Supplementary Information

PISR — A Search-Free Foundation Model for Symbolic Regression of Physical Laws

App. A. Benchmark construction, auditing, and leakage

The audited real-physics set is the union of three literature-derived physics collections (1,180 raw tasks), de-duplicated **by functional form** — a task is a duplicate if its truth surface reduces to the same expression after canonical variable renaming — yielding 797 unique tasks after removing 4 ill-posed items (target absent from the truth). The cross-domain set is 270 tasks after removing 1 unparseable item from 271. Auditing criteria (well-posed target, non-constant inputs, finite values) were applied *before* any model output was inspected.

Leakage audit. Fingerprinting every expression by its skeleton (numbers$\to$constant, variables$\to$canonical order), overlap between the test sets and the 400,000-formula retrieval library is 62% among trivial ($\leq$3-operator) skeletons — as expected for universal forms (`a·b`, `a/b`, `a+b·c`) — and 19% among complex (>3-operator) skeletons, where a match is an upper bound (same operator structure, not the same law). The named exclusive-solve laws (Shockley, Arrhenius, Nernst, Fowler–Nordheim, ...) have no skeleton match in the library. The ablation of Sec. 4.7 (decoder-only recovers 30.2% with no retrieval) confirms answers are generated, not looked up.

App. B. Exclusive-solve task lists

Tasks PISR recovers correctly while all four baselines fail: real-physics constituent samples 17/150 and 15/150 (vs all four baselines); merged audited sample 92/381 (vs PhyE2E + PSE); cross-domain 23/85; against GenSR, 0/300 baseline-correct on a 300-task cross-check. Full task-index lists and the per-task PISR/baseline outputs are provided as supplementary JSON, together with the reverse count (baseline-only solves) on the merged 797-task denominator.

App. C. Case-study data extraction

Turbulence: Johns Hopkins Turbulence Database `isotropic1024coarse` (longitudinal structure functions S_2, S_3) and `channel5200` (mean-velocity profile, `Re_τ = 5185`), queried via the givernylocal client; Reynolds-stress closure from the channel DNS profiles (`n = 41` wall-normal stations). Astronomy: NASA Exoplanet Archive TAP service (`pscomppars` host-star and planet tables) for equilibrium temperature, insolation and Kepler's third law; stellar catalogs for Stefan–Boltzmann. White-dwarf mass–radius: Bédard et al. (2020), VizieR J/ApJ/901/93 (183 objects, 0.29–1.32 M⊙). Epidemiology: SIR outbreak simulation. Recovered versus reference forms are tabulated in Table S1.

App. D. Baseline configurations and budgets

All baselines receive the identical point clouds and are run at matched compute. - **PySR** (genetic programming): `model = PySRRegressor(binary_operators=[+,—,*,/,pow], unary_operators=[exp,log,sqrt,` `timeout_in_seconds=30, maxsize=30)`; the full operator set is used, and variable names are sanitized to avoid Julia reserved-name collisions. - **PSE** (GPU-parallel symbolic enumeration): default enumeration budget, `n_symbol_layers = 2`. - **PhyE2E** (domain-specialized end-to-end transformer): core end-to-end transformer, single-shot decode. - **GenSR** (latent generative SR): conditional-VAE latent generator, sampled. Wall-clock is measured on idle single-GPU hardware; per-task cost is 12.25 s (PISR), ~30 s (PySR, PSE), ~90 s (GenSR), ~3 s (PhyE2E, single-shot).

App. E. Architecture and training hyperparameters

Latent dimension `d = 640`. **Formula encoder** `E_f`: AST graph network, 6 message-passing layers (≈5M parameters). **Data encoder** `E_d`: permutation-invariant pooled attention, 16 learned seed tokens (≈47M). **Flow**: DiT-style token-conditioned flow-matching field (≈63M), 20 Euler integration steps. **Retrieval**: library of 400,000 formula latents, `k = 8` nearest neighbours under cosine similarity. **Decoder**: 16 transformer layers + 4 conditioning layers (≈152M). Total ≈0.4B parameters. Training corpus ≈33.6M expressions with physics-derivation-graph supervision, three-term objective $L = \lambda_1 L\_contrast + \lambda_2 L\_recon + \lambda_3 L\_struct$. Contrastive temperature τ, loss weights (λ_1, λ_2, λ_3), optimizer, learning-rate schedule, batch size, and compute budget are listed in Supplementary Table S3.

App. F. Manifold geometry

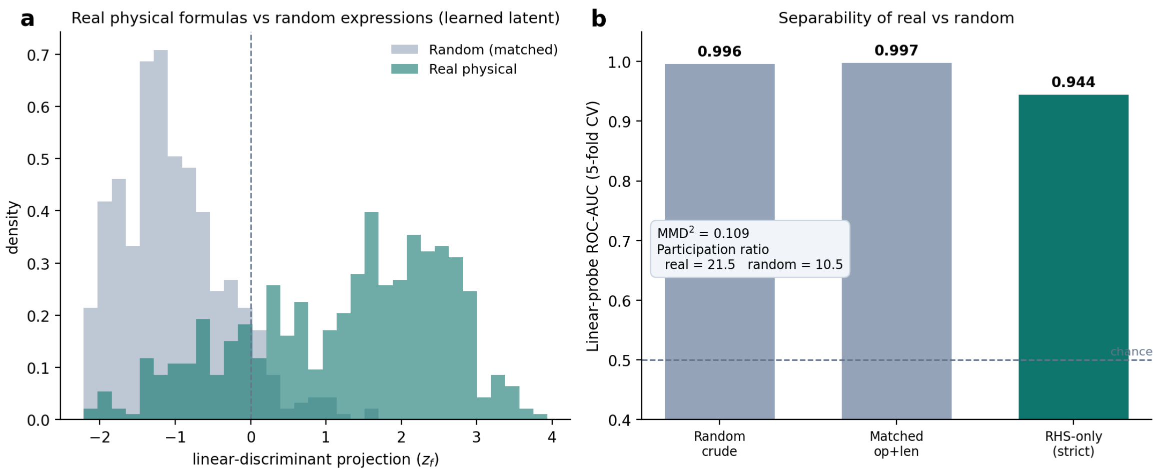
Real right-hand sides and structure-matched random expressions are embedded with the frozen formula encoder; separability is a five-fold cross-validated linear-probe ROC-AUC, distributional distance is the squared maximum-mean discrepancy (RBF kernel, median-heuristic bandwidth), and spread is the participation ratio of the embedded set's covariance eigenspectrum. The strictest control generates random expressions matched to the real set's operator-frequency and length distributions and removes the residual `target — f(x)` wrapper.

Table S4. Manifold separability by control (linear-probe five-fold ROC-AUC).

control on the random set	AUC	MMD^2	participation ratio (real / random)
unconstrained random expressions	0.996	—	—

control on the random set	AUC	MMD^2	participation ratio (real / random)
matched operator-frequency and length	0.997	—	—
matched statistics, residual wrapper removed (RHS-only)	**0.944**	0.109	21.5 / 10.5

Extended Data



Extended Data Fig. 1. Dedicated manifold-separability figure (`figures/Extended_Data_Figure_1.png`; shown below): (a) out-of-fold linear-discriminant projection of real physical formulas (teal) vs structure-matched random expressions (grey) — two visibly separated distributions; (b) linear-probe ROC-AUC for the unconstrained (0.996) and strict structure-matched (0.944) controls, with MMD^2 and participation ratios (Table S4).

Extended Data Table 1. Exclusive recoveries — the true form, and the algebraic form each baseline fits instead (actual outputs; the transcendental is replaced by a power/rational/$\sqrt{\ }$/trig surrogate).

law (domain)	true form	PySR	PhyE2E	PSE	GenSR	ours
Shockley diode (semi-cond.)	$I_s\, e^{-2\kappa d}$	$\surd$ (nested power in κ, d)	$\surd$-rational in V, d	$\sin(d)+1$, rational+trig	nested rational $1/(a+bd)$	✓
Arrhenius (kinetics)	$D_0\, e^{-Q/RT}$	D_0 atan$(\cdot)^{5.9}$	$D_0\, \surd$-poly.	$1-\cos(D_0)$	abs$(\cdot)$ + rational in D_0	✓
Nernst (electrochem.)	$\frac{RT}{zF} \ln \frac{c_o}{c_i}$	power-law ratio in c_o, c_i	$\surd$-rational	linear rational $(c_o - c_i)/(F^2 z^2)$	reciprocal $1/(a+bF)$	✓
Fowler–Nordheim (emission)	$\frac{E^2}{\phi} e^{-B\phi^{3/2}/E}$	nested power $+$ sinh	polynomial in E, ϕ	exp of a rational	polynomial $+$ buried exp	✓
Hill kinetics (biology)	$TF^n k_s/(K^n k_d)$	—	fixed-exp. $0.63/(K^2 \surd)$	timeout	nested rational	✓

Supplementary Tables

- **Table S5.** Structural accuracy across the three benchmarks (top-1), per method — PISR 51.3 / 71.1 / 39.5; PySR 43.7 / 55.2 / 31.9; PSE 33.1 / 50.7 / 32.8; GenSR 16.3 / 35.6 / 10.1; PhyE2E 14.4 / 33.7 / 2.5 (real-physics 797 / cross-domain 270 / Feynman 119). Visualized in Fig. 2b.
- **Table S1.** Recovery from raw data — law, domain, data source, recovered form (turbulence, astronomy, epidemiology, white dwarf).
- **Table S2.** Exclusive cross-domain solves — recovered form and each baseline's output (Hill kinetics, Cournot competition, Treynor ratio, Sharpe ratio).
- **Table S3.** Full training configuration (temperature, loss weights, optimizer, schedule, hardware).
- **Table S4.** Manifold separability by control (reproduced above).