

The Observer Effect in World Models: Invasive Adaptation Corrupts Latent Physics

Christian Internò^{1,2*} Jumpei Yamaguchi^{3,2*}
 Loren Amdahl-Cullen⁴ Markus Olhofer^{2†} David Klindt^{5†} Barbara Hammer^{1†}

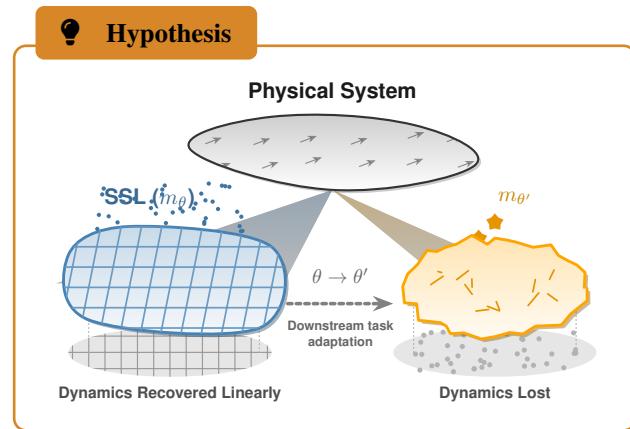
Abstract

Determining whether neural models internalize physical laws as world models, rather than exploiting statistical shortcuts, remains challenging, especially under out-of-distribution (OOD) shifts. Standard evaluations often test latent capability via downstream adaptation (e.g., finetuning or high-capacity probes), but such interventions can change the representations being measured and thus confound what was learned during self-supervised learning (SSL). We propose a non-invasive evaluation protocol, *PhyIP*. We test whether physical quantities are linearly decodable from frozen representations, motivated by the *linear representation hypothesis* (Nanda et al., 2023b). Across fluid dynamics and orbital mechanics, we find that when SSL achieves low error, latent structure becomes linearly accessible. *PhyIP* recovers internal energy and Newtonian inverse-square scaling on OOD tests (e.g., $\rho > 0.90$). In contrast, adaptation-based evaluations can collapse this structure ($\rho \approx 0.05$). These findings suggest that adaptation-based evaluation can obscure latent structures and that low-capacity probes offer a more accurate evaluation of physical world models.

1. Introduction

AI for natural sciences has evolved from computational acceleration to the construction of “World Models” (Ha & Schmidhuber, 2018; LeCun & Courant, 2022), aiming to synthesize vast observational data into representations that encode the governing physical laws of the system (Wang et al., 2023; Bommasani et al., 2022), rather than just statistical shortcuts (Geirhos et al., 2020).

*Equal contribution. †Co-advising. ¹Bielefeld University
²Honda Research Institute EU ³Tokyo Institute of Technology ⁴Simplex, Astera Institute ⁵Cold Spring Harbor Laboratory.
 Correspondence to: Christian Internò <christian.interno@uni-bielefeld.de>.



The aspiration is that a neural network, trained on diverse physical regimes, will implicitly learn the corresponding physical laws, analogous to the historical transition from describing motion (kinematics) to uncovering the forces that drive it (dynamics) (Schmidt & Lipson, 2009b).

However, despite high predictive fidelity within training distributions, neural models often fail to capture universal physical mechanisms (Coveney & Highfield, 2025; Motamed et al., 2025). If a model cannot distinguish universal laws from correlations, its scientific utility remains limited (Chen et al., 2022; Wang et al., 2023).

The prevailing methodology for adapting these models to new downstream tasks is ‘pretrain-then-fine-tune’, where the backbone is updated alongside a randomly initialized head for a specific downstream task (Wortsman et al., 2021; Mai et al., 2024). Recently, this paradigm has been extended to validate intrinsic knowledge. Invasive adaptation and high-capacity nonlinear probes are used to test whether inductive biases in models align with a postulated world model (Vafa et al., 2025; Belinkov, 2022a). However, theoretical works on probing warn that such high-capacity interventions often “learn the task” themselves rather than extracting it (Hewitt & Liang, 2019b; Ravichander et al., 2021), while feature distortion dynamics work suggests that, during adaptation, noisy gradients warp the backbone to fit random initializations rather than the task structure (Kumar et al., 2022; Trivedi et al., 2023).

This degrades physical representations, challenging the *Lin-*

ear Representation Hypothesis (LRH) (Nanda et al., 2023c; DiCarlo & Cox, 2007), which posits that features are encoded as linear directions in the activation space of capable models. Furthermore, when modeling continuous physical evolution (ranging from orbital ordinary differential equations (ODEs) to fluid partial differential equations (PDEs)), the structural equivalence between residual networks and Euler discretizations becomes critical (Chen et al., 2018b; Haber & Ruthotto, 2017). Since the model effectively learns to act as a numerical integrator for these dynamics (Chen et al., 2018a), invasive fine-tuning can disrupt the weights into non-physical regimes.

Bridging the gap between predictive fidelity and the evaluation of physical understanding requires a shift towards interpretable probing frameworks (Bereska & Gavves, 2024) and rigorous experimental design and control (Hewitt & Liang, 2019b; Belinkov, 2022b). We argue that apparent failures to encode physics are often not failures of learning but artifacts of the measurement process itself (i.e., the downstream adaptation task), which can distort the underlying representation (Santi et al., 2025; Belinkov, 2022a).

We propose the **Non-Invasive Physical Probe** (*PhyIP*), which is a mechanistic evaluation framework that uses time-independent linear readouts on frozen SSL representations to extract latent conserved quantities. These quantities are then distilled into interpretable formulas via symbolic regression (SR) (Biggio et al., 2021) and validated against strict control baselines for probes (Hewitt & Liang, 2019a). These controls are designed to rule out possible false positives and false negatives under Out-Of-Distribution (OOD) settings.

Consequently, we posit that adopting neural dynamics models as a *fixed measurement instrument* (Jaeger, 2001; Menocattini et al., 2026) is vital for valid scientific inquiry. In classical experimental design, the measuring tool should remain invariant relative to the subject to avoid confounding the observation with the instrument’s own adaptation (Mari et al., 2023).

We validate this hypothesis on high-fidelity simulations from the TheWell benchmark (Ohana et al., 2024): a *2D Turbulent Radiative Layer* (Stachenfeld et al., 2021), a *3D Red Supergiant* (Goldberg et al., 2022), and a *3D Supernova Explosion* (Hirashima et al., 2023). We find that low-error SSL models actively encode physical dynamics into linear subspaces across diverse regimes. In radiative turbulence, *PhyIP* recovers the internal energy law ($E \approx 1.5P$) with high precision. In the more complex *Red Supergiant* setting, it spontaneously develops a correction term for convective kinetic energy ($\sim \rho v_r^2$).

Conversely, we show that invasive methods such as non-linear probes (Belinkov, 2022a), last-layer fine-tuning

(LLFT) (Kirichenko et al., 2023a), and full fine-tuning via Inductive Bias Probes (IBP) (Vafa et al., 2025) can systematically mislead. In the supernova simulation, these methods report high accuracy despite SSL failure. Furthermore, in replicating the orbital mechanics experiment of Vafa et al. (2025), we provide a mechanistic analysis of this destructive intervention. We observe a collapse in representational similarity (CKA) (Kornblith et al., 2019) in deep blocks, where the optimizer minimizes loss by discarding time-varying features (speed, radius) and relying instead on constants (mass). This shows that physical knowledge was latent in the SSL model but corrupted by the measurement process itself, analogous to the “*observer effect*” (Sassoli de Bianchi, 2013; Heisenberg, 1927).

Our contributions:

1. **Non-Invasive Physical Probe (*PhyIP*)**: We introduce a framework to extract latent physical quantities from frozen representations without inducing distortion.
2. **Experimental Design**: We derive a bound linking SSL error (ϵ) and functional curvature (K_Φ) to linear decodability, enabling strict experimental control.
3. **Physics Recovery**: We successfully recover fundamental laws, including internal energy ($E \approx 1.5P$) and gravitational force ($F \propto 1/r^2$), across complex fluid and orbital benchmarks where invasive methods fail.
4. **Invasive Corruption Evidence**: These results demonstrate that adaptation acts as a destructive intervention that overwrite internal world models.

2. Preliminaries & Framework

To distinguish between a model that encodes physics and one that memorizes data, we must formalize the interaction between the physical dynamics, the neural architecture, and the probe as “measurement instrument”.

Data Generating Process (DGP): We assume the physical system is a time-dependent field $u(z, t)$ on a domain $\Omega \subseteq \mathbb{R}^D$ evolving via PDEs: $\frac{\partial u}{\partial t} = \mathcal{F}(u, \nabla u, \dots)$. To align with the discrete nature of computation, this continuous field is discretized into a finite-dimensional state vector $x(t) \in \mathcal{X} \subseteq \mathbb{R}^n$ (we also write x_t with a subscript to denote the functional dependence on time). Consequently, the evolution follows an autonomous ODE: $\dot{x}(t) = f(x(t))$ where $f : \mathcal{X} \rightarrow \mathbb{R}^n$ is the Lipschitz continuous vector field approximating the continuous dynamics.

Observational Data: The continuous DGP is observed at discrete time intervals Δt , yielding a dataset of trajectories \mathcal{T} . A trajectory $\tau \in \mathcal{T}$ is a sequence of states (x_0, x_1, \dots, x_T) where $x_{t+1} = x_t + \int_t^{t+\Delta t} f(x(s)) ds$. The learning task is to approximate this transition operator.

Physical Observables & Curvature: Let $\Phi : \mathcal{X} \rightarrow \mathbb{R}^k$ be

a target physical functional defining the quantity of interest $s = \Phi(x)$. We assume Φ is C^2 -smooth and define its *Curvature* K_Φ as the Lipschitz constant of the gradient $\nabla\Phi$ with respect to the Euclidean norm:

$$\|\nabla\Phi(x_a) - \nabla\Phi(x_b)\|_2 \leq K_\Phi \|x_a - x_b\|_2 \quad (1)$$

where $x_a, x_b \in \mathcal{X}$. K_Φ quantifies non-linearity in the input space. By Taylor’s theorem, the deviation of Φ from its linear tangent is at most $\frac{1}{2}K_\Phi\|\Delta x\|^2$. For linear functionals (e.g., Linear Momentum), $K_\Phi = 0$. For non-linear interactions (e.g., Gravitational Forces), $K_\Phi > 0$.

Neural Dynamics Model: Let m_θ be a model that maps an input trajectory $x_{0:t}$ to a latent representation h_t and a next-step prediction \hat{x}_{t+1} . The model is trained to minimize the self-supervised prediction error :

$$\epsilon = \mathbb{E}[\mathcal{L}_{\text{SSL}}(\hat{x}_{t+1}, x_{t+1})] \quad (2)$$

Where \mathcal{L}_{SSL} is typically the Mean Squared Error (MSE), quantifying the model’s ability to approximate the underlying physical transition operator.

Probe: A probe is a diagnostic function $P_W : \mathbb{R}^d \rightarrow \mathbb{R}^k$ with learnable parameters W . It maps the *frozen* latent representation h_t to the value of the target physical quantity:

$$s_{t+1} = \Phi(x_t). \quad (3)$$

2.1. Why a Linear Probe for Physics?

While the *Linear Representation Hypothesis* (DiCarlo & Cox, 2007) is well-documented for neural models trained on language data (Park et al., 2023; Nanda et al., 2023c), this property remains under-explored in models trained on physical systems. *We hypothesize that the linear encoding of physical dynamics is an emergent capability of a successful self-supervised optimization task.*

Most effective scientific models (e.g., Transformers, U-Nets, and Fourier Neural Operators) effectively model continuous dynamics via an Euler-like discretization (Haber & Ruthotto, 2017; Chen et al., 2018a). We formalize this *incremental residual prediction* property as follows: the model predicts a future state as an additive update: the model predicts a future state as an additive update $\hat{x}_{t+1} = x_t + g(h_t)$, where g is a learned decoder representing the state displacement. Complementarily, an *Origin-Preserving Decoder*: $g(\mathbf{0}) = \mathbf{0}$ ensures that a null latent state ($\mathbf{h}_t = \mathbf{0}$) corresponds to the identity map (no physical update), a property actively encouraged by zero-initialization (Goyal et al., 2017) and weight decay (He et al., 2015).

A model m_θ with low error ϵ must maintain a representation h that is locally consistent with the physical update Δx . We define the optimal linear probe P_{W^*} as the best first-order approximation mapping this latent space to the target

quantity update. The expected probe error in recovering the evolution of the physical target quantity $\Delta s = s(t+1) - s(t)$ is bounded by:

$$\begin{aligned} \mathbb{E} \left[\underbrace{|P_{W^*} h_t - \Delta s|^2}_{\text{Probe Error}} \right] &\leq \underbrace{C_1 \cdot \epsilon}_{\text{Modeling SSL Error}} \\ &+ C_2 \underbrace{[K_\Phi^2 \cdot \text{Var}(x)]}_{\text{Curvature Error}} + \underbrace{\mathcal{O}(\Delta t^4)}_{\text{Discretization Error}} \end{aligned} \quad (4)$$

See Section B for derivation. This inequality establishes the linear probe as a rigorous diagnostic tool. The error decomposes into two sources: 1) Modeling SSL Error (ϵ): If the model fails to predict the dynamics (high ϵ), the probe fails. 2) Curvature (K_Φ): If the physical dynamics is highly non-linear, a linear approximation naturally suffers.

Crucially, physical dynamics can be non-linear; for instance, in orbital mechanics, the force vector $\nabla\Phi(x)$, where x is a position trajectory, rotates as the planet moves. While a time-dependent probe (one that “rotates” its weights to match local gradients) might achieve lower error, it risks approximating the physics via its own capacity rather than measuring the representation (Hewitt & Liang, 2019b). Therefore, the probe optimal W^* must be a single constant matrix across all time steps. A probe success under these constraints serves as a possible *litmus test* for evaluation.

Takeaway 1: Probes as Fixed Instruments

The measurement instrument must be a **fixed, time-invariant linear operator** targeting the next state. Consequently, a successful readout under this strict condition serves as evidence that the SSL model has transformed the complex non-linear dynamics into a linearizable representation, consistent with the error bound derived in Eq. 4.

2.2. PhyIP: Non-Invasive Probe for Physics

The objective of our framework is to probe the internal geometry of a pre-trained SSL model m_θ on continuous physical trajectories without corrupting its learned representation. This three-stage process is summarized in Figure 1.

Feature Extraction and Linear Probing. To probe the geometry of the activation space, $\mathcal{H} \subseteq \mathbb{R}^d$, we train a linear probe, $P_W : \mathcal{H} \rightarrow \mathcal{Q}$, where $\mathcal{Q} \subseteq \mathbb{R}^k$ is the space of the physical quantity. The target dimension $k \geq 1$ depends on the quantity being probed; for example, $k = 1$ for a scalar (such as force magnitude) or $k = 2$ for a 2D vector (such as the force $\vec{F} = (F_x, F_y)$). Here, $h(t)$ is the internal activation vector (e.g., from the decoder block) of the pre-trained SSL model, m_θ . This activation is the result of the model processing Section 2.1, $h(t)$ represents the model’s internal “plan” to execute the update. Following Section 2.1,

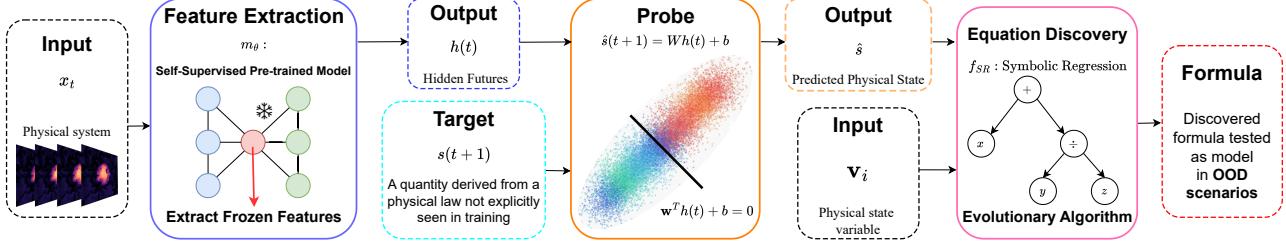


Figure 1. Overview of *PhyIP* framework. (1) **Feature Extraction:** We extract frozen activations, $h(t)$ from a SSL model. (2) **Linear Probing:** A linear probe is trained to predict a new physical quantity, $\hat{s}(t+1) = Wh(t) + b$. Its success on OOD tests indicates that \hat{s} is linearly encoded. (3) **Equation Discovery & Validation:** SR translates the probe’s function into a symbolic formula, $\hat{\Phi}_{SR}$, and tests it on OOD data to confirm its physical plausibility.

we probe Δs_t . The target is the physical state at the next time step, $\hat{s}(t+1)$, via the transformation:

$$\hat{s}(t+1) = Wh(t; m_\theta) + b \quad (5)$$

$\hat{s}(t+1) \in \mathcal{Q} \subseteq \mathbb{R}^k$ is the predicted quantity. The linear mapping matrix $W \in \mathbb{R}^{k \times d}$ and the bias vector $b \in \mathbb{R}^k$ transform the high-dimensional representation $h(t) \in \mathcal{H} \subseteq \mathbb{R}^d$ into the low-dimensional space \mathcal{Q} .

The probe’s optimal parameters, (W^*, b^*) , are found by minimizing a loss function (e.g., MSE) on a training set $D_{\text{train}} = \{(h_i, s_i)\}_{i=1}^N$. Here, the index i iterates over the N samples in the dataset and does not represent the time t . Each sample i is a pair created from a specific time step t in the simulation, such that $h_i = h(t)$ (the model activation at time t) and $s_i = s(t+1)$ (the ground-truth physical quantity at the next time step):

$$(W^*, b^*) = \arg \min_{W, b} \frac{1}{N} \sum_{i=1}^N \|s_i - (Wh_i + b)\|^2 \quad (6)$$

This loss function minimizes the Euclidean distance between the ground-truth quantity s_i and the probe’s linear prediction \hat{s}_i . The optimization process is constrained to learn only W and b (the linear map parameters), while the complex, nonlinear feature extraction of the underlying model m_θ (which produces h_i) remains frozen.

The probe’s success is then quantified by its generalization performance on an (OOD) test set (D_{OOD}). For our physical systems, D_{OOD} consists of simulations where key generative parameters (e.g., central star mass, gravitational constant, initial velocity, or boundary conditions) are sampled outside the distribution used for the self-supervised pre-training. A success here confirms that the linear encoding is a robust physical invariant, not a memorized correlation.

Equation Discovery. While an OOD-generalizing probe confirms a meaningful geometry, its learned mapping W remains opaque. To decode this geometry, we employ Symbolic Regression (SR). This step treats the probe’s output as the ground truth. The SR algorithm is given the physical

state variables x_i (e.g., position, pressure) as inputs and the probe’s predictions \hat{s}_i as targets. SR searches a space of symbolic expressions \mathcal{G} for an optimal formula $\hat{\Phi}_{SR}^*$:

$$\hat{\Phi}_{SR}^* = \arg \min_{\hat{\Phi} \in \mathcal{G}} \frac{1}{N} \sum_{i=1}^N \|\hat{s}_i - \hat{\Phi}(x_i)\|^2 \quad (7)$$

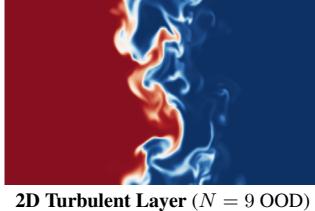
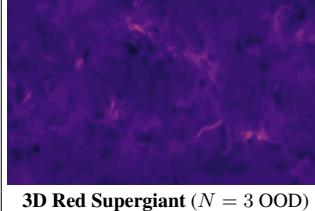
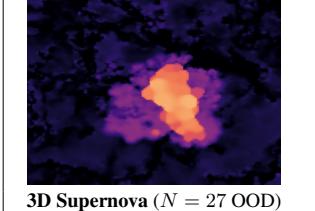
To ensure $\hat{\Phi}_{SR}^*$ represents a true physical principle, we treat it as a standalone physical hypothesis and evaluate its zero-shot generalization on D_{OOD} . A low loss on unseen simulation parameters serves as robust validation that the original model m_θ successfully encoded the governing law.

3. Probing Complex Fluid Dynamics Systems

Setup: To demonstrate the generality of *PhyIP*, we applied it to three high-dimensional fluid dynamics simulations from the TheWell benchmark (Ohana et al., 2024): a 2D Turbulent Radiative Layer (Stachenfeld et al., 2021), a 3D Red Supergiant Convective Envelope (Goldberg et al., 2022), and a 3D Supernova Explosion (Hirashima et al., 2023). The experimental setting satisfies the conditions of Section 2.1. We tested neural models—including U-Net (Ronneberger et al., 2015), UNetConvNext (Liu et al., 2022), FNO (Li et al., 2021b), and TFNO (Li et al., 2021a)—trained solely on self-supervised next-state prediction.

These models function as residual predictors, satisfying the setting discussed in Section 2. Using our non-invasive method *PhyIP*, a linear probe was trained on frozen activations (e.g., U-Net neck). To predict the global total internal energy E_{int} , we implemented the probe as a 1×1 Convolutional Layer (kernel size 1) followed by a global summation. This forces the probe to predict the local energy density contribution $(\rho u)_{i,j}$ at each voxel (i, j) using only the local latent vector $h_{i,j}$. The final prediction is the sum over the domain: $\hat{E}_{\text{int}} = \sum_{i,j} \text{Probe}(h_{i,j}) \Delta V$. To distill these probes into interpretable formulas, we restricted the PySR search space to basic arithmetic operators $\{+, -, \times, /\}$ with strict dimensional consistency constraints

Table 1. Comprehensive Probe Analysis on Fluid Dynamics. OOD evaluation of internal energy recovery. *PhyIP* (top) reliably extracts physical laws from frozen representations when the SSL error (ϵ) is low, recovering the ideal gas law ($E \approx 1.5P$) and kinetic corrections (ρv_r^2). The SN-3D results demonstrate that invasive probes (bottom) can hide collapse, whereas *PhyIP* correctly identifies it.

									
	2D Turbulent Layer ($N = 9$ OOD)			3D Red Supergiant ($N = 3$ OOD)			3D Supernova ($N = 27$ OOD)		
	Probe Task		SSL Task	Probe Task		SSL Task	Probe Task		SSL Task
Model / Method	MAPE ↓	$\rho \uparrow$	$\epsilon_{OOD} \downarrow$	MAPE ↓	$\rho \uparrow$	$\epsilon_{OOD} \downarrow$	MAPE ↓	$\rho \uparrow$	$\epsilon_{OOD} \downarrow$
I. PhyIP (Section 2.2)									
UNetConvNext	36.9 ±1.2	0.83 ±0.02	0.20 ±0.01	18.2 ±0.9	0.91 ±0.01	0.02 ±0.00	140.4 ±12.1	0.15 ±0.05	0.30 ±0.02
UNetClassic	42.3 ±2.1	0.71 ±0.04	0.26 ±0.02	25.6 ±1.5	0.69 ±0.03	0.09 ±0.01	135.9 ±10.5	0.08 ±0.02	0.40 ±0.03
FNO	76.0 ±5.3	0.61 ±0.06	0.49 ±0.04	95.1 ±4.2	0.22 ±0.08	0.05 ±0.01	95.3 ±8.1	0.18 ±0.04	0.36 ±0.02
TFNO	89.7 ±6.1	0.67 ±0.05	0.58 ±0.05	92.5 ±5.0	0.25 ±0.07	0.04 ±0.01	92.1 ±7.8	0.21 ±0.03	0.36 ±0.03
II. Baselines & Probes									
(i) Linear Probe on Raw Inputs	58.5 ±4.1	0.32 ±0.05	-	88.2 ±3.5	0.25 ±0.04	-	142.1 ±15.0	0.22 ±0.01	-
(ii) Time-Dependent Probe	22.1 ±1.0	0.88 ±0.02	-	15.4 ±0.8	0.95 ±0.01	-	93.2 ±11.5	0.64 ±0.05	-
(iii) MLP Probe (Belinkov, 2022a)	37.2 ±1.5	0.72 ±0.03	-	19.1 ±1.2	0.82 ±0.02	-	125.5 ±14.2	0.48 ±0.06	-
(iv) LL-FT (Kirichenko et al., 2023b)	32.5 ±2.8	0.80 ±0.04	-	23.4 ±4.1	0.80 ±0.05	-	65.0 ±13.1	0.59 ±0.04	-
(v) Full FT (IBP (Vafa et al., 2025))	41.2 ±3.5	0.81 ±0.04	-	21.1 ±6.4	0.80 ±0.03	-	18.3 ±18.5	0.71 ±0.01	-
SSL Input Vars	{ ρ, P, v_x, v_y }			{ $\rho, P, v_r, v_\theta, v_\phi$ }			{ $\rho, P, T, v_x, v_y, v_z$ }		
Probe Target	$E_{int} = \int 1.5P dV$ ($\gamma = 5/3$)			$E_{int} = \int \rho u dV$			$E_{int} = \int \rho u dV$		
Discovered Law ($\hat{\Phi}_{SR}$)	$E \approx 1.48 \cdot P$			$E \approx 1.45P + 0.42\rho v_r^2$			$E \approx 0.35 - \left[\frac{0.06}{(P+0.2)} \right]$		

(penalty 10^{12}) to prioritize simplicity over curve fitting.

Control Baselines & Invasive Probes. All probes are trained using supervised pairs derived exclusively from the in-Distribution (ID) SSL training set, ensuring strictly zero-shot evaluation on OOD regimes. We compare against: **i**) Raw Input Baseline: A linear regression on flattened raw input fields $x_t \in \mathbb{R}^{C \times H \times W}$. **ii**) Time-Dependent Probe: A linear probe $\{W_t\}$ optimized per time-step removing the curvature penalty (K_Φ) from our bound to quantify “curvature mismatch.” **iii**) Non-Linear MLP Probe (Belinkov, 2022a): A 3-layer MLP ($254 \rightarrow 32$) on frozen activations h_t . Finally, we test a **iv**) Last-Layer Fine-Finetuning (LL-FT) (Kirichenko et al., 2023b) and **v**) Full Fine-tuning Adaptation (all θ) via the IBP (Vafa et al., 2025). Success here despite high SSL error (ϵ) confirms the probe is *learning from scratch* rather than measuring the world model.

OOD Evaluation. We validated on 39 Out-of-Distribution (OOD) test sets, 9 unseen cooling rates for TRL-2D, 3 distinct stellar evolution phases for RSG-3D, and 27 novel environments varying ambient gas density and metallicity for Supernova. Results summarized in Table 1 reveal a divergence that provides empirical validation for Section 2.1.

For the **2D Turbulent Radiative Layer**, the U-Net architectures minimized the SSL test error ($\epsilon \approx 0.19$) significantly better than FNO models ($\epsilon \approx 0.50$) with the *PhyIP* linear encoding $\rho = 0.83$ and MAPE = 36.9. *PhyIP* symbolic

regression recovered the equation $E \approx 1.48 \cdot P$, which matches the constant (1.5 for $\gamma = 5/3$) with $< 2\%$ error. This success is visually confirmed in Figure 2 (top), where predictions form a linear cluster, and in Figure 2 (bottom), where the symbolic model (green) of UNetConvNext accurately forecasts the energy decay on unseen cooling rates. In contrast, higher-capacity or invasive baselines do not improve upon the linear readout: the MLP probe reaches $\rho = 0.72$ (MAPE = 37.2%), and both last-layer fine-tuning and full fine-tuning perform slightly worse than *PhyIP* on OOD.

The **3D Red Supergiant** simulation provides the strongest validation. The UNetConvNext achieved the lowest OOD prediction error ($\epsilon = 0.0201$), outperforming the FNO ($\epsilon = 0.0505$). As expected, this SSL mastery enabled precise *PhyIP*’s linear decoding ($\rho = 0.91$). Here, *PhyIP* shows that rather than just retrieving the static pressure law ($1.5P$), the probe recovered a composite symbolic expression $E \approx 1.45P + 0.42\rho v_r^2$. We validate this term via dimensional analysis, the quantity ρv^2 has the units of energy density (J/m^3), matching the units of Pressure (P). This confirms that the term is a valid correction for kinetic energy. As shown in Figure 2 (bottom), this “convective proxy” allows the symbolic formula of UNetConvNext to track the ground-truth energy evolution, while the scatter plot (Section 3, Center) confirms the precision of the *PhyIP* decoding. *PhyIP* probe reduces MAPE error by nearly $5\times$

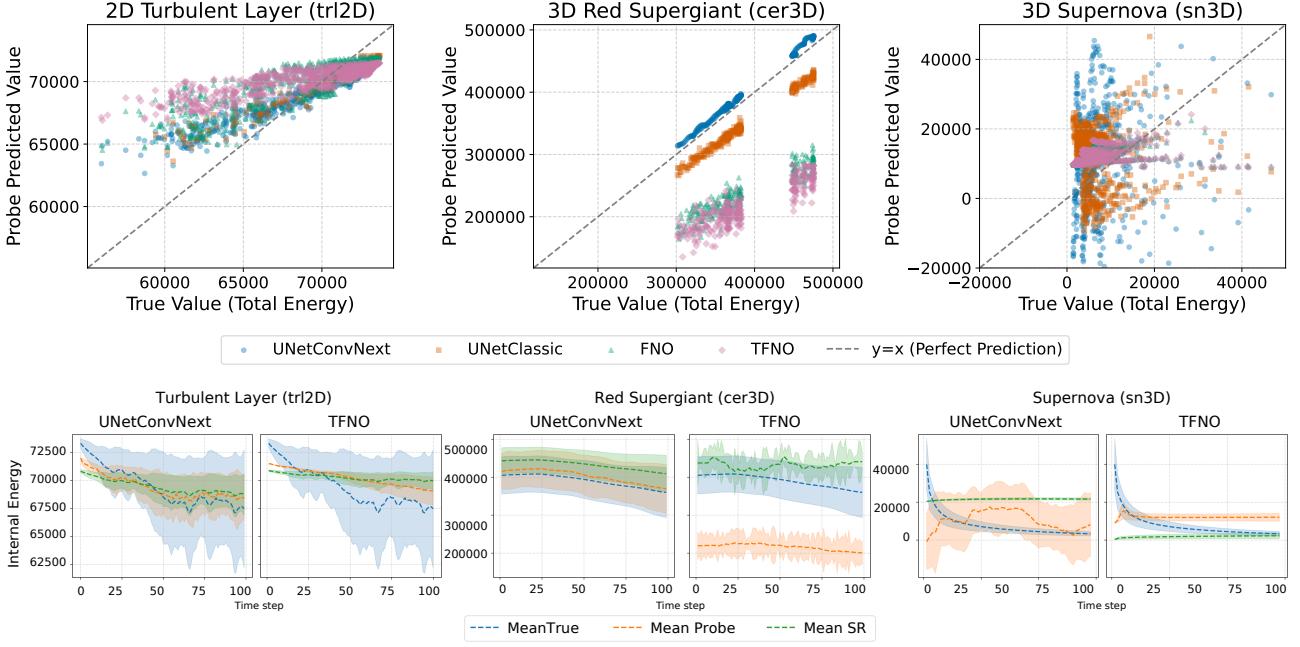


Figure 2. Probing Latent Physical Laws. **(Top)** The Non-Invasive Probe successfully extracts Total Internal Energy from frozen SSL representations in TRL-2D and RSG-3D (linear alignment), but correctly identifies representational collapse in the SN-3D experiment. **(Bottom)** Zero-shot generalization. Discovered symbolic formulas (Mean SR) accurately predict energy dynamics on unseen simulation parameters for the successful models

compared to the Raw Input Baseline ($\rho = 0.25$ and MAPE = 88.2%), proving high-level abstraction. The MLP Probe ($\rho = 0.82$ and MAPE = 19.1%) together with LL-FT and IBP fails to improve upon *PhyIP*, verifying that the relevant geometry is linearly encoded. Finally, by using the Time-Dependent Probe we gain better performance ($\rho = 0.95$ and MAPE = 15.4%) as expected.

Finally, the **3D Supernova** test proved challenging for all architectures, with SSL prediction error at a high $\epsilon \approx 0.30 - 0.37$. Empirically, no model when probed with *PhyIP* achieved a correlation above $\rho = 0.2$. Furthermore, *PhyIP* symbolic regression of the best performer model TFNO failed to find any physical law, fitting instead a spurious rational function $E \approx 0.35 - [\frac{0.06}{(P+0.2)}]$ that implies unphysical inverse scaling. This failure is visualized in Figure 2, where both the probe and symbolic formula completely fail to track the energy decay. **This failure exposes the danger of invasive probing:** while *PhyIP* probe correctly diagnoses this (MAPE = 140.4%), the full fine-tuning via IBP (Table 1, row v) achieves a deceptively low MAPE = 18.3% and $\rho = 0.71$. This massive discrepancy (140% \rightarrow 18% for MAPE and 0.21 \rightarrow 0.71) confirms that the invasive probes did not measure the model's knowledge but rather overwrote it, hallucinating competence where there was none.

Takeaway 2: Strict Experimental Control

In complex fluids, faithful evaluation requires strict experimental control. PhyIP succeeds only when SSL models the dynamics (low OOD ϵ), while invasive adaptation can mask backbone failure (SN-3D: MAPE 140% \rightarrow 18%) by learning the probe task itself and distorting otherwise valid linear structure (as seen in TRL-2D and RSG-3D).

4. The Confounding Nature of Invasive Probes

Motivation. While the Fluid Dynamics experiment (Section 3) demonstrate scalability, it is difficult to isolate the exact mechanism of invasive corruption. To provide a controllable test, we replicate the Orbital Mechanics experiment of Vafa et al. (2025) for the *inductive bias probe*. This setting allows us to: **(1)** directly compare *PhyIP* against invasive methods in a highly nonlinear task (force vector) and **(2)** perform a mechanistic analysis to point exactly when and how invasive probes overwrite knowledge.

Setup: A 109M parameter Transformer (m_θ), pre-trained (SSL) on orbital trajectories, is subjected to full-parameter fine-tuning on a small datasets where the output is the force vector at each point in the trajectory. The weights of the entire architecture θ are updated to θ'_i by minimizing the

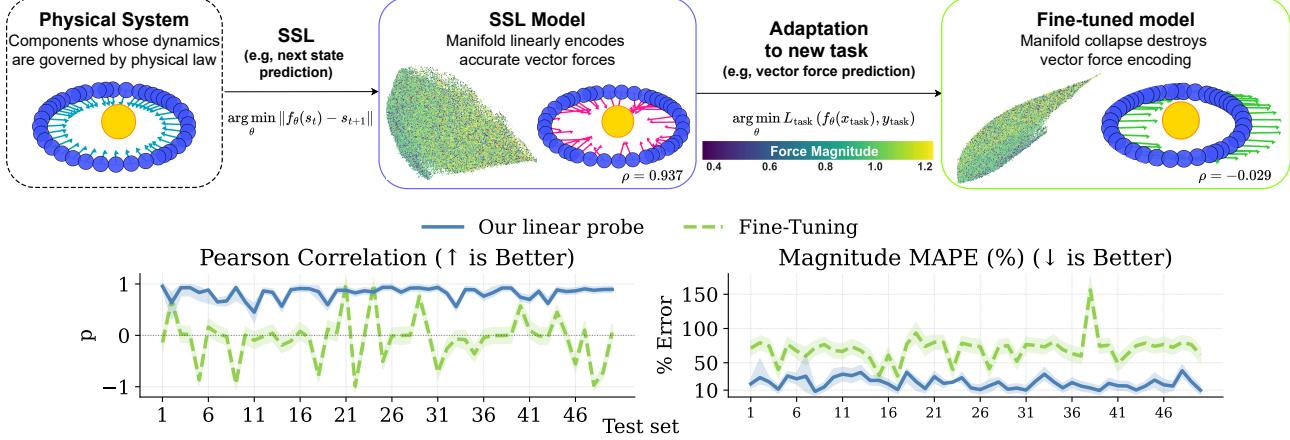


Figure 3. The Failure of Invasive Probing. Top: The orderly geometry of the SSL model (center) is destroyed by invasive fine-tuning (right), dropping correlation from $\rho = 0.94 \rightarrow -0.03$. Bottom: Quantitative Impact. This geometric destruction causes the invasive probe (Green) to fail erratically on OOD task, whereas our non-invasive probe (Blue) remains robust ($\rho > 0.8$).

task loss over a fine-tuning dataset $D_{\text{task},i}$. For each pair (x, s) , x represents the input trajectory and s the ground-truth physical target variable, the force vector

$$\theta'_i = \arg \min_{\theta} \sum_{(x,s) \in D_{\text{task},i}} \mathcal{L}_{\text{task}}(m_{\theta}(x), s) \quad (8)$$

The subsequent inductive bias analysis is performed on the fine-tuned model $m_{\theta'}$.

Representation Collapse: We investigate whether invasive fine-tuning acts as a *destructive intervention* that degrades the SSL model’s (m_{θ}) latent geometry. To establish a robust baseline, we employ the non-invasive probe *PhyIP* (Section 2.2) on the frozen activations of the decoder blocks . 9. We compare *PhyIP* against the fully adapted fine-tuned model via IBP. As shown in Figure 3, our probe maintains high performance ($\rho > 0.85$, MAPE < 30%) across 50 OOD tests, while the fine-tuned model exhibits erratic failure with MAPE spiking between 50% and 150% (see Appendix E for solar system validation results).

Table 2 confirms the non-triviality of the task: the Raw Input Probe fails (MAPE 65.2%). The Time-Dependent probe attains low error (MAPE 12.1% and $\rho = 0.96$) by adapting to local gradients, effectively bypassing the curvature term K_{Φ} in Equation (4). Introducing limited adaptation with LL-FT partially mitigates the degradation observed with the IBP (MAPE 45.3%, $\rho = 0.65$), but it still performs substantially worse than *PhyIP* (MAPE 24.5%, $\rho = 0.91$).

We selected blocks . 9 via a systematic analysis per block on the 50 OOD test set (Figure 4). While physical information is initially entangled, requiring the non-linear MLP ($d \rightarrow 254 \rightarrow 1$) probe to extract it, the representation ‘linearizes’ with depth, reaching maximal disentanglement at blocks . 9 ($\rho \approx 0.91$ and MAPE ≈ 0.25).

Table 2. Orbital Mechanics Baseline Analysis. Comparison of our Non-Invasive Probe against baselines on the 50 OOD test set (force vector task).

Method	OOD test sets	
	MAPE \downarrow	ρ (Pearson) \uparrow
I. Non-Invasive Probe (<i>PhyIP</i>)	24.5 \pm 4.2	0.91 \pm 0.02
II. Baselines & Controls		
(i) Linear Probe on Raw Inputs	65.2 \pm 4.5	0.45 \pm 0.03
(ii) Time-Dependent Probe	12.1 \pm 1.2	0.96 \pm 0.01
(iii) MLP Probe	22.5 \pm 3.0	0.88 \pm 0.04
(iv) LL-FT	45.3 \pm 5.1	0.65 \pm 0.06
(v) Full FT (IBP)	81.5 \pm 12.4	0.05 \pm 0.35

Mechanistic Analysis: Figure 3 visualizes the mechanism of this failure. Using 2D PaCMAP projections of activation manifolds (methodology in Section D), we observe that the SSL model (m_{θ}) encodes a gradient of ground-truth force magnitudes. In contrast, the fine-tuned model ($m_{\theta'}$) exhibits a modified manifold, confirming the modification of geometric structure. This collapse is driven by parameter shifts in the decoder’s attention and MLP layers (Section F). To quantify this, we analyze layer-wise activations $h^{(l)} \in \mathbb{R}^{B \times T \times d}$ for OOD test trajectories \mathcal{X} , where B is batch size, T sequence length, and $d = 768$.

We measure *Parameter Modifications*, quantified by the relative Frobenius norm of the weight change for each block $l : \delta^{(l)} = \frac{\|\theta'^{(l)} - \theta^{(l)}\|_F}{\|\theta^{(l)}\|_F}$ as Guo et al. (2020) shows. Second, we assess *Representational Drift* using Linear Centered Kernel Alignment (CKA) (Kornblith et al., 2019), which computes the $s_{\text{CKA}}^{(l)}$. Finally, we identify the specific physical concepts erased by this drift. We isolate the subset of neurons $S^{(l)}$ exhibiting the top 20% of parameter modifications $\delta_j = \|\mathbf{w}'_j - \mathbf{w}_j\|_2$ within the MLP projection layer.

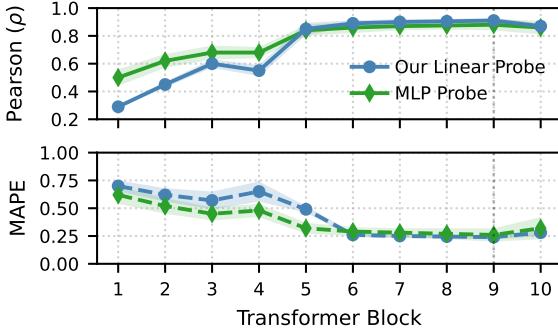


Figure 4. Probe Analysis per Block Layer-wise performance of Non-Invasive (*PhyIP*, Blue) vs. Non-Linear (MLP, Green) probes. While early layers MLP perform better than linear, the representation spontaneously *linearizes* in deep layers, peaking at Block 9 ($\rho \approx 0.91$).

We then compute the shift in the maximum Pearson correlation $\Delta\rho_k$ between the flattened activation history $\mathbf{h}_j \in \mathbb{R}^N$ and ground-truth physical vectors $\phi_k \in \mathbb{R}^N$:

$$\Delta\rho_k = \max_{j \in \mathcal{S}^{(l)}} |\text{corr}(\mathbf{h}'_j, \phi_k)| - \max_{j \in \mathcal{S}^{(l)}} |\text{corr}(\mathbf{h}_j, \phi_k)| \quad (9)$$

Applying these metrics reveals a distinct chain of corruption (Figure 5). Structural modifications are highly localized, while the global average is low, the attention and MLP layers of deep blocks (B5–B10) undergo significant shifts ($\delta^{(l)} \approx 0.10$). This structural modification directly drives functional collapse: while early layers remain stable ($s_{\text{CKA}} \approx 1.0$), the representations in these modified deep blocks diverge successfully ($s_{\text{CKA}} < 0.2$).

The results in Figure 5 show a targeted erasure of dynamic invariants: encoding for *Speed* and *Radius* drops significantly ($\Delta\rho \approx -0.15$), while static variables like *Mass* remain untouched. This confirms that fine-tuning specifically modifies parameters responsible for Hamiltonian dynamics ($\mathcal{K} \propto v^2, 1/r$) to minimize error on the narrow distribution.

Fine-tuning Dataset Analysis: The effectiveness of fine-tuning is heavily reliant on the dataset representativeness used for adaptation (Kumar et al., 2022). By replicating the dataset as in (Vafa et al., 2025), Figure 6 shows that the data for the Star Mass (m_2) is a single spike at 1.0 (Green). These discrepancies indicate that the fine-tuning dataset does not provide sufficient diversity to maintain the model’s general understanding of physics. This lack of support directly drives the optimizer to erase the now-unnecessary dynamic quantities (speed, radius) in favor of static shortcuts suitable for the narrow distribution (See to Section G for *Force Vectors* and *Force Magnitude* distribution analysis).

Symbolic Validation of Discovered Physics: We apply SR to distill the non-intrusive probe and IBP output into interpretable formulas as in Figure 1. When evaluated them

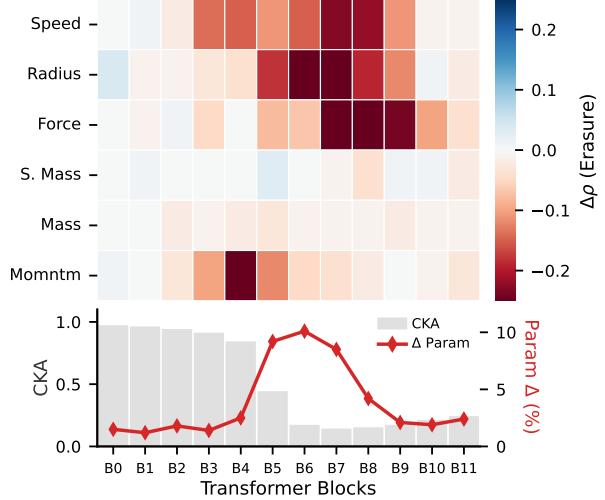


Figure 5. Mechanistic Origins of Erasure. **Top:** Heatmap of linear decodability change ($\Delta\rho$). Fine-tuning selectively erases dynamic variables (Speed, Radius) while preserving static one (Mass). **Bottom:** This collapse is driven by a parameter change causing a drop in representational similarity (CKA).

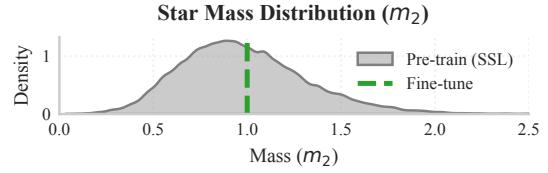


Figure 6. Narrow Data. Comparison of Star Mass (m_2) distributions. While the SSL pre-training data (Grey) covers a diverse physical range, the fine-tuning dataset (Green dashed) is a single point mass at $m_2 = 1.0$.

as a physical model on the 50 OOD orbital test sets (force vector task). Figure 7 shows non-invasive probe formula tracks the ground truth with high precision, whereas the formula extracted from the IBP remains erratic. The truth is the Newton Law (Newton, 1687): $F \propto \frac{m_1 m_2}{r^2}$. The formula discovered by the IBP (Vafa et al., 2025) is dominated by artifacts: $F \propto \underbrace{\left[\sin\left(\frac{1}{\sin(r-0.2)}\right) + 1.5 \right]}_{\text{Hallucinated Artifacts}} \cdot \underbrace{\frac{1}{r^{-1} + m_2}}_{\text{Distorted Decay}}$.

In contrast, the Non-Invasive Probe recovers the signal:

$$F \approx \underbrace{P(r, m_2)}_{\text{Residual Noise}} + \underbrace{\frac{1}{r^2}}_{\text{Recovered Law}}.$$

Although the non-invasive probe formula is still an approximation, it discovers a distinct additive term where the residual $P(r, m_2)$ vanishes, recovering the inverse-square law $F \approx 1/r^2$. The full list of discovered formulas is available in Appendix H (Table 4).

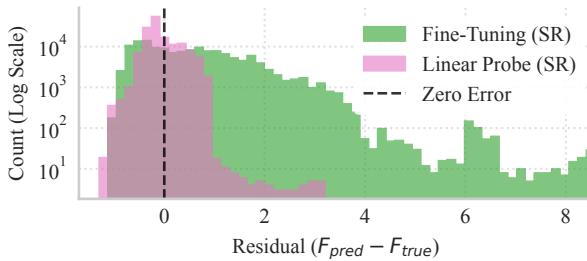


Figure 7. Symbolic Validation. Distribution of prediction errors on OOD data. Our non-invasive probe (Blue) achieves high precision compared to the erratic invasive baseline (Green).

Mechanistic Confirmation. The successful recovery of the $1/r^2$ term provides functional validation for the mechanistic analysis in Section 4. As visualized in the erasure heatmap (Figure 5), the neurons encoding Radius (r) were preserved in the SSL model but *specifically erased* during fine-tuning. Because the fine-tuned model lost the internal representation of r , it physically *could not* express the correct inverse-square law, defaulting to heuristics.

Takeaway 3: Erasure of Encoded Dynamics

Invasive probes do not merely measure inductive bias—they overwrite it: fine-tuning induces representational drift that selectively erases dynamic state variables (e.g., speed, radius) to fit narrow downstream data. Reliable world-model evaluation must preserve the backbone.

5. Conclusion

AI for scientific discovery has reached a critical stage (Wang et al., 2023). As models scale, the challenge shifts from training to correctly interpreting the latent knowledge they have internalized (Vafa et al., 2024; Mencattini et al., 2026). Distinguishing whether neural dynamics models internalize physical laws as world models (Vafa et al., 2024) or merely rely on statistical shortcuts (Geirhos et al., 2020) is computationally difficult, as standard invasive protocols often act as interventions that distort the underlying representation. To address this, we introduced the non-invasive *PhyIP* framework to evaluate the intrinsic physics of Self-Supervised Learning (SSL) models without inducing feature distortion (Kumar et al., 2022).

Empirically, our non-invasive approach reveals physical structures that standard invasive methods miss. On complex benchmarks from “TheWell” (Ohana et al., 2024), we precisely recovered the internal energy law ($E \approx 1.5P$) in radiative turbulence (Stachenfeld et al., 2021) and identified emergent correction terms for convective kinetic energy ($\sim \rho v_r^2$) in stellar simulations ($\rho > 0.90$) (Goldberg et al.,

2022). Furthermore, by replicating orbital mechanics experiments (Vafa et al., 2025), we successfully recovered the inverse-square law with high fidelity ($\rho \approx 0.91$), whereas invasive adaptation probes reported near-zero correlation ($\rho \approx 0.05$).

Conversely, we show that invasive probes—including nonlinear MLP probes (Belinkov, 2022a), Last-Layer Fine-Tuning (LLFT) (Kirichenko et al., 2023a), and Inductive Bias Probes (IBP) (Vafa et al., 2025) can act as destructive interventions. Rather than passively measuring latent knowledge, they overwrite the representation (Kumar et al., 2022). Our mechanistic analysis confirms that these optimizers systematically suppress complex time-varying features (e.g., speed, radius) to exploit simple constant identifiers (e.g., mass), effectively “hallucinating” competence or erasing physical laws to fit narrow data (Mukhoti et al., 2024; Geirhos et al., 2020).

In concurrent work, Liu et al. (2026) identify key inductive biases; specifically spatial smoothness (continuous regression), stability (noise injection), and temporal locality (context restriction); that enable Transformers to learn Newtonian physics with perfect fidelity ($R^2 \approx 1$). While they demonstrate that explicitly enforcing these constraints guarantees the acquisition of exact physical models, our work offers a complementary perspective focused on evaluation. We find that even without these additional inductive biases, standard SSL approximately encodes physical dynamics into linear subspaces. Although these latent representations may not always reach the perfect precision of constrained models, their linear extractability confirms that the core physical laws are already emerging. This validates the promise of general-purpose foundation models (Bommasani et al., 2022): that broad physical understanding can emerge implicitly from data scale and diversity.

We hope this work encourages the adoption of neural models as *fixed measurement instruments* (Peters et al., 2016) and the use of non-invasive protocols to distinguish true machine learning failures from artifacts of adaptation. Ultimately, our findings suggest that *Scientific AI requires not just better models, but better instruments to measure them*.

Limitations and Future Work: We identify three primary constraints. Linear probes prevent the probe from solving the physics independently but limit performance on highly non-linear entangled representations, where physical quantities may be encoded in more complex, non-linear geometries. Future research should investigate subspace-constrained or weight-preserving adaptation protocols. These methods would aim to acquire new task-specific capabilities while strictly protecting the linear physical invariants—such as conservation laws—already internalized by the model.

References

- Belinkov, Y. Probing classifiers: Promises, shortcomings, and advances. *Computational Linguistics*, 48(1):207–219, 2022a.
- Belinkov, Y. Probing classifiers: Promises, shortcomings, and advances. *Computational Linguistics*, 48(1):207–219, March 2022b. doi: 10.1162/coli_a_00422. URL <https://aclanthology.org/2022.cl-1.7/>.
- Bereska, L. and Gavves, E. Mechanistic interpretability for ai safety – a review, 2024. URL <https://arxiv.org/abs/2404.14082>.
- Biggio, L., Bendinelli, T., Neitz, A., Lucchi, A., and Parascandolo, G. Neural symbolic regression that scales. In *International Conference on Machine Learning (ICML)*, 2021.
- Bommasani, R., Hudson, D. A., Adeli, E., Altman, R., Arora, S., von Arx, S., Bernstein, M. S., Bohg, J., Bosselet, A., Brunskill, E., Brynjolfsson, E., Buch, S., Card, D., Castellon, R., Chatterji, N., Chen, A., Creel, K., Davis, J. Q., Demszky, D., Donahue, C., Doumbouya, M., Durmus, E., Ermon, S., Etchemendy, J., Ethayarajh, K., Fei-Fei, L., Finn, C., Gale, T., Gillespie, L., Goel, K., Goodman, N., Grossman, S., Guha, N., Hashimoto, T., Henderson, P., Hewitt, J., Ho, D. E., Hong, J., Hsu, K., Huang, J., Icard, T., Jain, S., Jurafsky, D., Kalluri, P., Karamcheti, S., Keeling, G., Khani, F., Khattab, O., Koh, P. W., Krass, M., Krishna, R., Kuditipudi, R., Kumar, A., Ladak, F., Lee, M., Lee, T., Leskovec, J., Levent, I., Li, X. L., Li, X., Ma, T., Malik, A., Manning, C. D., Mirchandani, S., Mitchell, E., Munyikwa, Z., Nair, S., Narayan, A., Narayanan, D., Newman, B., Nie, A., Niebles, J. C., Nilforoshan, H., Nyarko, J., Ogut, G., Orr, L., Papadimitriou, I., Park, J. S., Piech, C., Portelance, E., Potts, C., Raghunathan, A., Reich, R., Ren, H., Rong, F., Roohani, Y., Ruiz, C., Ryan, J., Ré, C., Sadigh, D., Sagawa, S., Santhanam, K., Shih, A., Srinivasan, K., Tamkin, A., Taori, R., Thomas, A. W., Tramèr, F., Wang, R. E., Wang, W., Wu, B., Wu, J., Wu, Y., Xie, S. M., Yasunaga, M., You, J., Zaharia, M., Zhang, M., Zhang, T., Zhang, X., Zhang, Y., Zheng, L., Zhou, K., and Liang, P. On the opportunities and risks of foundation models, 2022. URL <https://arxiv.org/abs/2108.07258>.
- Chen, B., Huang, K., Raghupathi, S., Chandratreya, I., Du, Q., and Lipson, H. Automated discovery of fundamental variables hidden in experimental data. *Nature Computational Science*, 2(7):433–442, 2022.
- Chen, R. T., Rubanova, Y., Bettencourt, J., and Duvenaud, D. K. Neural ordinary differential equations. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2018a.
- Chen, R. T. Q., Rubanova, Y., Bettencourt, J., and Duvenaud, D. K. Neural ordinary differential equations. In Bengio, S., Wallach, H., Larochelle, H., Grauman, K., Cesa-Bianchi, N., and Garnett, R. (eds.), *Advances in Neural Information Processing Systems*, volume 31. Curran Associates, Inc., 2018b. URL https://proceedings.neurips.cc/paper_files/paper/2018/file/69386f6bb1dfed68692a24c8686939b9-Paper.pdf.
- Coveney, P. and Highfield, R. Ai needs physics more than physics needs ai, 2025. URL <https://arxiv.org/abs/2512.16344>.
- Cranmer, M., Greydanus, S., Hoyer, S., Battaglia, P., Spergel, D., and Ho, S. Lagrangian neural networks. *arXiv preprint arXiv:2003.04630*, 2020.
- DiCarlo, J. J. and Cox, D. D. Untangling invariant object recognition. *Trends in Cognitive Sciences*, 11(8):333–341, 2007. ISSN 1364-6613. doi: <https://doi.org/10.1016/j.tics.2007.06.010>. URL <https://www.sciencedirect.com/science/article/pii/S1364661307001593>.
- Geirhos, R., Jacobsen, J.-H., Michaelis, C., Zemel, R., Brendel, W., Bethge, M., and Wichmann, F. A. Shortcut learning in deep neural networks. *Nature Machine Intelligence*, 2(11):665–673, 2020. doi: 10.1038/s42256-020-00257-z.
- Goldberg, J. A., Jiang, Y.-F., and Bildsten, L. Numerical simulations of convective three-dimensional red supergiant envelopes. *The Astrophysical Journal*, 929(2):156, apr 2022. doi: 10.3847/1538-4357/ac5ab3. URL <https://doi.org/10.3847/1538-4357/ac5ab3>.
- Goyal, P. et al. Accurate, large minibatch sgd: Training imagenet in 1 hour. *arXiv preprint arXiv:1706.02677*, 2017.
- Greydanus, S., Dzamba, M., and Yosinski, J. Hamiltonian neural networks. In *Advances in Neural Information Processing Systems (NeurIPS)*, volume 32, 2019.
- Guo, Y., Codella, N. C., Karlinsky, L., Codella, J. V., Smith, J. R., Saenko, K., Rosing, T., and Feris, R. A broader study of cross-domain few-shot learning. In *European Conference on Computer Vision*, pp. 124–141. Springer, 2020.
- Ha, D. and Schmidhuber, J. World models. 2018. doi: 10.5281/ZENODO.1207631. URL <https://zenodo.org/record/1207631>.

- Haber, E. and Ruthotto, L. Stable architectures for deep neural networks. *Inverse Problems*, 34(1):014004, dec 2017. doi: 10.1088/1361-6420/aa9a90. URL <https://doi.org/10.1088/1361-6420/aa9a90>.
- He, K., Zhang, X., Ren, S., and Sun, J. Deep residual learning for image recognition, 2015. URL <https://arxiv.org/abs/1512.03385>.
- Heisenberg, W. Über den anschaulichen inhalt der quantentheoretischen kinematik und mechanik. *Zeitschrift für Physik*, 43(3-4):172–198, 1927.
- Hewitt, J. and Liang, P. Designing and interpreting probes with control tasks. In Inui, K., Jiang, J., Ng, V., and Wan, X. (eds.), *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pp. 2733–2743, Hong Kong, China, November 2019a. Association for Computational Linguistics. doi: 10.18653/v1/D19-1275. URL <https://aclanthology.org/D19-1275/>.
- Hewitt, J. and Liang, P. Designing and interpreting probes with control tasks. In *Conference on Empirical Methods in Natural Language Processing (EMNLP)*, 2019b.
- Hirashima, K., Moriwaki, K., Fujii, M. S., Hirai, Y., Saitoh, T. R., Makino, J., and Ho, S. Surrogate modeling for computationally expensive simulations of supernovae in high-resolution galaxy simulations. *arXiv preprint arXiv:2311.08460*, 2023.
- Internò, C., Geirhos, R., Olhofer, M., Liu, S., Hammer, B., and Klindt, D. AI-generated video detection via perceptual straightening. In *The Thirty-ninth Annual Conference on Neural Information Processing Systems*, 2025. URL <https://openreview.net/forum?id=LsmUgStXby>.
- Jaeger, H. The “echo state” approach to analysing and training recurrent neural networks. 2001. URL <https://api.semanticscholar.org/CorpusID:15467150>.
- Kamienny, P.-A., d’Ascoli, S., Lample, G., and Charton, F. End-to-end symbolic regression with transformers. In *Advances in Neural Information Processing Systems (NeurIPS)*, volume 35, pp. 10269–10281, 2022.
- Karniadakis, G. E., Kevrekidis, I. G., Lu, L., Perdikaris, P., Wang, S., and Yang, L. Physics-informed machine learning. *Nature Reviews Physics*, 3(6):422–440, 2021.
- Kirichenko, P., Izmailov, P., and Wilson, A. G. Last layer re-training is sufficient for robustness to spurious correlations. In *The Eleventh International Conference on Learning Representations*, 2023a. URL <https://openreview.net/forum?id=Zb6c8A-Fghk>.
- Kirichenko, P., Izmailov, P., and Wilson, A. G. Last layer re-training is sufficient for robustness to spurious correlations, 2023b. URL <https://arxiv.org/abs/2204.02937>.
- Klindt, D., O’Neill, C., Reizinger, P., Maurer, H., and Molane, N. From superposition to sparse codes: interpretable representations in neural networks, 2025. URL <https://arxiv.org/abs/2503.01824>.
- Kornblith, S., Norouzi, M., Lee, H., and Hinton, G. Similarity of neural network representations revisited. In *International Conference on Machine Learning*, pp. 3519–3529. PMLR, 2019.
- Kügelgen, J. V., Sharma, Y., Gresele, L., Brendel, W., Schölkopf, B., Besserve, M., and Locatello, F. Self-supervised learning with data augmentations provably isolates content from style. In Beygelzimer, A., Dauphin, Y., Liang, P., and Vaughan, J. W. (eds.), *Advances in Neural Information Processing Systems*, 2021. URL https://openreview.net/forum?id=4pf_pOo0Dt.
- Kumar, A., Raghunathan, A., Jones, R. M., Ma, T., and Liang, P. Fine-tuning can distort pretrained features and underperform out-of-distribution. In *International Conference on Learning Representations*, 2022. URL <https://openreview.net/forum?id=UYneFzXSJWh>.
- LeCun, Y. and Courant. A path towards autonomous machine intelligence version 0.9.2, 2022-06-27. 2022. URL <https://api.semanticscholar.org/CorpusID:251881108>.
- Li, Z., Kovachki, N., Azizzadenesheli, K., Liu, B., Bhattacharya, K., Stuart, A., and Anandkumar, A. Fourier neural operator for parametric partial differential equations, 2021a. URL <https://arxiv.org/abs/2010.08895>.
- Li, Z., Kovachki, N. B., Azizzadenesheli, K., liu, B., Bhattacharya, K., Stuart, A., and Anandkumar, A. Fourier neural operator for parametric partial differential equations. In *International Conference on Learning Representations*, 2021b. URL <https://openreview.net/forum?id=c8P9NQVtmnO>.
- Liu, Z., Mao, H., Wu, C., Feichtenhofer, C., Darrell, T., and Xie, S. A convnet for the 2020s. *CoRR*, abs/2201.03545, 2022. URL <https://arxiv.org/abs/2201.03545>.
- Liu, Z., Sanborn, S., Ganguli, S., and Tolias, A. From kepler to newton: Inductive biases guide learned world models in transformers, 2026. URL <https://arxiv.org/abs/2602.06923>.

- Mai, Z., Chowdhury, A., Zhang, P., Tu, C.-H., Chen, H.-Y., Pahuja, V., Berger-Wolf, T., Gao, S., Stewart, C., Su, Y., and Chao, W.-L. Fine-tuning is fine, if calibrated. In *The Thirty-eighth Annual Conference on Neural Information Processing Systems*, 2024. URL <https://openreview.net/forum?id=XRJXKBeeTD>.
- Mari, L., Wilson, M., and Maul, A. *Philosophical Perspectives on Measurement*, pp. 81–121. Springer International Publishing, Cham, 2023. ISBN 978-3-031-22448-5. doi: 10.1007/978-3-031-22448-5_4. URL https://doi.org/10.1007/978-3-031-22448-5_4.
- Mencattini, T., Cadei, R., and Locatello, F. Exploratory causal inference in saence, 2026. URL <https://arxiv.org/abs/2510.14073>.
- Motamed, S., Culp, L., Swersky, K., Jaini, P., and Geirhos, R. Do generative video models understand physical principles?, 2025. URL <https://arxiv.org/abs/2501.09038>.
- Mukhoti, J., Gal, Y., Torr, P., and Dokania, P. K. Fine-tuning can cripple foundation models; preserving features may be the solution, 2024. URL <https://openreview.net/forum?id=VQ7Q6qdp0P>.
- Nanda, N., Chan, L., Lieberum, T., Smith, J., and Steinhardt, J. Progress measures for grokking via mechanistic interpretability, 2023a. URL <https://arxiv.org/abs/2301.05217>.
- Nanda, N., Lee, A., and Wattenberg, M. Emergent linear representations in world models of self-supervised sequence models, 2023b. URL <https://arxiv.org/abs/2309.00941>.
- Nanda, N. et al. Emergent linear representations in world models of self-supervised learning. *arXiv preprint arXiv:2309.00941*, 2023c.
- Newton, I. *Philosophiae naturalis principia mathematica*. Jussu Societatis Regiae ac Typis Josephi Streater, London, 1687. Annotated 1st Edition, Cambridge Digital Library. <https://cudl.lib.cam.ac.uk/view/PR-ADV-B-00039-00001/1>.
- Ohana, R., McCabe, M., Meyer, L., Morel, R., Agocs, F., Beneitez, M., Berger, M., Burkhardt, B., Dalziel, S., Fielding, D., et al. The well: a large-scale collection of diverse physics simulations for machine learning, 2024.
- Park, K., Choe, Y. J., and Veitch, V. The linear representation hypothesis and the geometry of large language models. In *Causal Representation Learning Workshop at NeurIPS 2023*, 2023. URL <https://openreview.net/forum?id=T0PoOJg8cK>.
- Peters, J., Bühlmann, P., and Meinshausen, N. Causal inference by using invariant prediction: Identification and confidence intervals. *Journal of the Royal Statistical Society Series B: Statistical Methodology*, 78(5):947–1012, 10 2016. ISSN 1369-7412. doi: 10.1111/rssb.12167. URL <https://doi.org/10.1111/rssb.12167>.
- Ravichander, A., Belinkov, Y., and Hovy, E. Probing the probing paradigm: Does probing accuracy reveal content or correlation? In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics*, 2021.
- Reizinger, P., Balestrieri, R., Klindt, D., and Brendel, W. Position: An empirically grounded identifiability theory will accelerate self supervised learning research. In *Forty-second International Conference on Machine Learning Position Paper Track*, 2025. URL <https://openreview.net/forum?id=ET6qJpl1Ei>.
- Ronneberger, O., Fischer, P., and Brox, T. U-net: Convolutional networks for biomedical image segmentation. In *International Conference on Medical image computing and computer-assisted intervention*, pp. 234–241. Springer, 2015.
- Santi, R. D., Vlastelica, M., Hsieh, Y.-P., Shen, Z., He, N., and Krause, A. Flow density control: Generative optimization beyond entropy-regularized fine-tuning. In *The Thirty-ninth Annual Conference on Neural Information Processing Systems*, 2025. URL <https://openreview.net/forum?id=JzCjNj1Sxi>.
- Sassoli de Bianchi, M. The observer effect. *Foundations of Science*, 18(2):213–243, 2013. doi: 10.1007/s10699-012-9298-3. URL <https://doi.org/10.1007/s10699-012-9298-3>.
- Schmidt, M. and Lipson, H. Distilling free-form natural laws from experimental data. *Science*, 324(5923):81–85, 2009a.
- Schmidt, M. and Lipson, H. Distilling free-form natural laws from experimental data. *Science*, 324(5923):81–85, 2009b. doi: 10.1126/science.1165893. URL <https://www.science.org/doi/abs/10.1126/science.1165893>.
- Spies, A. F., Edwards, W., Ivanitskiy, M. I., Skapars, A., Räuker, T., Inoue, K., Russo, A., and Shanahan, M. Transformers use causal world models in maze-solving tasks, 2025. URL <https://arxiv.org/abs/2412.11867>.
- Stachenfeld, K., Brandstetter, J., Pfaff, T., Hoi, S. C., Battaglia, P., and Kim, B. Learned coarse models for efficient turbulence simulation. *arXiv preprint arXiv:2112.15275*, 2021.

Teoh, J., Tomar, M., Ahn, K., Hu, E. S., Sharma, P., Islam, R., Lamb, A., and Langford, J. Next-latent prediction transformers learn compact world models, 2025. URL <https://arxiv.org/abs/2511.05963>.

Trivedi, P., Koutra, D., and Thiagarajan, J. J. A closer look at model adaptation using feature distortion and simplicity bias. In *The Eleventh International Conference on Learning Representations*, 2023. URL https://openreview.net/forum?id=wkg_b4-IwTZ.

Vafa, K., Chen, J. Y., Rambachan, A., Kleinberg, J., and Mullainathan, S. Evaluating the world model implicit in a generative model. In *The Thirty-eighth Annual Conference on Neural Information Processing Systems*, 2024. URL <https://openreview.net/forum?id=aVK4JFpegy>.

Vafa, K., Chang, P. G., Rambachan, A., and Mullainathan, S. What has a foundation model found? using inductive bias to probe for world models, 2025. URL <https://arxiv.org/abs/2507.06952>.

Wang, H., Fu, T., Du, Y., et al. Scientific discovery in the age of artificial intelligence. *Nature*, 620(7972):47–60, 2023. doi: 10.1038/s41586-023-06221-2. URL <https://doi.org/10.1038/s41586-023-06221-2>.

Wang, Y., Huang, H., Rudin, C., and Shaposhnik, Y. Understanding how dimension reduction tools work: An empirical approach to deciphering t-sne, umap, trimap, and pacmap for data visualization. *Journal of Machine Learning Research*, 22(201):1–73, 2021. URL <http://jmlr.org/papers/v22/20-1061.html>.

Wortsman, M., Ilharco, G., Kim, J. W., Li, M., Kornblith, S., Roelofs, R., Gontijo-Lopes, R., Hajishirzi, H., Farhadi, A., Namkoong, H., and Schmidt, L. Robust fine-tuning of zero-shot models. *arXiv preprint arXiv:2109.01903*, 2021. <https://arxiv.org/abs/2109.01903>.

Zimmermann, R. S., Sharma, Y., Schneider, S., Bethge, M., and Brendel, W. Contrastive learning inverts the data generating process. In Meila, M. and Zhang, T. (eds.), *Proceedings of the 38th International Conference on Machine Learning*, volume 139 of *Proceedings of Machine Learning Research*, pp. 12979–12990. PMLR, 18–24 Jul 2021. URL <https://proceedings.mlr.press/v139/zimmermann21a.html>.

Appendix Contents

A	Additional Related Work	14
B	Formal Derivation of Equation (2)	15
C	Fluid Dynamics Experiment Setting	17
D	Manifold Visualization Methodology	18
E	Analysis of Solar System Replication	19
F	Parameter Change Analysis	20
G	Full Fine-tuning Data Distribution Analysis	20
H	Symbolic Formula Comparison	21

A. Additional Related Work

We situate our work at the intersection of mechanistic interpretability and the dynamics of transfer learning for AI-driven scientific discovery.

AI for Physics. The discovery of physical laws from data typically relies on two paradigms: enforcing laws via architectural priors, such as Hamiltonian/Lagrangian Neural Networks (HNNs/LNNs) (Greydanus et al., 2019; Cranmer et al., 2020; Karniadakis et al., 2021), or post-hoc Symbolic Regression (SR) (Schmidt & Lipson, 2009a). While SR methods attempt to learn symbolic expressions directly from high-dimensional inputs (Biggio et al., 2021; Kamienny et al., 2022), they often struggle with the curse of dimensionality. However, recent findings suggest that standard SSL suffices to identify dynamics without physics-specific constraints (Chen et al., 2022; Reizinger et al., 2025). Internò et al. (2025) observe that physically consistent dynamics from natural video emerge as linear “straight” trajectories in pre-trained latent spaces, whereas AI-generated video creates curved latent trajectories due to physical artifact implausibility.

Inductive Biases and Causal Discovery. The mechanism by which generic foundation models capture physical laws remains a subject of intense debate. Vafa et al. (2025) utilized “Inductive Bias Probes” to audit models for physical compliance, concluding that standard Transformers achieve high predictive accuracy but fail to internalize fundamental forces. Responding to this, Liu et al. (2026) argue that this failure stems from a lack of architectural constraints; they distinguish between “Keplorian” world models (curve-fitting) and “Newtonian” models (causal forces), demonstrating that the latter only emerge when specific biases are enforced. However, recent mechanistic interpretability studies suggest that causal structures may emerge naturally without such constraints (Nanda et al., 2023b). Spies et al. (2025) identified latent “World Models” in maze-solving Transformers, while Mencattini et al. (2026) demonstrated that causal effects can be explicitly recovered from frozen foundation models using Sparse Autoencoders (Klindt et al., 2025), effectively disentangling the “treatment” variables from unstructured representations.

Self-Supervised Learning and Emergent World Models. The capability of next-token prediction to induce compact belief states remains a subject of active debate. Teoh et al. (2025) argue that in generic discrete domains, this objective is theoretically insufficient without auxiliary losses. This view is challenged by empirical findings in game playing (Nanda et al., 2023b) and mechanistic analysis (Nanda et al., 2023a), which demonstrate that Transformers develop linear representations

of “world models”—solely from the predictive objective. From a theoretical standpoint, work on identifiability in SSL (Zimmermann et al., 2021; Kügelen et al., 2021) suggests that it can provably recover ground-truth latent factors given sufficient data diversity. In the context of continuous physics dynamics, we align with the view that residual networks function as discretizations of Ordinary Differential Equations (Haber & Ruthotto, 2017; Chen et al., 2018a).

B. Formal Derivation of Equation (2)

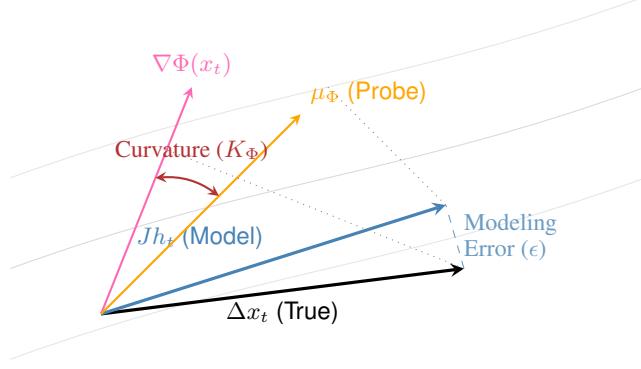


Figure 8. Geometry of the Linear Probe Error Bound. The total error stems from two mismatches: 1) **Modeling Error:** The deviation of the model’s linearized update (Jh_t) from the true physical update (Δx_t). 2) **Curvature Error:** The angular deviation between the local physical gradient ($\nabla\Phi_t$) and the global mean gradient (μ_Φ) used by the fixed linear probe. High curvature (K_Φ) increases this angle.

We derive the error bound for the linear probe, explicitly handling the approximation errors induced by the non-linearity of the physical functional and the neural decoder.

B.1. Proof Strategy: The Double Linearization

Our goal is to bound the error of a linear probe P_{W^*} mapping the latent representation h_t directly to the physical quantity update Δs_t . To do this, we decompose the true non-linear transition into two approximating linear steps:

1. **Physics Linearization:** We approximate the curved physical law $\Delta s_t = \Phi(x_{t+1}) - \Phi(x_t)$ via its local gradient $\nabla\Phi(x_t)$. The error in this step depends on the *Physical Curvature* K_Φ .
2. **Model Linearization:** We approximate the non-linear residual decoder $g(h_t)$ via its Jacobian J . The error in this step is bounded by the *SSL Prediction Error* ϵ .

The total probe error is derived by bounding the mismatch between the fixed linear probe W^* (which must average over the state space) and these local linearizations.

B.2. Setup and Definitions

Let the state space be $\mathcal{X} \subset \mathbb{R}^n$. The physical functional is $\Phi : \mathcal{X} \rightarrow \mathbb{R}^k$, such that $s_t = \Phi(x_t)$. The target update is $\Delta s_t = s_{t+1} - s_t$. The model predicts $\hat{x}_{t+1} = x_t + g(h_t)$, where $g(\mathbf{0}) = \mathbf{0}$.

- **Physics Dynamics:** $x_{t+1} = x_t + \Delta x_t$, where $\Delta x_t \approx \Delta t \cdot f(x_t)$.
- **SSL Objective:** $\mathbb{E}[\|\hat{x}_{t+1} - x_{t+1}\|^2] \leq \epsilon$.
- **Linear Probe:** $P_{W^*}(h_t) = W^* h_t$, where $W^* \in \mathbb{R}^{k \times d}$.

B.3. Step-by-Step Derivation

Step 1: Linearizing the Physical Law (Φ). Assuming Φ is C^2 -smooth, we expand the physical update around the current state x_t :

$$\Delta s_t = \nabla \Phi(x_t)^\top \Delta x_t + R_\Phi(x_t, \Delta t) \quad (10)$$

where the remainder is bounded by the curvature constant K_Φ (the Lipschitz constant of $\nabla \Phi$):

$$\|R_\Phi\| \leq \frac{1}{2} K_\Phi \|\Delta x_t\|^2 \in \mathcal{O}(\Delta t^2) \quad (11)$$

Step 2: Linearizing the Model Decoder (g). Since the decoder is origin-preserving, we linearize around the null latent $h_t = \mathbf{0}$:

$$g(h_t) = J h_t + R_g(h_t) \quad (12)$$

where $J = \nabla g(\mathbf{0}) \in \mathbb{R}^{n \times d}$ is the Jacobian. The SSL error constraint $\mathbb{E}[\|g(h_t) - \Delta x_t\|^2] \leq \epsilon$ implies that the linear term $J h_t$ approximates the true physical update Δx_t up to the training error and higher-order terms. Specifically, $\mathbb{E}[\|J h_t - \Delta x_t\|^2] \approx \epsilon$.

Step 3: Defining the Optimal Fixed Probe. A linear probe W^* must correspond to a single, time-independent matrix. The optimal choice is the projection of the *expected global gradient* onto the decoder's tangent space:

$$W^* = \mu_\Phi^\top J \quad (13)$$

where $\mu_\Phi = \mathbb{E}_x[\nabla \Phi(x)] \in \mathbb{R}^{n \times k}$ is the mean gradient of the functional over the state distribution.

Step 4: Error Decomposition. We analyze the squared error of the probe prediction against the true update:

$$\mathcal{L}_{\text{probe}} = \mathbb{E}[\|W^* h_t - \Delta s_t\|^2] \quad (14)$$

Substituting the linearizations from Steps 1 and 2, and adding/subtracting the term $\nabla \Phi(x_t)^\top J h_t$ (the local linear approximation):

$$\begin{aligned} \|W^* h_t - \Delta s_t\| &= \|\mu_\Phi^\top J h_t - (\nabla \Phi(x_t)^\top \Delta x_t + R_\Phi)\| \\ &= \underbrace{\|(\mu_\Phi - \nabla \Phi(x_t))^\top J h_t + \nabla \Phi(x_t)^\top (J h_t - \Delta x_t)\|}_{\text{Term A: Curvature Mismatch}} - R_\Phi \end{aligned} \quad (15)$$

Using the inequality $(a + b + c)^2 \leq 3(a^2 + b^2 + c^2)$, we bound the expectation:

Analyzing Term A (Curvature Mismatch): This term measures the error of using the *average* gradient μ_Φ instead of the *local* gradient $\nabla \Phi(x_t)$. Applying Cauchy-Schwarz:

$$\mathbb{E}[\text{Term A}^2] \leq \mathbb{E}[\|\mu_\Phi - \nabla \Phi(x_t)\|^2 \|J h_t\|^2] \quad (16)$$

We bound the step size $\|J h_t\|^2 \leq C_{\text{step}}$. The remaining term is strictly the variance of the gradient, $\text{Var}(\nabla \Phi(x))$. Using the Lipschitz property of the gradient (Curvature K_Φ):

$$\text{Var}(\nabla \Phi(x)) \leq K_\Phi^2 \cdot \text{Var}(x) \quad (17)$$

Thus, $\mathbb{E}[\text{Term A}^2] \leq C_{\text{step}} K_\Phi^2 \text{Var}(x)$.

Analyzing Term B (Modeling Error): This term measures the failure of the model to produce the correct state update.

$$\mathbb{E}[\text{Term B}^2] \leq \sup_x \|\nabla \Phi(x)\|^2 \cdot \mathbb{E}[\|J h_t - \Delta x_t\|^2] \leq C_{\text{grad}} \cdot \epsilon \quad (18)$$

Final Bound. Combining terms, we obtain the final inequality:

$$\mathbb{E} [\|P_{W^*} h_t - \Delta s_t\|^2] \leq C_1 \cdot \epsilon + C_2 [K_\Phi^2 \cdot \text{Var}(x)] + \mathcal{O}(\Delta t^4) \quad (19)$$

C. Fluid Dynamics Experiment Settings

To validate the generality of our non-invasive probing framework, we applied it to three complex, high-dimensional fluid dynamics simulations from *The Well* benchmark (Ohana et al., 2024). We tested multiple neural simulator architectures—UNet (Ronneberger et al., 2015), UNetConvNext (Liu et al., 2022), FNO (Li et al., 2021b), and TFNO (Li et al., 2021a)—all trained on a self-supervised next-step prediction task.

Our objective was to determine if these models implicitly learned the conservation of total internal energy (E_{int}) purely from observing state transitions. We extracted frozen activations $h(t)$ from the bottleneck (U-Nets) or the final spectral block (FNOs) and trained a linear probe to predict $E_{\text{int}}(t+1)$.

C.1. 2D Turbulent Radiative Layer (TRL-2D)

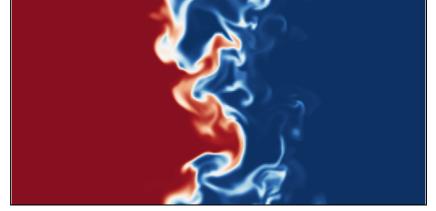
This simulation models a 2D slice of a stellar atmosphere or accretion disk, governed by compressible magnetohydrodynamics (MHD) with radiative transfer. It captures the interplay between magnetic turbulence and radiative cooling.

Governing Equations. The system evolves according to:

$$\frac{\partial \rho}{\partial t} + \nabla \cdot (\rho v) = 0 \quad (20)$$

$$\frac{\partial(\rho v)}{\partial t} + \nabla \cdot (\rho v v + P) = 0 \quad (21)$$

$$\frac{\partial E}{\partial t} + \nabla \cdot ((E + P)v) = -\frac{E}{t_{\text{cool}}} \quad (22)$$



where the internal energy is defined by the ideal gas law: $E = P/(\gamma - 1)$ with $\gamma = 5/3$.

Figure 9. TRL-2D Simulation (Stachenfeld et al., 2021)

Task & Probe Configuration.

- **SSL Input:** $\{\rho, v_x, v_y, P\}$ at time t . Resolution: 128×128 .
- **Probe Target:** The total internal energy $E_{\text{int}} = \int_{\Omega} \frac{P}{\gamma-1} dV$.

C.2. 3D Red Supergiant Convective Envelope (RSG-3D)

This dataset simulates the outer convective envelope of a red supergiant star, governed by compressible hydrodynamics with radiative transfer. It features strong convective upflows and buoyancy-driven turbulence.

Governing Equations.

$$\frac{d\rho}{dt} = -\rho \nabla \cdot V \quad (23)$$

$$\frac{d^2 \mathbf{r}}{dt^2} = -\frac{\nabla P}{\rho} + \mathbf{a}_{\text{visc}} - \nabla \Phi_{\text{grav}} \quad (24)$$

$$\frac{du}{dt} = -\frac{P}{\rho} \nabla \cdot V + \frac{\Gamma - \Lambda}{\rho} \quad (25)$$

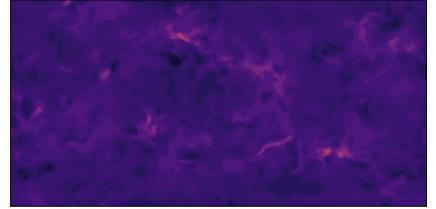


Figure 10. RSG-3D Simulation (Goldberg et al., 2022)

Here, u is specific internal energy, Φ_{grav} is gravitational potential, and Γ, Λ represent radiative heating/cooling.

Task & Probe Configuration.

- **SSL Input:** $\{\rho, P, v_r, v_\theta, v_\phi\}$ at time t . Resolution: $64 \times 64 \times 64$.
- **Probe Target:** Total internal energy $E_{\text{int}} = \int_{\Omega} \rho u dV$. Note that specific energy u is not an input; the probe must implicitly derive u from P and ρ via the equation of state.

C.3. 3D Supernova Explosion (SN-3D)

A simulation of a core-collapse supernova, involving extreme relativistic velocities, shock waves, and a simplified nuclear burning network.

Governing Equations. The system includes the standard conservation of mass and momentum, but energy is dominated by nuclear terms:

$$\frac{\partial E}{\partial t} + \nabla \cdot \dots = -cG_r^0 - \rho V \cdot \nabla \Phi \quad (26)$$

$$\frac{\partial I}{\partial t} + \mathbf{c} \mathbf{n} \cdot \nabla I = S(I, \mathbf{n}) \quad (27)$$

where S is the source term from the nuclear burning network, creating extreme non-linearities.

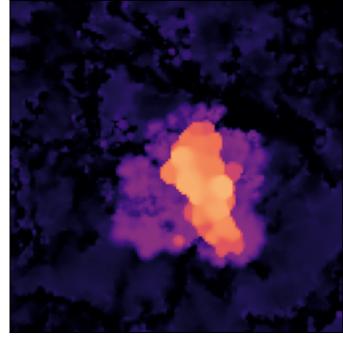


Figure 11. SN-3D Simulation (Hirashima et al., 2023)

Task & Probe Configuration.

- **SSL Input:** $\{\rho, P_{\text{gas}}, T, v_x, v_y, v_z\}$. Resolution: $64 \times 64 \times 64$.
- **Probe Target:** Internal energy density $\epsilon = \rho u$, derived from a lookup table of nuclear equations of state (EOS), not a simple ideal gas law.

C.4. Data Partitioning

To ensure the validity of OOD generalization claims, we enforced a strict separation of datasets following setting described in (Ohana et al., 2024).

1. **SSL Training Set (D_{train}):** Used *only* for pre-training the backbone model m_θ .
2. **Probe Training Set (D_{probe}):** A strictly In-Distribution (ID) subset held out from D_{train} . The probe learns the mapping W on this data. *Crucially, the probe never sees OOD data during training.*
3. **OOD Test Sets (D_{OOD}):** Completely distinct simulations with physical parameters (e.g., cooling rates, stellar mass) unseen in either D_{train} or D_{probe} .

C.5. Implementation and Reproducibility Details

We detail the training hyper-parameters for the SSL backbone, the non-invasive probes, and the invasive baselines in Table 3.

SSL Training. All models were trained using the AdamW optimizer with a cosine annealing schedule. Training was performed on $2 \times$ NVIDIA H100 GPUs.

Probe Training. The linear probes were trained using Ridge Regression (L2 regularization) on the frozen representations.

Invasive Baselines.

- **MLP Probe:** A non-linear probe composed of two dense layers ($d \rightarrow 254 \rightarrow 1$) with ReLU activation. Trained using Adam on the frozen representation.
- **Full Fine-tuning (IBP):** The entire backbone m_θ is unfrozen and updated to predict the physical target exactly as in (Vafa et al., 2025).

D. Manifold Visualization Methodology

The manifold visualizations in Figure 3 (Top) are generated to contrast the internal representation geometry of the pre-trained SSL model (m_θ) and the fine-tuned model ($m_{\theta'}$). The process, as implemented in the provided analysis script, is as follows:

- 1. *Data Sampling:* We sample 50,000 data points from the validation set. For each point, we compute and store its corresponding ground-truth force magnitude ($\|\vec{F}\| = \sqrt{F_x^2 + F_y^2}$), which is used to color the final plot.

Table 3. Dataset & Hyperparameter Specifications. Comparison of the Newtonian Two-Body task (Vafa et al., 2025) and The Well benchmarks (Ohana et al., 2024).

Feature	Two-Body (Newton)	TRL-2D	RSG-3D	SN-3D
Spatial Resolution	2 (x, y coords)	384×128	$256 \times 128 \times 256$	64^3
Total Trajectories	10×10^6	90	29	740
Data Partitioning (Train / Probe / OOD)				
SSL Train (N_{SSL})	10×10^6	72	23	592
Probe Train (N_{Probe})	10,000	9	3	74
OOD Test Set	5 Galaxies	9 Cooling Rates	3 Phases	27 Env. Vars
Probe & Adaptation Hyperparameters				
Optimizer	AdamW	AdamW	AdamW	AdamW
Batch Size	64	32	8	8
Ridge Reg. (α)	1.0	1.0	1.0	1.0
MLP Hidden Dim	254	254	254	254
Adaptation LR	1e-3	1e-3	1e-3	1e-3

- 2. *Activation Extraction*: We process these 50,000 inputs through both the frozen pre-trained model and the fine-tuned model. Using a PyTorch forward hook, we extract the high-dimensional activation vectors (embedding dimension $d = 768$) from the *input* to the 10th decoder block ('blocks.9'). This layer is chosen to match the layer used for linear probing. This step yields two distinct sets of high-dimensional representations: $h_{ssl} \in \mathbb{R}^{50000 \times 768}$ from the SSL model and $h_{ft} \in \mathbb{R}^{50000 \times 768}$ from the fine-tuned model.
- 3. *Dimensionality Reduction*: To visualize these 768-dimensional manifolds in 2D, we apply the PaCMAP dimensionality reduction algorithm (Wang et al., 2021). Crucially, we fit the PaCMAP algorithm *independently* to each set of activations. This generates two 2D projections, $Z_{ssl} = \text{PaCMAP}(h_{ssl})$ and $z_{ft} = \text{PaCMAP}(h_{ft})$. While other algorithms like t-SNE and UMAP were considered, we selected PaCMAP for its superior ability to preserve both *local* and *global* data structure.

E. Solar System Experiment

We replicated the true solar system setup from Vafa et al. (2025) to evaluate OOD generalization on real orbital dynamics. As shown in Figure 12, our non-invasive probe (blue) consistently outperforms the invasive fine-tuning baseline (green), which exhibits erratic behavior across the planetary suite.

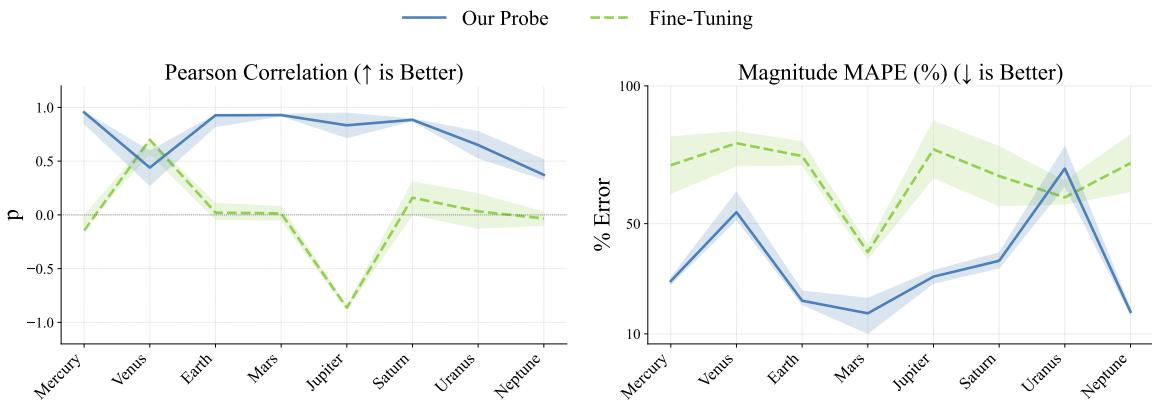


Figure 12. Solar System Generalization Analysis. Comparison of force vector prediction performance. The invasive probe (Green) fails systematically, exhibiting high variance and negative correlations. The non-invasive probe (Blue) remains robust, though it highlights specific OOD challenges for Venus and Uranus.

Outlier Analysis (Venus & Uranus). While the non-invasive probe generally succeeds, we identify two distinct failure modes. For *Venus*, the probe's Pearson correlation drops to $\rho \approx 0.4$ with a corresponding spike in error (MAPE > 50%).

Conversely, for *Uranus*, the probe maintains high linearity ($\rho \approx 0.8$) but suffers from calibration error ($\text{MAPE} \approx 70\%$). This suggests that while the model correctly encodes the *directionality* of the force at Uranus’s distance, the magnitude scaling at the outer solar system boundary drifts from the training distribution.

Systemic Failure of Invasive Probing. In contrast, the failure of the invasive fine-tuning probe is not merely an issue of precision, but of fundamental physical correctness. For Jupiter, the invasive probe exhibits a strong *negative* Pearson correlation ($\rho \approx -0.9$). This indicates that the fine-tuning process has inverted the vector field, effectively predicting a repulsive force rather than an attractive one. For the inner planets, the invasive probe shows near-zero correlation ($\rho \approx 0.0$), confirming the mechanistic finding in Section 4 that dynamic invariants (like the distance-force relationship) are erased during adaptation.

Our non-invasive probe maintains high performance ($\rho > 0.85$) on these same planets, confirming that the correct physical world model exists in the backbone but is destroyed by the invasive measurement.

F. Parameter Change Analysis

To quantify the invasiveness of the fine-tuning process, we analyze the magnitude of weight modifications across the transformer architecture. We compute the layer-wise relative change using the Frobenius norm:

$$\delta^{(l)} = \frac{\|\theta'^{(l)} - \theta^{(l)}\|_F}{\|\theta^{(l)}\|_F} \quad (28)$$

where $\theta^{(l)}$ represents the weights of layer l in the pre-trained SSL model, and $\theta'^{(l)}$ represents the weights after fine-tuning on the inductive bias task.

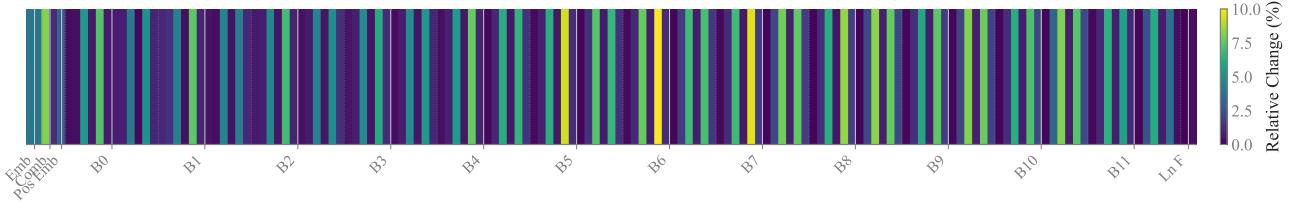


Figure 13. **Global Invasiveness:** Parameter heatmap showing modification concentrated in deep layers.

As illustrated in Figure 13, the modification is highly non-uniform.

- **Early Layers (Blocks 0–4):** Exhibit high stability ($\delta^{(l)} < 0.02$), indicating that the basic feature extraction for orbital trajectories remains largely intact.
- **Deep Layers (Blocks 5–10):** Show a sharp spike in parameter modification ($\delta^{(l)} > 0.10$), particularly in the MLP projections and Self-Attention output matrices.

This concentrated modification in the deeper blocks corroborates our “Erasure” hypothesis. In hierarchical models, deep layers typically encode high-level semantic variables and dynamic rules (?). The fact that the optimizer selectively targets these layers suggests it is overwriting the model’s high-level physics engine (the “World Model”) to replace it with the shallow heuristics required by the narrow fine-tuning distribution.

G. Full Fine-tuning Data Distribution Analysis

We analyze the distributional shift between the self-supervised pre-training dataset (D_{SSL}) and the downstream fine-tuning dataset (D_{task}).

As hypothesized in Section 4, invasive adaptation on narrow distributions encourages the model to discard complex physical dependencies in favor of statistical shortcuts. Figure 14 illustrates this discrepancy across key variables:

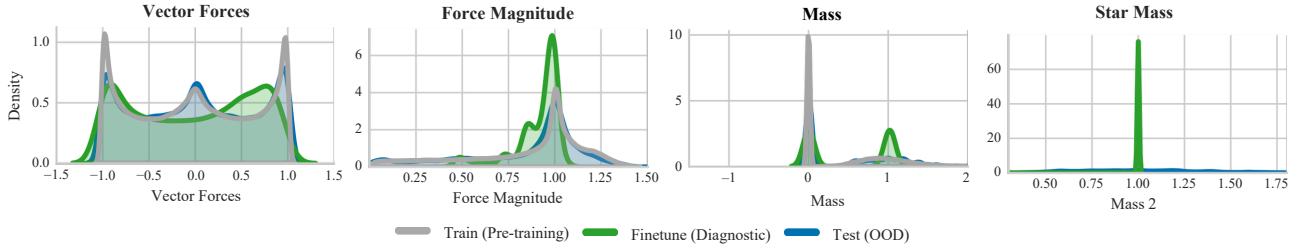


Figure 14. Data Distribution Shift. Comparison of physical variables between SSL (Grey) and Fine-tuning (Green). The fine-tuning data collapses to narrow regimes (e.g., single star mass), inducing simplicity bias.

- **Star Mass (m_2):** While the SSL pre-training data covers a continuous range of stellar masses, the fine-tuning dataset collapses to a single point mass at $m_2 = 1.0$. This lack of variance removes the incentive for the model to maintain m_2 as an active variable, leading to the “erasure” observed in our mechanistic analysis.
- **Force Magnitude ($\|\vec{F}\|$):** The fine-tuning dataset exhibits a highly peaked distribution centered around $\|\vec{F}\| \approx 1.0$, failing to cover the heavy tails of high-force interactions or low-force interactions. This restricts the optimizer’s ability to learn the full inverse-square law ($1/r^2$), as gradients are dominated by a specific force regime.
- **Force Vectors (\vec{F}):** The distribution of vector components in the fine-tuning set differs significantly from the isotropic distribution seen during pre-training, potentially overfitting the model to specific orbital orientations rather than learning rotation-invariant physics.

H. Symbolic Formula Comparison

Table 4 presents the symbolic equations discovered for the gravitational force magnitude (F) via Symbolic Regression (SR). We compare the equations extracted from the Invasive Fine-Tuning baseline (IBP) (Vafa et al., 2025) against those recovered by *PhyIP*. The SR algorithm (PySR) attempts to fit the scalar force magnitude $\|\vec{F}\|$ using the state variables r (distance), m_1 (planet mass), and m_2 (star mass). The search was constrained to standard arithmetic and trigonometric (+, -, *, /, sin, cos)

Table 4. Symbolic Equation Discovery. Comparison of discovered laws. The Invasive Probe fits spurious correlations (nested sines), while the Non-Invasive Probe recovers the structural $1/r^2$ dependence.

Source	Complexity	Discovered Symbolic Equation
Ground Truth	-	$\ \vec{F}\ \propto \frac{m_1 m_2}{r^2}$
Invasive FT (IBP) <i>(Baseline)</i>	High	$\ \vec{F}\ \propto \left[\sin\left(\frac{1}{\sin(r - 0.24)}\right) + 1.45 \right] \cdot \frac{1}{1/r + m_2}$ <i>(Fails to isolate $1/r^2$; relies on high-freq artifacts)</i>
Non-Invasive (Ours)	Low (Rank 1)	$\ \vec{F}\ \approx \frac{1}{1/r + 1.16}$
Non-Invasive (Ours)	Med (Rank 2)	$\ \vec{F}\ \approx \sin(\sin(\sin(r \cdot 0.07) + 0.48))$
Non-Invasive (Ours)	Best (Rank 3)	$\ \vec{F}\ \approx \underbrace{\frac{0.1}{r^2}}_{\text{Recovered Physics}} + \underbrace{\sin(\dots)}_{\text{Residual Noise}}$

To ensure reproducibility, we generated 7 candidate equations using PySR’s simulated annealing. The “Best” equation reported in Table 4 was selected using the Pareto frontier of the `score` metric (minimizing MSE while penalizing complexity).