WHAT DRIVES THE PREDICTION? HOW NEURAL MODELS RELY ON PHYSICAL INFORMATION

Anonymous authors
Paper under double-blind review

ABSTRACT

Physical quantities such as energy and momentum can be decoded accurately from a neural model even when models differ sharply in their response to the same input change. We study this gap between physical representation and predictive reliance. Matched Invariant Interchange (MII) compares activation exchanges selected for their correlation with a physical quantity against random exchanges of the same size. A complementary regression task varies the availability of a target-correlated hint during training and removes its informativeness at evaluation. Our contributions are: *i) Decodability–coupling dissociation*: across five dynamics systems and five exchange sizes per system, high decodability coexists with a maximum absolute MII gap of 0.0068. *ii) Distribution-dependent reliance*: across five invariant-regression systems, informative-hint and decorrelated-hint training both yield decode $R^2 = 0.99$–1.00, yet mean absolute output–target correlation after hint permutation is 0.09 versus approximately 1.00. Decorrelated-hint training also increases sensitivity to invariant information relative to hint information. *iii) Output-local coupling*: in pretrained Poseidon-B, the decoder MII gap is 0.893 for the spatial integral of one velocity component, $\int u$, and 0.027 for kinetic energy. The comparisons distinguish readable physical information, training-dependent signal reliance, and sensitivity arising near the output. They support evaluating physical representations through matched interventions and behavior after a predictive training signal changes.

1 INTRODUCTION

Neural dynamics models predict how physical systems evolve. Energy and momentum provide interpretable quantities for examining their internal representations. A linear probe, fitted to recover a physical quantity from hidden activations, measures its accessibility. How does this readable information relate to the computation producing a prediction? When a training objective improves physical decodability, this distinction determines whether the gain reflects more readable information, a changed prediction, or both.

We introduce *Matched Invariant Interchange (MII)*, a controlled diagnostic that compares invariant-selected and random activation exchanges using the same donor examples and number of units. A donor supplies activations to a recipient example; the frozen model then predicts from the modified representation. Comparing how strongly each prediction's physical quantity tracks the donor's quantity measures the selection advantage. A complementary regression task supplies the physical state and a target-correlated hint, offering a shortcut to the target. Changing the hint's training correlation tests how data shapes reliance; permuting it at evaluation tests the resulting behavior.

The hint comparison links training-dependent intervention sensitivity to behavior despite similar probe scores. The Poseidon-B comparison links the largest selective effect to a linear decoder readout.

Known closed-form expressions give probe targets and readouts on predicted states a shared physical definition. Intervention responses therefore come from the frozen predictor's physical output, independently of the fitted probe. The hint task reads its scalar prediction directly. These choices separate accessible information, selective output response, and behavior under an input shift.

Our contributions establish three connected phenomena:

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

i) Decodability–coupling dissociation. Across pendulum, LC circuit, two-body spring, 2D central force, and 1D wave systems, high decodability coexists with a maximum MII gap of 0.0068 for $k \in \{1, 2, 4, 8, 16\}$ exchanged units. Objective changes improve some probe scores while preserving small response shifts; pendulum MLP, GRU, and Transformer models also show high decodability with small probe-direction correlations. This separates gains in readable physical information from gains in selective intervention response.

ii) Distribution-dependent reliance. Changing the training hint's informativeness changes relative invariant-versus-hint sensitivity with the regression task fixed. Both conditions encode the invariant at $R^2 = 0.99$–1.00, yet mean output correlation under hint permutation changes from 0.09 to approximately 1.00. Decorrelated-hint training increases sensitivity to invariant information relative to hint information, distinguishing reliance between models with similar probe scores.

iii) Output-local coupling. Selective effects concentrate near the output and depend on readout algebra. In 158M-parameter Poseidon-B, the decoder gap for $\int u$ reaches 0.893, compared with 0.027 for kinetic energy. Comparing layers and quantities helps choose intervention sites and interpret strong responses in relation to the output map.

2 Related work

Physical structure and predictive use. Linear probes measure accessible information (Alain & Bengio, 2017, Hewitt & Manning, 2019, Belinkov, 2022), with control tasks distinguishing representation quality from probe behavior (Hewitt & Liang, 2019). SciNet exposes physical variables through a designed bottleneck (Iten et al., 2020); symbolic and sparse methods recover equations (Brunton et al., 2016, Udrescu & Tegmark, 2020, Lemos et al., 2023, Cranmer et al., 2020b), while trajectory analysis and Siamese networks discover invariants (Liu & Tegmark, 2021, Wetzel et al., 2020). Hamiltonian and Lagrangian networks constrain the prediction architecture (Greydanus et al., 2019, Cranmer et al., 2020a), and Noether Networks learn conservation objectives used inside prediction (Alet et al., 2021). These methods recover physical structure or build it into prediction. For a frozen predictor, the remaining question is how exchanging readable information affects its physical output.

Intervention geometry and output readouts. Activation patching and causal abstraction link internal variables to behavior (Vig et al., 2020, Geiger et al., 2021, Meng et al., 2022); alignment methods learn intervention subspaces (Geiger et al., 2024, Wu et al., 2023). Studies of Othello, programs, and spatial information examine structured internal representations (Li et al., 2023, Nanda et al., 2023, Jin & Rinard, 2024, Gurnee & Tegmark, 2024). Superposition and dictionary-learning studies examine features distributed across neurons (Elhage et al., 2022, Bricken et al., 2023, Templeton et al., 2024). In video encoders, Joseph et al. (2026) characterize a high-dimensional motion-direction code and steer its decoded direction. Our measurement follows the exchanged activations through the frozen predictor to a physical output readout. Matching intervention size and donors tests the selection advantage; varying layer and readout tests how output structure shapes that advantage.

Shortcut reliance and behavioral evaluation. Shortcut learning asks which predictive regularities training favors (Geirhos et al., 2020). World-model evaluations based on consistency and adaptation test behavior beyond latent recoverability (Vafa et al., 2024; 2025). Physics-inspired attention changes the architecture itself (Maitra, 2026). Our hint task fixes architecture and target while changing a competing signal's training correlation. Relative intervention sensitivity and behavior under hint permutation expose differences between models with near-identical invariant decodability.

3 Study design and measurements

3.1 Three comparisons connect representation to prediction

We compare physical quantities across dynamics systems, hint informativeness within regression, and layers within a pretrained PDE model to connect encoding, reliance, and output response.

The five systems vary the dynamics, physical quantity, and state structure. Pendulum combines nonlinear dynamics with a non-quadratic potential; LC provides linear dynamics with quadratic energy. The two-body spring supports both energy and linear-momentum readouts. Central force adds

bilinear angular momentum, while the wave equation extends the comparison to a spatial field with energy aggregated across locations. Their known expressions keep probe targets and physical readouts explicit (Appendix B). Kepler adds energy and angular-momentum probe-direction observations. The pendulum MLP, GRU, and Transformer comparison tests architecture dependence within one physical task.

Conservation penalties, an auxiliary head, and rollout training test whether rewarding physical structure changes decodability and output response together. The hint experiment fixes architecture, optimizer, and regression target within each system while changing the input–target correlation. Hint permutation at evaluation tests generalization to that shift. Poseidon holds the pretrained model fixed while varying layer and readout; kinetic energy serves as a physical diagnostic whose evolution depends on the flow dynamics.

3.2 Decodability and output readouts

A dynamics model f_θ predicts $\hat{x}_{t+1}$ from state x_t, with hidden activation $h_\ell(x) \in \mathbb{R}^m$. For a physical quantity ϕ, the next-state readout is

$$g(x) = \phi(f_\theta(x)). \tag{1}$$

This readout measures a physical quantity in the predicted next state. In the hint task, the target is the standardized quantity $z(\phi(x))$, where z subtracts the mean and divides by the standard deviation. The readout $g(x, c) = \hat{y}$ is then the model's scalar prediction, used to test which input signal supports invariant regression.

Decodability is measured with a linear ridge probe on standardized activations $\tilde{h}$:

$$R^2_{\text{dec}} = 1 - \frac{\sum_i (\phi(x_i) - \hat{\phi}^{\text{CV}}_{-i}(\tilde{h}_i))^2}{\sum_i (\phi(x_i) - \bar{\phi})^2}, \qquad w = \arg\min_w \|\Phi - \tilde{H}w\|_2^2 + \lambda_r \|w\|_2^2. \tag{2}$$

The probe uses five-fold cross-validation on standardized activations. A shuffled-label probe repeats the fit after permuting the quantity labels, providing a control for spurious predictability. High R^2_{dec} means that a linear readout captures the physical quantity accurately; the intervention analysis then asks how a change to those activations propagates through the trained predictor.

3.3 Matched Invariant Interchange

MII pairs a physical selection rule with a matched random control and reads both responses through the frozen predictor. For a donor assignment π, select the k units most correlated with the quantity and interchange their activations:

$$\mathcal{U}_k = \text{top-}\, k_j |\, \text{corr}(h^{(j)}, \phi)|, \qquad h'_i[\mathcal{U}_k] = h_{\pi(i)}[\mathcal{U}_k]. \tag{3}$$

All other units retain the recipient's values. A random set $\mathcal{R}_k$ of the same size is patched using the same donor assignment. The next-state statistic is

$$\tau(\mathcal{S}) = \text{corr}(g(x_i; h'_i[\mathcal{S}]), \phi(x_{\pi(i)})), \qquad \Delta_k = \tau(\mathcal{U}_k) - \tau(\mathcal{R}_k). \tag{4}$$

The *MII gap* Δ_k (the selective coupling gap in the figures) measures the extra donor correlation obtained by choosing invariant-correlated units instead of a random set of the same size. For example, $\Delta_k = 0.0068$ means that invariant-based selection adds 0.0068 correlation units over the random control. Matching exchange size and donors isolates the selection advantage.

A complementary *probe-direction interchange* exchanges the component along the linear probe's normalized weight vector $\hat{u} = w/\|w\|_2$. In the standardized activation coordinates used by the probe, it is

$$\tilde{h}'_i = \tilde{h}_i + \hat{u}\hat{u}^\top(\tilde{h}_{\pi(i)} - \tilde{h}_i). \tag{5}$$

This operation replaces one projected component with the donor's value and preserves orthogonal components. It tests a weighted combination of units; MII instead exchanges the selected units themselves.

For comparisons involving output transformations, we also use the interchange slope

$$\beta = \frac{\text{cov}(\Delta g, \Delta\phi)}{\text{var}(\Delta\phi)}, \quad \Delta g = g(x_i; h'_i) - g(x_i), \quad \Delta\phi = \phi(x_{\pi(i)}) - \phi(x_i). \tag{6}$$

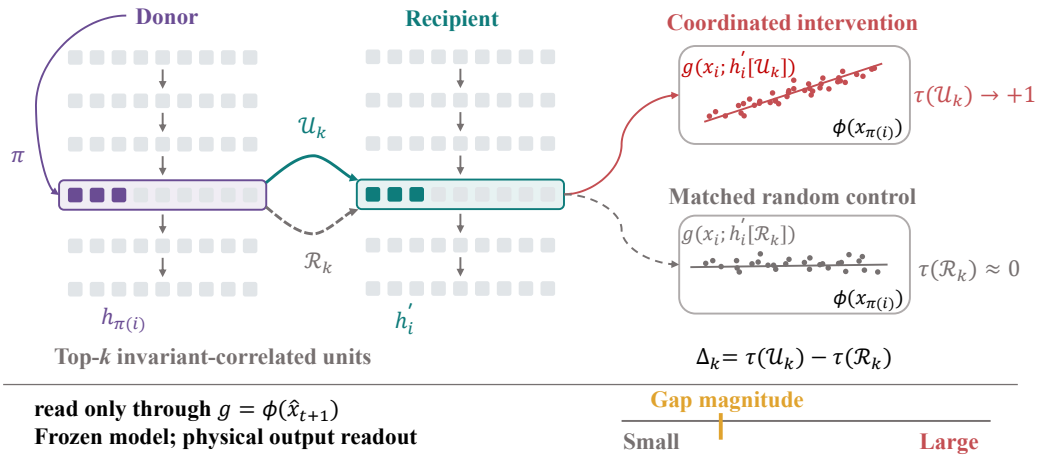


Figure 1: **Matched Invariant Interchange (MII).** The same donor supplies an invariant-selected unit set and a random set of equal size. Each patched activation is passed through the frozen remainder of the model, and the physical quantity is evaluated on its prediction. Comparing the resulting donor correlations yields Δ_k. Probe-direction interchange provides a complementary test along the linear probe weight vector.

Algorithm 1: Matched Invariant Interchange (MII)

1. Record clean activations H, quantities Φ, and physical output readouts g_0.
2. Evaluate five-fold ridge and shuffled-label probes on standardized activations. Reject a probe when $R^2_{\text{shuffle}} > 0.15$.
3. Assign donors with a random derangement. For each exchange size k, select $\mathcal{U}_k$ by Eq. 3 and sample a random set $\mathcal{R}_k$ of equal size.
4. Using the same donors, run the selected-unit and random-unit swaps separately through the frozen model. Evaluate the physical quantity on each predicted state.
5. Return both transfer correlations and their difference Δ_k for every exchange size. Summarize the effects across seeds.

The slope measures output change per unit of donor–recipient quantity difference. Its magnitude therefore depends on the readout's units. We use τ for transfer correlation, Δ_k for the MII gap, and β for this change-based slope.

Reading effect sizes. A positive Δ_k favors invariant-selected units; a negative gap favors the random control. We compare values across exchange sizes: the five-system maximum is 0.0068, while the largest Poseidon decoder gap is 0.893. Seed summaries and random-direction tests characterize uncertainty and direction specificity.

4 Decodability–Coupling Dissociation

Across five systems, physical quantities decode accurately, yet selecting their correlated units yields nearly the same donor correlation as selecting random units. High probe scores therefore coexist with little advantage from choosing those units for exchange. The largest $|\Delta_k|$ is 0.0068, and the mean absolute gap is 0.0014 over 25 combinations of system and exchange size. Table 1 places the probe and intervention measurements side by side; Figure 2A shows that the separation persists as the intervention expands from one to 16 units.

A physical quantity may be spread across several units. Increasing k tests whether a wider correlated set reveals the selective response missed by a narrow intervention. The flat curves show that expanding the set from one to 16 units preserves its small advantage over the matched control.

Table 1: Next-state decodability and intervention response. τ_{dir} is the probe-direction transfer correlation; $\max_k |\Delta_k|$ summarizes MII over $k \in \{1, 2, 4, 8, 16\}$. Dashes mark quantities evaluated only with probe-direction interchange. The lower block gives ten-seed uncertainty estimates and random-direction permutation tests.

System	Invariant	R^2_{dec}	τ_{dir}	$\max_k \lvert\Delta_k\rvert$
Pendulum	energy	1.00	-0.01	0.002
LC circuit	energy	0.97	0.00	0.003
Spring	energy	0.97	0.00	0.003
Central force (2D)	angular momentum	0.99	$+0.02$	0.007
Wave PDE	energy	0.99	0.00	0.002
Kepler ($1/r^2$)	energy	0.80	-0.02	—
Kepler ($1/r^2$)	angular momentum	0.88	$+0.04$	—
Statistics (10 seeds, 300 permutations):				
Pendulum	energy	—	-0.010 $[-0.019, +0.001]$	$p \approx 0.73$ (n.s.)
Central force	angular momentum	—	-0.012 $[-0.023, -0.001]$	$p \approx 0.71$ (n.s.)

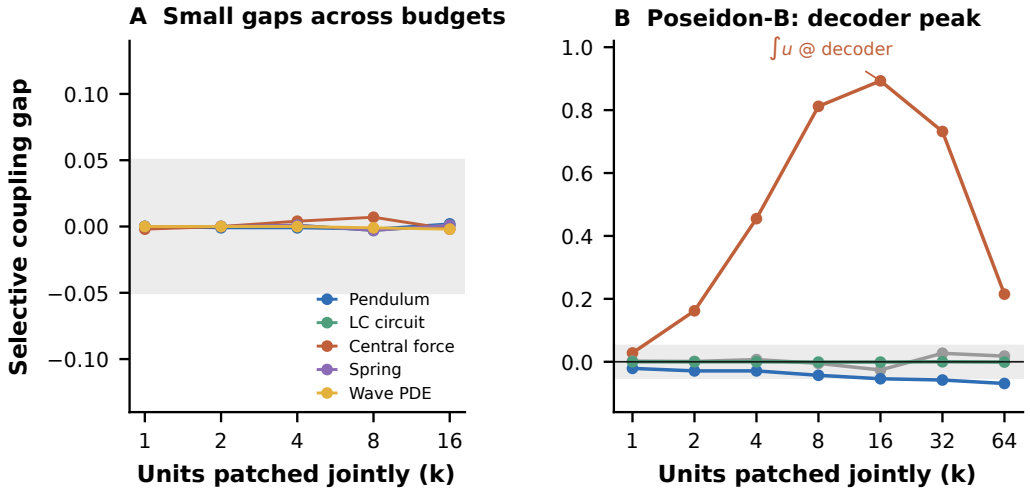


Figure 2: **Selective coupling across exchange sizes.** (A) Invariant-selected and random-unit swaps yield nearly equal donor correlations in all five simulated systems; all 25 gaps lie within 0.007 of zero. (B) The Poseidon decoder response for $\int u$ grows over intermediate exchange sizes and peaks at $k = 16$. Reading each curve across k distinguishes a persistent small gap from a selective effect concentrated at a particular intervention size.

The architecture comparison supports the same distinction: pendulum energy decodes at $R^2 = 1.00$ in the MLP and 0.98 in both the GRU and Transformer, with probe-direction correlations -0.01, $+0.02$, and $+0.02$. Kepler energy and angular momentum provide further probe-direction observations (Table 1).

Random-direction controls address whether small responses are specific to the invariant direction. Ten-seed estimates give pendulum $\tau = -0.010$ with 95% interval $[-0.019, +0.001]$ and central force -0.012 with $[-0.023, -0.001]$. The random-direction permutation tests give $p \approx 0.73$ and $p \approx 0.71$, respectively. For central force, the negative interval indicates a small inverse association; the random-direction test places this response within the variation produced by control directions.

4.1 Objective changes preserve the dissociation

Rewarding physical structure during training improves some probe scores while leaving intervention responses nearly unchanged. Across eight objective conditions, the mean intervention-response shift is -0.0003, and the reported intervals include zero. Table 2 pairs each probe-score change with its

Table 2: Conservation-oriented objectives can improve decodability while preserving small response shifts (4–5 seeds). The output penalty uses $\lambda = 10$. Every decode change compares the modified objective with its own baseline; the mean response shift across eight conditions is -0.0003.

Manipulation	System	Response shift	Decode shift
Output-path conservation	Pendulum	-0.001	≈ 0
Output-path conservation	LC circuit	-0.000	≈ 0
Output-path conservation	Spring	-0.000	≈ 0
Output-path conservation	Central force	-0.000	$0.715 \rightarrow 0.775$
Output-path conservation	Wave PDE	-0.000	$0.294 \rightarrow 0.313$
Auxiliary invariant head	Pendulum	-0.000	—
Output-path, single step	Wave PDE	-0.000	$0.438 \rightarrow 0.551$
K=12-step rollout	Pendulum	-0.000	≈ 0
Mean shift		**-0.0003**	

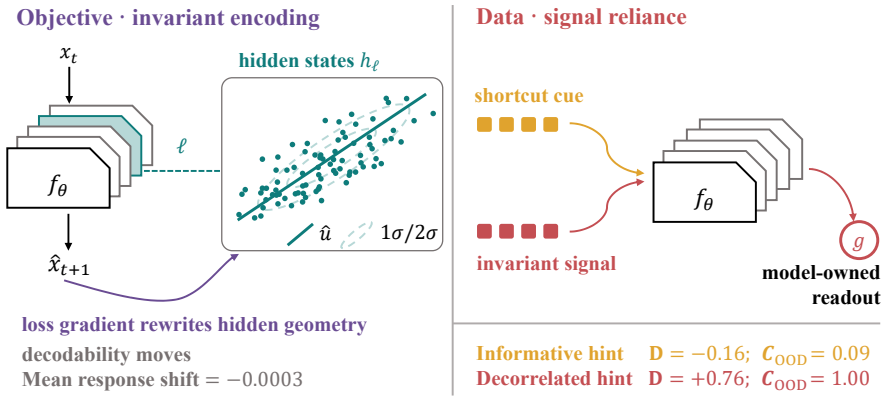


Figure 3: **Encoding and reliance respond to different training changes.** The objective comparison tests whether conservation-oriented training moves probe scores and intervention responses together. The hint comparison changes the availability of a competing signal while fixing the regression task, then relates the invariant-versus-hint contrast D to behavior with permuted hints. Together, the two comparisons distinguish improvements in encoding from changes in signal reliance.

intervention-response shift. The conservation penalty is

$$\mathcal{L}_\lambda = \|\hat{x}_{t+1} - x_{t+1}\|^2 + \lambda \|\phi(\hat{x}_{t+1}) - \phi(x_t)\|^2. \tag{7}$$

Table 2 compares this output penalty at $\lambda = 10$ on five systems, a pendulum auxiliary invariant head, a wave single-step variant, and a pendulum $K = 12$ rollout variant. These conditions test whether adding physical supervision through different training objectives changes the intervention response.

For central force the decode score rises from 0.715 to 0.775; for the wave conditions it rises from 0.294 to 0.313 and from 0.438 to 0.551. The largest increase is 0.113 in the single-step wave condition, whose response shift rounds to zero.

The rollout condition adds a behavioral check. At $\lambda = 10$, rollout MSE rises from 0.0015 to 0.0328, approximately $22\times$, and energy drift rises from 0.050 to 0.065. At this penalty strength, rollout predictions become less accurate and their energy drift increases. Predictive error and drift assess dynamics behavior; the probe measures accessibility of the physical quantity.

5 Distribution-Dependent Reliance

Changing the training hint's informativeness separates behavior while invariant decodability stays near ceiling. The preceding dynamics experiments alter the incentive to encode physics; here, controlled invariant regression alters the usefulness of a competing input signal. Its two routes to the target—the

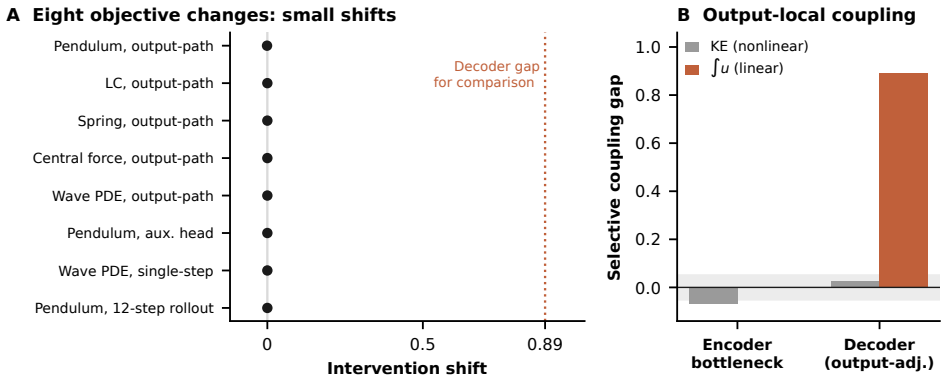


Figure 4: **Objective changes and output-local coupling.** (A) Eight objective modifications produce small response shifts relative to their respective baselines. The dotted 0.893 marker gives the scale of the Poseidon decoder gap shown in (B). (B) Comparing each quantity across layers, and the two quantities at the decoder, localizes the large selective response to the linear functional near the output. The two panels show complementary experimental comparisons.

physical state and an informative hint—let us compare relative intervention sensitivity with behavior after the hint loses its predictive value.

Task and distribution intervention. Each input includes physical state x and hint $c = z(\phi(x)) + \varepsilon$, with additive Gaussian noise ε. The target is $z(\phi(x))$. Shortcut-trained models receive the informative hint; decorrelated-hint models receive a permutation of the target values across the batch. Both groups are evaluated with permuted hints. There are four seeds per system and condition. This shift tests reliance on an input correlation while leaving the physical target unchanged.

Two task-specific measurements. We measure held-out behavior with $C_{\mathrm{OOD}} = |\operatorname{corr}(\hat{y}, \phi)|$, the absolute correlation between prediction and physical target under hint permutation. The intervention contrast is

$$D = \mathrm{dep}(\phi) - \mathrm{dep}(c), \tag{8}$$

where each dep is the transfer correlation after probe-direction interchange along the corresponding probe direction, read from $\hat{y}$. Thus D compares invariant and hint directions; Δ_k compares invariant-selected and random unit sets. A larger D indicates greater relative sensitivity to exchanging invariant information than to exchanging hint information.

Decodability remains high while behavior separates. The invariant decodes at $R^2 = 0.99$–1.00 in both groups, and the hint also decodes at 0.91–1.00. Yet mean C_{OOD} is 0.09 for shortcut-trained groups and approximately 1.00 for decorrelated-hint groups. The corresponding mean D values are -0.16 and $+0.76$. Table 3 shows the same ordering within every system: decorrelated-hint training produces a substantially larger D. In central force and wave, shortcut-trained values are already slightly positive (0.04 and 0.07), then rise to 0.68 and 0.94. The within-system increase in D is consistent across all five systems.

Table 3: Training conditions separate reliance and behavior across five systems (four seeds each). D: invariant-versus-hint transfer contrast; C_{OOD}: absolute output–target correlation after hint permutation. Both conditions retain high decode R^2.

	Shortcut-trained		Decorrelated hint		
System	D	C_{OOD}	D	C_{OOD}	$R^2_{\mathrm{dec}}(\phi)$ both
Pendulum	-0.19	0.14	$+0.80$	1.00	0.99–1.00
LC circuit	-0.35	0.09	$+0.69$	1.00	0.99–1.00
Central force	$+0.04$	0.03	$+0.68$	1.00	0.99–1.00
Spring	-0.35	0.09	$+0.69$	1.00	0.99–1.00
Wave PDE	$+0.07$	0.08	$+0.94$	0.98	0.99
Mean	-0.16	**0.09**	$+0.76$	**1.00**	

Across the ten group means, D and C_{OOD} correlate at $r = 0.94$; leaving out one system gives $r = 0.93$–0.96. This summarizes the separation between training conditions, which persists after

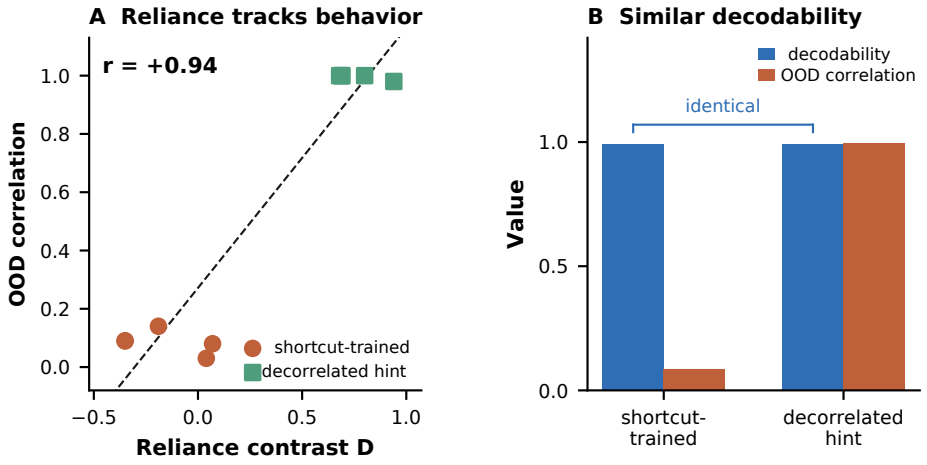


Figure 5: **Similar decodability, different reliance.** (A) Each point is a system–training-condition mean. The invariant-versus-hint contrast D separates the two training groups and correlates with held-out output correlation at $r = 0.94$ across ten group means. (B) Both groups have near-ceiling probe scores, while their behavior separates under hint permutation. Compare the two training conditions within each physical system.

excluding any one system. The per-system comparisons connect the change in relative sensitivity to the behavioral consequence of removing an informative training shortcut.

Here, the behavioral outcome is the absolute output–target correlation after hint permutation. Values near one indicate a strong linear association with the physical target under that input change.

6 Output-local coupling

The largest MII gap in pretrained Poseidon-B occurs at the decoder for a linear physical readout: 0.893, compared with 0.027 for kinetic energy at the same site. Alongside training-dependent reliance, layer position and the output quantity shape the measured response.

Why these two sites and quantities? Poseidon-B (Herde et al., 2024) supplies a large pretrained operator model with frozen weights. The encoder site has 768 channels and the decoder site 96. Comparing them tests whether a response strengthens nearer the field prediction. Here u and v are the two velocity components, and $\int u$ is the spatial integral of the first component. Comparing it with $\mathrm{KE} = \int \frac{1}{2}(u^2 + v^2)$ tests a linear versus quadratic physical quantity at each site. Both use the same predicted field: $\int u$ sums one velocity component, whereas kinetic energy sums squared velocities.

Table 4: Poseidon-B: encoder (768 channels) and decoder (96). KE is quadratic; $\int u$ is linear. Δ_{k^*} is the signed MII gap at the largest absolute gap.

Site	Quantity	R^2_{dec}	τ_{dir}	Δ_{k^*}
Encoder	KE	0.916	$+0.06$	-0.069
Encoder	$\int u$	0.998	-0.01	-0.001
Decoder	KE	0.982	$+0.09$	$+0.027$
Decoder	$\int u$	1.000	$+0.36$	$\mathbf{+0.893}$

The decoder gap peaks at an intermediate exchange size. For decoder $\int u$, Δ_k rises through 0.028, 0.162, 0.455, 0.812, and 0.893 for $k = 1, 2, 4, 8, 16$, then decreases to 0.732 and 0.215 at $k = 32, 64$. Decoder KE has maximum absolute gap 0.027. Encoder KE reaches -0.069, a negative gap favoring the random control. Table 4 gives the signed gap at the exchange size with the largest absolute response for each layer–quantity pair.

The measured gap declines after $k = 16$. At the full-exchange endpoint, swapping all m units makes $\mathcal{U}_m = \mathcal{R}_m$, so $\Delta_m = 0$ by construction. Selection advantage is therefore assessed with subsets of

units. Probe-direction interchange also yields a positive decoder correlation of 0.36 for $\int u$: both operations produce a positive response for this quantity at the decoder.

Why the readout matters. A local linear output map $f(h) = Ah + b$ and linear readout $\phi(x) = a^\top x$ give

$$g(h + \delta h) - g(h) = a^\top A \delta h. \tag{9}$$

Consequently, alignment with an output-sensitive direction can produce a large physical readout change directly. For a differentiable nonlinear model, the same argument applies locally with its Jacobian. Kinetic energy weights velocity perturbations by the current velocity field, so its sensitivity depends on both the state and the perturbation. The linear functional has a fixed readout direction, making alignment near the output easier to interpret.

Table 5 tests output construction in the small models. Slopes are -1.02 for linear momentum, -0.05 for spring energy, and $+0.60$ for a separately trained $\operatorname{asinh}(P)$ head. Each slope combines the exchanged information with its output map and units.

Table 5: Output construction and interchange slope β. The transformed head uses its own readout scale.

Quantity	Output form	β
Momentum	Linear	-1.02
Spring energy	Quadratic	-0.05
Momentum	$\operatorname{asinh}(P)$	$+0.60$

The transformation $\operatorname{asinh}(P) = \log(P + \sqrt{1 + P^2})$ is approximately linear near zero and compresses large momentum magnitudes. Its derivative $1/\sqrt{1 + P^2}$ makes this change in output scale explicit. The added head learns a new activation-to-output map, so its response jointly reflects that learned map and the nonlinear output scale.

7 Discussion and conclusion

One representation, different predictive roles. Decodability–coupling dissociation separates accessible information from selective sensitivity. Distribution-dependent reliance connects the training hint to the signal favored by prediction. Output-local coupling links a strong response to its layer and physical readout.

Implications for model and objective selection. A conservation-oriented objective is useful when it improves the property required by the application. The objective comparison makes those properties explicit: a probe gain measures more accessible encoding, rollout error measures predictive quality, and energy drift measures a physical diagnostic. The rollout comparison shows why encoding, prediction error, and conservation should be evaluated separately. The hint comparison adds a behavioral test: breaking a shortcut while preserving the physical target distinguishes models with almost identical probe scores.

A practical analysis sequence. Specify the physical quantity, target, layer, and output readout. Measure linear decodability with a shuffled-label control. Compare invariant-selected and matched random interventions over several exchange sizes, recording the component responses and their difference. Then evaluate the behavior the representation is meant to support under a relevant input shift. When a strong response appears near the output, compare quantities at the same layer and inspect the readout algebra.

Scope. The MII comparisons span five dynamics systems, the architecture comparison three pendulum backbones, the hint comparison five regression systems, and the depth comparison one pretrained PDE model. Distribution effects are established within invariant regression; objective changes address dynamics prediction. MII tests correlated units, with learned rotations and nonlinear subspaces providing natural extensions.

Conclusion. Accurate physical encoding can coexist with little selective intervention advantage. Removing an informative training shortcut changes relative signal reliance and preserves strong target association under hint permutation. Large coupling effects depend on layer and output quantity.

Reproducibility Statement

Section 3 defines the probe, activation interchanges, output readouts, and comparison statistics. Algorithm 1 specifies the MII procedure and matched control. Appendix A gives normalization and uncertainty details; Appendix B specifies the physical quantities and model families; Appendix D describes the pretrained-model evaluation. Tables and figure captions identify the exchange sizes and seed counts associated with the reported comparisons.

Ethics Statement

This study analyzes physical simulations and a pretrained PDE model. It involves no human-subject experiments or personal data. Our distinction between decodability, selective coupling, and task behavior supports careful interpretation of claims about learned physical representations.

AI Use Statement

AI tools were used to assist with code development and draft writing.

References

Guillaume Alain and Yoshua Bengio. Understanding intermediate layers using linear classifier probes. In *International Conference on Learning Representations (ICLR), Workshop Track*, 2017. URL https://arxiv.org/abs/1610.01644.

Ferran Alet, Dylan Doblar, Allan Zhou, Joshua Tenenbaum, Kenji Kawaguchi, and Chelsea Finn. Noether networks: Meta-learning useful conserved quantities. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2021. URL https://proceedings.neurips.cc/paper/2021/hash/886ad506e0c115cf590d18ebb6c26561-Abstract.html.

Peter W. Battaglia, Razvan Pascanu, Matthew Lai, Danilo Jimenez Rezende, and Koray Kavukcuoglu. Interaction networks for learning about objects, relations and physics. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2016. URL https://proceedings.neurips.cc/paper/2016/hash/3147da8ab4a0437c15ef51a5cc7f2dc4-Abstract.html.

Yonatan Belinkov. Probing classifiers: Promises, shortcomings, and advances. *Computational Linguistics*, 48(1):207–219, 2022. URL https://aclanthology.org/2022.cl-1.7/.

Johannes Brandstetter, Daniel E. Worrall, and Max Welling. Message passing neural PDE solvers. In *International Conference on Learning Representations (ICLR)*, 2022. URL https://arxiv.org/abs/2202.03376.

Trenton Bricken, Adly Templeton, Joshua Batson, Brian Chen, Adam Jermyn, Tom Conerly, Nick Turner, Cem Anil, Carson Denison, Amanda Askell, Robert Lasenby, Yifan Wu, Shauna Kravec, Nicholas Schiefer, Tim Maxwell, Nicholas Joseph, Zac Hatfield-Dodds, Alex Tamkin, Karina Nguyen, Brayden McLean, Josiah E. Burke, Tristan Hume, Shan Carter, Tom Henighan, and Christopher Olah. Towards monosemanticity: Decomposing language models with dictionary learning. *Transformer Circuits Thread*, 2023. URL https://www.transformer-circuits.pub/2023/monosemantic-features/index.html.

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. doi: 10.1073/pnas.1517384113. URL https://pmc.ncbi.nlm.nih.gov/articles/PMC4839439/.

Miles Cranmer, Sam Greydanus, Stephan Hoyer, Peter Battaglia, David Spergel, and Shirley Ho. Lagrangian neural networks. In *ICLR Workshop on Deep Differential Equations*, 2020a. URL https://arxiv.org/abs/2003.04630.

Miles Cranmer, Alvaro Sanchez-Gonzalez, Peter Battaglia, Rui Xu, Kyle Cranmer, David Spergel, and Shirley Ho. Discovering symbolic models from deep learning with inductive biases. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2020b. URL https://proceedings.neurips.cc/paper/2020/hash/c9f2f917078bd2db12f23c3b413d9cba-Abstract.html.

Nelson Elhage, Tristan Hume, Catherine Olsson, Nicholas Schiefer, Tom Henighan, Shauna Kravec, Zac Hatfield-Dodds, Robert Lasenby, Dawn Drain, Carol Chen, Roger Grosse, Sam McCandlish, Jared Kaplan, Dario Amodei, Martin Wattenberg, and Christopher Olah. Toy models of superposition. *Transformer Circuits Thread*, 2022. URL https://www.transformer-circuits.pub/2022/toy_model/.

Atticus Geiger, Hanson Lu, Thomas Icard, and Christopher Potts. Causal abstractions of neural networks. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2021. URL https://arxiv.org/abs/2106.02997.

Atticus Geiger, Zhengxuan Wu, Christopher Potts, Thomas Icard, and Noah D. Goodman. Finding alignments between interpretable causal variables and distributed neural representations. In *Proceedings of the Third Conference on Causal Learning and Reasoning*, volume 236 of *Proceedings of Machine Learning Research*, pp. 160–187. PMLR, 2024. URL https://proceedings.mlr.press/v236/geiger24a.html.

Robert Geirhos, Jörn-Henrik Jacobsen, Claudio Michaelis, Richard Zemel, Wieland Brendel, Matthias Bethge, and Felix A. Wichmann. Shortcut learning in deep neural networks. *Nature Machine Intelligence*, 2(11):665–673, 2020. doi: 10.1038/s42256-020-00257-z. URL https://www.nature.com/articles/s42256-020-00257-z.

Sam Greydanus, Misko Dzamba, and Jason Yosinski. Hamiltonian neural networks. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2019. URL https://arxiv.org/abs/1906.01563.

Wes Gurnee and Max Tegmark. Language models represent space and time. In *International Conference on Learning Representations (ICLR)*, 2024. URL https://proceedings.iclr.cc/paper_files/paper/2024/hash/0a6059857ae5c82ea9726ee9282a7145-Abstract-Conference.html.

Seungwoong Ha and Hawoong Jeong. Discovering conservation laws from trajectories via machine learning. *arXiv preprint arXiv:2102.04008*, 2021. URL https://arxiv.org/abs/2102.04008.

Maximilian Herde, Bogdan Raonić, Tobias Rohner, Roger Käppeli, Roberto Molinaro, Emmanuel de Bézenac, and Siddhartha Mishra. Poseidon: Efficient foundation models for PDEs. In *Advances in Neural Information Processing Systems (NeurIPS)*, volume 37, pp. 72525–72624, 2024. doi: 10.52202/079017-2311. URL https://proceedings.neurips.cc/paper_files/paper/2024/hash/84e1b1ec17bb11c57234e96433022a9a-Abstract-Conference.html.

John Hewitt and Percy Liang. Designing and interpreting probes with control tasks. In *Proceedings of EMNLP-IJCNLP*, 2019. URL https://aclanthology.org/D19-1275/.

John Hewitt and Christopher D. Manning. A structural probe for finding syntax in word representations. In *Proceedings of NAACL-HLT*, 2019. URL https://aclanthology.org/N19-1419/.

Raban Iten, Tony Metger, Henrik Wilming, Lídia del Rio, and Renato Renner. Discovering physical concepts with neural networks. *Physical Review Letters*, 124(1):010508, 2020. doi: 10.1103/PhysRevLett.124.010508. URL https://link.aps.org/doi/10.1103/PhysRevLett.124.010508.

Charles Jin and Martin Rinard. Emergent representations of program semantics in language models trained on programs. In *International Conference on Machine Learning (ICML)*, 2024. URL https://proceedings.mlr.press/v235/jin24e.html.

Sonia Joseph, Quentin Garrido, Randall Balestriero, Matthew Kowal, Thomas Fel, Shahab Bakhtiari, Blake A. Richards, and Mike Rabbat. Interpreting physics in video world models. In *International Conference on Machine Learning (ICML)*, 2026. URL https://arxiv.org/abs/2602.07050.

Pablo Lemos, Niall Jeffrey, Miles Cranmer, Shirley Ho, and Peter Battaglia. Rediscovering orbital mechanics with machine learning. *Machine Learning: Science and Technology*, 4(4):045002, 2023. doi: 10.1088/2632-2153/acfa63. URL https://doi.org/10.1088/2632-2153/acfa63.

Kenneth Li, Aspen K. Hopkins, David Bau, Fernanda Viégas, Hanspeter Pfister, and Martin Wattenberg. Emergent world representations: Exploring a sequence model trained on a synthetic task. In *International Conference on Learning Representations (ICLR)*, 2023. URL https://arxiv.org/abs/2210.13382.

Zongyi Li, Nikola Kovachki, Kamyar Azizzadenesheli, Burigede Liu, Kaushik Bhattacharya, Andrew Stuart, and Anima Anandkumar. Fourier neural operator for parametric partial differential equations. In *International Conference on Learning Representations (ICLR)*, 2021. URL https://openreview.net/forum?id=c8P9NQVtmnO.

Ziming Liu and Max Tegmark. Machine learning conservation laws from trajectories. *Physical Review Letters*, 126(18):180604, 2021. doi: 10.1103/PhysRevLett.126.180604. URL https://link.aps.org/doi/10.1103/PhysRevLett.126.180604.

Lu Lu, Pengzhan Jin, Guofei Pang, Zhongqiang Zhang, and George Em Karniadakis. Learning nonlinear operators via DeepONet based on the universal approximation theorem of operators. *Nature Machine Intelligence*, 3(3):218–229, 2021. doi: 10.1038/s42256-021-00302-5. URL https://www.nature.com/articles/s42256-021-00302-5.

Kingsuk Maitra. Momentum attention: The physics of in-context learning and spectral forensics for mechanistic interpretability. *arXiv preprint arXiv:2602.04902*, 2026. URL https://arxiv.org/abs/2602.04902.

Kevin Meng, David Bau, Alex Andonian, and Yonatan Belinkov. Locating and editing factual associations in GPT. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2022. URL https://proceedings.neurips.cc/paper_files/paper/2022/hash/6f1d43d5a82a37e89b0665b33bf3a182-Abstract-Conference.html.

Neel Nanda, Andrew Lee, and Martin Wattenberg. Emergent linear representations in world models of self-supervised sequence models. In *BlackboxNLP Workshop at EMNLP*, 2023. URL https://aclanthology.org/2023.blackboxnlp-1.2/.

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. doi: 10.1016/j.jcp.2018.10.045. URL https://www.sciencedirect.com/science/article/pii/S0021999118307125.

Alvaro Sanchez-Gonzalez, Jonathan Godwin, Tobias Pfaff, Rex Ying, Jure Leskovec, and Peter W. Battaglia. Learning to simulate complex physics with graph networks. In *International Conference on Machine Learning (ICML)*, 2020. URL https://proceedings.mlr.press/v119/sanchez-gonzalez20a.html.

Adly Templeton, Tom Conerly, Jonathan Marcus, Jack Lindsey, Trenton Bricken, Brian Chen, Adam Pearce, Craig Citro, Emmanuel Ameisen, Andy Jones, Hoagy Cunningham, Nicholas L. Turner, Callum McDougall, Monte MacDiarmid, Alex Tamkin, Esin Durmus, Tristan Hume, Francesco Mosconi, C. Daniel Freeman, Theodore R. Sumers, Edward Rees, Joshua Batson, Adam Jermyn, Shan Carter, Chris Olah, and Tom Henighan. Scaling monosemanticity: Extracting interpretable features from Claude 3 Sonnet. *Transformer Circuits Thread*, 2024. URL https://transformer-circuits.pub/2024/scaling-monosemanticity/.

Silviu-Marian Udrescu and Max Tegmark. AI Feynman: A physics-inspired method for symbolic regression. *Science Advances*, 6(16):eaay2631, 2020. doi: 10.1126/sciadv.aay2631. URL https://pmc.ncbi.nlm.nih.gov/articles/PMC7159912/.

Keyon Vafa, Justin Y. Chen, Ashesh Rambachan, Jon Kleinberg, and Sendhil Mullainathan. Evaluating the world model implicit in a generative model. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2024. URL https://proceedings.neurips.cc/paper_files/paper/2024/hash/2f6a6317bada76b26a4f61bb70a7db59-Abstract-Conference.html.

Keyon Vafa, Peter G. Chang, Ashesh Rambachan, and Sendhil Mullainathan. What has a foundation model found? Using inductive bias to probe for world models. In *Proceedings of the 42nd International Conference on Machine Learning*, volume 267 of *Proceedings of Machine Learning Research*, pp. 60727–60747. PMLR, 2025. URL https://proceedings.mlr.press/v267/vafa25a.html.

Jesse Vig, Sebastian Gehrmann, Yonatan Belinkov, Sharon Qian, Daniel Nevo, Yaron Singer, and Stuart Shieber. Investigating gender bias in language models using causal mediation analysis. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2020. URL https://proceedings.neurips.cc/paper/2020/hash/92650b2e92217715fe312e6fa7b90d82-Abstract.html.

Sebastian J. Wetzel, Roger G. Melko, Joseph Scott, Maysum Panju, and Vijay Ganesh. Discovering symmetry invariants and conserved quantities by interpreting siamese neural networks. *Physical Review Research*, 2(3):033499, 2020. doi: 10.1103/PhysRevResearch.2.033499. URL https://doi.org/10.1103/PhysRevResearch.2.033499.

Zhengxuan Wu, Atticus Geiger, Thomas Icard, Christopher Potts, and Noah D. Goodman. Interpretability at scale: Identifying causal mechanisms in Alpaca. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2023. URL https://proceedings.neurips.cc/paper_files/paper/2023/hash/f6a8b109d4d4fd64c75e94aaf85d9697-Abstract.html.

A Intervention details

Probe and activation coordinates. The linear probe fits the physical quantity from standardized activations using five-fold ridge regression. Layer comparisons evaluate how readily the quantity can be recovered at different sites. A shuffled-label fit provides the corresponding control, with the algorithm rejecting a probe when its shuffled score exceeds 0.15.

For probe-direction interchange, the exchanged component lies along $\hat{u} = w/\|w\|_2$. Write the activation standardization as $h = \mu + S\tilde{h}$, where S contains the feature scales. Expressed in raw activation coordinates, Eq. 5 gives the perturbation $S\hat{u}\hat{u}^\top S^{-1}(h_{\pi(i)} - h_i)$. This coordinate conversion preserves the direction defined by the standardized probe. The model then processes the modified activation through its remaining layers.

Matched comparisons. MII uses identical donor assignments for the invariant-selected and random unit sets. Both sets contain k units. The comparison therefore fixes intervention size and exchanged examples while varying the selection rule. Probe-direction controls instead vary the direction of the exchanged projection. These two controls correspond to the two intervention families defined in Section 3.

Effect summaries. The MII analysis compares the two transfer correlations through Δ_k. Since each correlation lies in $[-1, 1]$, their difference lies in $[-2, 2]$. Curves over k expose both the strongest selection advantage and its sensitivity to intervention size. The ten-seed summaries for pendulum and central force additionally provide bootstrap intervals and comparisons with 300 random-direction draws. The interval addresses uncertainty across seeds; the random-direction comparison addresses the specificity of the selected direction.

B SYSTEM SPECIFICATIONS AND CONSERVED-QUANTITY DEFINITIONS

Table 6 lists every dynamical system in the study with its state dimension, invariant class, and the closed-form conserved quantity used as ground truth, the stated target for both the decode fit (Eq. 2) and the invariant readout. The ODE systems use fixed-step RK4 integration at $\Delta t = 0.02$, with reported energy drift 6.5×10^{-6} over the evaluated horizon. Unless noted, mechanical constants are normalized to unit mass, length, and coupling.

Table 6: **Dynamical systems and their conserved quantities.** Closed-form physical quantities used for probe fitting and output evaluation.

System	Dim	Class	Conserved quantity ϕ
simple pendulum	2	energy (nonlinear)	$H = \frac{1}{2}\omega^2 - g\cos\theta$
LC circuit	2	energy (nonlinear)	$E = \frac{1}{2}Li^2 + \frac{1}{2C}q^2$
two-body spring/collision	4	linear mom. (linear)	$P = \sum_i m_i v_i$
2D central force	4	angular mom. (bilinear)	$L = x\,v_y - y\,v_x$
1D wave equation	—	energy (nonlinear)	$E = \sum_i \left(\frac{1}{2}w_i^2 + \frac{1}{2}c^2(\partial_x u)_i^2\right)$
Kepler two-body ($1/r^2$)	4	energy; ang. mom.	$E = \frac{1}{2}\|v\|^2 - \frac{1}{r};\ L = xv_y - yv_x$

Dynamics and dimensionality. The pendulum obeys $\dot\theta = \omega, \dot\omega = -g\sin\theta$; the LC system obeys $\dot q = i, \dot i = -q/(LC)$. Both central force and Kepler are described by $\ddot{\mathbf{r}} = -\mathbf{r}/\|\mathbf{r}\|^3$. The wave system uses $u_{tt} = c^2 u_{xx}$ with periodic boundaries, with $w = u_t$ denoting the velocity field.

Additional quantities. On the two-body spring we evaluate $P = \sum_i v_i$ and $E = \frac{1}{2}\sum_i v_i^2 + \frac{1}{2}(x_1 - x_2)^2$ under unit constants. On the wave field we also evaluate $P = \sum_i w_i$. The resulting quantities include linear, bilinear, and quadratic functions of the predicted coordinates, providing the output-construction comparisons in Table 5. The $\mathrm{asinh}(P)$ condition trains an additional head while freezing the representation, and thus is a changed readout experiment.

Architectures and training. The small-model experiments use MLP, GRU, and short-context Transformer backbones below 10^5 parameters, with state histories as inputs for sequence models. The pendulum comparison evaluates all three families. Networks are frozen during intervention so that each response reflects the trained predictor and the specified activation exchange.

C ARCHITECTURE COMPARISONS AND UNCERTAINTY

High decodability with small probe-direction transfer appears in three pendulum architectures. Table 7 compares these settings under the same energy readout. Table 8 gives seed-level uncertainty and random-direction comparisons for pendulum and central force.

Table 7: Pendulum architecture comparison. Energy remains highly decodable across feed-forward and sequence models, while probe-direction transfer correlations remain close to zero.

Architecture	Decode R^2	τ_{dir}	Seeds
MLP	1.00	-0.01	10
GRU	0.98	$+0.02$	5
Transformer	0.98	$+0.02$	5

Table 8: Ten-seed next-state estimates with bootstrap intervals and 300 random-direction control draws. The permutation comparison tests the specificity of the invariant direction.

System	τ (95% CI)	Permutation p
Pendulum	$-0.010\ [-0.019, +0.001]$	0.73
Central force	$-0.012\ [-0.023, -0.001]$	0.71

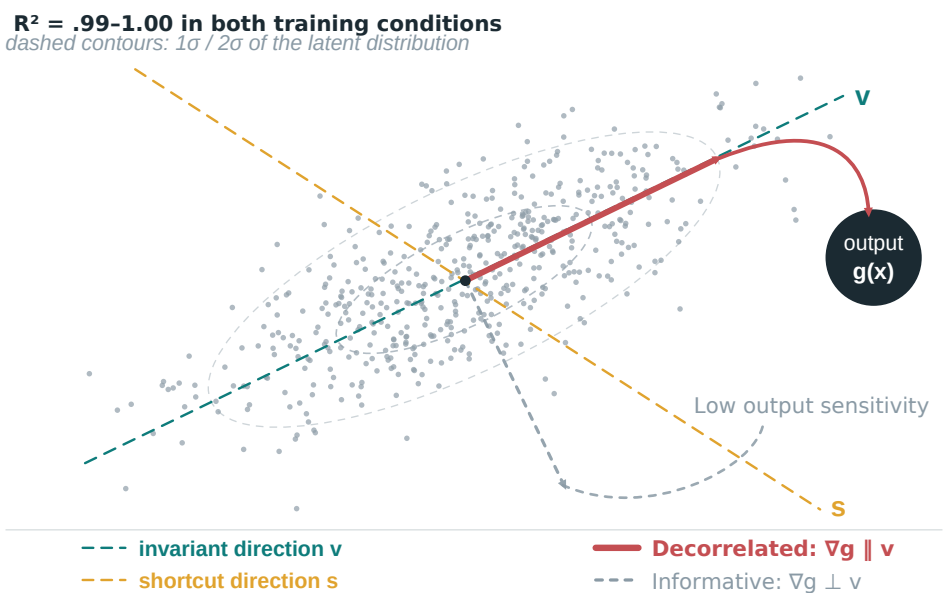


Figure 6: **Geometry of distribution-dependent reliance.** Both training conditions encode the physical quantity, while the output can be more sensitive to the hint direction or the invariant direction. The schematic connects this distinction to their different behavior under hint permutation; its geometry illustrates the interpretation rather than additional measurements.

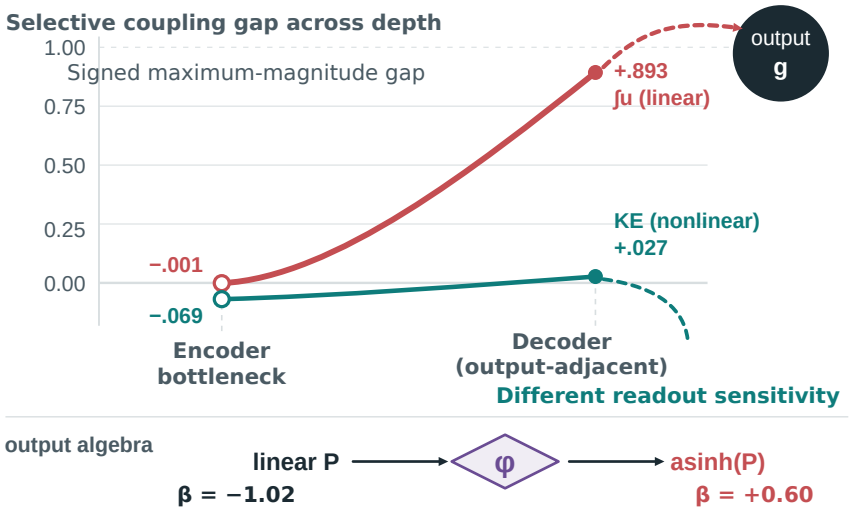


Figure 7: **Depth and output construction shape the measured response.** The upper panel compares signed maximum-magnitude gaps for the two Poseidon sites and quantities. The lower panel illustrates how a transformed output head changes the readout; its 0.60 value is the slope β from Table 5. Readout definition determines the statistic used in each comparison.

D PRETRAINED PDE MODEL AND PHYSICAL READOUTS

Model and measurement sites. Poseidon-B is a pretrained 158M-parameter scOT model. The encoder measurement uses `encoder.layers.3.blocks.7`, with 768 channels. The decoder measurement uses a 96-channel representation before the remaining upsampling and projection operations. We freeze the model for intervention and use lead time $t = 1.0$.

Constructed-field evaluation. The probe-direction depth comparison in Table 9 uses constructed divergence-free velocity fields. Spectral projection gives divergence RMS approximately 5×10^{-7}. Inputs are assembled as $[\rho, u, v, p]$ and normalized with scOT constants; the evaluation uses $N = 512$

pairs. NS-Sines supplies 512 simulated fluid-flow trajectories for a complementary evaluation. The constructed-field comparison measures responses to controlled physical inputs, while trajectory evaluation tests inputs drawn from simulated flow evolution.

The two readouts are kinetic energy, $\int \frac{1}{2}(u^2 + v^2)$, and the linear functional $\int u$. Both are computed from the predicted velocity field. Their different algebra allows a same-layer comparison of a quadratic and a linear physical quantity.

Table 9: Probe-direction depth comparison on constructed divergence-free fields. Decode R^2, slope β, and transfer correlation τ measure different aspects of the response. The decoder linear functional has the largest transfer correlation in this comparison.

Layer	Quantity	Decode R^2	β	τ
Encoder	KE	0.850	$+0.0032$	$+0.06$
Encoder	$\int u$	0.998	-0.0007	$+0.06$
Decoder	KE	0.936	$+0.0002$	-0.04
Decoder	$\int u$	1.000	$+0.1582$	$+0.36$

Sensitivity to output structure. A shallow residual-CNN comparison gives a linear-functional slope of 0.117 along the probe direction and 0.022 along a random direction. The larger slope along the probe direction quantifies its response relative to a random direction in this model. Table 9 uses change-based slopes and transfer correlations; the MII curves in the main text add a control matched for the number of exchanged units.

E What the comparisons add to prior work

Physical-structure methods differ in what they establish. Designed bottlenecks expose interpretable variables (Iten et al., 2020); Hamiltonian and Lagrangian architectures constrain prediction (Greydanus et al., 2019, Cranmer et al., 2020a); Noether Networks learn an inner-loop conservation objective that changes prediction (Alet et al., 2021). Symbolic and trajectory-based methods recover laws (Cranmer et al., 2020b, Udrescu & Tegmark, 2020, Brunton et al., 2016, Liu & Tegmark, 2021, Ha & Jeong, 2021). These approaches establish physical structure through representation design, prediction constraints, or equation recovery.

Causal abstraction and learned alignment explicitly study links between variables and computation (Geiger et al., 2021; 2024, Wu et al., 2023). MII makes the selection rule explicit, while the ridge probe uses the known physical quantity as its target. Othello studies already combine representation analysis with interventions (Li et al., 2023, Nanda et al., 2023), and video-physics analysis includes attention ablations and coordinated interventions (Joseph et al., 2026). The distinctive comparison here is between objective modifications, a controlled hint-distribution change, and physical-output readouts.

World-model evaluations based on consistency and adaptation (Vafa et al., 2024; 2025) probe behavior beyond accessible latent information. Our shortcut comparison connects a change in the training condition to both an intervention contrast and a behavioral outcome. Neural surrogates span physics-informed networks, neural operators, and graph simulators (Raissi et al., 2019, Li et al., 2021, Lu et al., 2021, Battaglia et al., 2016, Sanchez-Gonzalez et al., 2020, Brandstetter et al., 2022). These model families offer further settings in which decodability, selective coupling, and task behavior can be evaluated together.