OrthoReg: Orthogonal Regularization for Hybrid Symbolic-Neural Dynamical Systems

Till Richter **Niki Kilbertus**
Technical University of Munich
Helmholtz Munich
Munich Center for Machine Learning (MCML)
till.richter,niki.kilbertus@helmholtz-munich.de

Abstract

Dynamical systems are fundamental to modeling the natural world, yet modeling them involves a persistent trade-off: manually prescribed mechanistic models are interpretable by design but often overly simplistic and misspecified; in contrast, flexible data-driven neural methods lack physical insight. Hybrid modeling aims for the best of both worlds by combining a prescribed or symbolic, physics-based component with a flexible neural network. A critical challenge, however, is that the neural component may relearn mechanistic parts, yielding redundant and uninterpretable models, especially when the symbolic structure itself is discovered from data. Existing methods based on standard L^2 regularization rely on a projection argument that breaks when the symbolic component is learned through sparse discovery, allowing the neural augmentation to overlap with symbolic structure. We introduce **OrthoReg** (Orthogonal Regularization), which directly penalizes overlap between the symbolic and neural components, preventing symbolic structure from being absorbed by the neural residual. This yields a complementary decomposition: the symbolic part captures what the library can express, and the neural part captures what remains. On benchmark dynamical systems with partial library mismatch, OrthoReg improves symbolic recovery and out-of-distribution behavior.

1 Introduction

Dynamical systems modeling has long been a cornerstone across the sciences, especially for the natural and life sciences. Applications range from healthcare data [Choi et al., 2016, Hess et al., 2024, Seedat et al., 2022, Richter et al., 2026], climate modeling [Rolnick et al., 2022, Eyring et al., 2024], and power systems [Toubeau et al., 2019], just to name a few. However, it faces a fundamental trade-off: symbolic models, traditionally specified by hand, provide interpretability by design, but typically cannot capture complex unknown phenomena; flexible neural networks instead excel at fitting data from dynamical systems [Chen et al., 2018] but lack physical insight.

Hybrid modeling becomes essential when domain knowledge is partial: epidemiological models may capture core transmission dynamics but miss behavioral feedback mechanisms that modulate contact rates; mechanical systems follow known laws of motion but exhibit complex friction and damping effects not easily expressible in closed form; climate models encode fundamental physics but require data-driven corrections for sub-grid processes. In these scenarios, a purely symbolic approach underperforms due to missing phenomena, while purely neural models sacrifice the interpretability and physical consistency that domain experts require for scientific insight and decision-making. Hybrid modeling approaches [Rackauckas et al., 2020, Yin et al., 2021, Zou et al., 2024] combine physical priors (predetermined symbolic or parametrized expressions) with learned neural corrections expected to capture phenomena that are unknown or too complex to model directly. They promise the best of both worlds, but still require substantial prior knowledge in crafting the mechanistic part. This work focuses on sparse library-based discovery of the symbolic component, as in SINDy-style methods [Brunton et al., 2016]. In this regime, the symbolic component is fitted with a continuous

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

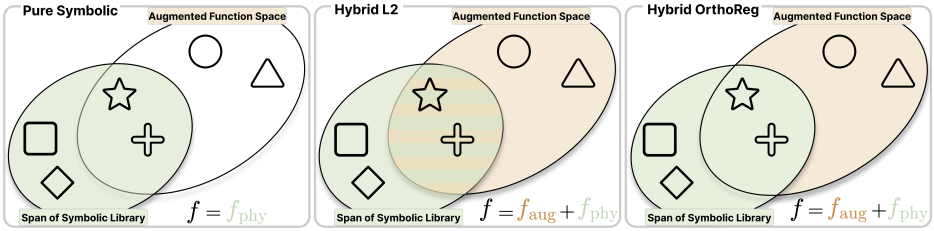

Figure 1: **Symbolic–neural decompositions under library mismatch.** The symbolic and augmented function spaces can overlap. Pure symbolic models are restricted to the symbolic span; standard hybrids may redundantly explain the shared region with both f_{phy} and f_{aug}; OrthoReg discourages this redundancy by pushing f_{aug} away from $\mathcal{F}_{\text{phy}}$, yielding a complementary decomposition.

sparsity penalty (e.g. L^1 regularization), where coefficient shrinkage leaves in-library residuals – the practically relevant case our analysis targets. Our goal is to make the symbolic and neural components complementary under the empirical inner product.

In their landmark paper, Yin et al. [2021] present the APHYNITY framework for hybrid dynamical systems modeling when the symbolic structure (but not exact parameter values) is known a priori. Like a large body of related existing work [Rackauckas et al., 2020, Mouli et al., 2024], APHYNITY decomposes the (autonomous) vector field of an ordinary differential equation (ODE) as $f = f_{\text{phy}} + f_{\text{aug}}$, where $f_{\text{phy}} \in \mathcal{F}_{\text{phy}} = \text{span}\{\phi_j\}_{j=1}^M$ captures dynamics within a predetermined library of "symbolic" functions $\{\phi_j\}_{j=1}^M$ (e.g., polynomials, trigonometric functions), while f_{aug} is supposed to capture the residual dynamics via flexible neural networks. When the symbolic structure is fixed and $\mathcal{F}_{\text{phy}}$ is a closed linear subspace, the two components are provably separated via simple L^2 regularization of f_{aug} (Proposition A.1). In this setting, optimizing out f_{aug} reduces to an L^2 projection onto the fixed subspace $\mathcal{F}_{\text{phy}}$, so the residual (and hence f_{aug}) is orthogonal to $\mathcal{F}_{\text{phy}}$, which underlies the analysis of Yin et al. [2021].

APHYNITY gives a projection-based view of hybrid modeling when the physical model class is fixed: the augmentation has a minimal norm subject to fitting the observed dynamics, and in finite-dimensional L^2 subspaces the residual is orthogonal to the physical component. This fixed-class picture changes when the symbolic component is learned through gradient-based sparse discovery. Continuous sparsity penalties, such as L^1-regularized library learning, shrink coefficients away from the unregularized projection and can leave residual structure inside the symbolic span. A neural augmentation can then relearn symbolic directions even when small in norm. OrthoReg targets this regime by directly penalizing empirical overlap between the augmentation and the symbolic library.

Example 1.1 (Counterexample: sparsity forces overlap)**.** Let $\mathcal{F}$ be a function space with inner product $\langle \cdot, \cdot \rangle_{\mathcal{D}}$ and induced norm $\| \cdot \|_{\mathcal{D}}$. Let $\phi_1, \phi_2 \in \mathcal{F}$ be orthonormal w.r.t. $\langle \cdot, \cdot \rangle_{\mathcal{D}}$ and $f = \phi_1 + \phi_2$. Consider

$$\min_{w \in \mathbb{R}^2,\, g \in \mathcal{F}} \; \|f - (w_1\phi_1 + w_2\phi_2 + g)\|_{\mathcal{D}}^2 + \lambda_2 \|g\|_{\mathcal{D}}^2 + \mu \|w\|_1.$$

If $0 < \mu < \frac{2\lambda_2}{1+\lambda_2}$, the unique minimizer satisfies $w_1 = w_2 = 1 - \frac{\mu(1+\lambda_2)}{2\lambda_2}$ and $g = \frac{\mu}{2\lambda_2}(\phi_1 + \phi_2)$, hence $g \notin \text{span}\{\phi_1, \phi_2\}$ although $\text{span}\{\phi_1, \phi_2\}$ contains f.

Thus, hybrid dynamical models that include symbolic discovery require explicit mechanisms to prevent neural-symbolic overlap. Figure 1 illustrates this failure mode and how OrthoReg restores a clean symbolic–neural split.

In this work, we introduce a theoretically grounded and practically effective method to learn hybrid dynamical systems, where the mechanistic component is discovered from data via symbolic regression while ensuring that the residual neural component remains orthogonal to the symbolic part. The key contribution is to bridge the gap between theoretical projection theory arguments and practical optimization to leverage both symbolic and data-driven modeling. Concretely, we provide:

- **Theoretical analysis** of symbolic–neural overlap under sparse library-based discovery: L^2 regularization can induce overlap between the two components (Theorem A.2), whereas OrthoReg yields a clean in-library / orthogonal-complement split under the empirical inner product (Theorem B.2).
- An **algorithmic solution** realizing the orthogonality constraint within the joint sparse-discovery objective.

- **Empirical validation** on benchmark dynamical systems with partial library mismatch, showing improved out-of-distribution generalization and sparse symbolic recovery. [1]

2 Related Work

We review data-driven methods for learning governing dynamics and the interpretability–expressiveness trade-off motivating OrthoReg.

(Dynamic) symbolic regression. Symbolic regression recovers closed-form expressions from data using genetic programming [Koza, 1994, Schmidt and Lipson, 2009], neural-guided search [Petersen et al., 2021, Udrescu and Tegmark, 2020], or sparse library-based regression such as SINDy [Brunton et al., 2016]. Sparse-library methods are highly interpretable when the true dynamics lie in the candidate set, but their expressiveness is limited by library design. Recent work improves symbolic discovery through physical constraints such as unit consistency [Tenachi et al., 2023], transformer-based symbolic generation [Lample and Charton, 2020, Biggio et al., 2021, Valipour et al., 2021, Kamienny et al., 2022, Vastl et al., 2024], extensions to ODEs and trajectory data [Becker et al., 2023, d'Ascoli et al., 2024, Sun et al., 2023], and methods for noisy, sparse, distributional, high-dimensional, or guided discovery settings [Qian et al., 2022, Dakhmouche et al., 2025, Liang and Yang, 2025, Tian et al., 2025, Hu et al., 2025]. However, when relevant effects are not representable by a compact symbolic library, purely symbolic models either fail or lose interpretability, motivating hybrid decompositions with a symbolic core and a flexible residual.

Physics-informed neural networks. PINNs [Raissi et al., 2019] embed differential equations as soft constraints and Universal ODEs [Rackauckas et al., 2020] parameterize unknown vector-field components with neural networks; surveys [Cuomo et al., 2022, Hao et al., 2022] position these as central scientific-ML paradigms. Both rely on a fixed symbolic part and do not address overlap when the symbolic component is itself discovered under sparsity.

Neural, symbolic, and hybrid methods. Hybrid methods combine symbolic interpretability with neural expressiveness. Gray-box discovery has been studied for PDEs via sparse regression [Rudy et al., 2017] and physics-informed symbolic objectives [Chen et al., 2021, Kiyani et al., 2023]. For ODEs, APHYNITY formalizes $f = f_{\text{phy}} + f_{\text{aug}}$ but assumes fixed symbolic structure [Yin et al., 2021]; related work alternates neural learning with symbolic distillation [Grigorian et al., 2024] or studies PINN/PIKAN optimization for gray-box identification [Daryakenari et al., 2026]. In related feedforward-control and nonlinear system-identification work, orthogonality between prescribed physics-based and learned additive components is enforced explicitly, either via a projection-based regularizer [Kon et al., 2022, Györök et al., 2025] or by an orthogonal-by-construction parametrization [Györök et al., 2026]. Pure neural ODE models can encode sparsity [Aliee et al., 2022], conservation/manifold constraints [Greydanus et al., 2019, Matsubara and Yaguchi, 2022, White et al., 2023], or meta-learned physics–ML trade-offs [Mouli et al., 2024], but generally lack explicit symbolic recovery. In contrast, OrthoReg targets the overlap problem that arises when the symbolic component is itself *discovered* jointly with a neural residual.

3 Background and Problem Setting

Let $\mathcal{F}$ be a class of measurable vector fields $f : \mathbb{R}^n \to \mathbb{R}^n$ with finite $L^2(\nu)$ norm, where ν is the data-generating measure on states. We use the population inner product $\langle f, g\rangle_\nu = \mathbb{E}_{x\sim\nu}[f(x)^\top g(x)]$ and the empirical semi-inner product $\langle f, g\rangle_{\mathcal{D}} = \frac{1}{N}\sum_{i=1}^{N} f(x_i)^\top g(x_i)$ on the sampled states $\mathcal{D} = \{x_i\}_{i=1}^N$, with induced norm and seminorm $\|\cdot\|_\nu$ and $\|\cdot\|_{\mathcal{D}}$.

For a closed linear subspace $\mathcal{F}_{\text{phy}} \subseteq \mathcal{F}$, we denote the corresponding population projection by $P^{\nu}_{\mathcal{F}_{\text{phy}}}$; empirically, $P^{\mathcal{D}}_{\mathcal{F}_{\text{phy}}}$ denotes the projection under $\|\cdot\|_{\mathcal{D}}$, understood up to equivalence on $\mathcal{D}$.

The functions $f \in \mathcal{F}$ are interpreted as vector fields of autonomous, first-order differential equations

$$\frac{\mathrm{d}x}{\mathrm{d}t} = f(x), \qquad \text{with solution trajectories } x : \mathbb{R} \to \mathbb{R}^n .$$

[1] Code: `https://github.com/richtertill/OrthoReg`.

Following prior work [Yin et al., 2021, Rackauckas et al., 2020], we assume an additive decomposition

$$f = f_{\text{phy}} + f_{\text{aug}}, \qquad f_{\text{phy}} \in \mathcal{F}_{\text{phy}},\ f_{\text{aug}} \in \mathcal{F}$$

of vector fields of interest into a "physical" (or symbolic/mechanistic) component and an "augmented" (or neural/residual) component. The space $\mathcal{F}_{\text{phy}} \subseteq \mathcal{F}$ of candidate symbolic components is typically restricted to functions that can be represented in closed form using known functions, so as to be amenable to direct interpretation and dissemination by humans.

Most existing methods assume f_{phy} to be either known exactly, or to be given as a parametric family in which only a (usually small) set of parameters is unknown. Practically, this is often implemented via a linear combination of non-linear basis functions:

$$\mathcal{F}_{\text{phy}} = \left\{ \sum_{j=1}^{M} \alpha_j \phi_j \;\middle|\; \alpha_j \in \mathbb{R} \text{ for } j \in \{1, \dots, M\} \right\}, \tag{1}$$

where the dictionary functions $\phi_j : \mathbb{R}^n \to \mathbb{R}^n$ are fixed. The dynamics governing most real systems are not perfectly described by such simple closed-form expressions, but contain higher-order effects or complex interactions that are rarely captured by simple interpretable mathematical expressions. To capture such residual effects, the augmentation $f_{\text{aug}} \in \mathcal{F}$ is supposed to be flexible and expressive, albeit potentially not easily interpretable. Hence, a natural choice to represent f_{aug} is via flexible function approximators such as neural networks, giving rise to the term "neural component." This is the premise of the hybrid-modeling line we build on: APHYNITY [Yin et al., 2021] and Universal ODEs [Rackauckas et al., 2020] augment a mechanistic core with a neural component precisely because real dynamics contain effects that no compact symbolic form captures. We adopt this premise; OrthoReg addresses the complementary question of keeping whatever residual is used from reabsorbing the symbolic core. Crucially, the neural component should *only capture effects that cannot be captured by the symbolic component*.

In the current formulation, one could simply set $f_{\text{aug}} \equiv f$ and $f_{\text{phy}} \equiv 0$. However, this would undermine the entire idea of hybrid modeling. When the physical model class is fixed, Yin et al. [2021] provide thorough theoretical guarantees showing that a relatively simple norm-based regularization scheme is sufficient to ensure that f_{aug} "only captures what is necessary, but not more." A simplified vector-field version of this norm-regularized principle is

$$\min_{f_{\text{phy}} \in \mathcal{F}_{\text{phy}}, f_{\text{aug}} \in \mathcal{F}} \ \|f - f_{\text{phy}} - f_{\text{aug}}\|_{\mathcal{D}}^2 + \lambda_2 \|f_{\text{aug}}\|_{\mathcal{D}}^2 \,. \tag{2}$$

For fixed f_{phy}, writing $r = f - f_{\text{phy}}$, the minimizer over f_{aug} is

$$\hat{f}_{\text{aug}} = \frac{1}{1+\lambda_2} r = \frac{1}{1+\lambda_2}(f - f_{\text{phy}}).$$

Substituting back gives

$$\min_{f_{\text{aug}} \in \mathcal{F}} \left[\|f - f_{\text{phy}} - f_{\text{aug}}\|_{\mathcal{D}}^2 + \lambda_2 \|f_{\text{aug}}\|_{\mathcal{D}}^2 \right] = \frac{\lambda_2}{1+\lambda_2} \|f - f_{\text{phy}}\|_{\mathcal{D}}^2.$$

Thus, for $\lambda_2 > 0$, optimizing over f_{phy} is equivalent to empirical least-squares projection onto $\mathcal{F}_{\text{phy}}$. If $\mathcal{F}_{\text{phy}}$ is a closed linear subspace, the Hilbert-space projection theorem gives $P^{\mathcal{D}}_{\mathcal{F}_{\text{phy}}}(f)$ up to $\mathcal{D}$-equivalence, and the residual $f - P^{\mathcal{D}}_{\mathcal{F}_{\text{phy}}}(f)$, hence $\hat{f}_{\text{aug}}$, is orthogonal to $\mathcal{F}_{\text{phy}}$ under $\langle \cdot, \cdot \rangle_{\mathcal{D}}$ [Lax, 2014]. APHYNITY extends this projection view beyond linear subspaces via existence and uniqueness guarantees under *proximinality* and *Chebyshevness* [Yin et al., 2021]; see Section A.

A natural extension beyond a fully known f_{phy}, or the structure being known up to a small set of parameters, is to discover the symbolic component itself from a larger dictionary, as in SINDy [Brunton et al., 2016]. After fixing the candidate basis functions $\{\phi_j\}_{j=1}^M$ we select only a small support set $S \subset \{1, \dots, M\}$ of basis functions that enter the expression with non-zero coefficients. The induced function space is

$$\mathcal{F}_{\text{phy}}(S) := \text{span}\{\phi_j \mid j \in S\}.$$

In practice, the set S is fitted via sparse regression methods (e.g., L1 regularization $\|\cdot\|_1$ or more involved iterated sparse regressions as in SINDy) to encourage small supports S.

While at first this appears to be a natural extension to APHYNITY, on closer inspection this changes the projection geometry underlying its argument. When sparse discovery is implemented through continuous shrinkage penalties, the symbolic fit is no longer an unregularized projection onto the selected span; the residual can therefore retain components in the symbolic library, which the neural augmentation may relearn (cf. Figure 1 and Example 1.1).

This is the fundamental gap our work addresses: expressive data-driven sparse symbolic discovery requires additional mechanisms to keep symbolic and neural components from overlapping. The closest faithful adaptation of APHYNITY to this setting is therefore a sparse symbolic component combined with an L^2-regularized neural augmentation, which we call the "Hybrid L^2" baseline.

4 Method: OrthoReg for Hybrid Modeling

We observe state trajectories of an unknown dynamical system and aim to learn an autonomous vector field $f : \mathbb{R}^n \to \mathbb{R}^n$ such that $\dot{x} = f(x)$. Following the hybrid modeling setup in Section 3, we model

$$\hat{f}(x) = \hat{f}_{\mathrm{phy}}(x; w) + \hat{f}_{\mathrm{aug}}(x; \vartheta), \;\; \hat{f}_{\mathrm{phy}}(x; w) = \sum_{j=1}^{M} w_j \, \phi_j(x), \tag{3}$$

where $\{\phi_j\}_{j=1}^M$ is a fixed symbolic library, w is sparse, and $\hat{f}_{\mathrm{aug}}(\cdot; \vartheta)$ is a flexible neural augmentation.

Learning from trajectories. The data consists of state observations over time (trajectories). Depending on the setting, derivative targets may or may not be available. We therefore consider two standard data-fit losses for learning $\hat{f}$. If (approximate) derivatives $y_i \approx \dot{x}(t_i)$ are available (e.g., from a simulator or numerical differentiation), we can fit $\hat{f}$ by vector-field regression

$$\mathcal{L}_{\mathrm{vf}}(w, \vartheta) = \frac{1}{N} \sum_{i=1}^{N} \left\| y_i - \hat{f}(x_i) \right\|^2. \tag{4}$$

If only states are observed, we fit by one-step prediction. Let $\mathcal{T}$ be the set of observed one-step transitions $(x_t, x_{t+1}, \Delta t_t)$ (potentially with irregular Δt_t) and let $\Psi_{\Delta t}(\cdot; \hat{f})$ denote a one-step ODE solver (e.g., explicit Euler $\Psi_{\Delta t}(x; \hat{f}) = x + \Delta t \, \hat{f}(x)$ or a Runge–Kutta method). The one-step loss is

$$\mathcal{L}_{\mathrm{step}}(w, \vartheta) = \frac{1}{|\mathcal{T}|} \sum_{(x_t, x_{t+1}, \Delta t_t) \in \mathcal{T}} \left\| x_{t+1} - \Psi_{\Delta t_t}(x_t; \hat{f}) \right\|^2. \tag{5}$$

In the following, $\mathcal{L}_{\mathrm{fit}}$ denotes either $\mathcal{L}_{\mathrm{vf}}$ or $\mathcal{L}_{\mathrm{step}}$. We train with $\mathcal{L}_{\mathrm{vf}}$ (state-only training is also compatible with OrthoReg; see Section J), as derivative-space training was empirically more stable, and evaluate in both derivative and state space. OrthoReg uses the empirical inner product $\langle \cdot, \cdot \rangle_{\mathcal{D}}$ over observed states $\mathcal{D} = \{x_i\}_{i=1}^N$, defined in Section 3.

4.1 OrthoReg Objective

OrthoReg encourages a non-redundant decomposition by penalizing correlation between the neural augmentation and the symbolic library. Concretely, we penalize the (squared) empirical inner products

$$\mathcal{L}_{\mathrm{reg}}^{\perp}(\vartheta) = \lambda \sum_{j=1}^{M} \left\langle \hat{f}_{\mathrm{aug}}(\cdot; \vartheta), \phi_j \right\rangle_{\mathcal{D}}^2, \tag{6}$$

so that $\mathcal{L}_{\mathrm{reg}}^{\perp} = 0$ implies $\hat{f}_{\mathrm{aug}} \perp \mathrm{span}\{\phi_1, \ldots, \phi_M\}$ w.r.t. $\langle \cdot, \cdot \rangle_{\mathcal{D}}$. Let $\mu \geq 0$ control symbolic sparsity and $\lambda \geq 0$ the orthogonality strength, then the full OrthoReg training objective is

$$\min_{w, \vartheta} \; \mathcal{L}_{\mathrm{fit}}(w, \vartheta) + \mu \|w\|_1 + \mathcal{L}_{\mathrm{reg}}^{\perp}(\vartheta), \tag{7}$$

The guarantee in Theorem 4.1 is tied to the additive objective in Equation (7): the orthogonality penalty must act on the additive residual $\hat{f}_{\mathrm{aug}}$ at the same empirical inputs as the data-fit. Staged or

alternating updates of the same objective fall within the same formal scope when they reach a solution satisfying the theorem assumptions, whereas unconstrained residual fitting after shrinkage-based symbolic learning and compositional architectures of the form $T_w + N_\vartheta \circ T_w$ do not (Proposition C.1 and Section C).

4.2 Implementation and Computational Considerations

Implementation. Algorithm 1 (Section H) summarizes training. We optimize the continuous L^1-regularized objective with Adam and report sparsity after thresholding small coefficients. The orthogonality penalty costs $\mathcal{O}(B\,n\,M)$ per batch of size B; in mini-batch training, finite-batch squaring of the inner-product estimate introduces a small bias, so our experiments use full- or large-batch estimates unless stated otherwise. Hyperparameter selection, feature scaling, and the autonomous-augmentation caveat for non-autonomous data are described in Sections F and G.8.

4.3 Theoretical Guarantees

We first show that L^2 regularization does not prevent – and can in fact induce – overlap between symbolic and neural components under sparse selection (Theorem A.2, Section A). The OrthoReg objective restores control via two structural results: a penalty bound showing that empirical overlap is controlled at the optimum and vanishes as $\lambda \to \infty$, and a triangular error decomposition (the in-library / orthogonal-complement split) that, under the empirical inner product, serves as a formal lens for the symbolic–neural decomposition.

Theorem 4.1 (Penalty bound for empirical orthogonality)**.** *Assume $\lambda > 0$ and that Equation* (7) *admits a global minimizer (w^*, ϑ^*). Let $w_{\mathrm{sym}} \in \arg\min_w \mathcal{L}_{\mathrm{fit}}(w, \vartheta_0) + \mu\|w\|_1$ denote the best pure-symbolic fit at $\hat{f}_{\mathrm{aug}} \equiv 0$ (parameters ϑ_0). Then*

$$\sum_{j=1}^{M} \langle \hat{f}_{\mathrm{aug}}(\cdot; \vartheta^*), \phi_j \rangle_{\mathcal{D}}^2 \le \tfrac{1}{\lambda}\big(\mathcal{L}_{\mathrm{fit}}(w_{\mathrm{sym}}, \vartheta_0) + \mu\|w_{\mathrm{sym}}\|_1\big),$$

so empirical orthogonality is enforced as $\lambda \to \infty$.

Proof idea. The orthogonality term is a quadratic penalty: comparing the optimum to the baseline $\hat{f}_{\mathrm{aug}} \equiv 0$ yields a $1/\lambda$ bound on constraint violation. A full proof is given in Theorem B.1. □

In isolation, the large-λ statement is close to immediate; we include it as a completeness check that the penalty controls what it is designed to control, not as a central claim. The substantive results are the L^2-failure result (Theorem A.2) and the error decomposition below (Theorem B.2).

Error decomposition. Beyond the penalty bound, orthogonality has a concrete geometric consequence for approximation. In the idealized vector-field regression setting, and when the learned augmentation is empirically orthogonal to $\mathcal{F}_{\mathrm{phy}}$, the squared error admits an orthogonal decomposition:

$$\|f - (\hat{f}_{\mathrm{phy}} + \hat{f}_{\mathrm{aug}})\|_{\mathcal{D}}^2 = \|P^{\mathcal{D}}_{\mathcal{F}_{\mathrm{phy}}}(f) - \hat{f}_{\mathrm{phy}}\|_{\mathcal{D}}^2 + \|f - P^{\mathcal{D}}_{\mathcal{F}_{\mathrm{phy}}}(f) - \hat{f}_{\mathrm{aug}}\|_{\mathcal{D}}^2, \tag{8}$$

Thus, the symbolic error is the gap to the in-library projection $P^{\mathcal{D}}_{\mathcal{F}_{\mathrm{phy}}}(f)$, while the neural error is the gap to the empirical orthogonal-complement component $f - P^{\mathcal{D}}_{\mathcal{F}_{\mathrm{phy}}}(f)$; see Theorem B.2. Under the empirical inner product, this decomposition serves as a formal lens for the symbolic–neural split: each component's error is the gap to its half of the orthogonal direct sum.

Comparison to L^2 regularization. L^2 regularization controls only the *magnitude* of the augmentation through the decomposition

$$\|\hat{f}_{\mathrm{aug}}\|_{\mathcal{D}}^2 = \|P^{\mathcal{D}}_{\mathcal{F}_{\mathrm{phy}}}(\hat{f}_{\mathrm{aug}})\|_{\mathcal{D}}^2 + \|\hat{f}_{\mathrm{aug}} - P^{\mathcal{D}}_{\mathcal{F}_{\mathrm{phy}}}(\hat{f}_{\mathrm{aug}})\|_{\mathcal{D}}^2,$$

where the equality follows from orthogonal decomposition in inner product spaces [Rudin, 1987]. Even when the total is small, the component $P^{\mathcal{D}}_{\mathcal{F}_{\mathrm{phy}}}(\hat{f}_{\mathrm{aug}})$ can be non-zero, allowing neural–symbolic overlap. When $\mathcal{F}_{\mathrm{phy}}$ is learned through sparsity constraints, this overlap can occur even in realizable settings, motivating explicit orthogonality regularization (Section A).

Empirical-to-population transfer. OrthoReg controls overlap under the empirical inner product used for training. A formal population guarantee for the learned augmentation would require uniform convergence over the augmentation hypothesis class; we leave this to future work (Section J).

5 Experiments

Table 1: Damped Pendulum results. Scale factors $\mathrm{e}k$ inside parentheses are powers of ten; x/$\dot{x}$ are state/derivative spaces, $\pm$ denotes one sample standard deviation.

Metric	Pure	L^2	OrthoReg
In-Distribution Performance			
MSE($\dot{x}_{\mathrm{ID}}$, e-2) (↓)	$\mathbf{1.27 \pm 0.07}$	1.41 ± 0.05	1.89 ± 0.01
MSE(x_{ID}, e-2) (↓)	0.74 ± 0.43	$\mathbf{0.31 \pm 0.02}$	3.65 ± 0.36
MSE($\dot{x}_{\mathrm{ID,ext}}$, e$-2$) (↓)	$\mathbf{1.33 \pm 0.07}$	1.49 ± 0.06	2.12 ± 0.02
Out-of-Distribution Performance			
MSE($\dot{x}_{\mathrm{OOD,T2}}$, e1) (↓)	194.46 ± 128.93	50.02 ± 24.23	$\mathbf{1.31 \pm 0.65}$
MSE($\dot{x}_{\mathrm{OOD,T3}}$, e$-2$) (↓)	$\mathbf{2.83 \pm 0.11}$	3.13 ± 0.09	3.78 ± 0.05
System Identification Quality			
F1 (↑)	0.43 ± 0.09	0.61 ± 0.03	$\mathbf{0.93 \pm 0.15}$
#Terms (↓)	11.4 ± 2.9	6.8 ± 0.4	$\mathbf{3.6 \pm 1.3}$

We evaluate OrthoReg across five dynamical systems of different complexities: a modified damped pendulum, a Duffing oscillator, a Lotka–Volterra predator-prey system, a time-modulated SIR epidemiological model, and a damped Helmholtz oscillator. Our evaluation focuses on two complementary metrics: (i) trajectory accuracy measured by mean-squared error (MSE) on derivatives and integrated states on in-distribution, out-of-distribution, and extrapolation states; and (ii) symbolic recovery quality measured by F1 score and the number of symbolic terms, where F1 is scored against the in-library ground-truth support (Section G). We compare three approaches: pure symbolic regression (SINDy), an L^2-regularized hybrid, and an OrthoReg-regularized hybrid. We report along four axes consistent with multi-axis evaluation principles in scientific ML [Wyder et al., 2025]: initial-condition extrapolation, parametric extrapolation, regime/basin extrapolation, and robustness under irregular sampling and observation noise. OOD-T2 denotes initial-condition extrapolation and OOD-T3 parametric extrapolation (Section G.2). Each experiment is repeated over five stochastic runs; complete experimental details are provided in Section G.

Reading the results by regime. OrthoReg is a decomposition regularizer, not a general accuracy booster, and the systems stress different regimes by design: clear, stable gains in symbolic recovery where the out-of-library dynamics are separable from the library (pendulum; Helmholtz, Section F.4); the sparsest support on Duffing, where state-space errors favor the pure symbolic model; similar performance where the library already suffices (Lotka–Volterra); and a sparser, lower-false-positive support under severe misspecification (SIR). Section F.5 breaks F1 into precision, recall and false-positive rate per system.

5.1 Damped Pendulum: Missing Dynamics

The modified damped pendulum extends the classical damped pendulum [Kharkongor and Mahato, 2018] with three forcing terms that lie outside the symbolic dictionary:

$$\ddot{\theta} + \alpha\dot{\theta} + \omega_0^2 \sin\theta \;=\; \beta_1 \cos(3\theta) + \beta_2 e^{-\theta^2} + \beta_3 \tanh(\dot{\theta}), \tag{9}$$

with $(\beta_1, \beta_2, \beta_3) = (0.3, 0.25, 0.15)$. The dictionary used by the symbolic component is $\mathrm{PolynomialLibrary}(\mathrm{degree} = 2) \oplus \mathrm{FourierLibrary}(n_{\mathrm{freq}} = 1)$, which contains $\sin\theta, \cos\theta, \sin\dot{\theta}, \cos\dot{\theta}$ and the monomials of $(\theta, \dot{\theta})$ of degree one and two; the three forcing terms above lie outside this dictionary by construction.

Table 1 reports performance under medium missing dynamics ($\beta_1 = 0.3, \beta_2 = 0.25, \beta_3 = 0.15$ in Equation (9)). Derivative and state MSE quantify trajectory fit, and F1 measures symbolic recovery against ground-truth terms. OrthoReg yields the strongest symbolic recovery (highest F1, fewest terms) and the lowest derivative error under initial-condition extrapolation (OOD-T2), about

$40\times$ below L^2 and $150\times$ below Pure, at the cost of higher in-distribution error. Under parametric extrapolation (OOD-T3), where coefficients are refit on the shifted system, the three methods are close, with the lowest error for Pure. This trade-off is consistent with the symbolic / orthogonal-residual split established by Theorem B.2: enforcing the split sacrifices some in-distribution fit in exchange for components that transfer to unseen initial conditions.

5.2 Comparison with PINN- and UODE-Style Baselines

Table 2: Baseline comparison on the modified damped pendulum; all methods share the library, L^1 weight, network and refit.

Metric	PINN-style	UODE-style	OrthoReg
MSE($\dot{x}_{\mathrm{ID},\ \mathrm{e}-2}$) (↓)	0.54 ±0.02	**0.34 ±0.01**	1.89 ±0.01
MSE($\dot{x}_{\mathrm{OOD,T2},\ \mathrm{e}1}$) (↓)	**0.14 ±0.00**	2.03 ±0.03	1.31 ±0.65
MSE($\dot{x}_{\mathrm{OOD,T3},\ \mathrm{e}-2}$) (↓)	1.05 ±0.02	**0.43 ±0.03**	3.78 ±0.05
F1 (↑)	0.79 ±0.06	0.56 ±0.03	**0.93 ±0.15**

We also compare to PINN-style [Raissi et al., 2019] and Universal-ODE-style [Rackauckas et al., 2020] couplings of the same library and network in Table 2 (details in Section G.6). Both fit in-distribution more closely, and neither keeps the network out of the library span; OrthoReg gives up some of this fit for the most accurate support (F1 0.93 vs. 0.79 and 0.56). On OOD-T2 it lies between the two (1.31 vs. 0.14 for PINN-style and 2.03 for UODE-style, $\times 10^1$). Canonical PySINDy STLSQ (same library and splits as the main experiments) is summarized in Section G.6: it improves symbolic support recovery relative to the Adam L^1 pure-symbolic rows when thresholding is the sparsity mechanism, but unstable rollouts in partial mismatch show why residual capacity remains necessary.

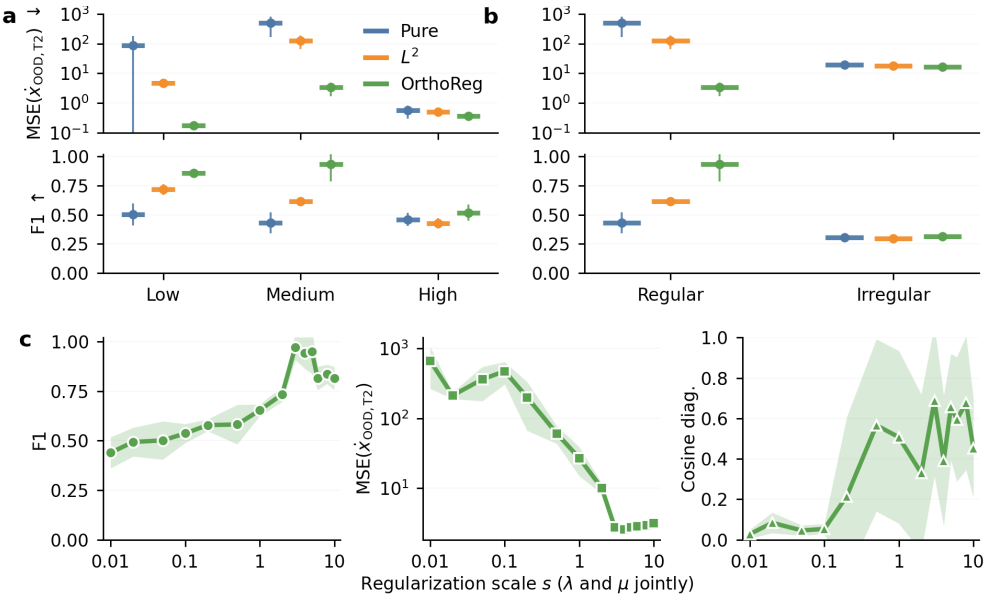

Figure 2: **Ablations.** (a, b) Top: normalized OOD-T2 derivative MSE; bottom: F1. (a) OrthoReg is most effective under partial library mismatch. (b) Irregular sampling degrades all methods, while OrthoReg retains competitive OOD behavior. (c) Intermediate regularization scale (λ and μ scaled jointly, Section G.7) gives the best trade-off between symbolic recovery and OOD-T2 error; right panel shows a residual–symbolic cosine diagnostic for OrthoReg.

5.3 Ablations

We ablate three factors that probe when and why orthogonal regularization is beneficial: (i) the degree to which the true dynamics exceed the symbolic library, (ii) robustness to non-ideal (irregular) observation times, and (iii) the strength of the orthogonality penalty. Figure 2 summarizes the main trends. Complete quantitative results and additional ablation analyses are reported in Section E.

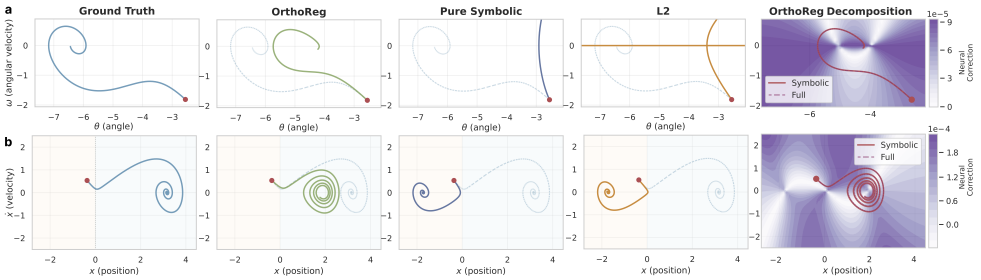

Figure 3: (a) Damped pendulum in (θ, ω), with $\omega := \dot{\theta}$. (b) Duffing in $(x, \dot{x})$ with shaded basins. Solid: inferred; dashed: ground truth. OrthoReg captures global trends; pure symbolic and L^2 distort the dynamics. Rightmost panels show the OrthoReg decomposition, highlighting the neural residual.

Dataset difficulty (Figure 2a). OrthoReg's gains under initial-condition extrapolation (OOD-T2) are largest in the low- and intermediate-missing regimes, where the system partially exceeds the library; in high-missing regimes ($\beta_1 = \beta_2 = \beta_3 = 2.0$) all methods degrade with OrthoReg retaining a modest advantage. Under parametric extrapolation (OOD-T3) the methods are close, with Pure lowest in the low and intermediate regimes. Full results in Section E.1.

Sampling scheme (Figure 2b). Under irregular sampling (time points jittered by Uniform$[-0.1, 0.1]$ vs. a regular grid with $\Delta t \approx 0.01$), absolute performance drops across all methods, but OrthoReg retains a marginal F1 advantage and the lowest OOD-T2 derivative error. Full breakdown in Section E.2.

Regularization strength (Figure 2c). OrthoReg has an intermediate operating range: weak λ under-enforces the orthogonality penalty, while overly strong λ constrains the augmentation. λ and μ are chosen per system and method from small sweeps. Right: cosine diagnostic for OrthoReg; regularization strength is assessed by F1 and OOD error.

5.4 Duffing Oscillator: Regime-Dependent Generalization

Table 3: Duffing oscillator. Training on the positive basin; OOD-T2 tests cross-basin (Regime-Ext), OOD-T3 tests parameter shifts in the negative basin (Param-Ext).

Metric	Pure	L^2	OrthoReg
Predictive Performance			
MSE(x_{ID}, e-1) (↓)	**4.75 ±0.13**	5.36 ±0.14	5.59 ±0.09
MSE($x_{\text{OOD,T2}}$, e-1) (↓)	**6.77 ±1.32**	8.85 ±0.49	9.08 ±0.19
MSE($x_{\text{OOD,T3}}$, e0) (↓)	**1.00 ±0.30**	1.77 ±0.08	1.79 ±0.06
System Identification Quality			
F1 (↑)	0.40 ±0.05	0.59 ±0.04	**0.63 ±0.05**
#Terms (↓)	11.0 ±1.6	3.8 ±0.4	**3.4 ±0.5**

We study the unforced Duffing oscillator, a multistable system whose cross-basin behavior probes whether models capture global structure beyond the observed basin [Göring et al., 2024], and train on trajectories from the positive basin only. Table 3 shows that OrthoReg attains the strongest symbolic recovery (highest F1, fewest terms), whereas the pure symbolic model attains the lowest state-space error in-distribution, across basins (OOD-T2) and under parameter shifts (OOD-T3); the two hybrids are close to each other on all three (details in Section F.2). On this system, the benefit of orthogonality is therefore in the recovered structure rather than in trajectory error.

5.5 Cross-System Validation

As auxiliary stress tests beyond the pendulum and Duffing, we evaluate OrthoReg on a Lotka–Volterra predator-prey system (coupled polynomial dynamics), a time-modulated SIR model (severe library mismatch and state-dependent timescales), and a damped Helmholtz oscillator whose corrugation term lies outside the library; full results are in Section F. We read LV as a low-headroom sanity check (Table 7: methods are comparable, with a small F1 gain for OrthoReg) and SIR as a misspecification stress test (Table 8: OrthoReg trades ID and OOD-T3 accuracy for the lowest OOD-T2 error and the

sparsest symbolic recovery – highest F1 and fewest terms). On Helmholtz, OrthoReg recovers the full in-library support in every run, matching the pure symbolic model, while the L^2 hybrid drops a true term in most runs (Section F.4). The pattern is consistent with the symbolic–neural split.

5.6 Summary

Across systems, OrthoReg is most effective in the regime it is designed for: partial library misspecification, where the symbolic library captures part of the dynamics but leaves structured residual effects. In this setting, OrthoReg improves symbolic recovery and reduces derivative error under initial-condition extrapolation, while accepting some loss in in-distribution fit; under parametric extrapolation with refit coefficients the methods perform comparably. The gains persist under irregular sampling, but diminish under heavy observation noise and when the missing dynamics are strongly entangled with library features (Section E.5). Where recall is not pinned at one, the support trades recall for precision in a regime-dependent way (lowest recall and lowest false-positive rate on SIR, full recall without false positives on Helmholtz); this operating point is set by the L^1 penalty and the data geometry rather than by the orthogonality penalty itself (Section F.5). These results support the view of OrthoReg as a decomposition regularizer: it is not a universal accuracy booster, but a mechanism for preserving a meaningful symbolic component while assigning non-symbolic effects to the residual.

6 Conclusion

Hybrid symbolic–neural models promise interpretable structure together with flexible residual modeling, but this promise depends on a clean division of labor between the two components. We showed that this division becomes fragile when the symbolic component is discovered through sparse library learning: standard L^2 regularization controls the magnitude of the neural residual, but not whether it relearns symbolic structure. We introduced **OrthoReg**, which makes complementarity explicit by penalizing empirical overlap between the neural augmentation and the symbolic library.

Our analysis shows that projection-based regularization can induce symbolic–neural overlap under sparse discovery (Theorem A.2), whereas OrthoReg controls empirical orthogonality violation at global minimizers (Theorem 4.1) and yields an in-library / orthogonal-complement error split (Theorem B.2). Empirically, OrthoReg improves symbolic recovery and out-of-distribution behavior in regimes of partial library mismatch, where the library captures part of the dynamics but leaves structured residual effects.

Several limitations remain. Our guarantees are empirical-inner-product statements for additive models with fixed differentiable libraries and continuous sparsity relaxations; derivative supervision, feature scaling, and noise sensitivity also matter in practice (Section J). A key next step is to extend orthogonality-regularized residual learning to broader symbolic-discovery pipelines, including canonical SINDy selection–refit, genetic-programming, and transformer-based symbolic regression. Real epidemiological time series, where the mechanistic core is trusted but behavioral and policy effects are observed without a closed-form model, are the natural concrete target. Two adaptations would be the starting point: concatenating time with the state input to handle non-autonomous effects (Section F), and state-space training (Equation (5)) or noise-aware differentiation to handle noisy observations (Section E.5). More broadly, OrthoReg points toward hybrid scientific models in which approximate mechanistic vocabularies and flexible residuals can coexist without erasing the symbolic explanation.

Acknowledgements

We thank Manuel Lubetzki and Alessandro Palma for helpful discussions and constructive feedback on this manuscript. This work has been supported by the German Federal Ministry of Education and Research (Grant: 01IS24082). This work has been supported by the European Union (ERC, DYNAMICAUS, 101221985). The authors declare no competing interests.

References

Hananeh Aliee, Till Richter, Mikhail Solonin, Ignacio Ibarra, Fabian Theis, and Niki Kilbertus. Sparsity in continuous-depth neural networks. *Advances in Neural Information Processing Systems*, 35:901–914, 2022.

Sören Becker, Michal Klein, Alexander Neitz, Giambattista Parascandolo, and Niki Kilbertus. Predicting ordinary differential equations with transformers. In *International conference on machine learning*, pages 1978–2002. PMLR, 2023.

Luca Biggio, Tommaso Bendinelli, Alexander Neitz, Aurelien Lucchi, and Giambattista Parascandolo. Neural symbolic regression that scales. In *International conference on machine learning*, pages 936–945. PMLR, 2021.

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.

Ricky T. Q. Chen, Yulia Rubanova, Jesse Bettencourt, and David K Duvenaud. Neural ordinary differential equations. In S. Bengio, H. Wallach, H. Larochelle, K. Grauman, N. Cesa-Bianchi, and R. Garnett, editors, *Advances in Neural Information Processing Systems*, volume 31. Curran Associates, Inc., 2018.

Zhao Chen, Yang Liu, and Hao Sun. Physics-informed learning of governing equations from scarce data. *Nature communications*, 12(1):6136, 2021.

Edward Choi, Mohammad Taha Bahadori, Jimeng Sun, Joshua Kulas, Andy Schuetz, and Walter Stewart. Retain: An interpretable predictive model for healthcare using reverse time attention mechanism. *Advances in neural information processing systems*, 29, 2016.

Salvatore Cuomo, Vincenzo Schiano Di Cola, Fabio Giampaolo, Gianluigi Rozza, Maziar Raissi, and Francesco Piccialli. Scientific machine learning through physics–informed neural networks: Where we are and what's next. *Journal of Scientific Computing*, 92(3):88, 2022.

Ramzi Dakhmouche, Ivan Lunati, and Hossein Gorji. Robust symbolic regression for dynamical system identification. *Transactions on Machine Learning Research*, 2025.

Nazanin Ahmadi Daryakenari, Khemraj Shukla, and George Em Karniadakis. Representation meets optimization: Training PINNs and PIKANs for gray-box discovery in systems pharmacology. *Computers in Biology and Medicine*, 201:111393, 2026.

Stéphane d'Ascoli, Sören Becker, Philippe Schwaller, Alexander Mathis, and Niki Kilbertus. ODE-Former: Symbolic regression of dynamical systems with transformers. In *International Conference on Learning Representations*, volume 2024, pages 21943–21976, 2024.

Veronika Eyring, William D Collins, Pierre Gentine, Elizabeth A Barnes, Marcelo Barreiro, Tom Beucler, Marc Bocquet, Christopher S Bretherton, Hannah M Christensen, Katherine Dagon, et al. Pushing the frontiers in climate modelling and analysis with machine learning. *Nature Climate Change*, 14(9):916–928, 2024.

Niclas Alexander Göring, Florian Hess, Manuel Brenner, Zahra Monfared, and Daniel Durstewitz. Out-of-domain generalization in dynamical systems reconstruction. In *Forty-first International Conference on Machine Learning*, 2024.

Samuel Greydanus, Misko Dzamba, and Jason Yosinski. Hamiltonian neural networks. *Advances in neural information processing systems*, 32, 2019.

Gevik Grigorian, Sandip V George, Sam Lishak, Rebecca J Shipley, and Simon Arridge. A hybrid neural ordinary differential equation model of the cardiovascular system. *Journal of the Royal Society Interface*, 21(212), 2024.

Bendeguz Mate Györök, Jan H. Hoekstra, Johan Kon, Tamas Peni, Maarten Schoukens, and Roland Toth. Orthogonal projection-based regularization for efficient model augmentation. In Necmiye Ozay, Laura Balzano, Dimitra Panagou, and Alessandro Abate, editors, *Proceedings of the 7th Annual Learning for Dynamics & Control Conference*, volume 283 of *Proceedings of Machine Learning Research*, pages 166–178. PMLR, 04–06 Jun 2025.

Bendegúz M. Györök, Maarten Schoukens, Tamás Péni, and Roland Tóth. Orthogonal-by-construction augmentation of physics-based input–output models. *IFAC Journal of Systems and Control*, 35:100376, 2026. ISSN 2468-6018.

Zhongkai Hao, Songming Liu, Yichi Zhang, Chengyang Ying, Yao Feng, Hang Su, and Jun Zhu. Physics-informed machine learning: A survey on problems, methods and applications. *arXiv preprint arXiv:2211.08064*, 2022.

Konstantin Hess, Valentyn Melnychuk, Dennis Frauen, and Stefan Feuerriegel. Bayesian neural controlled differential equations for treatment effect estimation. In *International Conference on Learning Representations*, volume 2024, pages 52586–52622, 2024.

Herbert W Hethcote. The mathematics of infectious diseases. *SIAM review*, 42(4):599–653, 2000.

Jiao Hu, Jiaxu Cui, and Bo Yang. Learning interpretable network dynamics via universal neural symbolic regression. *Nature Communications*, 16(1):6226, 2025.

Pierre-Alexandre Kamienny, Stéphane d'Ascoli, Guillaume Lample, and François Charton. End-to-end symbolic regression with transformers. *Advances in Neural Information Processing Systems*, 35:10269–10281, 2022.

D Kharkongor and Mangal C Mahato. Resonance oscillation of a damped driven simple pendulum. *European Journal of Physics*, 39(6):065002, 2018.

Elham Kiyani, Khemraj Shukla, George Em Karniadakis, and Mikko Karttunen. A framework based on symbolic regression coupled with extended physics-informed neural networks for gray-box learning of equations of motion from data. *Computer Methods in Applied Mechanics and Engineering*, 415:116258, 2023.

Johan Kon, Dennis Bruijnen, Jeroen Van De Wijdeven, Marcel Heertjes, and Tom Oomen. Physics-guided neural networks for feedforward control: An orthogonal projection-based approach. In *2022 American Control Conference (ACC)*, pages 4377–4382. IEEE, 2022.

John R Koza. Genetic programming as a means for programming computers by natural selection. *Statistics and computing*, 4(2):87–112, 1994.

Adam J Kucharski, Timothy W Russell, Charlie Diamond, Yang Liu, John Edmunds, Sebastian Funk, Rosalind M Eggo, Fiona Sun, Mark Jit, James D Munday, et al. Early dynamics of transmission and control of covid-19: a mathematical modelling study. *The lancet infectious diseases*, 20(5): 553–558, 2020.

Guillaume Lample and François Charton. Deep learning for symbolic mathematics. In *International Conference on Learning Representations*, 2020.

Peter D Lax. *Functional analysis*. John Wiley & Sons, 2014.

Senwei Liang and Haizhao Yang. Finite expression method for solving high-dimensional partial differential equations. *Journal of Machine Learning Research*, 26(138):1–31, 2025.

Takashi Matsubara and Takaharu Yaguchi. FINDE: Neural differential equations for finding and preserving invariant quantities, 2022.

S Chandra Mouli, Muhammad Alam, and Bruno Ribeiro. Metaphysica: Improving OOD robustness in physics-informed machine learning. In *The Twelfth International Conference on Learning Representations*, 2024.

Brenden K Petersen, Mikel Landajuela Larma, Terrell N. Mundhenk, Claudio Prata Santiago, Soo Kyung Kim, and Joanne Taery Kim. Deep symbolic regression: Recovering mathematical expressions from data via risk-seeking policy gradients. In *International Conference on Learning Representations*, 2021.

Zhaozhi Qian, Krzysztof Kacprzyk, and Mihaela van der Schaar. D-CODE: Discovering closed-form ODEs from observed trajectories. In *International Conference on Learning Representations*, 2022.

Christopher Rackauckas, Yingbo Ma, Julius Martensen, Collin Warner, Kirill Zubov, Rohit Supekar, Dominic Skinner, Ali Ramadhan, and Alan Edelman. Universal differential equations for scientific machine learning. *arXiv preprint arXiv:2001.04385*, 2020.

Maziar Raissi, Paris Perdikaris, and George E Karniadakis. Physics-informed neural networks: A deep learning framework for solving forward and inverse problems involving nonlinear partial differential equations. *Journal of Computational physics*, 378:686–707, 2019.

Till Richter, Weixu Wang, Alessandro Palma, and Fabian J Theis. Generative models of cell dynamics: from Neural ODEs to flow matching. *Communications Biology*, 2026.

David Rolnick, Priya L Donti, Lynn H Kaack, Kelly Kochanski, Alexandre Lacoste, Kris Sankaran, Andrew Slavin Ross, Nikola Milojevic-Dupont, Natasha Jaques, Anna Waldman-Brown, et al. Tackling climate change with machine learning. *ACM Computing Surveys (CSUR)*, 55(2):1–96, 2022.

Walter Rudin. *Real and complex analysis*. McGraw-Hill, Inc., 1987.

Samuel H Rudy, Steven L Brunton, Joshua L Proctor, and J Nathan Kutz. Data-driven discovery of partial differential equations. *Science advances*, 3(4):e1602614, 2017.

Michael Schmidt and Hod Lipson. Distilling free-form natural laws from experimental data. *science*, 324(5923):81–85, 2009.

Nabeel Seedat, Fergus Imrie, Alexis Bellot, Zhaozhi Qian, and Mihaela van der Schaar. Continuous-time modeling of counterfactual outcomes using neural controlled differential equations. In Kamalika Chaudhuri, Stefanie Jegelka, Le Song, Csaba Szepesvari, Gang Niu, and Sivan Sabato, editors, *Proceedings of the 39th International Conference on Machine Learning*, volume 162 of *Proceedings of Machine Learning Research*, pages 19497–19521. PMLR, 17–23 Jul 2022.

Fangzheng Sun, Yang Liu, Jian-Xun Wang, and Hao Sun. Symbolic physics learner: Discovering governing equations via Monte Carlo tree search. In *The Eleventh International Conference on Learning Representations*, 2023.

Wassim Tenachi, Rodrigo Ibata, and Foivos I Diakogiannis. Deep symbolic regression for physics guided by units constraints: toward the automated discovery of physical laws. *The Astrophysical Journal*, 959(2):99, 2023.

Yuan Tian, Wenqi Zhou, Michele Viscione, Hao Dong, David S Kammer, and Olga Fink. Interactive symbolic regression with co-design mechanism through offline reinforcement learning. *Nature Communications*, 16(1):3930, 2025.

Ryan J. Tibshirani. The lasso problem and uniqueness. *Electronic Journal of Statistics*, 7:1456–1490, 2013. doi: 10.1214/13-EJS815.

Jean-François Toubeau, Jérémie Bottieau, François Vallée, and Zacharie De Grève. Deep learning-based multivariate probabilistic forecasting for short-term scheduling in power markets. *IEEE Transactions on Power Systems*, 34(2):1203–1215, 2019. doi: 10.1109/TPWRS.2018.2870041.

Silviu-Marian Udrescu and Max Tegmark. AI Feynman: A physics-inspired method for symbolic regression. *Science advances*, 6(16):eaay2631, 2020.

Mojtaba Valipour, Bowen You, Maysum Panju, and Ali Ghodsi. SymbolicGPT: A generative transformer model for symbolic regression. *arXiv preprint arXiv:2106.14131*, 2021.

Martin Vastl, Jonáš Kulhánek, Jiří Kubalík, Erik Derner, and Robert Babuška. SymFormer: End-to-end symbolic regression using transformer-based architecture. *IEEE Access*, 12:37840–37849, 2024.

Alistair White, Niki Kilbertus, Maximilian Gelbrecht, and Niklas Boers. Stabilized neural differential equations for learning dynamics with explicit constraints. *Advances in Neural Information Processing Systems*, 36:12929–12950, 2023.

Philippe Wyder, Judah Goldfeder, Alexey Yermakov, Yue Zhao, Stefano Riva, Jan Williams, David Zoro, Amy Rude, Matteo Tomasetto, Joe Germany, Joseph Bakarji, Georg Maierhofer, Miles Cranmer, and Nathan Kutz. Common Task Framework For a Critical Evaluation of Scientific Machine Learning Algorithms. In *Advances in Neural Information Processing Systems 38 (NeurIPS 2025)*, pages 188884–188901. Neural Information Processing Systems Foundation, Inc. (NeurIPS), 2025. doi: 10.52202/085713-5677. URL `https://doi.org/10.52202/085713-5677`.

Shengbin Ye and Meng Li. Ab initio nonparametric variable selection for scalable symbolic regression with large p. In *Proceedings of the 42nd International Conference on Machine Learning*, volume 267 of *Proceedings of Machine Learning Research*, pages 72041–72062. PMLR, 2025. URL `https://proceedings.mlr.press/v267/ye25f.html`.

Yuan Yin, Vincent Le Guen, Jérémie Dona, Emmanuel De Bézenac, Ibrahim Ayed, Nicolas Thome, and Patrick Gallinari. Augmenting physical models with deep networks for complex dynamics forecasting. *Journal of Statistical Mechanics: Theory and Experiment*, 2021(12):124012, 2021.

Bob Junyi Zou, Matthew E Levine, Dessi P. Zaharieva, Ramesh Johari, and Emily Fox. Hybrid^2 neural ODE causal modeling and an application to glycemic response. In *Forty-first International Conference on Machine Learning*, 2024.

A Why L^2 Regularization Fails

Here we analyze when and why L^2 regularization guarantees orthogonality. We show that APHYNITY's approach works for fixed symbolic components (Proposition A.1), but fails when symbolic structure is learned jointly with sparsity constraints (Theorem A.2).

Problem Setup. Throughout this appendix, $\mathcal{F}$ is a linear function space (closed under sums and scalar multiplication) sufficiently expressive to contain $f - f_{\text{phy}}$ for the $f_{\text{phy}} \in \mathcal{F}_{\text{phy}}$ at hand; this is what allows the projection arguments below. Consider learning a hybrid model $\hat{f} = \hat{f}_{\text{phy}} + \hat{f}_{\text{aug}}$ where:

- $\hat{f}_{\text{phy}}(x; w) = \sum_{j=1}^{M} w_j \phi_j(x)$ with $w \in \mathbb{R}^M$ (symbolic component).
- $\hat{f}_{\text{aug}}(x; \vartheta)$ is a neural network (neural component); the family $\{\hat{f}_{\text{aug}}(\cdot; \vartheta)\}_\vartheta$ admits a parameter ϑ_0 for which $\hat{f}_{\text{aug}}(\cdot; \vartheta_0) \equiv 0$ (e.g. all final-layer weights set to zero for an MLP without an additive skip connection at the output).

The standard L^2-regularized loss is:

$$\mathcal{L}_{\text{L}^2}(w, \vartheta) = \underbrace{\|f - \hat{f}_{\text{phy}} - \hat{f}_{\text{aug}}\|_{\mathcal{D}}^2}_{\text{reconstruction}} + \underbrace{\mu\|w\|_1}_{\text{sparsity}} + \underbrace{\lambda_2 \|\hat{f}_{\text{aug}}\|_{\mathcal{D}}^2}_{L^2 \text{ regularization}} \tag{10}$$

When L^2 Regularization Guarantees Orthogonality. We first establish the positive result: L^2 regularization works as intended when the symbolic component is fixed. This is a key result of [Yin et al., 2021].

Proposition A.1 (Fixed symbolic hypothesis class)**.** *Assume $\mathcal{F}_{\text{phy}} = \text{span}\{\phi_1, \dots, \phi_M\}$ is fixed, f_{aug} is optimized over $\mathcal{F}$, and $\lambda_2 > 0$. Consider*

$$\min_{f_{\text{phy}} \in \mathcal{F}_{\text{phy}},\, f_{\text{aug}} \in \mathcal{F}} \|f - f_{\text{phy}} - f_{\text{aug}}\|_{\mathcal{D}}^2 + \lambda_2 \|f_{\text{aug}}\|_{\mathcal{D}}^2.$$

Then any minimizer satisfies $f_{\text{phy}} = P^{\mathcal{D}}_{\mathcal{F}_{\text{phy}}}(f)$ and $f_{\text{aug}} = \frac{1}{1+\lambda_2}(f - f_{\text{phy}})$, hence $f_{\text{aug}} \perp \mathcal{F}_{\text{phy}}$.

Proof. For fixed f_{phy}, the objective is quadratic in f_{aug} and is minimized, up to $\mathcal{D}$-equivalence, by $f_{\text{aug}} = \frac{1}{1+\lambda_2}(f - f_{\text{phy}})$. Plugging back reduces the problem to

$$\min_{f_{\text{phy}} \in \mathcal{F}_{\text{phy}}} \|f - f_{\text{phy}}\|_{\mathcal{D}}^2,$$

whose minimizer is the orthogonal projection $P^{\mathcal{D}}_{\mathcal{F}_{\text{phy}}}(f)$. The residual $f - f_{\text{phy}}$ is orthogonal to $\mathcal{F}_{\text{phy}}$, and scaling preserves orthogonality. □

When L^2 Regularization Fails. The situation changes when we learn the symbolic structure itself. The following theorem shows that joint optimization breaks the orthogonality guarantee.

Theorem A.2 (L^2 failure with sparse symbolic learning)**.** *There exist f and a fixed symbolic library $\{\phi_j\}_{j=1}^M$ such that the minimizer of Equation* (10) *is* not *orthogonal, that is, $\hat{f}^*_{\text{aug}} \not\perp \mathcal{F}_{\text{phy}}$, even though $f \in \mathcal{F}_{\text{phy}}$.*

Proof. Let ϕ_1, ϕ_2 be orthonormal w.r.t. $\langle \cdot, \cdot \rangle_{\mathcal{D}}$ and set $f = \phi_1 + \phi_2$, so $f \in \mathcal{F}_{\text{phy}} := \text{span}\{\phi_1, \phi_2\}$. Consider the (function-space) version of Equation (10), with g standing in for $\hat{f}_{\text{aug}}$ to lighten notation:

$$\min_{w \in \mathbb{R}^2,\, g \in \mathcal{F}} \|f - (w_1\phi_1 + w_2\phi_2 + g)\|_{\mathcal{D}}^2 + \mu\|w\|_1 + \lambda_2 \|g\|_{\mathcal{D}}^2.$$

For fixed w, the unique minimizer in g is $g(w) = \frac{1}{1+\lambda_2}(f - w_1\phi_1 - w_2\phi_2)$, and the objective reduces to the Lasso problem

$$\min_{w \in \mathbb{R}^2} \frac{\lambda_2}{1+\lambda_2} \|f - (w_1\phi_1 + w_2\phi_2)\|_{\mathcal{D}}^2 + \mu\|w\|_1.$$

Since ϕ_1, ϕ_2 are orthonormal and $f = \phi_1 + \phi_2$, the problem decouples and for $0 < \mu < \frac{2\lambda_2}{1+\lambda_2}$ the unique minimizer is

$$w_1^* = w_2^* = 1 - \tfrac{\mu(1+\lambda_2)}{2\lambda_2}.$$

Thus

$$g^* = g(w^*) = \tfrac{\mu}{2\lambda_2}(\phi_1 + \phi_2),$$

so $\langle g^*, \phi_j \rangle_{\mathcal{D}} = \frac{\mu}{2\lambda_2} \neq 0$ for $j \in \{1, 2\}$. Hence $g^* \notin \mathcal{F}_{\text{phy}}$ although $f \in \mathcal{F}_{\text{phy}}$. □

B Theoretical Guarantees of OrthoReg

Having established that L^2 regularization can fail under joint optimization with sparsity constraints, we now develop the theory of orthogonal regularization. We show that explicit orthogonality penalties yield quantitative control of neural–symbolic overlap and provide useful geometric structure.

The Orthogonality Constraint. OrthoReg solves:

$$\min_{w,\vartheta} \ \mathcal{L}_{\text{fit}}(w, \vartheta) + \mu \|w\|_1 + \lambda \sum_{j=1}^{M} \langle \hat{f}_{\text{aug}}(\cdot; \vartheta), \phi_j \rangle_{\mathcal{D}}^2, \tag{11}$$

where $\mathcal{L}_{\text{fit}}$ is any non-negative data-fit term (e.g., Equations (4) and (5)). Unlike L^2 regularization, OrthoReg penalizes the squared inner products $\langle \hat{f}_{\text{aug}}, \phi_j \rangle_{\mathcal{D}}^2$ rather than the total norm $\|\hat{f}_{\text{aug}}\|_{\mathcal{D}}^2$.

Penalty method: orthogonality at optimum.

Theorem B.1 (Penalty bound at optimum)**.** *Assume $\lambda > 0$. Let (w^*, ϑ^*) be a global minimizer of Equation* (11) *and let*

$$w_{\text{sym}} \in \arg\min_{w} \ \mathcal{L}_{\text{fit}}(w, \vartheta_0) + \mu \|w\|_1$$

be the best pure symbolic fit, where $\hat{f}_{\text{aug}}(\cdot; \vartheta_0) \equiv 0$. Then

$$\sum_{j=1}^{M} \langle \hat{f}_{\text{aug}}(\cdot; \vartheta^*), \phi_j \rangle_{\mathcal{D}}^2 \leq \frac{1}{\lambda} \Big(\mathcal{L}_{\text{fit}}(w_{\text{sym}}, \vartheta_0) + \mu \|w_{\text{sym}}\|_1 \Big).$$

In particular, for any sequence of global minimizers with $\lambda \to \infty$, the empirical orthogonality violations converge to zero.

Proof. By optimality of (w^*, ϑ^*), we have

$$\begin{aligned} &\mathcal{L}_{\text{fit}}(w^*, \vartheta^*) + \mu \|w^*\|_1 + \lambda \sum_{j=1}^{M} \big\langle \hat{f}_{\text{aug}}(\cdot; \vartheta^*), \phi_j \big\rangle_{\mathcal{D}}^2 \\ &\leqslant \ \mathcal{L}_{\text{fit}}(w_{\text{sym}}, \vartheta_0) + \mu \|w_{\text{sym}}\|_1, \end{aligned}$$

where $\hat{f}_{\text{aug}}(\cdot; \vartheta_0) \equiv 0$ by definition. Dropping the non-negative first two terms on the left yields the stated bound. The final claim follows by sending $\lambda \to \infty$. □

Remark (uniqueness up to $\mathcal{D}$-equivalence). The empirical $\langle \cdot, \cdot \rangle_{\mathcal{D}}$ is a semi-inner product on $\mathcal{F}$: functions agreeing on $\mathcal{D}$ are identified. The minimizers in Proposition A.1 and Theorem 4.1 are therefore unique only up to $\mathcal{D}$-equivalence; this matters most when $\mathcal{D}$ does not separate the augmentation hypothesis class on $\mathcal{X}$, and is mild in our reported regimes where the empirical samples are dense.

Approximation Quality. Beyond enforcing orthogonality, we want to understand what orthogonality buys in terms of approximation structure. We decompose the error into symbolic and orthogonal-complement components with the orthogonal decomposition:

Theorem B.2 (Orthogonal error decomposition)**.** *Let $\mathcal{F}_{\text{phy}}$ be a closed linear subspace of $(\mathcal{F}, \langle\cdot,\cdot\rangle_{\mathcal{D}})$ and let $P^{\mathcal{D}}_{\mathcal{F}_{\text{phy}}}$ denote the orthogonal projection onto $\mathcal{F}_{\text{phy}}$ with respect to $\langle\cdot,\cdot\rangle_{\mathcal{D}}$. For any $f_{\text{phy}} \in \mathcal{F}_{\text{phy}}$ and any $f_{\text{aug}} \in \mathcal{F}$ satisfying $f_{\text{aug}} \perp \mathcal{F}_{\text{phy}}$, we have*

$$\begin{aligned}\|f - (f_{\text{phy}} + f_{\text{aug}})\|^2_{\mathcal{D}} = \|P^{\mathcal{D}}_{\mathcal{F}_{\text{phy}}}(f) - f_{\text{phy}}\|^2_{\mathcal{D}} \\ + \|f - P^{\mathcal{D}}_{\mathcal{F}_{\text{phy}}}(f) - f_{\text{aug}}\|^2_{\mathcal{D}}.\end{aligned}$$

Proof. Write $f = P^{\mathcal{D}}_{\mathcal{F}_{\text{phy}}}(f) + (f - P^{\mathcal{D}}_{\mathcal{F}_{\text{phy}}}(f))$ with $P^{\mathcal{D}}_{\mathcal{F}_{\text{phy}}}(f) \in \mathcal{F}_{\text{phy}}$ and $f - P^{\mathcal{D}}_{\mathcal{F}_{\text{phy}}}(f) \perp \mathcal{F}_{\text{phy}}$. Then

$$\begin{aligned}f - (f_{\text{phy}} + f_{\text{aug}}) = \big(P^{\mathcal{D}}_{\mathcal{F}_{\text{phy}}}(f) - f_{\text{phy}}\big) \\ + \big(f - P^{\mathcal{D}}_{\mathcal{F}_{\text{phy}}}(f) - f_{\text{aug}}\big),\end{aligned}$$

where the first term lies in $\mathcal{F}_{\text{phy}}$ and the second term lies in $\mathcal{F}^{\perp}_{\text{phy}}$, hence they are orthogonal. Expanding the squared norm yields the identity. □

B.1 Finite-Sample Analysis

Our theoretical analysis so far has operated under the empirical inner product. In finite-sample settings we enforce orthogonality empirically, and here we quantify how well empirical inner products approximate their population counterparts.

Empirical vs Population Orthogonality.

Definition B.3 (Orthogonality Gap)**.** Define the empirical inner product

$$\langle f, g\rangle_{\mathcal{D}} = \frac{1}{N}\sum_{i=1}^{N} f(x_i)^{\top} g(x_i),$$

and the population inner product

$$\langle f, g\rangle_{\nu} = \mathbb{E}_{x\sim\nu}[f(x)^{\top} g(x)].$$

The orthogonality gap is $|\langle f, g\rangle_{\mathcal{D}} - \langle f, g\rangle_{\nu}|$.

The following concentration result applies to fixed functions f and g, that is, functions that do not depend on the specific sample $\mathcal{D}$.

Theorem B.4 (Finite-Sample Guarantee)**.** *Let $f, g : \mathcal{X} \to \mathbb{R}^n$ be fixed functions and assume $|f(x)^{\top} g(x)| \le R$ almost surely. Let $\mathcal{D} = \{x_1, \ldots, x_N\}$ be i.i.d. samples from distribution ν. Then with probability at least $1 - \delta$:*

$$|\langle f, g\rangle_{\mathcal{D}} - \langle f, g\rangle_{\nu}| \le R\sqrt{\frac{2\log(2/\delta)}{N}}.$$

Proof. Define random variables $Z_i = f(x_i)^{\top} g(x_i)$ for $i = 1, \ldots, N$, where $x_i \sim \nu$ are i.i.d. and satisfy $Z_i \in [-R, R]$ almost surely. Then

$$\langle f, g\rangle_{\mathcal{D}} = \frac{1}{N}\sum_{i=1}^{N} Z_i, \qquad \langle f, g\rangle_{\nu} = \mathbb{E}[Z_i].$$

Hoeffding's inequality states that for i.i.d. random variables $Z_i \in [a, b]$,

$$\mathbb{P}\left(\left|\frac{1}{N}\sum_{i=1}^{N} Z_i - \mathbb{E}[Z_i]\right| \ge \epsilon\right) \le 2\exp\left(-\frac{2N\epsilon^2}{(b-a)^2}\right).$$

Applying this with $a = -R$, $b = R$ and solving

$$2\exp(-2N\epsilon^2/4R^2) = \delta$$

yields

$$\epsilon = R\sqrt{2\log(2/\delta)/N}.$$

□

Corollary B.5 (Population Orthogonality)**.** *If we enforce $\langle \hat{f}_{\text{aug}}, \phi_j \rangle_{\mathcal{D}} = 0$ empirically, then under the assumption that $\hat{f}_{\text{aug}}$ is fixed independently of $\mathcal{D}$, we have*

$$|\langle \hat{f}_{\text{aug}}, \phi_j \rangle_{\nu}| \leq R\sqrt{\frac{2\log(2/\delta)}{N}}$$

with probability $1 - \delta$.

Intuition. Each inner product

$$\langle \hat{f}_{\text{aug}}, \phi_j \rangle_{\mathcal{D}} = \frac{1}{N}\sum_{i=1}^{N} \hat{f}_{\text{aug}}(x_i)^\top \phi_j(x_i)$$

is an average of N bounded random variables. Hoeffding's inequality for bounded variables $Z_i \in [-R, R]$ yields concentration around the population mean at rate $O(1/\sqrt{N})$.

Remark. Since $\hat{f}_{\text{aug}}$ is trained on the same sample used to evaluate $\langle \hat{f}_{\text{aug}}, \phi_j \rangle_{\mathcal{D}}$, the independence assumption is violated. A population-level guarantee for the *learned* augmentation therefore requires uniform convergence over the augmentation hypothesis class (e.g. Rademacher or VC-style bounds), and not single-function concentration; we discuss this in Section J.

B.2 Monte Carlo Approximation

In training we estimate the empirical inner products in the orthogonality penalty using minibatches. We quantify the resulting stochastic approximation error.

Minibatch estimator. Let $\mathcal{D} = \{x_i\}_{i=1}^N$ and fix a constraint ϕ_j. For a minibatch $\mathcal{B} \subseteq \{1, \ldots, N\}$ of size B, define

$$\langle \hat{f}_{\text{aug}}, \phi_j \rangle_{\mathcal{B}} := \frac{1}{B}\sum_{i \in \mathcal{B}} \hat{f}_{\text{aug}}(x_i)^\top \phi_j(x_i),$$

$$\langle \hat{f}_{\text{aug}}, \phi_j \rangle_{\mathcal{D}} := \frac{1}{N}\sum_{i=1}^{N} \hat{f}_{\text{aug}}(x_i)^\top \phi_j(x_i).$$

Assume $|\hat{f}_{\text{aug}}(x)^\top \phi_j(x)| \leq R$ for all $x \in \mathcal{D}$.

Lemma B.6 (Minibatch deviation)**.** *If $\mathcal{B}$ is sampled uniformly (with replacement) from $\mathcal{D}$, then for any $\delta \in (0, 1)$,*

$$\mathbb{P}\Big(\big|\langle \hat{f}_{\text{aug}}, \phi_j \rangle_{\mathcal{B}} - \langle \hat{f}_{\text{aug}}, \phi_j \rangle_{\mathcal{D}}\big| \geq \epsilon\Big) \leq 2\exp\left(-\frac{B\epsilon^2}{2R^2}\right),$$

and equivalently, with probability at least $1 - \delta$,

$$\big|\langle \hat{f}_{\text{aug}}, \phi_j \rangle_{\mathcal{B}} - \langle \hat{f}_{\text{aug}}, \phi_j \rangle_{\mathcal{D}}\big| \leq R\sqrt{\frac{2\log(2/\delta)}{B}}.$$

The same scaling holds for sampling without replacement (up to a finite-population correction factor).

Implication for optimization. Lemma B.6 shows that the minibatch inner products concentrate around their full-data counterparts at rate $\mathcal{O}(B^{-1/2})$. In the default configuration the orthogonality penalty is evaluated on the full training set each step (all trajectories fit into one batch). For completeness, we record a standard concentration bound for minibatch estimation.

C When Residual Pipelines Can Reintroduce Symbolic Overlap

We revisit the setting of Example 1.1. Let $\phi_1, \phi_2 \in \mathcal{F}$ be orthonormal with respect to $\langle \cdot, \cdot \rangle_{\mathcal{D}}$, let $f = \phi_1 + \phi_2$, and let the Stage-1 symbolic model be fitted with an L^1 penalty of weight $\mu \in (0, 2)$. The orthonormal Lasso solution is

$$\hat{w}_j = 1 - \mu/2, \qquad j = 1, 2,$$

and therefore leaves the empirical residual

$$r := f - \sum_{j=1}^{2} \hat{w}_j \phi_j = \frac{\mu}{2}(\phi_1 + \phi_2) \in \operatorname{span}\{\phi_1, \phi_2\}.$$

Thus, the residue created by sparsity regularization lies exactly in the symbolic span. The question is whether a neural residual fitted after, or through, the symbolic component is forced to remain orthogonal to this span.

Proposition C.1 (Unconstrained residual learning and compositional pipelines can retain symbolic overlap)**.** *In the setting above, the following hold.*

1. ***Unconstrained residual learning.*** *If w is fitted by Lasso and an unconstrained augmentation $g \in \mathcal{F}$ is subsequently fitted to the residual, then*

$$\hat{g} = r, \qquad \langle \hat{g}, \phi_j \rangle_{\mathcal{D}} = \mu/2, \qquad j = 1, 2.$$

2. ***Compositional architecture with an input-space penalty.*** *Consider a compositional architecture*

$$\hat{f}(x) = (I + N_\vartheta)(T_w(x)) = T_w(x) + N_\vartheta(T_w(x)), \qquad T_w(x) = \sum_{j=1}^{2} w_j \phi_j(x).$$

For this simplified example, T_w denotes the symbolic component itself; in compositional models with a residual structure $\hat{f} = T_w + N_\vartheta \circ T_w$, T_w is the physics map onto which N_ϑ composes. Define the effective additive augmentation at the Lasso solution $\hat{w}$

$$\widetilde{N}_\vartheta(x) := N_\vartheta(T_{\hat{w}}(x)).$$

Assume that, at the Lasso solution $\hat{w}_j = 1 - \mu/2$, the $2N$ points

$$\{x_i\}_{i=1}^{N} \cup \{T_{\hat{w}}(x_i)\}_{i=1}^{N}$$

are all distinct. Then there exists a sufficiently expressive N_ϑ such that

$$\hat{f}(x_i) = f(x_i), \qquad N_\vartheta(x_i) = 0, \qquad \widetilde{N}_\vartheta(x_i) = r(x_i)$$

for all $i = 1, \ldots, N$. Consequently,

$$\sum_i N_\vartheta(x_i)^\top \phi_j(x_i) = 0, \qquad j = 1, 2, \quad \textit{but} \quad \langle \widetilde{N}_\vartheta, \phi_j \rangle_{\mathcal{D}} = \langle r, \phi_j \rangle_{\mathcal{D}} = \mu/2.$$

Thus, an OrthoReg-style penalty applied to N_ϑ at the original inputs $\{x_i\}$ does not control the additive overlap of the effective augmentation $N_\vartheta \circ T_{\hat{w}}$.

Proof. For unconstrained residual learning, orthonormality gives the usual soft-thresholding solution $\hat{w}_j = (1 - \mu/2)_+$. Since $\mu \in (0, 2)$, this gives $\hat{w}_j = 1 - \mu/2$, and an unconstrained augmentation fits the residual exactly:

$$\hat{g} = f - \sum_{j=1}^{2} \hat{w}_j \phi_j = \frac{\mu}{2}(\phi_1 + \phi_2).$$

Taking empirical inner products with ϕ_1 and ϕ_2 gives $\langle \hat{g}, \phi_j \rangle_{\mathcal{D}} = \mu/2$.

For the compositional case, exact fit at the Lasso coefficients requires

$$N_\vartheta(T_{\hat{w}}(x_i)) = f(x_i) - T_{\hat{w}}(x_i) = r(x_i).$$

This constrains N_ϑ at the transported points $T_{\hat{w}}(x_i)$, not at the original data points x_i. Since the $2N$ points are distinct, a sufficiently expressive N_ϑ can interpolate both

$$N_\vartheta(T_{\hat{w}}(x_i)) = r(x_i) \quad \text{and} \quad N_\vartheta(x_i) = 0.$$

The empirical penalty on $N_\vartheta(x_i)$ is then zero, while the effective additive augmentation satisfies

$$\widetilde{N}_\vartheta(x_i) = N_\vartheta(T_{\hat{w}}(x_i)) = r(x_i),$$

and hence $\langle \widetilde{N}_\vartheta, \phi_j \rangle_{\mathcal{D}} = \mu/2$. □

Orthogonality-regularized residual learning. The failure above is not caused by staging itself, but by fitting the residual without an orthogonality constraint. If Stage 2 instead minimizes

$$\|r - g\|_{\mathcal{D}}^2 + \lambda \sum_j \langle g, \phi_j \rangle_{\mathcal{D}}^2,$$

then, on the orthonormal toy, $\langle \hat{g}, \phi_j \rangle_{\mathcal{D}} = (\mu/2)/(1+\lambda) \to 0$ as $\lambda \to \infty$. Orthogonalization therefore suppresses symbolic–neural overlap whether imposed jointly or in an explicitly regularized residual stage; it does not remove the L^1 shrinkage bias on w, so the shrinkage residue is left as fit error rather than absorbed by the neural component.

D Computational Complexity

We briefly discuss the additional computational overhead introduced by the orthogonality penalty.

Per-iteration cost. Let B be the number of states evaluated per step (all time points of all trajectories in the batch), n the state dimension, and M the number of library terms. For the vector-field regression loss used in the main experiments, computing the penalty requires evaluating $\phi_j(x_i)$ and forming inner products $\hat{f}_{\mathrm{aug}}(x_i)^\top \phi_j(x_i)$ for all $i \in [B]$ and $j \in [M]$, which costs $\mathcal{O}(B\,n\,M)$ arithmetic operations in addition to the usual network forward/backward passes; backpropagation has the same order. Rollout or sequence-prediction losses multiply this cost by the number of evaluated transitions or time steps.

Timing protocol. Wall-clock per epoch is measured as the median over a fixed window of training epochs after a short warm-up, on the modified damped pendulum (one batch of all 1000 training trajectories, $n = 2$, $M = 18$), with identical optimizer state and data loaders across configurations. The reported 5–15% overhead is the relative increase of the OrthoReg median over the Hybrid L^2 median across five seeds, isolating the cost of the orthogonality term (the inner products g_j of Algorithm 1, Section H); overhead grows linearly in M at fixed (B, n).

Memory requirements. The additional memory is dominated by storing the evaluated library features on the batch, which scales as $\mathcal{O}(B\,n\,M)$ floats. For example, per 1000 evaluated states with $n = 2$ and $M = 18$ this corresponds to approximately $3.6{\times}10^4$ scalars (about 144 KB in float32); on the SIR system ($n = 3$, $M = 45$) the same calculation gives roughly 540 KB per 1000 states. Both scale linearly with B and stay below the network activations cached for backpropagation, which also scale with B.

Mini-batch gradient bias. The penalty $\sum_j \langle \hat{f}_{\mathrm{aug}}, \phi_j \rangle_{\mathcal{D}}^2$ is a squared mean of bounded random variables, so the standard one-batch unbiased estimator of $\nabla_\vartheta \langle \hat{f}_{\mathrm{aug}}, \phi_j \rangle_{\mathcal{D}}$ (which would multiply two independent mini-batch estimates) is replaced in our implementation by the squared single-batch average. Under stochastic gradient descent this introduces an additive bias of order $\mathcal{O}(1/B)$ relative to the full-data quantity (exactly the variance of the per-batch inner-product estimate); the full-batch evaluation used in the main experiments (B equal to the full training set, see Section G) makes this bias zero by construction. A two-batch unbiased estimator is a drop-in alternative for memory-constrained settings.

Hardware and total compute. All reported experiments were run on an internal HPC cluster, allocating one general-purpose GPU per job, ≈ 64 GB of RAM, and a small CPU allocation (typically ≤ 8 cores). The workloads are intentionally small—autonomous ODEs in $n \leq 4$ states, MLPs with two hidden layers (width 32–200), 2000 Adam epochs, and batches of up to 1024 trajectories (the full training set of 1000 trajectories)—so that OrthoReg remains CPU-tractable; we used GPUs primarily to accelerate the neural ODE component, which dominates wall-clock time. Each individual training run completes within a small number of GPU-minutes; the full reported grid (main systems, ablations, and cross-system validation, each repeated over five seeds) accounts for the bulk of the project's compute, and preliminary tuning and discarded runs add a similar order-of-magnitude overhead on top.

E Additional Ablations

E.1 Full Dataset Difficulty Ablation

Table 4: Dataset difficulty ablation across missing dynamics regimes (mean effect strength absent from the symbolic library); MSE normalized by the squared norm of the target signal. OrthoReg trades ID accuracy for higher F1 and lower OOD-T2 error, with the largest OOD-T2 gains at low and medium difficulty; under OOD-T3 the methods are close.

Metric	Pure	L^2	OrthoReg
Low Difficulty ($\beta_1 = 0.1, \beta_2 = 0.08, \beta_3 = 0.05$)			
MSE($\dot{x}_{\mathrm{ID}}$, e-4) (↓)	$\mathbf{7.77 \pm 1.64}$	10.99 ± 0.83	26.82 ± 5.00
MSE($\dot{x}_{\mathrm{OOD,T2}}$, e0) (↓)	88.41 ± 91.81	4.66 ± 0.01	$\mathbf{0.18 \pm 0.04}$
MSE($\dot{x}_{\mathrm{OOD,T3}}$, e$-3$) (↓)	$\mathbf{1.53 \pm 0.26}$	2.01 ± 0.00	2.79 ± 0.09
F1 (↑)	0.50 ± 0.09	0.72 ± 0.05	$\mathbf{0.86 \pm 0.00}$
Medium Difficulty ($\beta_1 = 0.3, \beta_2 = 0.25, \beta_3 = 0.15$)			
MSE($\dot{x}_{\mathrm{ID}}$, e-3) (↓)	$\mathbf{9.26 \pm 0.53}$	10.31 ± 0.40	13.84 ± 0.08
MSE($\dot{x}_{\mathrm{OOD,T2}}$, e0) (↓)	492.39 ± 326.46	126.65 ± 61.35	$\mathbf{3.33 \pm 1.65}$
MSE($\dot{x}_{\mathrm{OOD,T3}}$, e$-2$) (↓)	$\mathbf{1.15 \pm 0.05}$	1.27 ± 0.04	1.54 ± 0.02
F1 (↑)	0.43 ± 0.09	0.61 ± 0.03	$\mathbf{0.93 \pm 0.15}$
High Difficulty ($\beta_1 = 2.0, \beta_2 = 2.0, \beta_3 = 2.0$)			
MSE($\dot{x}_{\mathrm{ID}}$, e-2) (↓)	3.91 ± 0.13	3.90 ± 0.11	$\mathbf{3.86 \pm 0.09}$
MSE($\dot{x}_{\mathrm{OOD,T2}}$, e$-1$) (↓)	5.66 ± 2.68	5.10 ± 1.32	$\mathbf{3.59 \pm 0.88}$
MSE($\dot{x}_{\mathrm{OOD,T3}}$, e$-2$) (↓)	4.49 ± 0.17	4.46 ± 0.10	$\mathbf{4.43 \pm 0.09}$
F1 (↑)	0.46 ± 0.05	0.43 ± 0.04	$\mathbf{0.52 \pm 0.07}$

All OOD rows report normalized derivative MSE; OOD-T2 is evaluated on the trained model and OOD-T3 after the refit (Section G.2).

Table 4 reports the full dataset difficulty ablation across regimes of increasing missing dynamics (controlled by the mean effect strength β of terms absent from the symbolic library; cf. β_i in Equation (9)). In the main paper, Figure 2a summarizes the central trends in out-of-distribution error and symbolic recovery (F1). The complete table reports in-distribution derivative accuracy ($\dot{x}_{\mathrm{ID}}$), initial-condition extrapolation ($\dot{x}_{\mathrm{OOD,T2}}$) and parametric extrapolation after the refit ($\dot{x}_{\mathrm{OOD,T3}}$).

Low difficulty ($\beta_1 = 0.1, \beta_2 = 0.08, \beta_3 = 0.05$). When the true dynamics are close to the symbolic library, all methods achieve low in-distribution errors (Table 4, first block), and the hybrid component is not strongly required. Under parametric extrapolation the pure symbolic model is most accurate, consistent with regularization being least critical when the library is largely sufficient. Under initial-condition extrapolation, however, OrthoReg remains substantially more robust, indicating that enforcing a clean split can mitigate extrapolation failures even when the library mismatch is small.

Medium difficulty ($\beta_1 = 0.3, \beta_2 = 0.25, \beta_3 = 0.15$). This regime shows the most pronounced separation between methods. OrthoReg simultaneously improves initial-condition extrapolation and symbolic recovery, yielding markedly lower $\dot{x}_{\mathrm{OOD,T2}}$ error and substantially higher F1 than both the unregularized baseline and L^2 regularization (Table 4, second block); under parametric extrapolation the three methods are within a factor of 1.4. Conceptually, the orthogonality penalty constrains the augmentation to capture residual structure without redundantly re-learning library terms, stabilizing identification.

High difficulty ($\beta_1 = 2.0, \beta_2 = 2.0, \beta_3 = 2.0$). When the missing dynamics strongly exceed the representational capacity of the library, all methods deteriorate in symbolic recovery and OOD performance (Table 4, third block), reflecting that the symbolic–neural split becomes intrinsically harder to recover. Nevertheless, OrthoReg retains a modest advantage on OOD evaluation and F1. In this regime, once the residual dominates, orthogonality alone is insufficient to recover interpretable structure and the symbolic–neural split becomes harder to identify.

Overall, Table 4 complements Figure 2a by showing that OrthoReg's strongest gains arise when the system *partially* exceeds the library: small mismatches do not demand strong regularization, whereas extreme mismatches limit recoverability for all methods.

Table 5: Sampling scheme ablation (regular vs. irregular); MSE normalized by the squared norm of the target signal. All methods degrade under irregular sampling; OrthoReg retains a small relative advantage on F1.

Metric	Pure	L^2	OrthoReg
Regular Sampling			
MSE($\dot{x}_{\mathrm{ID},\ \mathrm{e}-3}$) (↓)	$\mathbf{9.26 \pm 0.53}$	10.31 ± 0.40	13.84 ± 0.08
MSE($\dot{x}_{\mathrm{OOD,T2},\ \mathrm{e}0}$) (↓)	492.39 ± 326.46	126.65 ± 61.35	$\mathbf{3.33 \pm 1.65}$
MSE($\dot{x}_{\mathrm{OOD,T3},\ \mathrm{e}-2}$) (↓)	$\mathbf{1.15 \pm 0.05}$	1.27 ± 0.04	1.54 ± 0.02
F1 (↑)	0.43 ± 0.09	0.61 ± 0.03	$\mathbf{0.93 \pm 0.15}$
Irregular Sampling			
MSE($\dot{x}_{\mathrm{ID},\ \mathrm{e}0}$) (↓)	3.52 ± 0.00	3.52 ± 0.00	$\mathbf{3.52 \pm 0.00}$
MSE($\dot{x}_{\mathrm{OOD,T2},\ \mathrm{e}1}$) (↓)	1.95 ± 0.15	1.82 ± 0.13	$\mathbf{1.62 \pm 0.08}$
MSE($\dot{x}_{\mathrm{OOD,T3},\ \mathrm{e}0}$) (↓)	3.85 ± 0.00	3.85 ± 0.00	$\mathbf{3.85 \pm 0.00}$
F1 (↑)	0.30 ± 0.01	0.30 ± 0.01	$\mathbf{0.31 \pm 0.01}$

E.2 Full Sampling Scheme Ablation

Table 5 reports the full sampling scheme ablation across regular and irregular observation settings. Irregular sampling substantially increases derivative and OOD prediction errors and reduces symbolic recovery across all methods, consistent with increased solver error and gradient variance under non-uniform time steps (see Section G). Despite this degradation, OrthoReg retains a marginal advantage in symbolic recovery (F1) compared to baseline methods. These results complement the trends shown in Figure 2b and indicate that orthogonal regularization remains effective for symbolic recovery beyond idealized sampling assumptions.

E.3 Regularization-strength sweep

Table 6 lists the in-distribution derivative error and F1 behind Figure 2c for the pendulum (medium difficulty), as a function of the regularization scale s that multiplies λ and μ jointly (Section G.7). At small s the hybrid fits the in-distribution derivatives about as well as the pure symbolic model (1.27×10^{-2}, Table 1); as s grows over three orders of magnitude, in-distribution error increases overall, while F1 rises to an interior optimum and then declines. A smooth in-distribution curve with an interior F1 optimum indicates that the regularization acts as a trade-off between in-distribution fit and decomposition, rather than a sign of difficult optimization.

Table 6: Regularization-strength sweep on the pendulum (OrthoReg). Mean ± standard deviation over runs, separate runs from Table 1.

Scale s	MSE($\dot{x}_{\mathrm{ID},\ \mathrm{e}-2}$)	F1	Runs
0.01	1.24 ± 0.10	0.44 ± 0.08	5
0.02	1.34 ± 0.01	0.49 ± 0.07	5
0.05	1.29 ± 0.06	0.50 ± 0.10	5
0.1	1.25 ± 0.05	0.54 ± 0.05	4
0.2	1.36 ± 0.08	0.58 ± 0.03	5
0.5	1.50 ± 0.01	0.58 ± 0.10	5
1	1.62 ± 0.09	0.65 ± 0.03	5
2	1.70 ± 0.02	0.73 ± 0.04	5
3	1.86 ± 0.02	0.97 ± 0.06	5
4	1.89 ± 0.00	0.94 ± 0.08	5
5	1.96 ± 0.07	0.95 ± 0.11	5
6	1.99 ± 0.03	0.81 ± 0.06	5
8	2.20 ± 0.23	0.84 ± 0.05	5
10	2.27 ± 0.02	0.81 ± 0.06	5

E.4 Monte Carlo Sampling Ablation

We investigate the impact of Monte Carlo sampling on model performance by varying the number of training samples from 100 to 5000. Figure 4 shows the performance across different sample sizes for missing dynamics $\beta_1 = 0.8, \beta_2 = 0.6, \beta_3 = 0.4$.

F1 is stable across sample sizes for OrthoReg, consistent with the $\mathcal{O}(N^{-1/2})$ concentration of the empirical inner products that the OrthoReg penalty operates on (Theorem B.4).

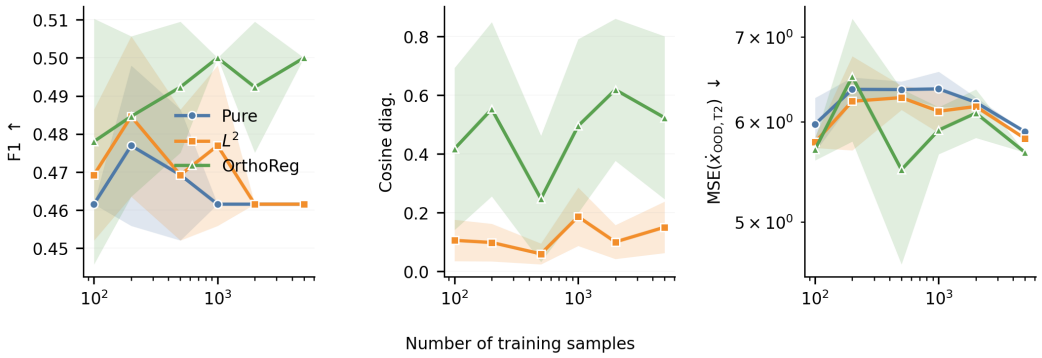

Figure 4: Monte Carlo sampling ablation study ($\beta = (0.8, 0.6, 0.4)$). Performance is shown across different sample sizes (100–5000) for F1 score, OOD-T2 MSE, and a cosine diagnostic $\left|\langle \hat{f}_{\text{aug}}, \hat{f}_{\text{phy}} \rangle_{\mathcal{D}}\right| / (\|\hat{f}_{\text{aug}}\|_{\mathcal{D}} \|\hat{f}_{\text{phy}}\|_{\mathcal{D}})$ between the learned augmentation and the assembled symbolic prediction $\hat{f}_{\text{phy}} = \sum_j \hat{w}_j \phi_j$ on the empirical training distribution. The cosine takes values in $[0, 1]$: smaller values mean $\hat{f}_{\text{aug}}$ and $\hat{f}_{\text{phy}}$ are more directionally separated under the empirical inner product, larger values mean they share direction. We report it to visualize the concentration result of Theorem B.4: at rate $\mathcal{O}(N^{-1/2})$ the empirical inner products that drive the OrthoReg penalty concentrate around their full-data values. The cosine is a within-method stability diagnostic for OrthoReg; it is built from the empirical inner products that appear in the OrthoReg training penalty $\sum_j \langle \hat{f}_{\text{aug}}, \phi_j \rangle_{\mathcal{D}}^2$ but is not the penalty itself.

E.5 Noise Robustness Ablation

We investigate the robustness of OrthoReg, L^2 hybrid, and Pure Symbolic methods to measurement noise in the training data. Gaussian noise with standard deviation $\varepsilon \in \{0.0, 0.01, 0.05, 0.1, 0.2, 0.5\}$ is added to the state observations after trajectory generation, simulating realistic measurement uncertainty.

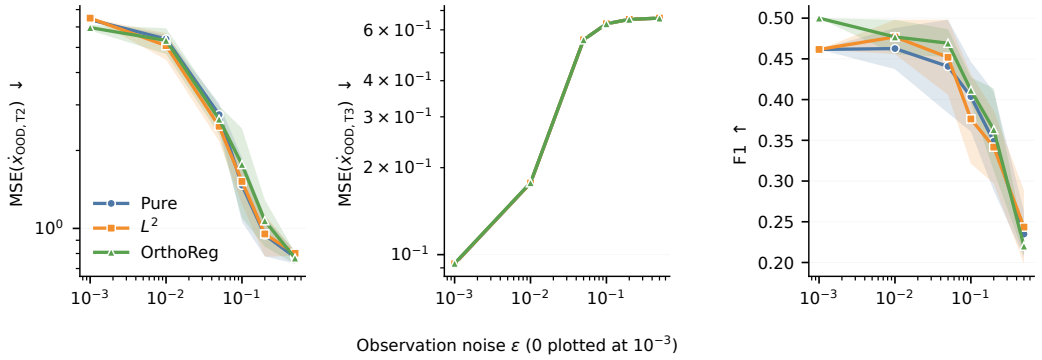

Figure 5: Noise robustness study on the damped pendulum system: normalized OOD-T2 and OOD-T3 derivative MSE and F1 against the observation-noise level (mean $\pm$ standard deviation over runs; $\varepsilon = 0$ is plotted at 10^{-3}). At zero noise OrthoReg has the highest F1 and a slightly lower OOD-T2 error, but it does not provide special robustness to noise compared to the baselines; the three methods are indistinguishable on OOD-T3 and converge on all metrics at high noise ($\varepsilon \geq 0.1$).

Figure 5 shows that all three methods respond similarly to increasing noise levels. At zero noise OrthoReg attains the highest F1 and a slightly lower OOD-T2 derivative error, but these differences vanish as noise increases. At noise levels $\varepsilon \geq 0.1$, all methods converge to similar performance, with OOD-T3 derivative MSE increasing from ~ 0.1 (zero noise) to ~ 0.65 (high noise) and F1 scores dropping from ~ 0.5 to ~ 0.2 across all methods.

Overall, orthogonal regularization does not materially improve robustness to measurement noise in this setting; performance converges across methods at higher noise levels. Improving robustness likely requires changes to the estimation procedure (e.g., denoising or noise-aware differentiation) rather than additional separation constraints. At high noise ($\varepsilon \gtrsim 0.1$), the smoothed-finite-difference

targets used to compute $y_i \approx \dot{x}(t_i)$ are themselves dominated by amplified noise (Savitzky–Golay window of length 5), so the convergence of all methods at this regime partly reflects target corruption rather than failure of the regularizer itself; an analytic-derivative ablation would disentangle the two and is left to future work (Section J).

F Cross-System Validation

The main experiments (Section 5) evaluate OrthoReg on the modified pendulum and Duffing systems. Here we assess transfer to additional systems that introduce different sources of library mismatch, focusing on Lotka–Volterra (temporal modulation) and SIR (state dependence and memory). Together, these systems test distinct mismatch regimes: structured feature mismatch (Duffing), non-autonomous perturbations (Lotka–Volterra), and strongly misspecified state-dependent residual structure (SIR).

Autonomous augmentation with non-autonomous data. Both LV and SIR include explicit time-dependent factors in the ground-truth dynamics, while the augmentation $\hat{f}_{\text{aug}}(x; \vartheta)$ in Equation (7) takes only the state x as input. An autonomous augmentation can therefore fit such factors only as a state-dependent envelope on the training horizon $[0, T]$; the OOD splits below test state extrapolation rather than time extrapolation. Concatenating t with x in the network input is a one-line architectural change that would close this gap.

F.1 Lotka–Volterra System: Temporal Coupling

We evaluate OrthoReg on a modified predator-prey system with temporally modulated and state-dependent interactions. The dynamics are:

$$\frac{dx}{dt} = \alpha x - \beta xy + \varepsilon_1 x \left(\sin(\omega_{\text{fast}} t)\cos(\omega_{\text{fast}} xy)\right) \times \sin\big(\omega_{\text{slow}}(x+y)\big),$$
$$\frac{dy}{dt} = \delta xy - \gamma y + \varepsilon_2 y \left(\sin(\omega_{\text{fast}} t)\cos(\omega_{\text{fast}} xy)\right) \times \sin\big(\omega_{\text{slow}}(x+y)\big) \sin\Big(\frac{x}{y + c_{\text{reg}}}\Big).$$

Here, ε_1 and ε_2 control the strength of missing dynamics not captured by the symbolic feature library, while c_{reg} is a small constant that prevents division by zero. The runs reported below use $(\varepsilon_1, \varepsilon_2) = (3.0, 2.0)$ (see the per-system reproducibility values in Section G). The augmented terms introduce high-frequency temporal modulation, state-dependent coupling, and asymmetric predator-prey interactions; the explicit $\sin(\omega_{\text{fast}} t)$ factors make the system non-autonomous (Section F).

Table 7: Lotka–Volterra results including derivative metrics; MSE normalized by the squared norm of the target signal. Scale factors shown inside the metric parentheses indicate the power of ten applied to the reported values (e.g., e-2). Uncertainty terms reported as ± 0.0 fall below the displayed precision after rescaling. Abbreviations: MSE = mean squared error; ID = in-distribution; OOD = out-of-distribution.

Metric	Pure	L^2	OrthoReg
Predictive Performance			
MSE($\dot{x}_{\text{ID}}$, e-2) (↓)	1.64 ±0.00	1.64 ±0.00	1.64 ±0.00
MSE($\dot{x}_{\text{OOD,T2}}$, e-2) (↓)	7.20 ±0.01	7.14 ±0.02	**7.03 ±0.02**
MSE($\dot{x}_{\text{OOD,T3}}$, e-2) (↓)	**1.21 ±0.00**	1.22 ±0.00	1.22 ±0.00
System Identification Quality			
F1 (↑)	0.39 ±0.02	0.40 ±0.00	**0.43 ±0.01**
#Terms (↓)	16.6 ±0.9	16.0 ±0.0	**14.8 ±0.4**

Results and Analysis. Across all predictive metrics in Table 7, the three methods perform similarly in this configuration, indicating that the chosen library and training distribution already capture most of the dynamics relevant to our evaluation protocol. OrthoReg yields a small but consistent gain in identification quality (F1: 0.43 vs. 0.39) while selecting slightly fewer terms. We therefore view Lotka–Volterra as a "low headroom" setting: it serves as a sanity check that OrthoReg does not degrade performance when the decomposition problem is comparatively easy, rather than as a regime with large expected improvements.

F.2 Duffing oscillator: multistability and cross-basin generalization

This appendix complements the main-text results in Table 3 by detailing the experimental setup and the origin of the cross-basin generalization challenge. Our Duffing configuration and the motivation for the basin split follow Göring et al. [2024].

System and regimes. We consider the unforced Duffing oscillator in first-order form,

$$\dot{x} = y, \tag{12}$$

$$\dot{y} = ay - x(b + cx^2), \tag{13}$$

with parameters $[a, b, c] = [-\frac{1}{2}, -1, \frac{1}{10}]$, placing the dynamics in a bistable regime with two symmetric basins of attraction at $x = \pm\sqrt{-b/c}$ (and $y = 0$). Göring et al. [2024] use this multistable setting to study out-of-domain generalization for learned dynamical-system models.

Why cross-basin generalization is challenging. When training data are restricted to a single basin, the model never observes trajectories from the opposite basin. For multistable flows that are *not topologically transitive* on the test domain, Göring et al. [2024] formalize that the resulting OOD generalization problem is not strictly learnable for universal-approximator hypothesis classes (Theorem 4.2), and empirically illustrate this behavior on the Duffing system.

Library mismatch. The shared symbolic library (polynomial degree $2 \oplus$ Fourier($n = 1$); Section G) does not contain the cubic term cx^3, so any method using this library must either (i) approximate the missing nonlinearity through other terms (pure symbolic) or (ii) represent it through the neural augmentation (hybrid models).

Effect of orthogonal regularization. OrthoReg encourages the neural augmentation to be (approximately) orthogonal to the span of library terms on the training distribution, encouraging the symbolic component to retain the structure expressible by the library while pushing the neural component toward dynamics not captured by it. Empirically, this is associated with sparser and more stable identification compared to L^2-regularized hybrids, while cross-basin state error is lowest for the pure symbolic model (Table 3).

Numerical note on basin crossings. Although the unforced system is basin-confining in the idealized continuous dynamics, numerical integration and initial conditions close to the separatrix can yield occasional apparent basin crossings. In the qualitative visualization we therefore treat trajectories as illustrative, and we generate/inspect trajectories to ensure the intended train-test basin split is respected (cf. Section G for integration details).

F.3 SIR: State Dependence and Memory

System Design. We extend the classical normalized SIR model with a multiplicative *tweaked* term per compartment that combines a state-dependent bilinear coupling, an exponentially decaying time envelope, and a state-modulated frequency. Concretely, with $\omega_S = \omega_{\text{fast}}(1 + S)$, $\omega_I = \omega_{\text{fast}}(1 + 2I)$, $\omega_R = \omega_{\text{slow}}(1 + R)$,

$$\frac{dS}{dt} = -\beta \frac{SI}{S + I + R} \; + \; \varepsilon_1 \, S \, I \, \sin(\omega_S t) \, e^{-t/\tau_{\text{memory}}} \; \tanh\big(\omega_{\text{slow}}(I + R)\big), \tag{14}$$

$$\frac{dI}{dt} = +\beta \frac{SI}{S + I + R} - \gamma \, I \; + \; \varepsilon_2 \, I \, R \, \cos(\omega_I t) \, e^{-t/\tau_{\text{memory}}} \; \sin\big(\omega_{\text{fast}} \, \tfrac{SI}{S+I+c}\big), \tag{15}$$

$$\frac{dR}{dt} = +\gamma \, I \; + \; \varepsilon_3 \, S \, R \, \sin(\omega_R t) \, e^{-t/\tau_{\text{memory}}} \; \cos\big(\omega_{\text{slow}} \, \tfrac{SR}{R+c}\big), \tag{16}$$

with parameters $\beta = 2.0$, $\gamma = 1.0$, $(\varepsilon_1, \varepsilon_2, \varepsilon_3) = (0.3, 0.25, 0.2)$, $\omega_{\text{fast}} = 3.0$, $\omega_{\text{slow}} = 1.5$, $\tau_{\text{memory}} = 2.0$, $c = 10^{-6}$ (Section G). Each tweaked term is added (sign $+$) to the corresponding standard SIR derivative, so $\varepsilon_i \to 0$ recovers the classical normalized SIR model exactly. The base rates β, γ are constant; all additional state dependence enters through the tweaked residual

Table 8: SIR results under severe library mismatch. OrthoReg improves OOD-T2 extrapolation and sparsity, but trades off ID and OOD-T3 accuracy.

Metric	Pure	L^2	OrthoReg
Predictive Performance			
MSE($\dot{x}_{\mathrm{ID},\ \mathrm{e}-5}$) (↓)	$\mathbf{2.12 \pm 0.73}$	31.42 ± 4.93	90.64 ± 7.56
MSE($\dot{x}_{\mathrm{OOD,T2},\ \mathrm{e}-3}$) (↓)	6.89 ± 1.86	3.84 ± 2.85	$\mathbf{1.45 \pm 0.22}$
MSE($\dot{x}_{\mathrm{OOD,T3},\ \mathrm{e}-5}$) (↓)	$\mathbf{3.04 \pm 0.49}$	48.74 ± 7.02	147.07 ± 3.78
System Identification Quality			
F1 (↑)	0.14 ± 0.03	0.15 ± 0.03	$\mathbf{0.20 \pm 0.06}$
#Terms (↓)	40.2 ± 3.1	20.4 ± 3.6	$\mathbf{6.8 \pm 3.3}$

terms [Hethcote, 2000, Kucharski et al., 2020]. As in LV, the explicit time factors make this a non-autonomous stress test for the autonomous augmentation (Section F).

The SIR system demonstrates a challenging regime where the symbolic library has a severe mismatch with the true dynamics, a large number of candidate terms are present, and the tweaked term combines state-modulated frequencies $\sin(\omega_k(\cdot)t)$ with a decaying time envelope $e^{-t/\tau_{\mathrm{memory}}}$ that the autonomous augmentation can absorb only as a state-dependent envelope (Section F). While all methods are challenged by this setup, OrthoReg retains an advantage under initial-condition extrapolation (OOD-T2) and in system identification (highest F1, fewest terms), while the pure symbolic model is most accurate under parametric extrapolation (OOD-T3).

F.4 Helmholtz oscillator: a fifth, low-headroom system

System design. We add a damped Helmholtz oscillator with a corrugated potential,

$$\ddot{x} = -\delta\dot{x} - kx - qx^2 + A\cos(\Omega x), \tag{17}$$

with $\delta = 0.06$, $k = 1.6$, $q = 0.16$, $A = 2.0$ and $\Omega = 7$, written as a first-order system in $(x_0, x_1) = (x, \dot{x})$. The symbolic library is the same $\mathrm{PolynomialLibrary}(\mathrm{degree} = 2) \oplus \mathrm{FourierLibrary}(n_{\mathrm{freq}} = 1)$ used for the other two-state systems (9 features per equation, 18 coefficient slots); the corrugation $A\cos(\Omega x)$ with $\Omega = 7$ lies outside it and has to be carried by the residual. The in-library ground truth has four terms: x_1 in the first equation and x_1, x_0, x_0^2 in the second. Unlike limit-cycle systems, the damped single-well dynamics explore a full-dimensional region of state space, so the symbolic coefficients are well identified rather than near-collinear.

Table 9: Helmholtz oscillator. In-library support scores (Section F.5); mean ± standard deviation over five runs.

Metric	Pure	L^2	OrthoReg
MSE($\dot{x}_{\mathrm{ID}}$)	0.957 ± 0.000	$\mathbf{0.202 \pm 0.423}$	0.957 ± 0.001
F1	1.00 ± 0.00	0.89 ± 0.06	1.00 ± 0.00
Recall	1.00 ± 0.00	0.80 ± 0.11	1.00 ± 0.00
FPR	0.00 ± 0.00	0.00 ± 0.00	0.00 ± 0.00
#Terms	4.0 ± 0.0	3.2 ± 0.4	4.0 ± 0.0

Results. Helmholtz is a low-headroom system in the sense of Section F.1: the pure symbolic model already recovers the full in-library support, so the question is whether adding a neural residual destroys this good fit. OrthoReg does not: it recovers all four true terms with no false positives in every run and matches the pure symbolic model in-distribution, whereas the L^2 hybrid drops a true term in four of five runs (Table 9), in exchange for a lower in-distribution error in those runs.

Summary. Taken together, the cross-system experiments show that OrthoReg remains most useful when the library captures part of the dynamics but leaves structured residuals. The gains are modest in low-headroom settings such as LV, confined to symbolic recovery on Duffing, and mixed under the strongly misspecified SIR variant, where OrthoReg improves OOD-T2 and sparsity but sacrifices ID and OOD-T3 accuracy.

F.5 Precision, recall and full-support scores

The F1 scores in the main tables are computed against the in-library ground-truth support (Section G). Table 10 decomposes them into precision, recall and false-positive rate (FPR, the fraction of in-library non-terms that are selected), and additionally reports recall and F1 against the *full* generating support, which also counts the K out-of-library terms as true terms. Since out-of-library terms cannot be selected, full-support precision equals in-library precision and full-support recall is capped at $P/(P+K)$ for every method, where P is the number of true in-library terms; the ranking by full-support F1 is therefore driven by precision and recall on the library. All scores are computed per run from the same runs as the main tables.

Table 10: Support recovery per system: in-library precision, recall and FPR, and recall and F1 against the full generating support. Mean $\pm$ standard deviation over runs.

Metric	Pure	L^2	OrthoReg
Pendulum (3 true in-library terms of 18; $K = 3$ out-of-library)			
Precision	0.28 ± 0.07	0.44 ± 0.03	0.90 ± 0.22
Recall	1.00 ± 0.00	1.00 ± 0.00	1.00 ± 0.00
FPR	0.56 ± 0.19	0.25 ± 0.03	0.04 ± 0.09
Full-support recall	0.50 ± 0.00	0.50 ± 0.00	0.50 ± 0.00
Full-support F1	0.35 ± 0.06	0.47 ± 0.02	0.63 ± 0.07
Duffing (3 true in-library terms of 18; $K = 1$ out-of-library)			
Precision	0.26 ± 0.03	0.53 ± 0.07	0.60 ± 0.09
Recall	0.93 ± 0.15	0.67 ± 0.00	0.67 ± 0.00
FPR	0.55 ± 0.09	0.12 ± 0.03	0.09 ± 0.04
Full-support recall	0.70 ± 0.11	0.50 ± 0.00	0.50 ± 0.00
Full-support F1	0.37 ± 0.05	0.51 ± 0.03	0.54 ± 0.04
Lotka–Volterra (4 true in-library terms of 18; $K = 2$ out-of-library)			
Precision	0.24 ± 0.01	0.25 ± 0.00	0.27 ± 0.01
Recall	1.00 ± 0.00	1.00 ± 0.00	1.00 ± 0.00
FPR	0.90 ± 0.06	0.86 ± 0.00	0.77 ± 0.03
Full-support recall	0.67 ± 0.00	0.67 ± 0.00	0.67 ± 0.00
Full-support F1	0.35 ± 0.01	0.36 ± 0.00	0.38 ± 0.01
SIR (4 true in-library terms of 45; $K = 3$ out-of-library)			
Precision	0.08 ± 0.02	0.09 ± 0.02	0.18 ± 0.09
Recall	0.80 ± 0.21	0.45 ± 0.11	0.25 ± 0.00
FPR	0.90 ± 0.07	0.45 ± 0.08	0.14 ± 0.08
Full-support recall	0.46 ± 0.12	0.26 ± 0.06	0.14 ± 0.00
Full-support F1	0.14 ± 0.03	0.13 ± 0.03	0.15 ± 0.03
Helmholtz (4 true in-library terms of 18; $K = 1$ out-of-library)			
Precision	1.00 ± 0.00	1.00 ± 0.00	1.00 ± 0.00
Recall	1.00 ± 0.00	0.80 ± 0.11	1.00 ± 0.00
FPR	0.00 ± 0.00	0.00 ± 0.00	0.00 ± 0.00
Full-support recall	0.80 ± 0.00	0.64 ± 0.09	0.80 ± 0.00
Full-support F1	0.89 ± 0.00	0.78 ± 0.06	0.89 ± 0.00

Precision–recall operating point. The orthogonality penalty is not a mechanism that removes symbolic terms: it constrains the neural residual to remain orthogonal to the library while coefficients and residual are fitted jointly, and sparsity itself is governed by the L^1 penalty. OrthoReg replaces the L^2 penalty on the residual with an orthogonality penalty; it does not add a term that suppresses symbolic coefficients. Changes in support therefore arise indirectly, and the resulting operating point depends on the degree and type of misspecification rather than being a fixed bias towards precision. On the pendulum, recall is pinned at 1 for all methods and the F1 gain is a precision gain. On Duffing, OrthoReg and the L^2 hybrid miss one of three true terms while Pure recovers nearly all of them, and precision compensates. On Lotka–Volterra all methods recover the full in-library support and OrthoReg selects the fewest spurious terms. On Helmholtz OrthoReg reaches recall 1 with no false positives in every run. Under severe misspecification (SIR), recall decreases for every method and OrthoReg has the lowest recall together with the lowest FPR; here its advantage is a sparser support, and we regard the loss of recall as a genuine limitation in this regime.

Relation to variable pre-screening. In symbolic regression with variable pre-screening, such as PAN [Ye and Li, 2025], a variable excluded before the search permanently removes the true equation from the downstream search space, so minimizing false negatives is the right objective. The hybrid

setting differs: no symbolic term is removed from the search space, symbolic coefficients and residual are optimized jointly, and dynamics that are not represented symbolically can still be captured by the residual. The asymmetry between false negatives and false positives emphasized for pre-screening therefore does not transfer directly, and the appropriate operating point depends on the intended downstream use of the hybrid model.

G Experimental Design and Implementation

Overview. All experiments follow a common protocol: we generate synthetic trajectories from known ground-truth dynamical systems, train models to predict the vector field from state observations, and evaluate predictive accuracy, symbolic recovery, and neural–symbolic separation. To isolate the effect of orthogonal regularization, OrthoReg and the hybrid baselines share the same model structure and training pipeline; they differ only in the form of the regularization applied.

G.1 Evaluation metrics

We report two complementary result metrics:

- **Predictive performance.** Mean-squared error (MSE) on derivatives, computed against the ground-truth vector field (available for all synthetic systems), and, when numerically stable, MSE of long-horizon trajectory rollouts in state space.
- **System identification.** F1 score comparing the recovered symbolic support against the ground-truth library terms after coefficient thresholding at 0.045.

A useful diagnostic is the exploratory cosine $|\langle \hat{f}_{\mathrm{aug}}, \hat{f}_{\mathrm{phy}} \rangle_{\mathcal{D}}| / (\|\hat{f}_{\mathrm{aug}}\|_{\mathcal{D}} \|\hat{f}_{\mathrm{phy}}\|_{\mathcal{D}})$ between the learned augmentation and the assembled symbolic prediction $\hat{f}_{\mathrm{phy}} = \sum_j w_j \phi_j$. This quantity is an indirect proxy for the OrthoReg training penalty $\sum_j \langle \hat{f}_{\mathrm{aug}}, \phi_j \rangle_{\mathcal{D}}^2$ (the two coincide up to a library reweighting only when w concentrates on a single term) and is numerically unstable when either component has small empirical norm, so we do not report it as a result and do not use it to rank methods.

G.2 In-distribution and out-of-distribution evaluation

We distinguish four evaluation regimes:

- **In-distribution (ID):** test trajectories drawn from the same distribution of initial conditions and time points as training;
- **In-distribution, time-extrapolated (ID,ext):** same initial-condition distribution as training, evaluated on an extended window $[0, T_{\mathrm{ext}}]$ with $T_{\mathrm{ext}} > T$ on the autonomous continuation of the trajectory; this stresses temporal extrapolation while keeping the IC distribution fixed;
- **OOD-T2 (initial-condition extrapolation, "IC-Ext").** trajectories initialized from a wider range of initial conditions than training (for Duffing, the opposite basin);
- **OOD-T3 (parametric extrapolation, "Param-Ext").** trajectories generated from perturbed system parameters while keeping the symbolic library fixed, probing extrapolation under model mismatch. Before OOD-T3 is scored, the trained model is refit for a short continuation of training on time-extended trajectories of the perturbed system; OOD-T2 and the ID splits are scored on the trained model.

Tables retain the OOD-T2/OOD-T3 column labels for compactness; the regime/basin axis corresponds to the Duffing cross-basin split (Section F.2). Hyperparameters, including λ, are chosen per system and method from small sweeps (Section G.8).

G.3 Data generation and derivative targets

Trajectories are generated by numerically integrating the ground-truth vector field using `scipy.integrate.odeint` (LSODA adaptive solver) and then sampled to obtain state observations $\{x(t_i)\}$. Models are trained using the vector-field regression objective (Equation (4)) with derivative targets $y_i \approx \dot{x}(t_i)$ computed via smoothed finite differences (PySINDy

`SmoothedFiniteDifference`, second-order accuracy with Savitzky–Golay smoothing, window length 5).

Temporal discretization. For the pendulum experiments we simulate trajectories over a horizon $T = 6.0$ with $n_T = 600$ time points, yielding a step size $\Delta t = T/(n_T - 1) \approx 0.010$. Other systems use their respective simulation horizons and discretizations as specified by the experimental configuration.

G.4 Model classes and symbolic library

Neural augmentation. Hybrid models use a multilayer perceptron that takes the library features as input, with two hidden layers, layer normalization and $\tanh$ activations. The hidden width is set per system; for the pendulum it is 32. The PINN- and UODE-style baselines in Table 2 use the same hidden layers and width, so the neural-capacity comparison is at parity. Network weights use orthogonal initialization (gain 0.01) with zero biases; symbolic coefficients use the PyTorch default.

Symbolic library. We use a fixed feature library combining a polynomial basis (degree 2, with cross-terms for multivariate systems) and a Fourier basis ($n_{\text{frequencies}} = 1$, i.e. $\sin$ and $\cos$ at the unit frequency). Throughout, M counts scalar feature–component pairs (basis features times state dimension n), so for the pendulum ($n = 2$) with poly-2 $\oplus$ Fourier-1 we have $M = 18$, and for SIR ($n = 3$) the same construction gives $M = 45$. Symbolic coefficients are optimized with an L^1 penalty and thresholded at 0.045 when computing support recovery metrics.

G.5 Optimization and effective batch size

All models are optimized using Adam. The derivative regression phase is trained for 2000 epochs with learning rate 0.0089. In the default Monte Carlo setup we use $n_{\text{samples}} = 1000$ trajectories, all of which fit into a single optimization batch; thus, each epoch effectively processes the full training set. When minibatching is used, the batch size is stated explicitly.

G.6 Baselines

Pure symbolic (SINDy-like). Sparse regression over the symbolic library with an L^1 penalty, trained using the same optimizer infrastructure as the hybrid methods to ensure comparability. The canonical SINDy baseline uses sequential thresholded least squares (STLSQ).

L^2-regularized hybrid. Joint optimization over symbolic coefficients w and neural parameters ϑ with loss

$$\mathcal{L} = \mathcal{L}_{\text{fit}}(w, \vartheta) + \mu\|w\|_1 + \lambda_2\|\hat{f}_{\text{aug}}\|^2_{\mathcal{D}}.$$

OrthoReg. Setup identical to the L^2 hybrid baseline except that the neural regularizer is replaced by

$$\lambda \sum_{j=1}^{M} \langle \hat{f}_{\text{aug}}, \phi_j \rangle^2_{\mathcal{D}}.$$

PINN- and UODE-style baselines. Both baselines use the same library, L^1 penalty ($\mu = 0.012$), threshold, network (two hidden layers of width 32, layer normalization, $\tanh$), epochs and OOD-T3 refit as OrthoReg, and differ only in how the library and the network are coupled (coupling weights at the implementation defaults). The PINN-style model fits a network $u_\psi(t, x)$ to the derivative targets with the physics residual $\|u_\psi - f_{\text{phy}}\|^2$, predicts with u_ψ and reads the support from f_{phy}. The UODE-style model predicts with $f_{\text{phy}} + f_{\text{aug}}$ and penalizes $0.1\,\mathbb{E}\|f_{\text{aug}}\|_2$. Intuitively, the PINN-style residual asks the library to explain everything the network fits, including the out-of-library forcing, which it approximates with extra terms, while the UODE-style penalty keeps the residual small but not orthogonal, so the residual can take over an in-library term, as in the L^2 hybrid; OrthoReg removes exactly this freedom. Baseline columns are means over five seeds; the OrthoReg column

Table 11: PySINDy STLSQ versus Pure (Adam L^1) and OrthoReg (Tables 1, 3, 7 and 8). †: divergent rollout or clamp.

Metric	STLSQ	Pure	OrthoReg
Pendulum (Equation (9))			
MSE($\dot{x}_{\mathrm{ID}}$, e-2)	**1.17 ±0.00**	1.27 ±0.07	1.89 ±0.01
F1	0.70 ±0.00	0.43 ±0.09	**0.93 ±0.15**
#Terms	9.0 ±0.0	11.4 ±2.9	**3.6 ±1.3**
Duffing (Section F.2)			
MSE(x_{ID}, e-1)	†	**4.75 ±0.13**	5.59 ±0.09
MSE($x_{\mathrm{OOD,T2}}$, e-1)	†	**6.77 ±1.32**	9.08 ±0.19
F1	0.62 ±0.00	0.40 ±0.05	**0.63 ±0.05**
#Terms	7.0 ±0.0	11.0 ±1.6	**3.4 ±0.5**
Lotka–Volterra (Section F.1)			
F1	0.36 ±0.00	0.39 ±0.02	**0.43 ±0.01**
SIR (Section F.3)			
F1	0.18 ±0.00	0.14 ±0.03	**0.20 ±0.06**

repeats Table 1. For reference, giving the baselines the true base pendulum equation instead of the library (information OrthoReg does not use) lowers their derivative error below OrthoReg's on every split, and a physics-free MLP likewise fits more closely; neither identifies a symbolic model from the library, which is the task OrthoReg addresses.

Canonical PySINDy STLSQ (selection–refit baseline). The Pure Symbolic rows in the main tables (Tables 1, 3, 7 and 8) use the same continuous L^1+Adam sparse-regression pipeline as the hybrid models (Section G.6), which gives a controlled comparison within one optimization stack. As a canonical SINDy-style selection–refit reference, we additionally run PySINDy `STLSQ` on the same symbolic libraries and train/ID/OOD splits: `SequentialThresholdedLeastSquares(threshold=0.045, normalize_columns=False)` with second-order finite differences (`FiniteDifference(order=2)`). Where trajectory rollouts diverged or hit the numerical clamp, state MSE is marked †; STLSQ standard deviations are zero because the splits are deterministic and the fits are bitwise identical across seeds.

The comparison underscores why OrthoReg targets partial library misspecification rather than purely symbolic reconstruction. STLSQ reduces shrinkage bias and improves support recovery relative to the continuous L^1+Adam Pure baseline on the pendulum and Duffing systems. However, because STLSQ has no residual component, out-of-library dynamics must be approximated within the fixed dictionary, which can lead to unstable rollouts even when derivative-fit or F1 diagnostics look reasonable. OrthoReg instead couples sparse symbolic recovery with a flexible residual discouraged from using library directions, yielding stronger pendulum recovery and, on Duffing, finite hybrid rollouts where STLSQ rollouts diverge. The LV and SIR results show the complementary limitation: when the library mismatch is severe or the evaluation is dominated by time-dependent perturbations, F1 alone is not a reliable summary of dynamical performance.

G.7 Per-system reproducibility values

For each system we list (i) the ground-truth parameters used for training trajectories, (ii) the perturbation used to generate the OOD-T3 split, and (iii) the simulation horizon and step size. Final regularization hyperparameters are listed at the end of the block.

- **Modified damped pendulum** (Equation (9), dataset `pendulum.yaml`):
  - parameters $\omega_0 = 1.0$, $\alpha = 0.2$, $(\beta_1, \beta_2, \beta_3) = (0.3, 0.25, 0.15)$;
  - OOD-T3 perturbation: all parameters scaled by 1.2;
  - $T = 6.0$, $n_T = 600$, $\Delta t \approx 0.010$.
- **Duffing oscillator** (Section F.2, dataset `duffing.yaml`):
  - parameters $a = -0.5$, $b = -1.0$, $c = 0.1$ (cubic term cx^3 absent from the dictionary);
  - OOD-T3 perturbation: $(a, b, c) \mapsto (1.2\,a,\ 1.2\,b,\ 2.0\,c)$;

- $T = 40$, $\Delta t = 0.01$.

- **Lotka–Volterra** (Section F.1, dataset `lotka_volterra.yaml`):
  - base parameters $(\alpha, \beta, \gamma, \delta) = (5, 2.5, 5, 2.5)$, modulation frequencies $\omega_{\text{fast}} = 10$, $\omega_{\text{slow}} = 2$; missing-dynamics weights $(\varepsilon_1, \varepsilon_2) = (3.0, 2.0)$ at the base weights (1.5, 1.0) scaled by 2.0;
  - OOD-T3 perturbation: all parameters scaled by 1.2;
  - $T = 6$, $\Delta t \approx 0.01$.
- **Tweaked SIR** (Section F.3, dataset `sir.yaml`):
  - base parameters $\beta = 2.0$, $\gamma = 1.0$; tweaked-term parameters $(\varepsilon_1, \varepsilon_2, \varepsilon_3) = (0.3, 0.25, 0.2)$, $\omega_{\text{fast}} = 3.0$, $\omega_{\text{slow}} = 1.5$, $\tau_{\text{memory}} = 2.0$;
  - OOD-T3 perturbation: all parameters scaled by 1.2;
  - $T = 10$, $\Delta t = 0.1$.
- **Helmholtz oscillator** (Section F.4, dataset `helmholtz.yaml`):
  - parameters $\delta = 0.06$, $k = 1.6$, $q = 0.16$, $A = 2.0$, $\Omega = 7$;
  - OOD-T2: initial-condition box scaled by 1.2; OOD-T3 perturbation: (k, q) scaled by 1.2, same initial conditions;
  - $T = 20$ with 800 time points, $\Delta t \approx 0.025$.
- **Final regularization hyperparameters** (selected per Section G.8):
  - L^2 baseline: base values $\lambda_2 = 0.005$ (on $\|\hat{f}_{\text{aug}}\|^2_{\mathcal{D}}$), symbolic sparsity $\mu = 0.001$;
  - OrthoReg: base values $\lambda = 0.005$ (on the inner-product penalty), symbolic sparsity $\mu = 0.003$;
  - both base values are multiplied by a common regularization scale s chosen per system and method: pendulum $s = 0.85$ (L^2) and $s = 4.0$ (OrthoReg); Duffing $s = 1.0$ and $s = 1.2$; Lotka–Volterra $s = 1.0$ for both; SIR $s = 0.5$ for both. For the Helmholtz oscillator (Section F.4) the L^2 hybrid uses $s = 30$, OrthoReg uses $\lambda = 1.5$ and $\mu = 0.003$, and the pure symbolic model uses $\mu = 0$. The regularization-strength ablation (Figure 2c) varies s, i.e. λ and μ jointly.

G.8 Regularization selection

We select regularization strengths (λ, μ) per system and method from small sweeps. The selected (λ, μ) are fixed across in-distribution and out-of-distribution evaluations; the same protocol is applied to the L^2 baseline with its own grid.

Penalty scale dependence on library norms. The orthogonality penalty $\lambda \sum_j \langle \hat{f}_{\text{aug}}, \phi_j \rangle^2_{\mathcal{D}}$ is scale-dependent in the library: rescaling $\phi_j \to c_j \phi_j$ rescales the per-feature contribution by c_j^2. The reported runs operate on bounded state ranges where the empirical norms $\|\phi_j\|_{\mathcal{D}}$ are comparable across j; for libraries with very different per-feature scales, standardizing features to unit empirical norm (and rescaling λ accordingly) is a robust drop-in.

H Algorithm

Algorithm 1: OrthoReg training (vector-field regression).
Input: samples $\{(x_i, y_i)\}_{i=1}^N$, library $\{\phi_j\}_{j=1}^M$, weights $\lambda, \mu \geq 0$.
Initialize $w \in \mathbb{R}^M$, neural parameters ϑ.
For each epoch and minibatch $\mathcal{B} \subset \{1, \ldots, N\}$ of size B:
1. $\hat{f}(x_i) = \sum_{j=1}^M w_j \, \phi_j(x_i) + \hat{f}_{\text{aug}}(x_i; \vartheta)$.
2. $L_{\text{fit}} = \frac{1}{B} \sum_{i \in \mathcal{B}} \|y_i - \hat{f}(x_i)\|^2$.
3. $g_j = \frac{1}{B} \sum_{i \in \mathcal{B}} \hat{f}_{\text{aug}}(x_i; \vartheta)^\top \phi_j(x_i), \quad L_{\text{orth}} = \lambda \sum_{j=1}^M g_j^2$.
4. $L_{\text{sparse}} = \mu \|w\|_1$.
5. Update (ϑ, w) by autodiff on $L_{\text{fit}} + L_{\text{orth}} + L_{\text{sparse}}$.

I LLM Usage Disclosure

Large language models were used as writing and programming assistants throughout the project, including for text polishing, code development, and improving the presentation of mathematical arguments. The research idea, theoretical results, experimental design, analyses, figures, and manuscript were produced by the authors.

J Limitations

Scope of the guarantee. Our guarantees apply to the additive decomposition $f = f_{\mathrm{phy}} + f_{\mathrm{aug}}$ under the empirical inner product used for training. They therefore do not directly cover compositional architectures such as $\hat{f} = T_w + N_\vartheta \circ T_w$, unconstrained residual-learning pipelines, or population correlations of the learned augmentation without an additional uniform-convergence argument (Proposition C.1 and Section C). A full population-level theory is left for future work.

Empirical-to-population transfer. The penalty bound (Theorem 4.1) and the error decomposition (Theorem B.2) are stated for the empirical inner product $\langle \cdot, \cdot \rangle_{\mathcal{D}}$. A bound on the population correlation $\langle \hat{f}_{\mathrm{aug}}, \phi_j \rangle_\nu$ for the *learned* augmentation cannot be obtained by single-function concentration alone (the function depends on $\mathcal{D}$); it would require uniform convergence over the augmentation hypothesis class (e.g. Rademacher complexity or VC-style arguments). A uniform-convergence statement is left to future work.

Symbolic-discovery regime. OrthoReg is formulated for a fixed, finite, differentiable library and a continuous sparsity relaxation optimized jointly with the neural augmentation. Selection–refit methods such as STLSQ, L^0 search, genetic programming, and transformer-based symbolic search operate in a different regime: after support selection, coefficients are typically refit by unregularized least squares, avoiding shrinkage bias on the retained terms. Combining such solvers with orthogonality-regularized residual learning is a natural extension, but it is not covered by the joint-objective certificate analyzed here. We do suspect the failure is more general: Example 1.1 isolates a coefficient-fitting step that trades exact refit for shrinkage, followed by a residual fit without decorrelation – not linearity per se. We prove this only for the Lasso case and flag the general case as a hypothesis, not a result.

Identifiability of the symbolic split. Orthogonality constrains $\hat{f}_{\mathrm{aug}}$, not w, so it cannot make collinear coefficients identifiable. Take $\phi_1 = \phi_2$ on $\mathcal{D}$ (unit norm) and $f = c\phi_1$ with $c > \mu/2$. The objective depends on w only through $s = w_1 + w_2$, and $\min_s (c - s)^2 + \mu|s|$ has the unique minimizer $s^* = c - \mu/2$, attained by the whole segment $\{w_1 + w_2 = s^*,\ w_1, w_2 \geq 0\}$. The fitted function is unique; the split is not. This is classical Lasso non-uniqueness under collinear columns [Tibshirani, 2013]; away from the exact tie, near-collinearity leaves a unique but ill-conditioned split that resampling perturbs. Identifiability of the split is thus a separate matter from what OrthoReg targets, which is function-space non-redundancy.

Dictionary-dependent interpretability and OOD scope. The recovered symbolic support is mechanistically meaningful only when the chosen library spans a scientifically relevant component of the true vector field. Theorem B.2 separates in-library and orthogonal-complement error under the empirical inner product; consequently, OrthoReg is most useful when missing dynamics are well separated from the library span, and less beneficial when the residual is strongly correlated with library functions on the observed distribution. OrthoReg therefore provides a complementary empirical decomposition, not recovery of governing laws absent from the library.

Derivative supervision and noise. The experiments use vector-field regression with derivative targets estimated from smoothed finite differences. This isolates the decomposition problem but makes high-noise settings sensitive to derivative-estimation error, as seen in Section E.5; in particular, at $\varepsilon \gtrsim 0.1$ the smoothed-FD targets (Savitzky–Golay window of length 5) themselves limit recoverability. State-prediction losses (Equation (5)) or noise-aware differentiation are natural extensions for sensor-level data.

Implementation. The reported implementation optimizes the continuous L^1-regularized objective with Adam and obtains exact sparsity by post-hoc thresholding. The orthogonality penalty is scale-dependent in the library features, so feature standardization is advisable for libraries with very different per-feature scales (Section G.8). The Pure-Symbolic rows use the same continuous sparsity pipeline as the hybrid models for controlled comparison; STLSQ-style pure-symbolic baselines are discussed in Section G.6.

NeurIPS Paper Checklist

1. **Claims**

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

   Answer: [Yes]

   Justification: The abstract and introduction state three claims: (i) under sparse library-based symbolic discovery, L^2 regularization can induce symbolic–neural overlap; (ii) OrthoReg yields a clean in-library / orthogonal-complement split; (iii) on benchmark dynamical systems with partial symbolic-library mismatch, OrthoReg improves symbolic recovery, separates components more cleanly than L^2, and improves out-of-distribution behavior in regimes where the library is partially misspecified. These are matched by Theorem A.2, Theorem 4.1, Theorem B.2, and the experiments in Section 5 and Sections E and F, with scope and caveats stated explicitly in Section 5 and Section J.

   Guidelines:

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

2. **Limitations**

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

   Answer: [Yes]

   Justification: The Conclusion flags the differentiability requirement for symbolic components (which excludes non-gradient methods such as sequential thresholding), and we discuss the additional limitations: dependence on a chosen library, scope of the empirical inner product, evaluation on synthetic ODEs only, and the fact that derivative-space training is used because it was empirically more stable than state-only training.

   Guidelines:

   - The answer [N/A] means that the paper has no limitation while the answer [No] means that the paper has limitations, but those are not discussed in the paper.
   - The authors are encouraged to create a separate "Limitations" section in their paper.
   - The paper should point out any strong assumptions and how robust the results are to violations of these assumptions (e.g., independence assumptions, noiseless settings, model well-specification, asymptotic approximations only holding locally). The authors should reflect on how these assumptions might be violated in practice and what the implications would be.
   - The authors should reflect on the scope of the claims made, e.g., if the approach was only tested on a few datasets or with a few runs. In general, empirical results often depend on implicit assumptions, which should be articulated.
   - The authors should reflect on the factors that influence the performance of the approach. For example, a facial recognition algorithm may perform poorly when image resolution is low or images are taken in low lighting. Or a speech-to-text system might not be used reliably to provide closed captions for online lectures because it fails to handle technical jargon.
   - The authors should discuss the computational efficiency of the proposed algorithms and how they scale with dataset size.
   - If applicable, the authors should discuss possible limitations of their approach to address problems of privacy and fairness.

- While the authors might fear that complete honesty about limitations might be used by reviewers as grounds for rejection, a worse outcome might be that reviewers discover limitations that aren't acknowledged in the paper. The authors should use their best judgment and recognize that individual actions in favor of transparency play an important role in developing norms that preserve the integrity of the community. Reviewers will be specifically instructed to not penalize honesty concerning limitations.

3. **Theory assumptions and proofs**

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

   Answer: [Yes]

   Justification: All theorems and propositions are numbered and cross-referenced. Main-text statements (Theorem 4.1, Theorem B.2) include a proof sketch; the formal versions (Proposition A.1, Theorem A.2, Theorem B.1, Theorem B.2, Proposition C.1) and full proofs are in Sections A to C, with assumptions stated in each theorem's preamble.

   Guidelines:

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

4. **Experimental result reproducibility**

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

   Answer: [Yes]

   Justification: All systems used in the paper are synthetic ODEs with explicit governing equations and trajectory-generation procedures (Section G); architectures, optimizer (Adam), learning rate, batch size, number of epochs, OOD-split protocol, baseline implementations, and hyperparameter-selection procedure are documented in Sections G, G.2, G.6 and G.8.

   Guidelines:

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

(a) If the contribution is primarily a new algorithm, the paper should make it clear how to reproduce that algorithm.
(b) If the contribution is primarily a new model architecture, the paper should describe the architecture clearly and fully.
(c) If the contribution is a new model (e.g., a large language model), then there should either be a way to access this model for reproducing the results or a way to reproduce the model (e.g., with an open-source dataset or instructions for how to construct the dataset).
(d) We recognize that reproducibility may be tricky in some cases, in which case authors are welcome to describe the particular way they provide for reproducibility. In the case of closed-source models, it may be that access to the model is limited in some way (e.g., to registered users), but it should be possible for other researchers to have some path to reproducing or verifying the results.

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

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

   Answer: [Yes]

   Justification: The code is publicly available at `https://github.com/richtertill/OrthoReg`. All datasets used in the paper are synthetic; the governing equations, integration scheme, noise model, and experimental protocol are documented in Section G. The repository contains the implementation, the five benchmark systems and the baselines.

   Guidelines:

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

6. **Experimental setting/details**

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

   Answer: [Yes]

   Justification: Optimizer (Adam), learning rate, number of epochs, network architectures and widths, batch size, and Monte Carlo sample count are reported in Section G; ID/OOD-T2/OOD-T3 split definitions are in Section G.2; baseline implementations (Pure Symbolic, Hybrid L^2, PINN-style, UODE-style) are in Section G.6; and the protocol for selecting λ is in Section G.8.

   Guidelines:

   - The answer [N/A] means that the paper does not include experiments.

- The experimental setting should be presented in the core of the paper to a level of detail that is necessary to appreciate the results and make sense of them.
- The full details can be provided either with the code, in appendix, or as supplemental material.

7. **Experiment statistical significance**

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

   Answer: [Yes]

   Justification: Each experiment is repeated over five stochastic runs with different random seeds (Section 5); reported $\pm$ values in tables are sample standard deviations across these runs, capturing variability from network initialization.

   Guidelines:

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

8. **Experiments compute resources**

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

   Answer: [Yes]

   Justification: Section D reports per-step complexity ($\mathcal{O}(B\,n\,M)$), the wall-clock overhead of OrthoReg over the L^2 baseline (5–15% on the pendulum, measured as a median across five seeds), the additional memory footprint (≈ 144 KB on the pendulum and ≈ 540 KB on SIR, per 1000 evaluated states), and the hardware used (one general-purpose GPU on an internal HPC cluster, ≈ 64 GB RAM, ≤ 8 CPU cores). Per-run wall-clock and an order-of-magnitude estimate of total project compute (including preliminary and discarded runs) are stated in the same paragraph; OrthoReg is also CPU-tractable since the workloads are small (autonomous ODEs, MLPs with two hidden layers (width 32–200), 2000 Adam epochs).

   Guidelines:

   - The answer [N/A] means that the paper does not include experiments.
   - The paper should indicate the type of compute workers CPU or GPU, internal cluster, or cloud provider, including relevant memory and storage.
   - The paper should provide the amount of compute required for each of the individual experimental runs as well as estimate the total compute.

- The paper should disclose whether the full research project required more compute than the experiments reported in the paper (e.g., preliminary or failed experiments that didn't make it into the paper).

9. **Code of ethics**

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

   Answer: [Yes]

   Justification: The work is foundational methodology for hybrid symbolic-neural ODE modeling, evaluated on synthetic dynamical systems generated by the authors. It involves no human subjects, no scraped or sensitive data, no released large pretrained model, and no foreseeable dual-use concern, and we have reviewed the NeurIPS Code of Ethics.

   Guidelines:

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

10. **Broader impacts**

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

    Answer: [N/A]

    Justification: The paper studies a regularizer for hybrid symbolic-neural ODE modeling on synthetic dynamical systems. It is foundational methodology with no immediate deployment context and no direct path to a specific positive or negative societal application; downstream impact will depend on the scientific domain in which a hybrid model is later deployed (e.g., epidemiology, climate, mechanics) rather than on OrthoReg itself.

    Guidelines:

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

11. **Safeguards**

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

Answer: [N/A]

Justification: We do not release any pretrained models, generative models, or scraped datasets. The artifacts associated with the paper are a small open-source codebase (`https://github.com/richtertill/OrthoReg`) and synthetic ODE benchmarks generated from explicit governing equations.

Guidelines:

- The answer [N/A] means that the paper poses no such risks.
- Released models that have a high risk for misuse or dual-use should be released with necessary safeguards to allow for controlled use of the model, for example by requiring that users adhere to usage guidelines or restrictions to access the model or implementing safety filters.
- Datasets that have been scraped from the Internet could pose safety risks. The authors should describe how they avoided releasing unsafe images.
- We recognize that providing effective safeguards is challenging, and many papers do not require this, but we encourage authors to take this into account and make a best faith effort.

12. **Licenses for existing assets**

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

Answer: [Yes]

Justification: We do not use third-party datasets; all benchmark systems are synthetic ODEs generated by us. Methodological precursors and baselines (APHYNITY [Yin et al., 2021], SINDy [Brunton et al., 2016], Universal ODE [Rackauckas et al., 2020], and other works compared against) are cited at first use. Standard open-source scientific-computing libraries are used under their respective permissive licenses; specific package versions are pinned in the released code.

Guidelines:

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

13. **New assets**

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

Answer: [Yes]

Justification: No new datasets or models are released. The OrthoReg implementation is released at `https://github.com/richtertill/OrthoReg` under the MIT license, with a README, dependency specifications and the synthetic-data generation code described in Section G.

Guidelines:

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

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

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

    Answer: [N/A]

    Justification: The paper involves no crowdsourcing and no research with human subjects.

    Guidelines:

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

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

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

    Answer: [N/A]

    Justification: The paper involves no research with human subjects, so no IRB approval is required.

    Guidelines:

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

16. **Declaration of LLM usage**

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

    Answer: [N/A]

    Justification: LLMs are not part of the core methodology of OrthoReg. They were used as writing and programming assistants throughout the project, including for text polishing, code development, and improving the presentation of mathematical arguments; the research idea,

theoretical results, experimental design, analyses, figures, and manuscript were produced by the authors. This is documented in the LLM Usage Disclosure appendix.

Guidelines:

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