 for HNN, FS-HNN, and LM-HNN. The diagnostics follow the Fourier-domain analysis of spectral bias in Rahaman et al. (2019) and the frequency principle of Xu et al. (2020), which examine the residual of a coordinate network by wavenumber over its input domain, and the link between the wavenumber content of the residual and the vector-field error follows the kernel analysis of Wang et al. (2022). LM-HNN has the smallest Hamiltonian residual and the smallest vector-field error on every system, by more than an order of magnitude on the pendulum systems and by a factor of three to nine on the chains, with the narrowest margin on the fast–slow oscillator. The improvement observed in the rollouts is therefore already present in the learned energy and its gradient before any numerical integration is performed. FS-HNN, which decomposes the Hamiltonian by temporal frequency band, does not consistently reduce either quantity relative to the conventional HNN, which is consistent with the limitation lying in the phase-space wavenumber content of the energy rather than in the temporal spectrum of the trajectories.

The effective wavenumber behaves differently. LM-HNN matches or exceeds the k_{eff} of the conventional HNN on every system even though both its residual and its gradient error are smaller. This follows from the definition. Since k_{eff} is the ratio of the RMS gradient error to the RMS residual, removing the smooth part of the residual lowers the denominator by a larger fraction than the numerator and shifts the ratio upward, so the quantity reports where the remaining error is concentrated rather than how large it is. The residual that LM-HNN leaves behind sits at finer scales of the energy than the residual of the HNN, which is the expected outcome once the low-wavenumber modes that a coordinate network fits first have been resolved and only the slowest-learned components remain (Eq. (4)). Because the gradient error of LM-HNN stays below that of the HNN on every system, this finer residual does not translate into a larger vector-field error, and k_{eff} is read jointly with σ_H and $\|\nabla\delta H\|_{\text{rms}}$ as Appendix A.3 notes.

Figure 3 examines the residual directly along a representative probe line, obtained by displacing a single site of the chain while the remaining coordinates are held fixed, for the unperturbed FPUT chain and for the perturbed chain with $k = 4$. On the unperturbed chain the residual of the HNN is smooth but far from zero along the whole line, while the residual of LM-HNN stays close to zero.

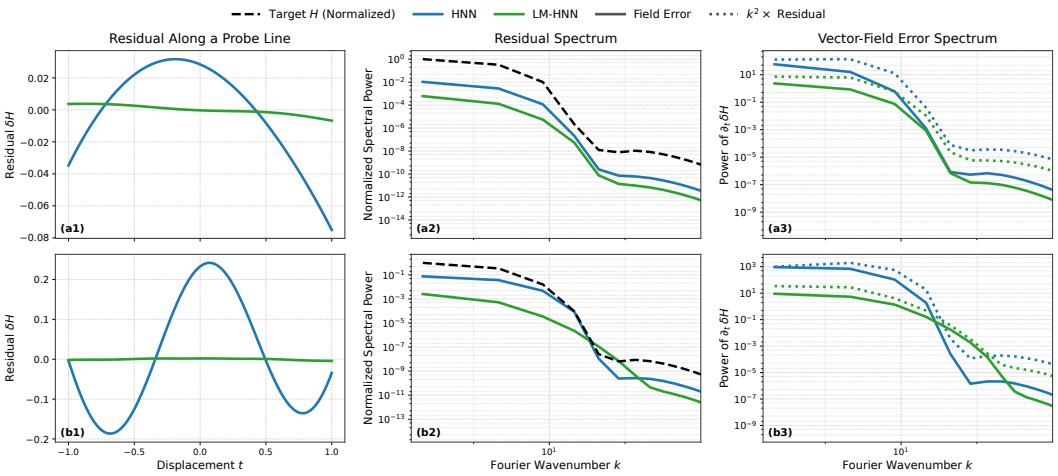

Figure 3: **Residual spectra along probe lines.** Left: the Hamiltonian residual δH along a line obtained by displacing one site, for HNN and LM-HNN on the standard **(a)** FPUT chain and the **(b)** perturbed FPUT chain with $k = 4$. Middle: power spectra of the target Hamiltonian and the residual of each model. Right: power spectrum of the corresponding vector-field error (solid) together with k^2 times the Hamiltonian-residual spectrum (dotted). Differentiation amplifies higher-wavenumber residuals, while LM-HNN reduces the residual over the spectral range occupied by the target.

The corresponding power spectra show that LM-HNN reduces the residual across the wavenumber band that carries the target Hamiltonian power rather than at one end of it. On the perturbed chain the residual of the HNN is itself an oscillation at the perturbation wavenumber, the signature of a component that the monolithic network has not resolved, and its residual spectrum concentrates its power near $k = 4$. LM-HNN removes this component and its residual remains flat along the line. The right column relates these residuals to the dynamics. The power spectrum of the directional derivative of the residual coincides with k^2 times the residual spectrum, which confirms along the probe lines the relation of Eq. (5) that underlies the effective wavenumber. Differentiation therefore weights the high-wavenumber part of the residual most heavily, and a residual that is small in energy but rapidly varying can dominate the vector-field error.

Taken together, the table and the probe-line spectra locate the advantage of LM-HNN at the level of the scalar Hamiltonian. The structured representation lowers the residual over the band occupied by the target, and because the vector field weights that band by k^2, the reduction appears in the gradient error before it appears in any rollout.

C.2 Controlled Spectral-Bias Study on the Perturbed FPUT Chain

The diagnostics of the previous subsection measure the residual that each model leaves on a fixed Hamiltonian. The perturbed FPUT chain of Section 5.2 instead varies the Hamiltonian itself, so that the wavenumber the model must resolve is prescribed rather than inferred. Adding the term $\frac{a}{k}(1 - \cos ks)$ of Eq. (29) to the bond potential inserts a force component $a \sin(ks)$ whose amplitude is fixed at $a = 0.5$ for every k, so the added structure carries the same dynamical weight at every wavenumber and differs only in how rapidly it varies with the bond strain. This subsection reports the quantities behind Figure 1(c,d), the derivative errors and recovery fractions of the two Hamiltonian models across the sweep in Table 11, and the rollout and peak errors of all four models in Tables 12 and 13.

Table 11 lists the derivative error of each model against the perturbed vector field, the error incurred by omitting the perturbation entirely, and the fraction of the perturbation force recovered. The omission error is nearly constant across k, as expected for a force of fixed amplitude, so it provides a common reference for how much of the added component each model has learned. Both models recover the $k = 1$ and $k = 2$ components with only a small loss, and the heat maps of Figure 1(c,d) show that both learn the lowest wavenumbers first. The two diverge from $k = 3$. The recovery of the conventional HNN falls steadily with k, and at $k = 6$ its derivative error approaches the omission

Table 11: **Spectral-bias sweep on the perturbed FPUT chain ($N = 8$).** We report the derivative and rollout errors of the HNN and LM-HNN against the perturbation wavenumber k, and the fraction of the perturbation force recovered, $1 - \mathrm{MSE}/\mathrm{MSE}_{\mathrm{ignored}}$. The HNN loses the perturbation between $k = 2$ and $k = 4$; LM-HNN recovers it up to $k = 6$.

Perturbation wavenumber k	0	1	2
HNN derivative MSE	$(6.0 \pm 0.1) \times 10^{-3}$	$(3.3 \pm 0.2) \times 10^{-3}$	$(1.1 \pm 0.1) \times 10^{-2}$
LM-HNN derivative MSE	$\mathbf{(5.6 \pm 0.9) \times 10^{-4}}$	$\mathbf{(4.9 \pm 1.2) \times 10^{-4}}$	$\mathbf{(2.1 \pm 0.5) \times 10^{-3}}$
HNN rollout MSE	$(2.0 \pm 0.1) \times 10^{-1}$	$(1.2 \pm 0.1) \times 10^{-1}$	$(2.0 \pm 0.2) \times 10^{-1}$
LM-HNN rollout MSE	$\mathbf{(4.9 \pm 0.6) \times 10^{-3}}$	$\mathbf{(5.0 \pm 1.1) \times 10^{-3}}$	$\mathbf{(1.2 \pm 0.3) \times 10^{-2}}$
Perturbation ignored (deriv. MSE)	–	1.1×10^{-1}	1.2×10^{-1}
HNN force recovered	–	97%	91%
LM-HNN force recovered	–	**100%**	**98%**
Perturbation wavenumber k	3	4	6
HNN derivative MSE	$(4.5 \pm 0.1) \times 10^{-2}$	$(8.6 \pm 0.1) \times 10^{-2}$	$(1.1 \pm 0.0) \times 10^{-1}$
LM-HNN derivative MSE	$\mathbf{(1.9 \pm 0.2) \times 10^{-4}}$	$\mathbf{(8.6 \pm 2.9) \times 10^{-4}}$	$\mathbf{(3.6 \pm 1.0) \times 10^{-3}}$
HNN rollout MSE	$(6.8 \pm 0.8) \times 10^{-1}$	$(1.0 \pm 0.0) \times 10^{0}$	$(1.3 \pm 0.1) \times 10^{0}$
LM-HNN rollout MSE	$\mathbf{(2.4 \pm 0.6) \times 10^{-3}}$	$\mathbf{(1.4 \pm 0.7) \times 10^{-2}}$	$\mathbf{(1.8 \pm 0.3) \times 10^{-1}}$
Perturbation ignored (deriv. MSE)	1.2×10^{-1}	1.2×10^{-1}	1.2×10^{-1}
HNN force recovered	62%	26%	10%
LM-HNN force recovered	**100%**	**99%**	**97%**

Table 12: **Rollout MSE on the perturbed FPUT ($N = 8$) with increasing k.** LM-HNN achieves the lowest error at every tested wavenumber. At $k = 4$, its rollout error remains below that of every competing baseline on the plain FPUT chain ($k = 0$), demonstrating a substantially extended effective spectral range.

k	HNN	SympNet	FS-HNN	LM-HNN (ours)
0	$(2.0\pm0.1)\times10^{-1}$	$(6.2\pm0.6)\times10^{-2}$	$(1.2\pm0.4)\times10^{-2}$	$\mathbf{(4.9\pm0.6)\times10^{-3}}$
2	$(2.0\pm0.2)\times10^{-1}$	$(6.3\pm1.0)\times10^{-2}$	$(1.7\pm0.4)\times10^{-1}$	$\mathbf{(1.2\pm0.3)\times10^{-2}}$
4	$(1.0\pm0.0)\times10^{0}$	$(1.2\pm0.4)\times10^{-1}$	$(6.7\pm2.5)\times10^{-1}$	$\mathbf{(1.4\pm0.7)\times10^{-2}}$
8	$(8.5\pm0.4)\times10^{-1}$	$(8.3\pm0.2)\times10^{-1}$	$(9.8\pm0.3)\times10^{-1}$	$\mathbf{(5.0\pm0.1)\times10^{-1}}$
16	$(4.2\pm0.2)\times10^{-1}$	$(3.4\pm0.5)\times10^{-1}$	$(4.9\pm0.4)\times10^{-1}$	$\mathbf{(1.8\pm0.0)\times10^{-1}}$

error, which means that within the training budget it has stopped resolving the component at all. LM-HNN recovers the perturbation nearly completely through $k = 6$, with derivative errors that remain well below those of the HNN at every wavenumber, and its recovery declines only beyond that range. The ordering in which wavenumbers are learned is the same in both models, which is the ordering spectral bias predicts, and the local representation moves through it fast enough that the effective spectral limit shifts to a higher wavenumber within the same budget.

The same transition governs long-horizon prediction. Table 12 reports the rollout error of all four models as k increases, and LM-HNN has the lowest error at every tested wavenumber. Through $k = 4$ its rollout error stays within a factor of three of its unperturbed value, which is below what the HNN and SympNet reach on the unperturbed chain. FS-HNN, which is accurate on the plain chain, loses an order of magnitude already at $k = 2$. The peak error of Table 13 tells the same story for the amplitude of the motion. LM-HNN stays below 1.2% through $k = 4$, whereas the peak errors of the baselines rise to several percent and reach roughly 10% at $k = 4$ for the HNN and FS-HNN, so the extrema of the motion, which are fixed by the level set of the Hamiltonian through the initial state, are preserved only by the model that resolves the added structure of the potential. At $k = 8$ and $k = 16$ the perturbation lies beyond the effective spectral limit of every model, LM-HNN included, and all errors rise well above their levels at $k \leq 4$, with LM-HNN remaining the lowest. The errors of all models are larger at $k = 8$ than at $k = 16$, which is consistent with the energy amplitude $2a/k$ of the perturbation shrinking with k, so that once the component is unresolved its net effect on the trajectory diminishes.

Across recovery fraction, derivative error, rollout error, and peak error, the sweep locates the same cutoff. The monolithic representation loses the added component between $k = 2$ and $k = 4$, while the local representation carries it to $k = 6$ and degrades only beyond. The location of the cutoff depends on the training budget, as the heat maps show, but where it falls for a given budget is set by the representation, since both models share the data, the budget, and the parameter count.

C.3 Accuracy Across Systems

We supplement the cross-system comparison in Section 5.4 with model sizes and representative rollouts from unseen initial conditions. Table 14 reports the parameter counts of HNN, SympNet, FS-HNN, and LM-HNN across the six benchmark systems. The models use comparable parameter budgets within each system, showing that LM-HNN's improvements in derivative and rollout accuracy are achieved with a similar number of trainable parameters.

Table 13: **Peak error on the perturbed FPUT ($N = 8$) with variable perturbation wavenumber k.** LM-HNN achieves the lowest peak error at every tested wavenumber and remains below 1.2% for $k \leq 4$.

k	HNN	SympNet	FS-HNN	LM-HNN
0	$4.86\pm0.13\%$	$2.12\pm0.11\%$	$2.99\pm0.85\%$	$\mathbf{0.83\pm0.06\%}$
2	$5.16\pm0.56\%$	$1.55\pm0.10\%$	$4.08\pm1.18\%$	$\mathbf{1.16\pm0.12\%}$
4	$11.37\pm0.27\%$	$3.16\pm0.87\%$	$9.22\pm1.19\%$	$\mathbf{0.69\pm0.20\%}$
8	$9.99\pm0.53\%$	$9.91\pm0.83\%$	$10.82\pm0.37\%$	$\mathbf{7.16\pm0.13\%}$
16	$7.58\pm0.48\%$	$6.85\pm0.59\%$	$8.06\pm0.61\%$	$\mathbf{4.77\pm0.05\%}$

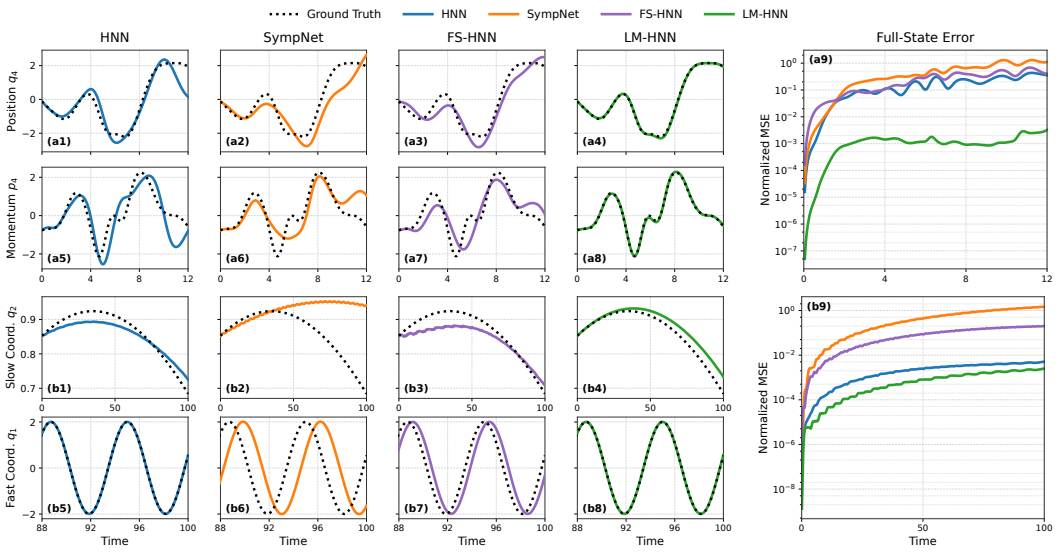

Figure 4: **Rollouts from unseen initial conditions.** **(a)** FPUT chain ($N = 8$), one position coordinate, one momentum coordinate, and the normalized full-state error over 240 steps. **(b)** fast–slow oscillator ($\varepsilon = 0.01$), the slow coordinate q_2 over the full horizon, the fast coordinate q_1 over the last two fast periods, and the normalized full-state error over 1000 steps. LM-HNN stays on the reference trajectory of the chain while the baselines accumulate phase and amplitude errors; on the fast–slow oscillator it has the lowest error over the whole horizon, following the slow drift most closely and keeping the fast phase, whereas the HNN drifts below the slow orbit in the middle of the horizon and SympNet and FS-HNN lose the fast phase.

Figure 4 (top) illustrates the evolution of prediction errors on the FPUT chain with $N = 8$. The trajectory spans 240 steps, extending beyond the 120-step horizon used in the quantitative comparison. LM-HNN closely follows the reference position and momentum coordinates, while the baselines develop visible phase and amplitude errors. Its normalized full-state error also remains substantially lower over the illustrated horizon. These trajectories show that the improvement extends to both the timing and extent of the predicted motion.

Figure 4 (bottom) examines the fast–slow oscillator with $\varepsilon = 0.01$ over 1000 steps. The slow coordinate q_2 shows the gradual evolution of the orbit, while the final two periods of q_1 reveal accumulated phase differences in the fast component. LM-HNN captures the main slow variation and remains closely aligned with the reference fast oscillation. HNN deviates from the slow orbit around the middle of the horizon, while SympNet and FS-HNN exhibit visible phase shifts in the fast component. The corresponding full-state error remains lower for LM-HNN over the rollout.

C.4 Hamiltonian Error and Rollout Generalization

We examine how the unseen-energy results in Section 5.3 relate to the accuracy of the learned Hamiltonian within the training range. Table 15 combines Hamiltonian and spectral diagnostics on

Table 14: **Parameter counts of LM-HNN and baselines.** The models are chosen to use comparable parameter budgets so that the accuracy comparison primarily reflects differences in representation rather than raw model capacity.

System	HNN	SympNet	FS-HNN	LM-HNN (ours)
Pendulum	8,828	8,832	8,072	8,834
Double pendulum	17,803	17,920	17,468	17,798
FPUT	26,971	26,880	26,468	26,892
FPUT-Toda	26,971	26,880	26,468	26,892
FPUT-Morse	26,971	26,880	26,468	26,892
Fast–slow	22,041	22,016	21,488	21,957

Table 15: **Hamiltonian and spectral diagnostics and rollout errors from unseen initial conditions inside and above the training energy range.** LM-HNN has smaller Hamiltonian residual, vector-field error, and rollout error across the tested systems. On several systems, much of its remaining residual power lies beyond the k_{99} bandwidth containing 99% of the target Hamiltonian power.

System	Model	σ_H	$\|\nabla\delta H\|_{\mathrm{rms}}$	k_{eff}
Pendulum	HNN	$(4.5\pm0.8)\times10^{-4}$	$(1.9\pm0.4)\times10^{-3}$	4.1 ± 0.1
	LM-HNN	$\mathbf{(2.7\pm0.9)\times10^{-5}}$	$\mathbf{(1.1\pm0.5)\times10^{-4}}$	4.2 ± 0.3
Double pendulum	HNN	$(1.5\pm0.2)\times10^{-2}$	$(7.1\pm0.7)\times10^{-2}$	4.8 ± 0.4
	LM-HNN	$\mathbf{(3.8\pm1.5)\times10^{-4}}$	$\mathbf{(3.4\pm1.2)\times10^{-3}}$	9.1 ± 0.4
FPUT (N=8)	HNN	$(1.7\pm0.1)\times10^{-1}$	$(3.1\pm0.0)\times10^{-1}$	1.8 ± 0.1
	LM-HNN	$\mathbf{(4.4\pm0.0)\times10^{-2}}$	$\mathbf{(9.4\pm0.7)\times10^{-2}}$	2.1 ± 0.2
Perturbed FPUT (k=16)	HNN	$(2.4\pm0.2)\times10^{-1}$	$(1.5\pm0.0)\times10^{0}$	6.3 ± 0.4
	LM-HNN	$\mathbf{(1.2\pm0.1)\times10^{-1}}$	$\mathbf{(1.4\pm0.0)\times10^{0}}$	12.6 ± 0.8
FPUT-Toda	HNN	$(2.7\pm0.1)\times10^{-1}$	$(4.9\pm0.1)\times10^{-1}$	1.8 ± 0.1
	LM-HNN	$\mathbf{(3.0\pm0.4)\times10^{-2}}$	$\mathbf{(8.4\pm0.4)\times10^{-2}}$	2.9 ± 0.3
FPUT-Morse	HNN	$(2.7\pm0.3)\times10^{-1}$	$(3.2\pm0.1)\times10^{-1}$	1.2 ± 0.1
	LM-HNN	$\mathbf{(3.2\pm0.5)\times10^{-2}}$	$\mathbf{(6.1\pm0.8)\times10^{-2}}$	1.9 ± 0.2

System	Model	δH power beyond k_{99}	Rollout (in range)	Rollout (above range)
Pendulum	HNN	$12.62\pm0.63\%$	$(2.7\pm0.6)\times10^{-5}$	$(1.3\pm0.2)\times10^{0}$
	LM-HNN	$87.14\pm7.02\%$	$\mathbf{(4.1\pm2.5)\times10^{-8}}$	$\mathbf{(2.7\pm2.4)\times10^{-8}}$
Double pendulum	HNN	$2.96\pm0.33\%$	$(4.3\pm5.0)\times10^{-2}$	$(1.7\pm0.6)\times10^{0}$
	LM-HNN	$87.97\pm7.35\%$	$\mathbf{(3.2\pm3.2)\times10^{-6}}$	$\mathbf{(8.1\pm5.3)\times10^{-6}}$
FPUT (N=8)	HNN	$0.85\pm0.02\%$	$(2.0\pm0.1)\times10^{-1}$	$(1.8\pm0.1)\times10^{0}$
	LM-HNN	$0.78\pm0.04\%$	$\mathbf{(4.9\pm0.6)\times10^{-3}}$	$\mathbf{(6.5\pm1.1)\times10^{-2}}$
Perturbed FPUT (k=16)	HNN	$22.11\pm0.99\%$	$(4.2\pm0.2)\times10^{-1}$	$(1.6\pm0.1)\times10^{0}$
	LM-HNN	$85.04\pm1.20\%$	$\mathbf{(1.8\pm0.0)\times10^{-1}}$	$\mathbf{(10.0\pm1.1)\times10^{-2}}$
FPUT-Toda	HNN	$0.76\pm0.01\%$	$(7.0\pm0.9)\times10^{-1}$	$(1.7\pm0.2)\times10^{0}$
	LM-HNN	$0.78\pm0.11\%$	$\mathbf{(6.7\pm4.7)\times10^{-3}}$	$\mathbf{(1.3\pm0.4)\times10^{-2}}$
FPUT-Morse	HNN	$0.86\pm0.01\%$	$(5.1\pm0.5)\times10^{-1}$	$(1.2\pm0.1)\times10^{0}$
	LM-HNN	$1.07\pm0.12\%$	$\mathbf{(1.5\pm0.4)\times10^{-3}}$	$\mathbf{(4.8\pm1.5)\times10^{-3}}$

validation states with normalized rollout errors from unseen initial conditions inside and above the training-energy range. Following Appendix A.3, the energy residual $\delta H = H_\theta - H$ is mean-centered before computing σ_H, removing the additive offset that is undetermined by derivative supervision.

Across the reported systems, LM-HNN achieves a smaller Hamiltonian residual σ_H and lower RMS vector-field error $\|\nabla\delta H\|_{\mathrm{rms}}$ than HNN, accompanied by lower rollout errors in both evaluation regimes. The accuracy improvement is therefore already visible in the learned energy and its derivatives before numerical integration. The association with lower rollout error above the training-energy range further connects the quality of the learned Hamiltonian to the generalization behavior observed in the main experiments.

The spectral diagnostics characterize the variation remaining in these residuals. On the pendulum systems and the perturbed FPUT chain, most of LM-HNN's residual power lies beyond k_{99}, the bandwidth containing 99% of the target Hamiltonian power. Together with the smaller residual magnitude, this is consistent with more accurate recovery of the dominant Hamiltonian structure. LM-HNN also has a larger effective residual wavenumber k_{eff} on several systems. As defined in Eq. (38), this quantity measures gradient error relative to residual magnitude, so a larger value can coexist with lower absolute energy and vector-field errors. These spectral quantities therefore describe the composition of the remaining error and should be interpreted alongside σ_H and $\|\nabla\delta H\|_{\mathrm{rms}}$.

Table 16: **Fast–slow oscillator ($\varepsilon = 0.01$) rollout error.** We report the rollout MSE over the 30 unseen test initial conditions after 200 and 1000 steps with $\Delta t = 0.1$, reported for the full state and for the slow pair (q_2, p_2). LM-HNN achieves the lowest error in every column, improving both the complete dynamics and the weak slow component.

		full state		slow pair	
Model	Params	200 steps	1000 steps	200 steps	1000 steps
HNN	22,041	$(2.2_{\pm 2.6})\times 10^{-3}$	$(8.7_{\pm 3.7})\times 10^{-3}$	$(1.4_{\pm 0.2})\times 10^{-1}$	$(1.6_{\pm 0.2})\times 10^{-1}$
SympNet	22,016	$(1.2_{\pm 0.6})\times 10^{-1}$	$(1.3_{\pm 0.6})\times 10^{0}$	$(5.8_{\pm 0.4})\times 10^{0}$	$(7.4_{\pm 0.7})\times 10^{0}$
FS-HNN	21,488	$(1.7_{\pm 0.4})\times 10^{-3}$	$(2.1_{\pm 0.9})\times 10^{-3}$	$(1.9_{\pm 0.5})\times 10^{-1}$	$(3.0_{\pm 1.4})\times 10^{-1}$
LM-HNN	21,957	$\mathbf{(2.3_{\pm 1.6})\times 10^{-4}}$	$\mathbf{(1.8_{\pm 0.5})\times 10^{-3}}$	$\mathbf{(2.8_{\pm 0.5})\times 10^{-2}}$	$\mathbf{(4.1_{\pm 1.1})\times 10^{-2}}$

C.5 Detailed Results on the Fast–Slow Oscillator

We examine prediction accuracy on the fast–slow system of Eq. (33), with $\varepsilon = 0.01$, by evaluating the full state and the slow pair (q_2, p_2) separately. Table 16 reports normalized rollout errors after 200 and 1000 steps with $\Delta t = 0.1$, using the initial conditions of the 30 held-out test trajectories. The separate slow-pair metric measures how accurately the gradual evolution of these variables is recovered, complementing the full-state metric, which also includes the fast oscillations.

LM-HNN achieves the lowest mean error for both the full state and the slow pair at both horizons. The improvement in full-state prediction is therefore accompanied by more accurate slow dynamics. Although errors increase as the rollout extends from 200 to 1000 steps, LM-HNN retains its advantage in both metrics. This complements the representative trajectories in Figure 4, where LM-HNN captures the main slow evolution while remaining closely aligned with the reference fast oscillation.

The comparison with FS-HNN is informative because that baseline explicitly separates trajectory information across temporal frequency bands. LM-HNN obtains lower mean errors for both state groups at both horizons, supporting the effectiveness of organizing the Hamiltonian around its physical interaction structure even when the resulting dynamics evolve on widely separated temporal scales.