PS-PINN: RECOVERING GOVERNING EQUATIONS AND READABLE SOLUTIONS FROM THE SAME NOISY DATA

Anonymous authors
Paper under double-blind review

ABSTRACT

A physical system observed over space and time can be described by its governing partial differential equation (PDE), a transferable rule, or by its solution, one readable trajectory of that rule. Existing methods recover one or the other from noisy data: equation discovery returns a formula for the rule but a neural-network solution, whereas neuro-symbolic solvers return a readable solution but require the rule to be given. Solvers that add a neural residual to a formula face a further problem: training scores only the sum of the two parts, so the residual can take over what the formula was meant to explain while the fit is unchanged. We present PS-PINN (Physics–Symbolic Physics-Informed Neural Network), which recovers the rule from the same noisy observations, enforces the estimate only when stated, and represents the solution as a sparse formula plus a neural residual. We show that the training objective never rewards keeping the formula, and we therefore fit the formula first and the residual second, with optional joint refinement, and report measures of how the two parts share the solution. Across five synthetic PDEs, this symbolic-first training keeps a formula several times more accurate than matched joint training at equal prediction error; the residual vanishes where a formula suffices and concentrates on fronts where it does not; and enforcing a recovered rule helps when its error is moderate and harms when it is not.

1 INTRODUCTION

A physical system observed over space and time admits two descriptions. The governing partial differential equation (PDE), such as $u_t = \nu u_{xx}$ for the spread of heat, states the rule the system obeys; it remains valid when the initial or boundary conditions change, but it reveals no behavior until it is solved. The solution $u(x, t)$ is one trajectory produced by that rule under one set of conditions; it can be read directly, but it does not transfer to other conditions. A scientist who has only the equation must simulate to see what happens, and one who has only the solution cannot predict a new scenario (Schmidt & Lipson, 2009; Brunton et al., 2016; Udrescu & Tegmark, 2020; Cranmer, 2023). This paper asks for both, in readable form, from one set of noisy measurements.

Two problems remain open. *First*, no method returns both descriptions from the same noisy data. Equation-discovery methods (Brunton et al., 2016; Rudy et al., 2017; Schaeffer, 2017; Messenger & Bortz, 2021; Stephany & Earls, 2024b) recover the rule but return the solution either not at all or as the weights of a neural network (Both et al., 2021; Chen et al., 2021). Neuro-symbolic solvers (Majumdar et al., 2022; 2023; Das et al., 2025; Liang & Yang, 2025; Oikonomou et al., 2025; Niu et al., 2026) recover a readable solution, but every one of them takes the rule as given and uses it as the training signal; none recovers the rule, and none says what to do when the only rule available is an estimate that may be wrong. *Second*, when the field contains a steep front, a formula alone cannot fit it and the solver must add a neural residual, as physics-informed symbolic networks (PINSN) (Majumdar et al., 2022) do. That representation is ambiguous. Training only checks the sum of the two parts, so the network is free to reproduce what a formula term would have contributed; the formula then loses that term while the total fit stays the same. The same ambiguity is documented for models that add a neural network to a physics-based model (Yin et al., 2021; Dona et al., 2022; Takeishi & Kalousis, 2021). Known remedies restrict the residual: keep it small (Yin et al., 2021), or forbid it from resembling any formula term (Richter & Kilbertus, 2026). Both presume that the correct residual is small and unlike the formula. That presumption fails where the formula itself is wrong in a structured way: at a sharp interface a smooth formula overshoots, and the residual must subtract

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.

that overshoot, which makes the correct residual both large and shaped like the formula (we observe this in Sec. 5.2). Two things remain untested: whether the formula can be retained by the order in which the two parts are trained rather than by a constraint on the residual, and how to tell whether it has been, since prediction error, which scores only the sum, cannot.

The solution itself faces a representation conflict. A closed-form solution, a weighted sum of a few known functions such as sines and exponentials, is accurate and readable when the field is smooth, but a short formula cannot resolve a steep front or a thin interface (Udrescu & Tegmark, 2020; Cranmer, 2023). A neural network can represent such features, but its output is a set of weights rather than a formula, and it tends to blur sharp features (Rahaman et al., 2019; Tancik et al., 2020; Wang et al., 2022; Krishnapriyan et al., 2021). The two representations fail in complementary situations, which motivates a hybrid: a readable *symbolic core* that carries the global structure and a small *neural corrector* that covers what the formula cannot.

A hybrid model, however, does not guarantee a hybrid explanation. The training loss scores only the sum of the two parts, while the sparsity penalty that keeps the formula short applies to the core alone. Moving structure from the core into the corrector therefore never raises the loss and usually lowers it, so an optimizer that trains both parts together tends to empty the core. Prediction error stays low while the readable part degrades, and the degradation is invisible to anyone who monitors accuracy alone. Which part ends up explaining the data is decided by the training procedure, not by the architecture.

We present PS-PINN (Physics–Symbolic Physics-Informed Neural Network), a framework that recovers both descriptions from the same noisy observations. A standard sparse-regression backend recovers the equation. The solution is represented as

$$\hat{u}(x,t) = \Phi(x,t)\boldsymbol{\alpha} + r_\theta(x,t), \tag{1}$$

where Φ is a dictionary of candidate functions, $\boldsymbol{\alpha}$ a sparse coefficient vector that defines the symbolic core, and r_θ a neural corrector. The core is fitted first, with the corrector held at zero; the corrector is then trained on what the core leaves unexplained. The recovered equation can optionally be enforced as a constraint on the solution, and we state per experiment whether it is. We evaluate on five synthetic PDEs: a standing wave and a spreading heat profile, whose solutions are short formulas; a fluid equation that develops a steep front (Burgers); a phase-separation equation with moving interfaces (Allen–Cahn); and a two-dimensional wave.

Contributions. (*i*) PS-PINN recovers the rule from the same noisy data instead of requiring it, and enforces the estimate during solution fitting only when stated, so that the benefit and the risk of enforcing a recovered rule can both be measured (Secs. 4.1, 4.3). (*ii*) Because the objective never rewards keeping the formula (Proposition 1), we fit the formula first with the residual switched off, then train the residual on what remains, and only then refine the two alternately and jointly (Sec. 4.4). (*iii*) We report the residual's share of the solution, its overlap with the formula, where it concentrates, and the formula's accuracy on its own, so that a collapsed formula is visible even when prediction error is unchanged (Sec. 4.2). (*iv*) Across five synthetic PDEs, joint training matched in budget, schedule and update rule reaches nearly the same prediction error as symbolic-first training on Allen–Cahn and Burgers, yet leaves a formula with 2.3 to 5.7 times the error, and the residual vanishes where a formula suffices and concentrates on fronts where it does not (Secs. 5.2–5.4). Enforcing an Allen–Cahn rule estimated under a prescribed structure lowers reconstruction error 2.5-fold relative to leaving it unenforced, whereas enforcing a biased rule would raise it by orders of magnitude.

2 Related Work

Interpretable governing-law discovery. Sparse identification on candidate libraries (Brunton et al., 2016; Schaeffer, 2017) established this line of research. DeepMoD (Both et al., 2021) added a neural surrogate with automatic differentiation for noisy data, and the PDE-READ / PDE-LEARN family (Stephany & Earls, 2022; 2024a;b) refined human-readable discovery with rational networks, joint response–coefficient training, and weak-form losses. The closest prior at the level of the delivered pair is PINN-SR (Chen et al., 2021), which jointly learns a parsimonious closed-form PDE and a full-field DNN surrogate through alternating optimization, but its continuous response remains encoded in network weights. UPINN (Podina et al., 2023) symbolifies an unknown differential-operator term under a partially known skeleton, while the solution stays a PINN surrogate. Recent discovery work strengthens the identification itself, for example through dimensional verification that

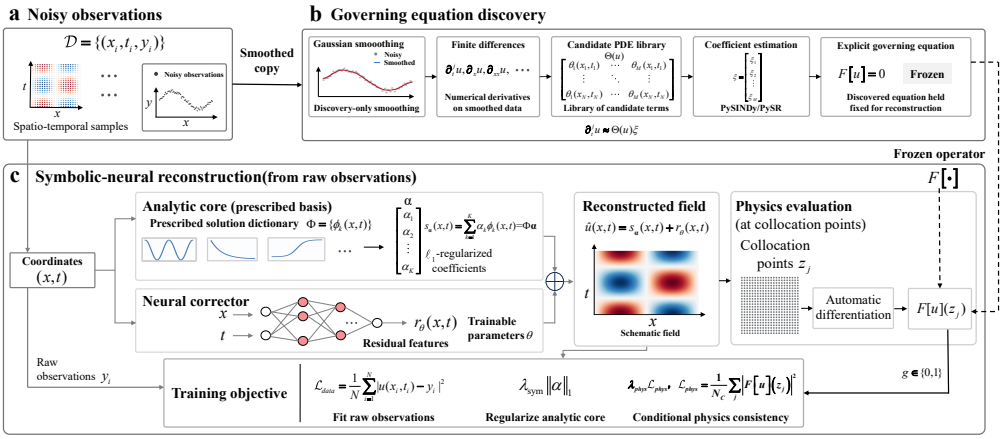


Figure 1: **PS-PINN recovers both descriptions from one dataset.** (a) Noisy observations. (b) Equation discovery on a smoothed copy: numerical derivatives, a library of candidate terms, sparse coefficient estimation, and an explicit equation that is then held fixed. (c) Solution reconstruction from the raw observations: an symbolic core on a prescribed dictionary plus a neural corrector, trained to fit the data and, when enabled, to satisfy the recovered equation at collocation points.

prunes the candidate library (Wang et al., 2025) and physical-hypothesis selection (Feng et al., 2025). These methods make the governing law readable; when a continuous realization is also returned, it is typically represented by a neural surrogate. PS-PINN retains the discovery objective and targets an interpretable structure for the realization as well. The discovery module is backend-agnostic; we instantiate it with PySINDy-based sparse regression and PySR-based symbolic regression.

Interpretable and neuro-symbolic PDE solutions. PINSN (Majumdar et al., 2022) is the direct predecessor of our representation. Under a known PDE it learns a solution by a physics-informed loss, first as a differentiable operator-grammar network (PISN) and then augmented with an MLP residual, $u_{\mathrm{PINSN}} = u_{\mathrm{PISN}} + F(x; \theta_{\mathrm{res}})$. Beyond the known-law setting, three differences matter. First, our symbolic core uses a sparsity-regularized combination of functions from an explicit analytic dictionary. The minimal Wave configuration yields a compact two-term expression, and other deployed configurations retain an explicit analytic component whose compactness depends on the dictionary and the learned coefficients (Appendix L), whereas grammar-expanded expressions are long and cumbersome by the authors' own account. Second, the residual's share is measured through ρ_{NN}, and the curriculum gives the symbolic component explanatory priority, and ρ_{NN} measures the residual branch's share. Third, the law itself is recovered from the same noisy observations and can optionally be coupled into training (Sec. 4.3). Post-hoc symbolic regression approaches (Majumdar et al., 2023; Das et al., 2025) first train a PINN under a prescribed PDE and then fit an explicit symbolic expression to the PINN-predicted solution, whereas PS-PINN retains the analytic core and neural corrector within a single representation, allowing the neural component to remain active where the symbolic dictionary is insufficient. DeSyR (Niu et al., 2026) adds physics-only coefficient refinement and gated verification for explicit solution recovery, again under a known operator. Its gate verifies a candidate solution against the known equation. These methods make the realization readable, but the governing law is prescribed rather than recovered from the same noisy observations.

Responsibility allocation in hybrid models. Staged and alternating optimization is itself common ground. PINN-SR alternates between network and coefficients (Chen et al., 2021), PDE-LEARN trains in burn-in/sparsify/fine-tune phases (Stephany & Earls, 2024a), and PINSN's hypernetwork variant trains symbolic-first, residual-second (Majumdar et al., 2022). OrthoReg (Richter & Kilbertus, 2026) shows that when the symbolic structure of a hybrid dynamical model is itself learned by sparse discovery, a neural augmentation can re-learn the symbolic component, and prevents this with an explicit orthogonality penalty on the vector field $f_{\mathrm{sym}} + f_{\mathrm{NN}}$. Our curriculum addresses the analogous allocation problem at the level of the PDE realization $\hat{u} = \Phi\boldsymbol{\alpha} + r_\theta$. Our curriculum assigns the symbolic branch primary responsibility for the PDE realization representation with a neural correction (Sec. 5.4), using analytic reference operators to isolate representation training from discovery error (Appendix C).

Discovered-physics reuse and reliability. Reusing a discovered equation during training is likewise not new. PINN-SR couples the discovered law and the field throughout (Chen et al., 2021), and R-DISCOVER alternates discovery with an embedding phase enforcing the current PDE (Du et al., 2024). Operator reliability of candidate equations also exists, including information-criterion ranking by parsimony and PINN-measured precision (Xu et al., 2023) and uncertainty-penalized selection (Thanasutives et al., 2024), and wrong physics can coexist with low training loss (McShannon & Dietrich, 2026). Whether a data-discovered operator is reliable enough to enforce as a training constraint is not decided automatically here: coupling is an experimental choice stated per experiment, and operator reliability estimation is outside the scope of this paper (Sec. 4.3).

3 Problem Statement

Let Ω be a spatial domain and $Q = \Omega \times (0, T]$ its space–time domain. We observe

$$\mathcal{D} = \{(x_i, t_i, y_i)\}_{i=1}^{N}, \qquad y_i = u^\star(x_i, t_i) + \epsilon_i, \tag{2}$$

with observation noise $\epsilon_i \sim \mathcal{N}(0, \sigma^2)$, and assume the underlying process obeys a PDE $\mathcal{F}^\star(u, u_t, \nabla u, \nabla^2 u, \ldots) = 0$ that can be written as $\partial_t^j u = \Theta(u)\,\boldsymbol{\xi}$ over a library of candidate terms $\Theta(u) = [u, \partial_x u, \partial_x^2 u, u u_x, \ldots]$ with a sparse coefficient vector $\boldsymbol{\xi}$ (Brunton et al., 2016; Schaeffer, 2017). From $\mathcal{D}$ alone the method must return two objects: an estimate $\hat{\mathcal{F}}$ of the rule and an estimate $\hat{u}$ of the solution.

The rule. Recover an explicit equation $\hat{\mathcal{F}}[u] = 0$ with numerical coefficients, e.g. $u_t + u u_x - \nu u_{xx} = 0$. It describes the dynamics and remains usable when initial or boundary conditions change.

The solution. Recover $\hat{u}(x, t)$ for the observed configuration in a form that can be read, not only evaluated. As in Equation (1), $\hat{u} = \Phi\boldsymbol{\alpha} + r_\theta$, where the *symbolic core* $\Phi\boldsymbol{\alpha}$ is a sparse combination of a prescribed dictionary Φ and the *neural corrector* r_θ is a small network. The dictionary is specified per benchmark and encodes prior knowledge about the solution's form (Appendix B); the method selects terms from it and does not search beyond it. We say the core is *retained* if (a) whenever the dictionary suffices, $\|r_\theta\| \to 0$ and $\hat{u}$ reduces to a formula, and (b) when it does not, the corrector stays small and concentrates where the dictionary fails, such as on a front. Neither property follows from the representation, since any function in the span of Φ can be produced by either part; whether the core is retained depends on training (Sec. 4.4).

Measuring the split. We quantify (a) and (b) with the *corrector share*

$$\rho_{\mathrm{NN}} = \frac{\|r_\theta\|_2^2}{\|\Phi\boldsymbol{\alpha}\|_2^2 + \|r_\theta\|_2^2}, \qquad \rho_{\mathrm{sym}} = 1 - \rho_{\mathrm{NN}}, \tag{3}$$

the fraction of the two parts' combined energy carried by the corrector, which property (a) requires to vanish. It compares the two parts to each other rather than to $\hat{u}$, because they need not be orthogonal; the *overlap* $\gamma = \cos\angle(\Phi\boldsymbol{\alpha}, r_\theta)$ records whether the corrector duplicates the core ($\gamma > 0$) or cancels part of it ($\gamma < 0$). Property (b) is measured by the *localization score* ρ_{shock}, the fraction of the corrector's energy inside a fixed band around the front. Finally, the *core-only error* is the error of $\Phi\boldsymbol{\alpha}$ alone against the clean field; it is the quantity that prediction error hides, since two models with equal $\|\hat{u} - u^\star\|$ can differ greatly in it. If a closed form exists we also report Expression Structural Accuracy (ESA), which scores the recovered formula against the true one. Metric definitions are given in Appendix F.

4 Methodology

4.1 Governing Equation Discovery

Direct differentiation amplifies observation noise. We therefore use a Gaussian-smoothed copy of the observations for derivative-based equation discovery, while reconstruction always trains against the original noisy observations. From the smoothed field we form $\partial_t u, \partial_x u, \partial_{xx} u, \ldots$ by finite differences, build $\Theta(u)$, and seek a sparse coefficient vector via

$$\hat{\boldsymbol{\xi}} \approx \arg\min_{\boldsymbol{\xi}} \; \left\|\partial_t^j u - \Theta(u)\boldsymbol{\xi}\right\|_2^2 + \lambda_{pde}\|\boldsymbol{\xi}\|_0 \tag{4}$$

using sequential thresholded least squares (Brunton et al., 2016). The ℓ_0 count avoids the shrinkage bias of ℓ_1 on retained coefficients, and STLSQ is a heuristic for this objective rather than an exact solver. The Burgers coefficient fit uses PySR (Cranmer, 2023). The recovered $\hat{\boldsymbol{\xi}}$ defines the delivered equation $\hat{\mathcal{F}}$ and is frozen thereafter (Appendix C). This is standard by design, because this paper's contribution lies in what is delivered alongside the equation rather than in the discovery algorithm.

4.2 Symbolic–Neural Solution Representation

The solution is represented as

$$\hat{u}(x,t) = \underbrace{\textstyle\sum_{k=1}^{K} \alpha_k \phi_k(x,t)}_{\Phi\boldsymbol{\alpha}} + \underbrace{r_\theta(x,t)}_{\text{neural correction}}, \tag{5}$$

with sparsity of $\boldsymbol{\alpha}$ encouraged through an ℓ_1 penalty. Where an implementation adds a trainable offset b_0, the reported core is $\Phi\boldsymbol{\alpha} + b_0$, with b_0 unpenalized and suppressed in the notation. The delivered core's sparsity varies by configuration, from two active terms on minimal Wave to 18 of 19 candidates on noisy Allen–Cahn (Appendices G, C). Neither pure representation suffices: $\hat{u} = r_\theta$ alone is a black box that smears sharp fronts under spectral bias, and $\hat{u} = \Phi\boldsymbol{\alpha}$ alone is limited to the span of the dictionary. The *symbolic core* represents global structure using the prescribed basis functions: periodicity, separability, exponential decay, polynomial and trigonometric shape, and the smooth bulk of the field. The *neural corrector* is trained to approximate residual features that are not adequately represented by the dictionary, including steep fronts and unresolved details; the curriculum encourages this allocation but does not guarantee spatial localization. Empirically the corrector's largest responses concentrate near fronts and interfaces (Sec. 5.2). The per-benchmark solution dictionaries are physics-informed and hand-specified (Appendix B).

4.3 Enforcing the Recovered Rule

If we choose to enforce the discovered rule, the solution branch minimizes

$$\mathcal{L} = \mathcal{L}_{\text{data}} + \lambda_{\text{sym}} \|\boldsymbol{\alpha}\|_1 + \lambda_{\text{phys}} \mathcal{L}_{\text{phys}}, \tag{6}$$

with $\mathcal{L}_{\text{data}} = \frac{1}{N} \sum_i |\hat{u}(x_i, t_i) - y_i|^2$ the fit to the raw observations and $\mathcal{L}_{\text{phys}} = \frac{1}{N_c} \sum_j |\mathcal{F}_c[\hat{u}](z_j)|^2$ the residual of a rule $\mathcal{F}_c$ at collocation points z_j. The first two terms are the maximum a posteriori objective under Gaussian noise and a Laplace prior (Appendix O); the Burgers pipeline uses a front-weighted data term and no final-stage ℓ_1 (Appendix E). Whether the physics term is active, and which $\mathcal{F}_c$ it uses, is stated per experiment: the *recovered* rule of Sec. 4.1, a rule with *prescribed terms* and estimated coefficients, or the *true* rule, used to study the solution branch apart from discovery error. With the physics term off, $\hat{\mathcal{F}}$ and $\hat{u}$ are both returned but nothing ties them together; we do not decide automatically whether a recovered rule is reliable enough to enforce (diagnostic in Appendix C).

4.4 Symbolic-First Training Curriculum

The objective of Equation (6) is non-convex, and the central failure mode of direct joint training is not inaccuracy but misallocation: the network's expressivity lets r_θ absorb structure intended for the symbolic core. This behavior is consistent with the structure of the objective. Its prediction terms depend only on the sum $\Phi\boldsymbol{\alpha} + r_\theta$, while the sparsity penalty acts on $\boldsymbol{\alpha}$ alone. Whenever some θ' realizes $r_{\theta'} = r_\theta + \Phi\boldsymbol{\alpha}$, emptying the core therefore lowers the objective by $\lambda_{\text{sym}}\|\boldsymbol{\alpha}\|_1$, strictly so whenever $\boldsymbol{\alpha} \neq 0$ (Appendix O). Because the optimization target does not identify the division of explanatory labor, the curriculum allocates responsibility before the components are trained jointly.

Proposition 1 (The objective does not identify the allocation)**.** *Let $\mathcal{L}(\boldsymbol{\alpha}, \theta)$ be the objective of Equation* (6)*. If for the current $(\boldsymbol{\alpha}, \theta)$ there is a θ' with $r_{\theta'} = r_\theta + \Phi\boldsymbol{\alpha}$, then $\mathcal{L}(\mathbf{0}, \theta') = \mathcal{L}(\boldsymbol{\alpha}, \theta) - \lambda_{\text{sym}}\|\boldsymbol{\alpha}\|_1$, with strict decrease whenever $\lambda_{\text{sym}} > 0$ and $\boldsymbol{\alpha} \neq \mathbf{0}$.*

Proof. The data and physics terms depend on $(\boldsymbol{\alpha}, \theta)$ only through $\hat{u} = \Phi\boldsymbol{\alpha} + r_\theta$, which is unchanged by the substitution, while the ℓ_1 term drops from $\lambda_{\text{sym}}\|\boldsymbol{\alpha}\|_1$ to zero. □

Whenever the network can represent such a θ', which an over-parameterized multilayer perceptron typically can, emptying the formula into the residual lowers the objective. The optimizer is therefore not merely indifferent to the allocation but pushed towards collapse, and the allocation actually

Table 1: **The symbolic-first curriculum at a glance** (1D Wave budgets; other budgets and objective deviations in Appendix E). Stages III–IV are optional refinement; "passes" are epochs over the data.

Stage	Trains	Frozen	Loss terms	Passes	λ_{phys}
I	$\boldsymbol{\alpha}$ (proximal LASSO)	θ ($r_\theta \equiv 0$)	data + ℓ_1	400	off
II	θ	$\boldsymbol{\alpha}$	data (+ physics if coupled)	500	$10^{-3} \to 1$ (step)
III	$\boldsymbol{\alpha} \leftrightarrow \theta$ (alternating)	—	data + ℓ_1 (+ physics)	$2 \times (100{+}150)$	1
IV	$\boldsymbol{\alpha}, \theta$ (jointly)	—	full objective	300	1

Table 2: **What each method delivers.** "Interpretable solution representation" asks whether the delivered solution is readable; for PS-PINN on shock problems it is an analytic backbone plus a localized correction, not a closed form. PINSN and DeSyR give readable solutions for a known PDE.

Method	Explicit PDE	Interpretable solution representation	Shock-capable	Continuous
PINN (oracle)	given	✗	✓	✓
PDE-FIND / SINDy	✓	✗	—	—
PySR (solution)	✗	✓	weak	✓
PINN-SR	✓	✗	✓	✓
PINSN	given	✓ (symbolic + residual)	✓	✓
DeSyR	given	✓ (symbolic)	—	✓
PS-PINN	✓	symbolic core + corrector	✓	✓

reached is decided by the training path. The curriculum fixes that path by allocating responsibility before the two parts are trained together.

Table 1 records what each stage trains, freezes, and optimizes. **Stage I** fits the symbolic component before the neural corrector is trained. With $r_\theta \equiv 0$, the convex LASSO is solved by proximal gradient (Daubechies et al., 2004; Beck & Teboulle, 2009); classical LASSO theory motivates the initialization (Tibshirani, 1996), though we do not assume its support-recovery conditions hold for every overcomplete dictionary. **Stage II** trains r_θ on the residual $y - \Phi\boldsymbol{\alpha}^{(0)}$, assigning the network the structure left unexplained by the initial core; an active physics weight steps to its final value partway through. **Stage III** alternates block-coordinate updates $\boldsymbol{\alpha} \leftrightarrow \theta$ (Tseng, 2001) at that weight, with the $\boldsymbol{\alpha}$ half-step proximal on the data term alone (Appendix E). The analyzed alternating scheme's convergence properties, at fixed λ_{phys}, are stated in Appendix O. **Stage IV** releases all parameters jointly (pseudocode and budgets: Appendix E). Stages I–II establish the allocation; Stages III–IV are an optional refinement, giving two operating points, *interpretability-first* (I+II) and *accuracy-refined* (I–IV), whose trade-off Sec. 5.4 measures.

5 Experiments and Results

Setup. The main benchmarks are 1D Wave ($u_{tt} = u_{xx}$, standing-wave data on a 128×128 grid), a symbolic-friendly regime whose solution is a low-complexity closed form, and viscous Burgers ($u_t + uu_x = \nu u_{xx}$, $\nu = 0.01/\pi$, 256×100, shock forming at $t \gtrsim 0.5$), a symbolic-hostile regime relative to the deployed dictionary and complexity budget (Appendix G). Additional PDEs (1D Heat, Allen–Cahn, 2D Wave) appear in Sec. 5.5. Observations carry additive Gaussian noise, scaled by $\mathrm{std}(u)$ for Burgers and absolute elsewhere. All methods receive identical noisy tensors, grids, seeds, and metrics. Main cells and controlled ablations use $n \geq 5$ seeds, diagnostics $n = 3$, as each table states (Appendices A, H). Symbolic recovery in these experiments selects terms from hand-specified, physics-informed solution dictionaries (Sec. 6, Appendix B). Baselines are PySR (Cranmer, 2023) for a pure symbolic solution, an oracle PINN with the true PDE hard-coded for a pure neural one, and PySINDy (Brunton et al., 2016) and DeepMoD (Both et al., 2021) for discovery.

5.1 Governing Equation and Solution Recovery

We first ask whether the rule and the solution can be returned from the same data and what enforcing the recovered rule is worth. One limit is intrinsic: a single standing-wave mode satisfies $u_{xx} \equiv -k^2 u$ exactly, so no method can identify a transferable rule from it. We therefore report the rule-identifiable cases in Table 3 and treat the others as tests of the solution branch alone.

Table 3: **Dual-artifact recovery from one noisy dataset** ($n \geq 3$ seeds per row; noisy Track-A rows $n=5$; Burgers pooled $n=10$). Track A covers dual-output recovery with controlled operator coupling; both artifacts are delivered, and selected configurations additionally couple a discovered or template-constrained operator, in some cases an intentionally imperfect one to quantify operator mismatch. Two-mode Wave rows select the operator among candidates; † rows use template-constrained operators with protocol-specified coupling, and single-mode Heat shares the Wave degeneracy (Appendix C). Track B stresses the representation; the law is not identifiable (single-mode Wave) or biased (Burgers), so no operator is coupled. ‡Imperfect operator coupled by protocol.

Benchmark	σ	Recovered equation	Coeff. err.	Coupling protocol	Hybrid RMSE
Track A: dual-output recovery with controlled operator coupling					
Two-mode Wave	0	$\hat{u}_{tt} = 1.00\, u_{xx}$	$<1\%$	coupled	3.3×10^{-4}
1D Heat†	0	$\hat{u}_t = 0.05001\, u_{xx}$	0.02%	coupled	1.4×10^{-3}
Allen–Cahn†	0	$\hat{u}_t = 0.01002\, u_{xx} + 1.000\, u - 1.000\, u^3$	$\leq 0.2\%$	coupled	9.0×10^{-3}
Allen–Cahn†	0.005–0.02	$\hat{u}_t = 0.012\, u_{xx} + 1.14\, u - 1.14\, u^3$	14–21%	coupled‡	$1.3\text{–}1.4 \times 10^{-2}$
Track B: representation stress under operator uncertainty					
Two-mode Wave	0.005–0.02	$\hat{u}_{tt} = 0.86\, u_{xx}$ (coefficient biased)	-14%	not coupled	$2\text{–}4 \times 10^{-4}$
Burgers	0.10	$u_t + u u_x = \hat{\nu} u_{xx}$ (structure correct; $\hat{\nu} \approx 3.2\nu$)	$+220\%$ (ν)	not coupled	6.9×10^{-3}
1D Wave (single-mode)	0.10	not identifiable from this trajectory	—	not coupled	3.8×10^{-3}

At $\sigma = 0$ every coefficient is recovered to within 1%, and enforcing the recovered Allen–Cahn rule lowers reconstruction RMSE threefold ($0.027 \to 0.009$). At $\sigma \in \{0.005, 0.01, 0.02\}$ the Allen–Cahn coefficients drift by 14–21%, yet enforcing them still lowers reconstruction RMSE 2.4–2.6× ($0.035 \to 0.013$) and core-only error ≈1.8× (per-seed pairing in Table 9). The recovered cores there are explicit but not compact: 18 of 19 candidates stay active. The ordering effect of Sec. 5.4 also holds with the recovered rule enforced: repeating the three schedules on Allen–Cahn at $\sigma = 0.01$ with each run's own discovered rule, the core-first and fit-refined schedules retain core-only errors of 0.068 and 0.090, while joint training at the same budget leaves 0.510 ($p = 3.5 \times 10^{-4}$), with all three indistinguishable in reconstruction (Appendix E).

Enforcing a wrong rule is harmful. The two-mode Wave rule recovered at the same noise levels is biased by ∼14%; it is left unenforced and the solution is unaffected (RMSE $2\text{–}4 \times 10^{-4}$), whereas enforcing it would raise RMSE by over two orders of magnitude. At $\sigma = 0.10$ the Burgers rule has the correct terms but $\hat{\nu} \approx 3.2\,\nu$, and the single-mode Wave rule is not identifiable; both run with the physics term off, so the remaining sections test the solution branch, and the recovered rule is reported alongside without being enforced.

5.2 Symbolic–Neural Decomposition

Accuracy alone would not justify the representation; the question is what the delivered solution is. Figure 2 shows the trained decomposition on a shared symmetric color scale, with the normalized neural-correction share ρ_{NN} of Equation (3) measured per benchmark. On **1D Wave** with the minimal two-term dictionary the corrector shrinks to a normalized share of $\rho_{\mathrm{NN}} \approx 3 \times 10^{-6}$ and its panel is visually empty. The deliverable is effectively the closed-form expression $\hat{u} \approx 1.0008 \sin(2\pi x) \cos(2\pi t)$, with remaining terms below 10^{-3}, which a reader can check directly. The formula as printed has RMSE 4.0×10^{-4}, consistent with the full core's 7.9×10^{-4} (Appendix G). On **Burgers** the deliverable is not a full closed form and we do not claim one. The symbolic core reconstructs a smooth backbone dominated by the initial-profile mode, with weak temporal corrections (core-only RMSE 0.278 against a field of unit amplitude; hybrid 0.0076), and the corrector carries the rest with $\rho_{\mathrm{NN}} = 0.21$. Its amplitude peaks on the shock trajectory and its energy concentrates there (3.5× enrichment; Figure 2), while the diffuse remainder corrects the temporal evolution the backbone under-resolves. On **Allen–Cahn** the corrector is small ($\rho_{\mathrm{NN}} = 0.021$, $\|r_\theta\|_\infty = 0.21$) and sits at the moving interfaces; in these experiments the neural component's largest responses concentrate where

Table 4: **Main benchmark comparison** ($\sigma = 0.10$, training domain; mean±s.d.). PS-PINN: four stages, uncoupled; Wave $K = 14$, $n = 5$; Burgers pooled $n = 10$. Baselines: $n = 5$. R^2: three-decimal rounding.

Method	RMSE ↓	R^2 ↑
Wave		
PySR	$\mathbf{0.0011}$±0.0003	1.000
PINN (oracle)	0.0100±0.0015	1.000
PS-PINN	0.0038±0.0011	1.000
Burgers		
PySR	0.0972±0.0148	0.974
PINN (oracle)	0.0169±0.0019	0.999
PS-PINN	$\mathbf{0.0069}$±0.0004	1.000

the symbolic basis loses expressivity. The allocation also degrades gracefully when the analytic prior is wrong: rebuilding the Wave dictionary at a mis-specified wavenumber $k^\star(1+\delta)$ moves ρ_{NN} by over four orders of magnitude ($7 \times 10^{-6} \to 0.29$ for $\delta\colon 0 \to 10\%$) as the corrector absorbs the share the basis no longer represents, and at $\delta{=}10\%$ the hybrid RMSE is one ninth of the core's (Appendix L).

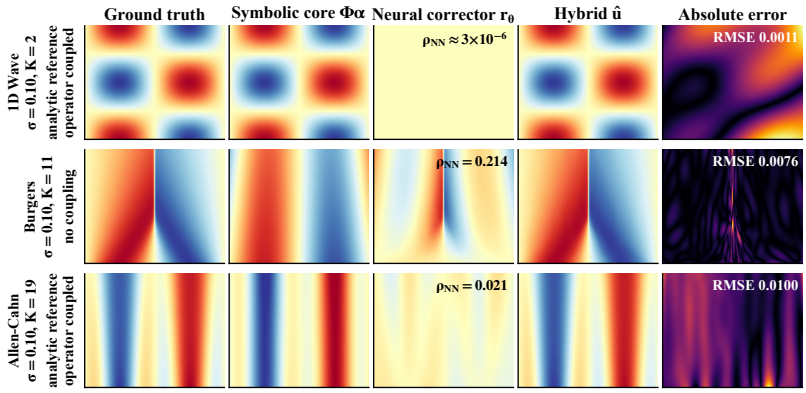


Figure 2: **The trained decomposition follows the analytic-first design** ($\sigma = 0.10$, deployed decomposition configurations; row labels: noise, dictionary size K, and coupling protocol). Each row presents the ground truth, symbolic core $\Phi\boldsymbol{\alpha}$, neural corrector r_θ, hybrid solution $\hat{u}$, and absolute error; the first four panels share a symmetric color scale within each row. 1D Wave (minimal-library diagnostic; RMSE protocol differs from Table 4): $\rho_{\mathrm{NN}} \approx 3 \times 10^{-6}$, so the corrector contributes negligibly and the delivered expression is a closed form within the supplied dictionary. Burgers: the corrector ($\rho_{\mathrm{NN}} = 0.21$) peaks on the shock trajectory, with $\rho_{\mathrm{shock}} = 0.69$ of its energy within 20% of the domain around the front. Allen–Cahn: a small corrector ($\rho_{\mathrm{NN}} = 0.021$) at the moving interfaces.

5.3 What Each Part Adds

Ablations retrain from scratch under the same schedule and budget with one part gated to zero throughout, so on Wave the network trains in 1100 of 1700 passes ($n{=}5$; Appendix J). On **Wave** the core alone matches the hybrid (0.0014 vs. 0.0016) and the corrector alone is $48\times$ worse (0.077). On **Burgers** the corrector alone matches the hybrid (0.0077) and the core alone is $38\times$ worse (0.29). The hybrid matches the better single part in each regime; what it adds on Burgers is not accuracy but the readable core delivered alongside. Fitting a core to an already trained network also yields a core and a corrector, but the sum is the network by construction: on Burgers this matches the training-time decomposition, whereas on two-term Wave it leaves a core $20\times$ worse and no closed form, because the network's error in the span of the dictionary passes into the core (Appendix K).

Forming the decomposition after training (fitting the core to a trained network) also returns a core and a corrector, but its total is that field by construction: it matches the curriculum on Burgers and, on minimal Wave, leaves a core $20\times$ worse and no closed form (Appendix K).

5.4 Effect of the Training Order

Table 5 tests whether the core survives when training order is the only thing that differs, with budget, physics schedule, data objective and, in the proximal arm, the coefficient update rule matched (all arms and tests in Appendix E; trajectories in Figure 3). Panel B is the main result. Joint training reaches prediction error within noise of the staged schedules, yet its core-only error is $5.5\times$ higher on Allen–Cahn (0.477 vs. 0.086) and $2.5\times$ higher on Burgers (0.690 vs. 0.279); across every joint arm of Table 12 the ratio ranges from 2.3 to 5.7. Matching the proximal coefficient update does not close the gap: the joint proximal arm ends with a still worse core, and the staged core remains better on Allen–Cahn ($p = 0.01$) and Burgers ($p = 0.02$). On two-term Wave the ordering gap is not significant ($p = 0.09$) while the update rule itself matters ($p = 0.04$; Table 11). This is the behavior Proposition 1 predicts: prediction error cannot see the reallocation, and only the training path prevents it.

Panel A shows the stages doing their intended jobs. Stage I sets the core; Stage II keeps it fixed while reconstruction improves 3–$25\times$; Stage III trades a little core accuracy for fit on Allen–Cahn and briefly worsens Burgers prediction; Stage IV lowers prediction error on both while degrading

Table 5: **Stage progression and training controls** ($\sigma = 0.10$, $n = 5$). RMSE is reported as mean±s.d. **A:** Successive checkpoints at increasing budgets; prediction uses the core at Stage I and the hybrid thereafter. **B:** Full-budget controls. Joint Adam uses the curriculum weight schedule (Allen–Cahn) or shock-weighted objective (Burgers). Full protocols and tests: Appendix E.

	Allen–Cahn		Burgers	
Stage / arm	Pred.	Core	Pred.	Core
A. Cumulative stages				
I	0.0681 ± 0.0001	0.0681 ± 0.0001	0.2785 ± 0.0011	0.2785 ± 0.0011
I+II	0.0216 ± 0.0100	0.0681 ± 0.0001	0.0111 ± 0.0020	0.2785 ± 0.0011
I+II+III	0.0100 ± 0.0025	0.0709 ± 0.0020	0.0138 ± 0.0013	0.2789 ± 0.0013
I–IV	0.0073 ± 0.0018	0.0860 ± 0.0137	0.0076 ± 0.0005	0.2789 ± 0.0013
B. Full-budget controls				
I+II, full	0.0110 ± 0.0031	0.0681 ± 0.0001	0.0077 ± 0.0005	0.2785 ± 0.0011
Joint, Adam	0.0088 ± 0.0026	0.477 ± 0.056	0.0138 ± 0.0012	0.690 ± 0.013
Joint, proximal	0.0134 ± 0.0042	2.8 ± 1.2	0.0160 ± 0.0028	1.38 ± 0.62

the Allen–Cahn core to 0.086. Given the full budget, the core-first schedule (I+II) is not statistically distinguishable from the fit-refined one (I–IV) in reconstruction on any tested benchmark and keeps an equal or better core (Table 12). The refinement stages are therefore an optional trade of core accuracy for fit, which is the choice between the two operating points of Sec. 4.4.

5.5 Additional PDE Benchmarks and Extrapolation

Trained on $[0, 1]^2$ and evaluated on $t \in [0, 5]$, $x \in [0, 2]$ (n=5), an oracle PINN given the true operator collapses outside the training box ($R^2 = -7.6 \pm 7.1$), while the selected symbolic core holds amplitude and phase across the $5\times$ horizon ($R^2 = 0.99998$). This extrapolation test uses a library with a periodicity prior near the true wavenumber. The selected symbolic core carries the observed transfer (Table 6; Appendix M). Breadth runs on 1D Heat, Allen–Cahn and 2D Wave ($\sigma = 0.10$, no operator coupling, n=5 each) reach RMSE 0.0034, 0.0402 and 0.0035, and coupling the analytic reference operator cuts the Allen–Cahn RMSE fivefold to 0.0075 (Appendix N).

Table 6: **Wave extrapolation** (R^2, mean±s.d.; $n = 5$, $\sigma = 0.10$). Training: $[0, 1]^2$; extended: $t \in [0, 5]$, $x \in [0, 2]$. Extended evaluates the hybrid or oracle PINN; Core evaluates the symbolic component on the extended domain. The dictionaries use a periodicity prior.

Model	Training	Extended	Core
PS-PINN ($K = 2$)	0.99999 ± 0.00002	0.9999 ± 0.0000	0.99999 ± 0.00001
PS-PINN ($K = 14$)	0.99998 ± 0.00002	0.9929 ± 0.0052	0.99998 ± 0.00001
Oracle PINN	0.9753 ± 0.0038	-7.59 ± 7.10	—

6 Discussion and Conclusion

On minimal-dictionary Wave the deliverable is effectively a closed-form expression within the supplied dictionary ($r_\theta \approx 0$); on Burgers it is a partially analytic representation.

The symbolic selection uses hand-specified dictionaries, including a Wave dictionary with a periodicity prior near the true wavenumber (Appendices B, L). The division of labor is thus relative to the deployed dictionary and training protocol. **Applicability.** The method suits a smooth bulk with localized hard-to-express regions; fully smooth closed forms are served equally well by pure symbolic regression, and turbulent fields are untested. **Limitations.** Noise is mainly Gaussian, benchmarks are synthetic 1D/2D PDEs, and operator coupling is specified per experiment rather than decided automatically. We do not claim the observations uniquely determine the selected decomposition.

PS-PINN delivers an explicit governing equation and an analytic-core-plus-corrector solution from the same observations. Matched training controls show that similar prediction error can accompany very different symbolic-core error, which motivates evaluating hybrid scientific models for both output fidelity and symbolic-core fidelity.

ETHICS STATEMENT

This work advances transparent scientific machine learning: models whose delivered equation and solution can both be read, checked, and transferred by domain scientists. We foresee no direct negative societal consequences; automated discovery tools should assist, not replace, human scientific judgment.

REPRODUCIBILITY STATEMENT

All benchmarks are synthetic and fully specified (Appendix A), the solution dictionaries are enumerated per benchmark (Appendix B), every experiment fixes and reports its random seeds, and hyperparameters, curriculum pseudocode, and compute are recorded in Appendices D, E, and P. Per-run records and driver scripts ship with the code release, except the archived single-seed Burgers generalization subsection in Appendix M.

AI USE STATEMENT

AI tools were used for manuscript preparation and figure refinement. All methodological decisions, experimental protocols, numerical results, analyses, and conclusions were independently verified by the authors, who take full responsibility for the content.

REFERENCES

Amir Beck and Marc Teboulle. A fast iterative shrinkage-thresholding algorithm for linear inverse problems. *SIAM journal on imaging sciences*, 2(1):183–202, 2009.

Christopher M. Bishop. *Pattern Recognition and Machine Learning*. Springer, 2006.

Gert-Jan Both, Subham Choudhury, Pierre Sens, and Remy Kusters. DeepMoD: Deep learning for model discovery in noisy data. *Journal of Computational Physics*, 428:109985, 2021.

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

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

Miles Cranmer. Interpretable machine learning for science with PySR and SymbolicRegression.jl. *arXiv preprint arXiv:2305.01582*, 2023.

Joy Das, Bivas Bhaumik, Soumen De, and Satyasaran Changdar. Physics-informed neural network with symbolic regression for deriving analytical approximate solutions to nonlinear partial differential equations. *Neural Computing and Applications*, 37(24):20205–20240, 2025.

Ingrid Daubechies, Michel Defrise, and Christine De Mol. An iterative thresholding algorithm for linear inverse problems with a sparsity constraint. *Communications on Pure and Applied Mathematics: A Journal Issued by the Courant Institute of Mathematical Sciences*, 57(11):1413–1457, 2004.

Jérémie Dona, Marie Déchelle, Patrick Gallinari, and Marina Levy. Constrained physical-statistics models for dynamical system identification and prediction. In *International Conference on Learning Representations (ICLR)*, 2022.

Mengge Du, Yuntian Chen, Longfeng Nie, Siyu Lou, and Dongxiao Zhang. Physics-constrained robust learning of open-form partial differential equations from limited and noisy data. *Physics of Fluids*, 36(5):057123, 2024. doi: 10.1063/5.0204187.

Ivar Ekeland and Roger Temam. *Convex analysis and variational problems*. SIAM, 1999.

Mingquan Feng, Yixin Huang, Yizhou Liu, Bofang Jiang, and Junchi Yan. PhysPDE: Rethinking PDE discovery and a physical HYpothesis selection benchmark. In *International Conference on Learning Representations (ICLR)*, 2025.

Aditi S. Krishnapriyan, Amir Gholami, Shandian Zhe, Robert M. Kirby, and Michael W. Mahoney. Characterizing possible failure modes in physics-informed neural networks. In *Advances in Neural Information Processing Systems (NeurIPS)*, volume 34, pp. 26548–26560, 2021. arXiv:2109.01050.

Senwei Liang and Haizhao Yang. Finite expression method for solving high-dimensional partial differential equations. *Journal of Machine Learning Research*, 26(138):1–31, 2025. arXiv:2206.10121.

Ritam Majumdar, Vishal Jadhav, Anirudh Deodhar, Shirish Karande, Lovekesh Vig, and Venkataramana Runkana. Physics informed symbolic networks. *arXiv preprint arXiv:2207.06240*, 2022. NeurIPS Workshop on the Symbiosis of Deep Learning and Differential Equations.

Ritam Majumdar, Vishal Jadhav, Anirudh Deodhar, Shirish Karande, Lovekesh Vig, and Venkataramana Runkana. Symbolic regression for PDEs using pruned differentiable programs. *arXiv preprint arXiv:2303.07009*, 2023. ICLR Workshop on Physics for Machine Learning.

David McShannon and Nicholas Dietrich. Silent failures in physics-informed neural networks: Parameter poisoning and the limits of loss-based validation. *arXiv preprint arXiv:2606.25151*, 2026.

Daniel A. Messenger and David M. Bortz. Weak SINDy for partial differential equations. *Journal of Computational Physics*, 443:110525, 2021. doi: 10.1016/j.jcp.2021.110525.

Pancheng Niu, Jun Guo, Qiaolin He, Jingcai Guo, and Yanchao Shi. Desyr: A decoupled symbolic recovery framework with pinn-guided structure search and physics-informed coefficient refinement. *arXiv preprint arXiv:2609.00530*, 2026.

Orestis Oikonomou, Levi Lingsch, Dana Grund, Siddhartha Mishra, and Georgios Kissas. Neuro-symbolic AI for analytical solutions of differential equations. *arXiv preprint arXiv:2502.01476*, 2025.

Trevor Park and George Casella. The bayesian lasso. *Journal of the american statistical association*, 103(482):681–686, 2008.

Lena Podina, Brydon Eastman, and Mohammad Kohandel. Universal physics-informed neural networks: Symbolic differential operator discovery with sparse data. In *International Conference on Machine Learning*, volume 202 of *Proceedings of Machine Learning Research*, pp. 27948–27956, 2023.

Nasim Rahaman, Aristide Baratin, Devansh Arpit, Felix Draxler, Min Lin, Fred Hamprecht, Yoshua Bengio, and Aaron Courville. On the spectral bias of neural networks. In *International conference on machine learning*, pp. 5301–5310. PMLR, 2019.

Till Richter and Niki Kilbertus. Orthoreg: Orthogonal regularization for hybrid symbolic-neural dynamical systems. *arXiv preprint arXiv:2606.19145*, 2026.

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

Hayden Schaeffer. Learning partial differential equations via data discovery and sparse optimization. *Proceedings of the Royal Society A: Mathematical, Physical and Engineering Sciences*, 473(2197), 2017.

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

Robert Stephany and Christopher Earls. PDE-READ: Human-readable partial differential equation discovery using deep learning. *Neural Networks*, 154:360–382, 2022.

Robert Stephany and Christopher Earls. PDE-LEARN: Using deep learning to discover partial differential equations from noisy, limited data. *Neural Networks*, 174:106242, 2024a.

Robert Stephany and Christopher Earls. Weak-PDE-LEARN: A weak form based approach to discovering PDEs from noisy, limited data. *Journal of Computational Physics*, 506:112950, 2024b.

Naoya Takeishi and Alexandros Kalousis. Physics-integrated variational autoencoders for robust and interpretable generative modeling. In *Advances in Neural Information Processing Systems (NeurIPS)*, volume 34, pp. 14809–14821, 2021. arXiv:2102.13156.

Matthew Tancik, Pratul Srinivasan, Ben Mildenhall, Sara Fridovich-Keil, Nithin Raghavan, Utkarsh Singhal, Ravi Ramamoorthi, Jonathan Barron, and Ren Ng. Fourier features let networks learn high frequency functions in low dimensional domains. *Advances in neural information processing systems*, 33:7537–7547, 2020.

Pongpisit Thanasutives, Takashi Morita, Masayuki Numao, and Ken-ichi Fukui. Adaptive uncertainty-penalized model selection for data-driven PDE discovery. *IEEE Access*, 12:13165–13182, 2024. doi: 10.1109/ACCESS.2024.3354819.

Robert Tibshirani. Regression shrinkage and selection via the lasso. *Journal of the Royal Statistical Society Series B: Statistical Methodology*, 58(1):267–288, 1996.

Paul Tseng. Convergence of a block coordinate descent method for nondifferentiable minimization. *Journal of optimization theory and applications*, 109(3):475–494, 2001.

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

Martin J Wainwright. Sharp thresholds for high-dimensional and noisy sparsity recovery using ℓ_1-constrained quadratic programming (lasso). *IEEE transactions on information theory*, 55(5): 2183–2202, 2009.

Liang Wang, Tingwei Ji, Yunzhe Huang, Hongjie Zhou, and Fangfang Xie. Physics-based discovery of governing equations from scarce and noisy data. *Physical Review E*, 111:065314, 2025. doi: 10.1103/d4tm-92vb.

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

Hao Xu, Junsheng Zeng, and Dongxiao Zhang. Discovery of partial differential equations from highly noisy and sparse data with physics-informed information criterion. *Research*, 6:0147, 2023. doi: 10.34133/research.0147.

Yuan Yin, Vincent Le Guen, Jérémie Dona, Emmanuel de Bézenac, Ibrahim Ayed, Nicolas Thome, and Patrick Gallinari. Augmenting physical models with deep networks for complex dynamics forecasting. In *International Conference on Learning Representations (ICLR)*, 2021. arXiv:2010.04456.

A PDE BENCHMARKS AND DATA GENERATION

All benchmarks are synthetic with known ground truth, so reconstruction error is measured against the noise-free field. Table 7 summarizes; details follow.

Table 7: Benchmark specifications.

PDE	Equation	Domain	Grid	Data source	Noise convention
1D Wave	$u_{tt} = u_{xx}$	$x, t \in [0, 1]$	128×128	analytic	absolute
Burgers	$u_t + uu_x = \nu u_{xx}$, $\nu = \frac{0.01}{\pi}$	$x \in [-1, 1]$, $t \in [0, 1]$	256×100	BDF solver	scaled by $\mathrm{std}(u)$
1D Heat	$u_t = 0.05\, u_{xx}$	$x, t \in [0, 1]$	128×128	analytic	absolute
Allen–Cahn	$u_t = \varepsilon^2 u_{xx} + u - u^3$, ε^2=0.01	$x \in [-1, 1)$, $t \in [0, 1]$	128×128	numerical	absolute
2D Wave	$u_{tt} = u_{xx} + u_{yy}$	$t \in [0, 1]$, $(x, y) \in [-1, 1]^2$	24^3	analytic	absolute

1D Wave. Exact standing-wave solution $u^\star(x, t) = \sin(2\pi x)\cos(2\pi t)$ sampled on the grid, with periodic boundary and analytic initial condition. **Burgers.** Initial condition $u(x, 0) = -\sin(\pi x)$ with homogeneous Dirichlet boundaries, integrated by the method of lines with an implicit BDF solver to handle stiffness. A sharp shock forms at $x = 0$ from $t \approx 0.5$. **1D Heat.** Separable decaying mode consistent with $u_t = 0.05\, u_{xx}$. **Allen–Cahn.** Metastable front dynamics in the $\tanh$-front regime, from $u(x, 0) = 0.53\sin(\pi x) - 0.47\sin(3\pi x)$ with periodic boundaries, integrated by sub-stepped explicit Euler on a 256×200 grid and subsampled to 128×128. **2D Wave.** Product standing mode $\sin(2\pi x)\sin(2\pi y)\cos(2\sqrt{2}\pi t)$ with unit speed.

Noise model. Clean fields are corrupted by additive Gaussian white noise $\xi \sim \mathcal{N}(0, 1)$: $u_{\text{noisy}} = u_{\text{clean}} + \sigma\,\mathrm{std}(u)\,\xi$ for Burgers and $u_{\text{noisy}} = u_{\text{clean}} + \sigma\xi$ for the others, so noise levels are comparable within a benchmark but not across benchmarks. All noise-robustness results in this paper are under this Gaussian assumption.

Sampling, seeds, evaluation. Training uses the full noisy grid; evaluation is training-domain reconstruction against the clean field (extrapolation protocols are separate, Appendix M). Every multi-seed cell fixes and reports its random seeds. Main-result and controlled-comparison cells use $n \geq 5$ (the $\sigma = 0.10$ Burgers PS-PINN cell pools two independent 5-run batches, $n = 10$, while the standard-dictionary Wave cells are $n = 5$); the operator diagnostic check and the library-sensitivity diagnostic use $n = 3$.

B SYMBOLIC SOLUTION DICTIONARIES

The solution dictionaries are hand-specified, physics-informed candidate sets. PS-PINN performs sparse selection and reconstruction over them, not unrestricted open-form discovery. In particular, the Wave dictionary encodes a periodicity prior with wavenumbers near the true mode. Every interpretability and extrapolation claim in the paper must be read with this prior in mind (Sec. 5.5).

1D Wave (full, 14 elements). Built from four trigonometric families instantiated at three near-identical wavenumbers $k_j \in \{6.2832,\ 6.28327,\ 6.28203\}$ around the true 2π:

$$\varphi_S^{(j)} = \sin(k_j x)\cos(k_j t), \quad \tilde{\varphi}_S^{(j)} = \sin(k_j t)\cos(k_j x), \quad \varphi_\pm^{(j)} = \sin(k_j(x \pm t)),$$

plus two off-frequency mixed terms $\sin(t + 8.375\,x)$ and $\sin(8.375\,t + x)$ (14 elements in total). Standing and traveling forms are linked by the product-to-sum identity $\varphi_S = \frac{1}{2}(\varphi_+ + \varphi_-)$, which makes this dictionary deliberately redundant. **1D Wave (minimal, 2 elements).** The traveling-wave pair $\sin(2\pi(t \pm x))$, an oracle-dependent controlled configuration, since the true wavenumber is given. **Burgers.** Polynomials up to degree two and trigonometric/hyperbolic candidates such as $\sin(\pi x)$ and $\tanh$ profiles. The shock front itself is intentionally not representable, which is what the neural corrector is for. **1D Heat.** Separable modes $\sin(kx)\,e^{-\alpha k^2 t}$ at several wavenumbers around the true 2π and nearby decay rates, plus generic low-order polynomial nuisance terms, so the sparse selection has to choose the correct mode rather than have it handed over. **Allen–Cahn.** Standing $\tanh$ fronts at the equilibria $u = \pm 1$ and moving fronts $\tanh(\alpha(x \pm vt))$ over a small grid of widths α and speeds v, plus low-order corrections. **2D Wave (9 elements).** Four product modes at the true wavenumber, $\{\sin(kx)\sin(ky),\ \cos(kx)\cos(ky)\} \times \{\cos, \sin\}(\omega t)$ with $k = 2\pi$ and $\omega = \sqrt{2}\,k$; one harmonic $\sin(2kx)\sin(2ky)\cos(2\omega t)$; and the low-order nuisance terms 1, x, y, t.

Table 8: Dictionary summary and encoded prior knowledge.

PDE	#basis	Families	Prior knowledge encoded
1D Wave (full)	14	standing + traveling trig	wavenumbers near true 2π; deliberate redundancy
1D Wave (minimal)	2	traveling trig	true wavenumber
Burgers	11	poly ≤ 2, trig, $\tanh$	smooth families only; no shock profile
1D Heat	43	$\sin\cdot\exp$ modes, poly	wavenumbers/decay near truth + nuisance terms
Allen–Cahn	19	$\tanh$ fronts, low-order	front regime; widths/speeds gridded
2D Wave	9	product trig modes, low-order	true wavenumber; one harmonic + nuisance terms

C Governing-Equation Discovery Details

Preprocessing. A Gaussian-smoothed copy of the observations ($\sigma_{\text{smooth}} = 2$ grid units) is used only for derivative-based discovery, while all solution training consumes the raw noisy tensor. This separation matters for interpreting the results, since it means the reconstruction quality of Sec. 5 is not an artifact of training on denoised data. **Derivatives and library.** Spatial and temporal derivatives are computed by central finite differences on the smoothed field, with boundary trimming. The candidate PDE library $\Theta(u)$ contains low-order derivatives and their pointwise products ($u, u_x, u_{xx}, uu_x, \ldots$), specified per benchmark. **Sparse regression.** Sequential thresholded least squares (STLSQ) with a pre-set magnitude threshold and column normalization, with coefficients rescaled to physical units after selection. On Burgers, the coefficient fit uses PySR's symbolic regression over the derivative features. The recovered $\hat{\boldsymbol{\xi}}$ is frozen after discovery. **Operator provenance per protocol.** Two distinct protocols use the physics residual, and they differ in provenance. (*i*) In the Track-A experiments of Table 3, the residual uses the operator recovered from the same observations, that is, a discovered or template-constrained operator is coupled with its estimated coefficients. This holds at $\sigma = 0$ for two-mode Wave, where the operator is selected among candidates. It also holds, under protocol-specified coupling, for the template-constrained Heat and Allen–Cahn rows at $\sigma = 0$ and for Allen–Cahn at $\sigma \in \{0.005, 0.01, 0.02\}$, whose drifted coefficients (14–21%) are enforced as recovered precisely to measure the value of enforcement (see below). On two-mode Wave at the same noise levels the biased operator is not coupled; the coupled counterfactual is reported only as the harm it would cause. (*ii*) In the separate high-noise ($\sigma = 0.10$) experiments the coupling varies by protocol and is stated per table. The breadth rows of Sec. 5.5 train with the physics term off for 1D Heat, 2D Wave, and Allen–Cahn; the single Allen–Cahn enforcement row uses the analytic reference operator ($\varepsilon^2 = 0.01$). The decomposition and curriculum experiments use the analytic reference operators for minimal Wave ($c^2 = 1$) and Allen–Cahn to isolate solution-representation behavior from discovery error, while Burgers trains with the coupling off. 1D Wave and Burgers use their data-discovered operators only where stated. The question of when a discovered operator is reliable enough to couple is outside the scope of this paper.

Operator diagnostic check. This diagnostic is reported only as a robustness characterization of the discovered operators and is intentionally decoupled from the coupling protocol of Sec. 5, which fixes coupling per experiment. It is oracle-free and pre-registered. Its thresholds were fixed before any labeled perturbation experiment and were not tuned afterwards, and no clean field or true coefficient enters it. The input is the noisy observations alone. The operator is re-identified under an ensemble of derivative-estimation mechanisms: Gaussian-smoothed central differences with $\sigma_{\text{smooth}} \in \{0, 1, 2, 4\}$ grid units, Savitzky–Golay local polynomials with windows $\{7, 11\}$ and order 4, and spectral derivatives with and without a half-Nyquist low-pass on periodic benchmarks. Two weak-form members, which test the PDE against compactly supported polynomial bumps on non-overlapping 24- and 32-point patches and so estimate no derivatives of the data, join the ensemble for the dispersion and support-instability statistics; identifiability is evaluated on the strong-form members only. Each member runs least squares on five random 80% row subsamples; the strong-form members are restricted to active-dynamics rows where the target magnitude is at least its interior median. Three statistics decide, with Θ denoting the PDE candidate feature matrix of the discovery regression (columns u_{xx}, u, u^3, ...), not the solution dictionary Φ of the main text. (*a*) *Identifiability*: the smallest eigenvalue of the unit-normalized candidate Gram matrix, minimized over the strong-form members, must satisfy $\lambda_{\min} \geq 0.05$. (*b*) *Propagated dispersion*: members whose own $\lambda_{\min}$ is below 0.05 are excluded, and the operator is flagged if fewer than three remain. With $m_{c,j}$ and $\text{IQR}_{c,j}$ the median and interquartile range of coefficient ξ_j over the five subsamples of member c, the dispersion of term j is $d_j = (\max_c m_{c,j} - \min_c m_{c,j}) + \text{median}_c\, \text{IQR}_{c,j}$, the cross-estimator range of medians plus the typical within-estimator spread, and $U_{\text{op}} = \sum_j d_j \, \|\Theta_{:,j}\|_2 / \|\text{target}\|_2 < 0.10$, with

column and target norms taken from the unsmoothed finite-difference system. (*c*) *Support instability*: term j is active in a retained subsample fit when its normalized contribution $|\xi_j| \, \|\Theta_{:,j}\|_2 / \|\mathrm{target}\|_2$ exceeds 0.05, and the fraction of fits whose active set differs from the most frequent one must satisfy $\rho \leq 0.20$. An operator passes the diagnostic only if all three hold. Its results on this paper's data are as follows. Two-mode Wave at $\sigma = 0$ passes ($\lambda_{\min} = 0.09$, $U_{\mathrm{op}} = 0.0$, $\rho = 0$). Two-mode Wave at $\sigma \in \{0.005\text{–}0.02\}$ fails by dispersion ($U_{\mathrm{op}} = 0.69\text{–}0.87$ across 15 runs, against a threshold of 0.10). Burgers at $\sigma = 0.10$ fails by dispersion ($U_{\mathrm{op}} = 2.6$; the preprocessing-induced $\hat{\nu}$ bias surfaces as estimator disagreement). Single-mode Heat is non-identifiable for $\sigma \leq 0.05$ ($\lambda_{\min} \leq 0.013$, and $\lambda_{\min} \approx 0$ at $\sigma = 0$, from the $u_{xx} \equiv -k^2 u$ degeneracy); at $\sigma = 0.10$ the noise lifts $\lambda_{\min}$ to the threshold and one of three seeds passes marginally ($\lambda_{\min} = 0.051$, $U_{\mathrm{op}} = 0.098$). Allen–Cahn fails by dispersion at every tested σ including 0 ($U_{\mathrm{op}} \geq 0.35$), because different smoothing scales distort the sharp interfaces differently. These results do not decide coupling. For instance, the noisy Allen–Cahn rows couple operators that the diagnostic flags, to measure what coupling a drifted operator still gains; coupling them improves reconstruction $2.4\text{–}2.6\times$ and the core itself $\approx 1.8\times$ ($0.14\text{–}0.16 \rightarrow 0.082\text{–}0.089$).

Same-run dual-artifact pairing (noisy Allen–Cahn enforcement). Table 9 pairs, per run, the estimated equation with the solution artifacts delivered by that same run at $\sigma = 0.01$. This is the representative noise level under a selection rule fixed in advance (median tested noise, all seeds shown). The delivered core is explicit and inspectable but not compact. In every seed, 18 of the 19 candidates stay active, led by $+0.42 \sin(\pi x)$ and $-0.39 \sin(3\pi x)$ with a spread of small $\tanh$-front terms (≤ 0.06 each, seed 0). Compactness claims in this paper are limited to the Wave configurations (Sec. 2, Appendix L). What this experiment demonstrates is the dual output and the enforcement gain, not a compact Allen–Cahn formula.

Table 9: Noisy Allen–Cahn dual-artifact pairing ($\sigma = 0.01$, all seeds; template $u_t = a\, u_{xx} + b\, u + c\, u^3$; discovered operator coupled by protocol). "Active" counts $|c_k| > 10^{-3} \max_k |c_k|$ over the de-standardized (physical-unit) coefficients, a reporting convention only. This reporting threshold does not modify $\boldsymbol{\alpha}$ during training, where the only sparsity operation is the proximal soft-threshold (Appendix E), and active-term counts under different reporting thresholds are not directly comparable.

Seed	Estimated PDE ($\hat{u}_t =$)	Active	Core RMSE	Hybrid RMSE	ρ_{NN}
0	$0.0121 u_{xx} + 1.154 u - 1.154 u^3$	18/19	0.075	0.012	0.025
1	$0.0121 u_{xx} + 1.146 u - 1.137 u^3$	18/19	0.081	0.014	0.030
2	$0.0121 u_{xx} + 1.143 u - 1.134 u^3$	18/19	0.092	0.012	0.033
3	$0.0121 u_{xx} + 1.146 u - 1.138 u^3$	18/19	0.109	0.016	0.042
4	$0.0120 u_{xx} + 1.142 u - 1.138 u^3$	18/19	0.092	0.012	0.032

D Network Architecture and Hyperparameters

All compared variants and controls share the same architecture, initialization scheme, and per-benchmark training budget; no per-method tuning is performed on top. Comparisons are always within one benchmark, so configuration-specific defaults never separate two arms of the same comparison.

E Symbolic-First Curriculum and Pseudocode

Sparsity mechanism, verified against the implementation. In the staged arms, the only sparsity operation applied during training is the proximal soft-threshold shown in Algorithm 1, namely $\mathrm{soft}(z, \eta\lambda_{\mathrm{sym}})$ in the Stage-I and Stage-III $\boldsymbol{\alpha}$ half-steps. The Burgers Stage I anneals it along a λ-path from $\lambda_{\max}$ down to $10^{-4}\lambda_{\max}$. No hard-threshold pruning is applied at any stage. Coefficients shrunk to zero remain in the optimizer's parameter list and may re-enter in Stages III–IV. The thresholds of Table 10 and Table 9 are reporting conventions only, used for active-term counts in logs and tables, and never modify $\boldsymbol{\alpha}$. The staged arms (four-stage, both two-stage arms, and the component ablation) call the same stage implementations, so the rule is identical across them. The Adam-based joint controls do not: they update $\boldsymbol{\alpha}$ with Adam alongside θ, the ℓ_1 term entering as a subgradient penalty, except that the Burgers matched-objective arm instead soft-thresholds $\boldsymbol{\alpha}$ once per epoch. The joint, proximal-$\boldsymbol{\alpha}$ arm instead applies the staged Stage-III half-step at every mini-batch. Appendix O analyzes an idealized alternating scheme with a fixed linear operator and fixed loss weights under

Table 10: Model and optimization hyperparameters for the minimal-dictionary Wave, 1D Heat, Allen–Cahn, 2D Wave, and all control and ablation arms. The standard $K{=}14$ Wave configuration behind Table 4 uses its own defaults (batch 1024, $\lambda_{\mathrm{sym}} = 10^{-4}$, stage passes $400/1000/5{\times}(50{+}200)/500$); the Burgers entries below use the same defaults.

Component	Setting
Neural corrector r_θ	MLP, 4 layers, 128 hidden units, $\tanh$ activations
Input/output scaling	inputs normalized to the unit cube; output unscaled
Symbolic coefficients $\boldsymbol{\alpha}$	zero-initialized; proximal (soft-threshold) updates
Optimizer / learning rate	Adam, 10^{-3} (both branches)
λ_{sym} (ℓ_1, standardized $\boldsymbol{\alpha}$)	10^{-5} (harness benchmarks; see caption)
Sparsity (training)	proximal soft-threshold only; no pruning; terms may re-enter
Active-term count (reporting only)	$\lvert c_k\rvert > 10^{-3}\max\lvert c\rvert$, de-standardized (App. E)
Data loss	MSE; Burgers: shock-weighted MAE (App. E)
Batch size	4096 (Wave, Heat, Allen–Cahn), 2048 (Burgers)
Collocation points	training-grid points (physics residual, when active)
λ_{phys} schedule	10^{-3}, stepped to 1.0 one fifth into Stage II
Stage passes, minimal Wave/Heat/AC	$400/500/2{\times}(100{+}150)/300$ (1700)
Stage passes, Burgers	$500/1000/5{\times}(50{+}200)/500$ (3250)
Stage passes, standard Wave	$400/1000/5{\times}(50{+}200)/500$ (3150)
Stage passes, 2D Wave (breadth runs)	$400/300/2{\times}(100{+}80)/200$ (1260)
Budgets	fixed per benchmark across arms (except two-stage, truncated)

Algorithm 1: PS-PINN symbolic-first curriculum; Stages III–IV optional

Input: noisy data $\mathcal{D}$; dictionary Φ; coupled operator $\mathcal{F}_c$ (if coupling is enabled)

Stage I (symbolic initialization): $r_\theta \equiv 0$; solve $\boldsymbol{\alpha}^{(0)} = \arg\min_{\boldsymbol{\alpha}} \frac{1}{2}\|y - \Phi\boldsymbol{\alpha}\|_2^2 + \lambda_{\mathrm{sym}}\|\boldsymbol{\alpha}\|_1$ by proximal gradient;

Stage II (neural correction): freeze $\boldsymbol{\alpha}^{(0)}$; train $\theta^{(0)} = \arg\min_\theta \|y - (\Phi\boldsymbol{\alpha}^{(0)} + r_\theta)\|_2^2$ ($+\,\lambda_{\mathrm{phys}}\mathcal{L}_{\mathrm{phys}}$ if active; λ_{phys} steps $10^{-3}\to 1$ after the first fifth of the stage);

Stage III (alternating refinement): **for** $k = 0, 1, 2, \ldots$ **do**

  $\boldsymbol{\alpha}^{(k+1)} \leftarrow$ proximal step on the data + ℓ_1 terms at $\theta^{(k)}$; `// targets` $y - r_{\theta^{(k)}}$`; no physics gradient`

  $\theta^{(k+1)} \leftarrow$ Adam steps on $\mathcal{L}(\boldsymbol{\alpha}^{(k+1)}, \cdot)$ at $\lambda_{\mathrm{phys}} = 1$;

Stage IV (joint fine-tuning): release $(\boldsymbol{\alpha}, \theta)$ jointly; minimize $\mathcal{L}$;

Output: discovered equation $\hat{\mathcal{F}}$ and solution $\hat{u} = \Phi\boldsymbol{\alpha} + r_\theta$

its stated block-solver and regularity assumptions. Using proximal updates in the implementation does not by itself verify those assumptions. In particular, raising λ_{phys} between steps changes the objective by $(\lambda_2 - \lambda_1)\mathcal{L}_{\mathrm{phys}} \geq 0$ at fixed parameters, so fixed-weight monotone descent does not extend across the weight increase. The analysis therefore does not establish monotone descent for the full finite-budget curriculum, its weight schedule, or joint fine-tuning, and the curriculum's value rests on the controlled comparisons of Tables 5 and 12.

Implementation deviations from Equation (6) and Algorithm 1. Three aspects of the deployed pipelines differ from the idealized description, and all reported results come from the deployed versions. (*i*) *Weight schedule.* When the physics term is active, λ_{phys} starts at 10^{-3} and steps to 1 after the first fifth of Stage II; Stages III–IV train at the full weight throughout, with no smooth ramp inside Stage III (the reported Burgers runs train with the physics term off; Appendix C). The linear-ramp joint control reported in Table 12 instead increases λ_{phys} linearly from 10^{-3} to 1 over the first 60% of its total passes, so it probes gradual weighting in general rather than replicating the curriculum's schedule. (*ii*) $\boldsymbol{\alpha}$ *half-steps.* In Stages I and III the $\boldsymbol{\alpha}$ update is a proximal (soft-threshold) gradient step on the data-plus-ℓ_1 terms against the current residual $y - r_\theta$; the physics residual contributes no gradient to $\boldsymbol{\alpha}$ and enters the alternation only through the θ half-step. When every dictionary element satisfies the enforced operator exactly, as in the minimal Wave dictionary where $\mathcal{N}\Phi = 0$, the physics term does not depend on $\boldsymbol{\alpha}$ and this coincides with a proximal step on the full objective; for the nonlinear Allen–Cahn residual it is a partial-objective surrogate. The exact block solves assumed in Appendix O are idealizations in this respect as well. (*iii*) *Burgers objective.*

The Burgers pipeline replaces the mean-squared data term with a shock-weighted loss, the mean absolute error plus twice the mean over the largest decile of absolute errors, in the neural updates of Stages II–IV, and its Stage-IV objective carries no ℓ_1 term (sparsity of $\boldsymbol{\alpha}$ comes from the Stage-I/III proximal steps); Wave and Allen–Cahn use the mean-squared data term and keep the ℓ_1 term in Stage IV. For the Burgers comparison reported in Table 5, the matched-objective joint arm uses the same shock-weighted data loss and the same soft-threshold operator as the staged arm, but not the same $\boldsymbol{\alpha}$ update rule: the staged $\boldsymbol{\alpha}$ half-steps are per-batch proximal gradient steps on the squared data term, with a λ-path in Stage I, whereas the joint arm updates $\boldsymbol{\alpha}$ with Adam on the shock-weighted loss and soft-thresholds once per epoch. The comparison therefore matches the data loss and the sparsity operator, not the optimizer or thresholding frequency; the joint, proximal-$\boldsymbol{\alpha}$ arm below matches those as well. The constant-weight joint recipe, which uses the squared data term with an ℓ_1 penalty, remains reported separately in Table 12.

Budget-matched joint control (constant weight). The control starts from the same data, noise sample, and model initialization, and minimizes the full objective of Equation (6) directly over $(\boldsymbol{\alpha}, \theta)$ from the first step with Adam on both blocks (the ℓ_1 term as a subgradient penalty), with all loss weights at their final values, no parameter freezing, and no ramp. Its budget matches the curriculum's total passes over the data (Allen–Cahn 1700, Burgers 3250 epochs), which also matches wall-clock within 15%. Per-stage full-grid snapshots record hybrid RMSE, symbolic-only RMSE, $\|r_\theta\|_\infty$, and the active-term count $|\mathrm{supp}(\boldsymbol{\alpha})|$. Figure 3 plots the trajectories and Table 12 the endpoints of every arm. The control matches hybrid RMSE within 1.1–1.3× while its symbolic part never leaves the amplitude scale of the field (Allen–Cahn) or ends worse than predicting zero (Burgers). Joint training reallocates the core's share to the network.

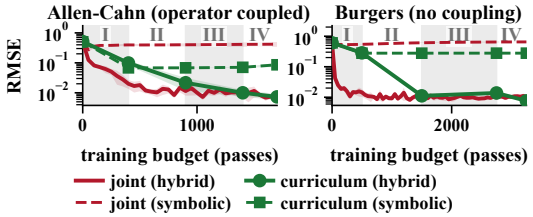


Figure 3: **Curriculum anatomy.** Stages I–IV shaded ($n = 5$). The joint control reaches similar hybrid error while retaining a poorer symbolic core. The Stage-I hybrid snapshot includes the untrained corrector; the Stage-I prediction in Table 5 uses the symbolic core alone.

Two-stage arms. The truncated two-stage arm stops after Stages I+II, at 900 of 1700 passes on minimal Wave and Allen–Cahn and 1500 of 3250 on Burgers (46–53% of the four-stage budget), and isolates the effect of omitting later refinement under reduced compute. Its symbolic-first, residual-second ordering is analogous to PINSN's staged variant (Majumdar et al., 2022) but is not a reproduction of it: it uses our dictionary, architecture, and data (Appendix H). We additionally evaluate a full-budget two-stage arm, which extends Stage II to the complete budget (Table 12). Because $\boldsymbol{\alpha}$ is frozen after Stage I, the extra Stage-II budget can improve only the corrector, so this arm keeps the Stage-I core. It is the full-budget two-stage control tested here, with the Stage-I budget fixed at its deployed value; two-stage schedules that give Stage I more of the budget, and could therefore change the core, were not tested. At matched budget it is not statistically distinguishable from the four-stage arm in reconstruction on any of the three benchmarks, while retaining an equal or better symbolic core. Stages III–IV are therefore not necessary for reconstruction; they define an optional refinement path that may trade core fidelity for prediction accuracy.

Joint training with proximal $\boldsymbol{\alpha}$ (pre-registered). The joint controls above update $\boldsymbol{\alpha}$ with Adam, whereas the staged arms take per-batch proximal steps, so they alone cannot separate stage ordering from the coefficient update rule. A further arm removes that difference. It trains $\boldsymbol{\alpha}$ and θ jointly from initialization, with no freezing and no symbolic-first phase. Each mini-batch takes the staged Stage-III proximal half-step on $\boldsymbol{\alpha}$, with the staged step size, threshold and frequency (plus the Stage-I offset rule for b_0 on Allen–Cahn and Wave), followed by one Adam step on θ with the loss, learning rate and physics schedule of the matched Adam-based joint arm (joint, curriculum schedule on Allen–Cahn and Wave; joint, matched objective on Burgers). The arm, the comparators, the effect definitions and the decision rule were registered before any run (n=5). Table 11 gives the two effects. On Allen–Cahn and Burgers the symbolic-first advantage survives matching the proximal update, with confidence intervals for the ordering effect that exclude zero. This is not late-stage drift alone:

the best core reached at any logged epoch of joint-proximal training (every 50 epochs) is 0.15–0.23 on Allen–Cahn and 0.40–0.61 on Burgers, against staged endpoints of 0.086 and 0.279 (a descriptive check, not a test). In the second half of training the two branches drift in lockstep on both benchmarks, with core RMSE and corrector RMS correlated at 0.94–1.00 per seed, while prediction RMSE ends at 0.013 and 0.016. This drift is consistent with the allocation ambiguity discussed in Appendix O: the prediction terms depend only on the sum, although the ℓ_1 penalty can distinguish reallocations, and the remark compares objective values without predicting this trajectory. On Allen–Cahn the proximal arm ends with an even worse core than the Adam-based one, so proximal updates do not by themselves protect the core. On minimal Wave the remaining ordering gap is not significant, whereas proximal updates improve significantly on Adam. The full-budget two-stage comparator gives the same conclusions. The arm still differs from the staged arms in ways listed in the registration before any run: the staged arms use stage-dependent θ learning rates that decrease from 10^{-3} across Stages II–IV, update $\boldsymbol{\alpha}$ with Adam in Stage IV, and on Burgers anneal a λ-path in Stage I and decay the Stage-II learning rate; against the Adam-based joint arm, θ keeps its own Adam state and, on Burgers, only θ gradients are clipped. The comparison therefore separates ordering from the $\boldsymbol{\alpha}$-update rule, not from every other optimization detail. What it supports is that, under these controls, the staged schedules retain a better core on Allen–Cahn and Burgers: matching the $\boldsymbol{\alpha}$-update rule does not eliminate the symbolic-first advantage there, although the update rule itself matters on minimal Wave.

Table 11: **Ordering versus $\boldsymbol{\alpha}$-update rule** (symbolic-core RMSE, $\sigma = 0.10$, n=5 per arm; 95% Welch intervals and two-sided Welch p). The ordering effect is the core RMSE of *joint, proximal* $\boldsymbol{\alpha}$ minus that of the four-stage arm, so a positive value means symbolic-first ordering helps. The update-rule effect is the core RMSE of *joint, Adam* $\boldsymbol{\alpha}$ (joint, curriculum schedule; Burgers: joint, matched objective) minus that of *joint, proximal* $\boldsymbol{\alpha}$, so a positive value means proximal updates help. The arms still differ in the pre-registered residual details listed in the text, so these effects separate ordering from the $\boldsymbol{\alpha}$-update rule only.

	four-stage	joint, proximal $\boldsymbol{\alpha}$	joint, Adam $\boldsymbol{\alpha}$	ordering effect	update-rule effect
Allen–Cahn	0.086	2.8	0.477	$+2.75$ $[+1.07, +4.43]$, $p = 0.01$	-2.36 $[-4.04, -0.68]$, $p = 0.02$
Burgers	0.279	1.38	0.690	$+1.10$ $[+0.24, +1.96]$, $p = 0.02$	-0.69 $[-1.55, +0.17]$, $p = 0.09$
Wave (minimal)	0.0014	0.0060	0.0180	$+0.0046$ $[-0.0012, +0.0103]$, $p = 0.09$	$+0.0120$ $[+0.0007, +0.0233]$, $p = 0.04$

Operator provenance, and joining the two evidence chains. The comparisons of Tables 5 and 12 use analytic reference operators (minimal Wave, Allen–Cahn) or train with the coupling off (Burgers), which isolates representation training from discovery error. The dual-output experiments of Table 3 instead enforce data-discovered coefficients. Reporting only those two would leave their product unsupported: whether the ordering effect survives when the enforced operator is itself recovered from the same noisy observations. It does. Repeating the three schedules on Allen–Cahn at $\sigma = 0.01$, with the operator discovered from each run's own observations and held fixed across arms (coefficients drift $+15$ to $+21\%$ from the true $(\varepsilon^2, 1, -1)$; n=5), reproduces the same pattern. The two staged arms retain the core (0.090 ± 0.012 for the four-stage curriculum, 0.0681 ± 0.0001 for the full-budget two-stage arm), while joint training at the same budget leaves a core of 0.510 ± 0.078, nearly six times worse ($p = 3.5 \times 10^{-4}$). All three arms are indistinguishable in reconstruction (0.0132 ± 0.0014, 0.0141 ± 0.0022 and 0.0122 ± 0.0009; against the curriculum, $p = 0.52$ for the two-stage arm and $p = 0.28$ for the joint arm), and the joint arm in fact attains the numerically best reconstruction of the three. This reallocation is therefore not an artifact of enforcing a clean operator, and prediction accuracy remains unable to reveal it.

F Metrics for Analytic-First Interpretability

Numerical accuracy. RMSE and MAE are computed against the clean field on the training grid; R^2 is the coefficient of determination.

Expression Structural Accuracy (ESA). For a predicted symbolic expression f_{pred} and ground truth f_{true},

$$\mathrm{ESA}(f_{\mathrm{pred}}, f_{\mathrm{true}}) = \tfrac{1}{3}\big(S_{\mathrm{op}} + S_{\mathrm{tok}} + S_{\mathrm{tree}}\big),$$

where S_{op} is the Jaccard overlap of operator sets (e.g. $\sin$, $+$, $\times$), S_{tok} the Jaccard overlap of lexical token sets (operators, variables, numeric constants), and S_{tree} a normalized parse-tree similarity. ESA refers only to solution expressions; PDE operator recovery is described directly in text.

Normalized neural-correction share. $\rho_{\mathrm{NN}} = \|r_\theta\|_2^2 \,/\, (\|\Phi\boldsymbol{\alpha}\|_2^2 + \|r_\theta\|_2^2)$, a cross-term-free share in $[0, 1]$ with $\rho_{\mathrm{sym}} = 1 - \rho_{\mathrm{NN}}$ holding exactly. Analytic-first behavior predicts $\rho_{\mathrm{NN}} \to 0$ on symbolic-

Table 12: **Every training schedule compared, at matched budget** ($\sigma = 0.10$; n is the seed count per arm). All arms of a benchmark share its data, noise realizations, initialization scheme, dictionary, operator provenance and coupling protocol (analytic reference operator coupled for minimal Wave and Allen–Cahn, no coupling for Burgers), and every arm except *two-stage, truncated* consumes the full per-benchmark budget (1700 passes for minimal Wave and Allen–Cahn, 3250 for Burgers). Panel B of Table 5 reports a matched-budget subset of these controls, including the full-budget two-stage arm, which gives the two-stage schedule the whole budget by extending Stage II. *Two-stage, truncated* stops after Stages I+II at 46–53% of that budget and is retained here only as a stage-ablation reference. *Joint, curriculum schedule* raises λ_{phys} with the curriculum's own stepped rule rather than a linear ramp, and *joint, matched objective* minimizes the Burgers shock-weighted loss and soft-thresholds $\boldsymbol{\alpha}$ once per epoch. These Adam-based joint arms also differ from the staged arms in the $\boldsymbol{\alpha}$ update rule. *Joint, proximal $\boldsymbol{\alpha}$* removes that difference: it trains from initialization with no freezing and no symbolic-first phase, updates θ with Adam, and updates $\boldsymbol{\alpha}$ with the staged arms' per-batch proximal step (pre-registered; Appendix E). *Post-hoc decomposition* trains a neural field first and fits the core to it afterwards (Appendix K). Welch two-sample tests of the four-stage curriculum against the full-budget two-stage arm give $p = 0.29$ (minimal Wave), $p = 0.08$ (Allen–Cahn) and $p = 0.83$ (Burgers), so at these seed counts the two schedules are not separated on reconstruction; the Allen–Cahn comparison has the largest effect size ($g = 1.2$) and is the only one that a larger study might resolve.

Arm	Hybrid RMSE	Symbolic-core RMSE	$\lvert\mathrm{supp}(\boldsymbol{\alpha})\rvert$	n
Allen–Cahn				
four-stage curriculum	0.0073±0.0018	0.086±0.014	18	5
two-stage, truncated	0.0226±0.0050	0.0681±0.0001	18	5
two-stage, full budget	0.0110±0.0031	0.0681±0.0001	18	5
joint, constant λ	0.0083±0.0012	0.419±0.076	18	5
joint, linear ramp	0.0094±0.0032	0.455±0.082	17	5
joint, curriculum schedule	0.0088±0.0026	0.477±0.056	17	5
joint, proximal α	0.0134±0.0042	2.8±1.2	18	5
post-hoc decomposition	0.0140±0.0039	0.0691±0.0009	18	5
Burgers				
four-stage curriculum	0.0076±0.0005	0.2789±0.0013	9	5
two-stage, truncated	0.0098±0.0023	0.2785±0.0011	10	5
two-stage, full budget	0.0077±0.0005	0.2785±0.0011	10	5
joint, constant λ	0.0102±0.0006	0.652±0.024	7	5
joint, matched objective	0.0138±0.0012	0.690±0.013	10	5
joint, proximal α	0.0160±0.0028	1.38±0.62	10	5
post-hoc decomposition	0.0074±0.0008	0.2771±0.0000	10	5
1D Wave (minimal)				
four-stage curriculum	0.0016±0.0010	0.0014±0.0006	2	5
two-stage, truncated	0.0023±0.0008	0.0014±0.0006	2	5
two-stage, full budget	0.0025±0.0012	0.0014±0.0006	2	5
joint, linear ramp	0.0028±0.0009	0.0146±0.0074	2	5
joint, curriculum schedule	0.0031±0.0011	0.0180±0.0082	2	5
joint, proximal α	0.0020±0.0005	0.0060±0.0041	2	5
post-hoc decomposition	0.0719±0.0042	0.028±0.012	2	5

friendly problems. Because $\|\hat{u}\|^2 = \|\Phi\boldsymbol{\alpha}\|^2 + \|r_\theta\|^2 + 2\langle\Phi\boldsymbol{\alpha}, r_\theta\rangle$, ρ_{NN} is a *branch* share and not a percentage of the total solution energy; the branches need not be orthogonal.

Branch overlap. $\gamma = \langle\Phi\boldsymbol{\alpha},\, r_\theta\rangle \,/\, (\|\Phi\boldsymbol{\alpha}\|_2\|r_\theta\|_2)$. Measured on the fields of Figure 2, $\gamma = +0.008$ on 1D Wave, $+0.026$ on Burgers, and -0.35 on Allen–Cahn, where the cross term contributes -11% of $\|\hat{u}\|^2$ because the corrector partially cancels the core's overshoot at the interfaces. Wave and Burgers branches are near-orthogonal. On Allen–Cahn the two branches interact, which is why we report γ alongside ρ_{NN} rather than treating the share as an energy percentage.

Shock-localization score. With $Q_{\mathrm{shock}} \subset Q$ a space–time band around the front, whose position at each time is located as $\arg\max_x |\partial_x u|$ of the clean field while its half-width is fixed ($w = 0.2$ on

Burgers' domain $[-1, 1]$; the gradient locates the front but does not set w),

$$\rho_{\text{shock}} = \frac{\|r_\theta\|^2_{2,Q_{\text{shock}}}}{\|r_\theta\|^2_{2,Q}},$$

so $\rho_{\text{shock}} \to 1$ means the corrector spends its energy only where the basis loses expressivity.

G Full Symbolic–Neural Decomposition Results

Figure 2 in the main text shows, for each benchmark, ground truth, symbolic core $\Phi\boldsymbol{\alpha}$, neural corrector r_θ, hybrid $\hat{u}$, and absolute error; the first four panels of each row share a symmetric color scale, and the error panel uses its own scale (deployed configurations, $\sigma = 0.10$, seed 0). Measured interpretability metrics (Appendix F):

The Wave core as printed. De-standardizing the trained coefficients of the same run gives $\hat{u}_{\text{sym}} = 0.5002\,\sin(2\pi(t{+}x)) - 0.5006\,\sin(2\pi(t{-}x)) + 0.0006$, which in product form reads $1.0008\,\sin(2\pi x)\cos(2\pi t) - 0.0004\,\sin(2\pi t)\cos(2\pi x) + 0.0006$. Evaluating the formula exactly as printed in the main text, that is $1.0008\,\sin(2\pi x)\cos(2\pi t)$ with the cross term and intercept dropped as below 10^{-3}, against the true solution on the same 128×128 grid gives RMSE 4.0×10^{-4}. Evaluating the full expression at the four-decimal coefficients printed above gives 7.5×10^{-4}, and the unrounded coefficients give 7.9×10^{-4}, matching the core-only RMSE of Table 13; the differences are rounding effects at the 10^{-4} level, and the printed single-term form is slightly better because dropping the spurious intercept removes a constant offset. The reported values use the same coefficients and evaluation grid as Figure 2.

Table 13: Decomposition metrics for the fields of Figure 2.

Benchmark	ρ_{NN}	$\|r_\theta\|_\infty$	core-only RMSE	hybrid RMSE
1D Wave (minimal)	3.1×10^{-6}	0.003	0.0008	0.0011
Burgers	0.214	0.879	0.278	0.0076
Allen–Cahn	0.021	0.210	0.078	0.0100

On Burgers, the front band of half-width 0.2 around the per-time front location (argmax of $|\partial_x u|$) covers 19.8% of the domain and holds $\rho_{\text{shock}} = 0.687$ of the corrector's energy, a $3.5\times$ enrichment over area. The corrector additionally carries a diffuse, low-amplitude component correcting the temporal evolution that the smooth symbolic backbone under-resolves. On Allen–Cahn the corrector concentrates at the moving interfaces with a 2.1% branch share.

What the Burgers core actually is, and why the front stays out of it. The complete delivered core of the Figure-2 configuration reads, with no terms omitted, as follows. The de-standardized coefficients come from a re-training of that configuration ($\sigma = 0.10$, seed 0); 10 of the 11 dictionary candidates are active, the constant candidate's coefficient is exactly zero after the ℓ_1 stage, and the core-only RMSE is 0.279 (0.278 in the plotted run, within GPU run-to-run variation).

$$\begin{aligned}\hat{u}_{\text{sym}} = \; & -0.918\,\sin(\pi x) + 0.566\,xt - 0.080\,\tanh(x{-}t) + 0.057\,\tanh(x) - 0.027\,t^2 \\ & + 0.016\,x - 0.015\,x^2 - 0.014\,t + 0.013\,e^{-t} + 0.011\,\tanh(x{-}0.5t) - 0.010.\end{aligned}$$

The trailing -0.010 is the output intercept produced by de-standardizing the standardized features. The model fits standardized columns, and mapping back to physical units contributes $-\sum_k \alpha_k \mu_k/\sigma_k$, so this value is not a dictionary term and is not counted in the $10/11$ support. The dominant term is the (decayed) initial profile $-\sin(\pi x)$ plus a weak spatio-temporal tilt xt. The remaining temporal terms (t^2, t, e^{-t}) are an order of magnitude smaller, and the traveling-front candidates receive only marginal weight. The core is not literally static, but its temporal content is far too weak to track the front. This is a faithful reflection of the dictionary's expressivity boundary, not a training accident. (*i*) The true front of this configuration is stationary at $x = 0$ by the odd symmetry of the initial condition, while the dictionary's front candidates $\tanh(x{-}t)$ and $\tanh(x{-}0.5t)$ move at speeds 1 and 0.5, the wrong trajectory. (*ii*) The only stationary front candidate, $\tanh(x)$, has $O(1)$ width against a true front width of $\sim 10^{-2}$. (*iii*) The dictionary contains no stationary-narrow-front $\times$ time-modulation product, so the amplitude decay and progressive steepening have nowhere to live in the span of Φ. The corrector consequently carries both the front and the temporal evolution, consistent with the main-text description and a concrete illustration of why the Burgers deliverable is a partially analytic representation rather than a closed form.

H COMPLETE BASELINE COMPARISONS

Fairness protocol. All baselines receive exactly the same noisy observation tensors as PS-PINN, with the same grids, noise realizations, and seeds, and are evaluated with the same metrics against the same clean fields. PySR and the oracle PINN are retrained per seed, and preprocessing is aligned, so no method trains on smoothed data. The PINN baseline is an oracle whose loss hard-codes the true PDE, while PS-PINN's operator comes from data. The prior information differs in kind rather than pointing one way. The oracle PINN holds the advantage on equation information, while PS-PINN uses physics-informed solution dictionaries, with the Wave library encoding candidate frequencies near the truth (Appendix B). **Comparison boundary.** A quantitative comparison with PINSN itself is not reported because a validated implementation was unavailable. We therefore claim no experimental advantage over PINSN, and the two-stage arms of Tables 5 and 12 are training-schedule controls on our own model, not PINSN results. Training budgets are matched per benchmark.

Table 14: Full baseline comparison at $\sigma = 0.10$ (training-domain reconstruction; mean±s.d.; PS-PINN four-stage with coupling off, n=5 on Wave and pooled n=10 on Burgers, baselines n=5 noise realizations).

Bench	Method	RMSE ↓	MAE ↓	R^2 ↑
1D Wave	PySR (symbolic)	$\mathbf{0.0011 \pm 0.0003}$	$\mathbf{0.0009 \pm 0.0002}$	**1.000**
	PINN (oracle)	0.0100 ± 0.0015	0.0081 ± 0.0014	1.000
	PS-PINN	0.0038 ± 0.0011	0.0031 ± 0.0010	1.000
Burgers	PySR (symbolic)	0.0972 ± 0.0148	0.0566 ± 0.0095	0.974
	PINN (oracle)	0.0169 ± 0.0019	0.0075 ± 0.0016	0.999
	PS-PINN	$\mathbf{0.0069 \pm 0.0004}$	$\mathbf{0.0051 \pm 0.0002}$	**1.000**

What the standard Wave configuration delivers. The Wave rows above use the standard K=14 dictionary; the corresponding four-stage evaluation uses coupling off and n=5 seeds per noise level. The corrector is negligible in magnitude: $\rho_{\mathrm{NN}} = 3.8 \times 10^{-5}$ at $\sigma = 0.10$, 1.4×10^{-5} at 0.05 and 1.1×10^{-5} at 0.01, so its share tracks the observation noise rather than any structural mismatch, and $\|r_\theta\|_\infty \leq 0.006$ throughout. The core, however, is not compact: nine of the fourteen candidates remain active in four of five seeds at $\sigma = 0.10$ (one seed selects three), and only eight of the fourteen dictionary slots agree on active/inactive status across seeds, a support agreement of 0.57. The two reasons the earlier "closed form" label was wrong are therefore independent, and only one of them concerns the corrector. The compact two-term expression of Sec. 5.2 belongs to the minimal-dictionary configuration of Figure 2, not to this one.

For the discovery-side baselines (PySINDy, DeepMoD), the recovered equations are the primary output. On noisy Burgers, PySINDy identifies an algebraic constraint rather than a valid PDE under steep front gradients, and DeepMoD recovers the general structure with redundant terms and underestimated diffusion; these are properties of the recovered equations and are not shown in the figure. Figure 4 shows the qualitative field behavior of the compared methods on the same data, with the discovery baselines' fields included for context; no quantitative solution metric is reported for them.

I NOISE ROBUSTNESS

Representative noise sweep for the two main benchmarks. The Wave rows are the four-stage curriculum on the standard K=14 dictionary with the coupling off (n=5 per cell, hybrid prediction); the Burgers rows are the four-stage curriculum with the coupling off (n=5 per cell, pooled n=10 at $\sigma = 0.10$). $\max|\hat{u}|$ is the reconstruction amplitude against the true value 1.0 (1D Wave only; Burgers has no analytic amplitude target).

Error grows smoothly with σ and R^2 stays at 1.000 throughout; no noise level is cherry-picked. The full σ grids for all five benchmarks use fixed, recorded per-cell seed lists.

J REPRESENTATION ABLATIONS

Strict same-protocol arms. All arms share identical architecture, data, noise realizations, seeds, curriculum stages, and budgets. A branch is disabled by multiplying its output by zero, so parameter

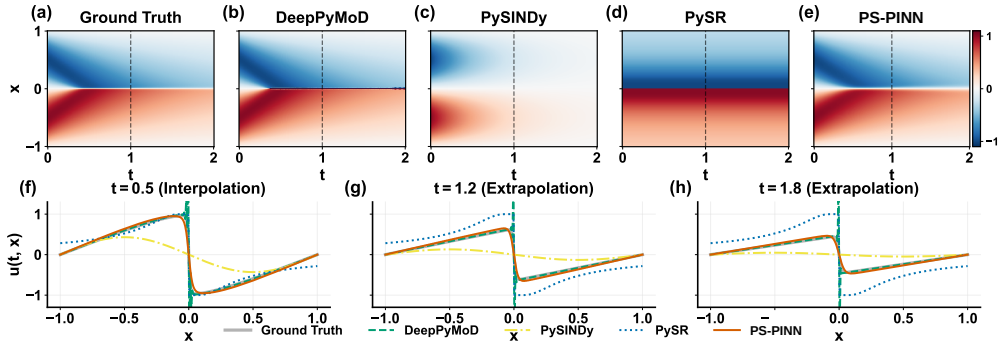


Figure 4: **Reconstruction paradigms on noisy Burgers data. (Top)** Fields; the dashed line at $t = 1.0$ separates training from extrapolation, where PySINDy (c) and PySR (d) turn non-physical. **(Bottom)** Snapshots: all methods interpolate (f); in extrapolation (g–h) the PySR closed form fails to propagate the front while PS-PINN tracks the ground truth. Single-seed qualitative protocol.

Table 15: PS-PINN across noise levels (mean±s.d.).

Bench / σ	RMSE ↓	R^2 ↑	max $\lvert\hat{u}\rvert$ ($\to 1$)
Wave, σ=0.01	0.0008 ± 0.0002	1.000 ± 0.000	1.000 ± 0.001
Wave, σ=0.05	0.0031 ± 0.0013	1.000 ± 0.000	1.002 ± 0.004
Wave, σ=0.10	0.0038 ± 0.0011	1.000 ± 0.000	1.002 ± 0.003
Burgers, σ=0	0.0011 ± 0.0001	1.000 ± 0.000	—
Burgers, σ=0.05	0.0039 ± 0.0002	1.000 ± 0.000	—
Burgers, σ=0.10	0.0069 ± 0.0004	1.000 ± 0.000	—

counts and the optimizer's parameter list are unchanged. Each arm is reinitialized with the same seed and trained independently from scratch, with the disabled branch gated throughout training rather than removed only at evaluation. The arms inherit the curriculum's stage schedule, so a branch trains only in the stages that update it. On Wave the network participates in 1100 of the 1700 nominal passes (Stages II–IV), while Stage I and the alternating α-halves update the core. The comparison therefore answers what removing a branch does under this curriculum, not what a stand-alone pure network could achieve under a training protocol designed for it. $n = 5$ seeds.

Table 16: Strict same-protocol component ablation ($\sigma = 0.10$, mean±s.d., n=5). Configurations: Wave = deployed minimal dictionary (K=2, analytic reference operator coupled); Burgers = deployed (K=11, no coupling). External baselines (PySR, oracle PINN) are *not* listed here: they differ in architecture, objectives, and tuning, and appear as baselines in Table 4. The Burgers hybrid value (0.0077) is from an independent re-training; its 0.0001 difference from the four-stage arm of Table 5 (0.0076) is within run-to-run variation.

Arm	Wave RMSE	Burgers RMSE	Wave ρ_{NN}	Burgers ρ_{NN}
Symbolic only (r_θ gated to 0)	$\mathbf{0.0014 \pm 0.0006}$	0.2924 ± 0.0003	0	0
Neural only (Φ gated to 0)	0.0773 ± 0.0071	$\mathbf{0.0077 \pm 0.0002}$	1	1
Hybrid	0.0016 ± 0.0010	0.0077 ± 0.0005	$\sim 10^{-5}$	0.22

Each regime has one pure representation that fails by $\sim 40\times$ (neural-only on Wave, symbolic-only on Burgers), and the hybrid matches the surviving specialist in both. On Wave this is because the hybrid effectively is its symbolic core ($r_\theta \to 0$); on Burgers it reaches parity with the same-protocol network while additionally delivering the analytic backbone and the measured decomposition. For reference, the separate constant-weight joint-training control of Table 12, which is not part of this same-protocol ablation, reaches a comparable Burgers total (0.0102 ± 0.0006) but collapses the symbolic core (0.652 against 0.279), so in these controls the interpretable outcome depends on symbolic-first training as well as on the architecture, while accuracy changes far less.

K Post-hoc Decomposition Baseline

The curriculum forms the decomposition *during* training. The most direct alternative forms it afterwards, and it also keeps a neural corrector, so the presence of a fallback branch cannot by itself distinguish the two. Concretely (n=5 seeds, $\sigma = 0.10$):

1. train a neural-only field f on the same noisy observations, with the same architecture, data loss and operator provenance as the deployed pipeline ($\boldsymbol{\alpha}$ gated to zero throughout);
2. fit a sparse core $s = \Phi\boldsymbol{\alpha}$ to f over the *same* dictionary, with the same proximal soft-threshold and the same λ_{sym};
3. deliver $s + (f - s)$.

The delivered total is then exactly f, and the corrector $r = f - s$ is whatever the dictionary cannot express. Compute is matched to the curriculum rather than merely bounded: the field receives the curriculum's total passes minus its Stage-I budget, and the core fit receives exactly the Stage-I budget, so the arm consumes 1700 passes on minimal Wave and Allen–Cahn and 3250 on Burgers, the same as every other full-budget arm of Table 12. The core is fitted to the trained field rather than to the raw observations, which is the stronger version of the baseline: its target is already denoised, so the comparison does not hand the curriculum an advantage that comes only from noise.

What the comparison reports. Core RMSE against the clean field, total RMSE, the active-term count $|\mathrm{supp}(\boldsymbol{\alpha})|$ under the reporting threshold of Appendix E, the corrector share ρ_{NN}, and cross-seed *support agreement*: the fraction of dictionary slots whose active/inactive status is shared by all five seeds, which measures whether the two procedures select a stable structure rather than a differently-wrong one per seed.

Result. All comparisons below are Welch two-sample tests on five seeds per arm, and every curriculum reference arm is at n=5. Because Stages III–IV are an optional refinement (Sec. 4.4), the baseline is compared against both operating points: interpretability-first (I+II, the full-budget two-stage arm) and accuracy-refined (I–IV).

The post-hoc total is exactly the pretrained field by construction, so decomposition cannot improve its total reconstruction error. The fitted core, however, is not bounded by the teacher's error against ground truth: projection onto the analytic dictionary may either improve or worsen core accuracy. On the minimal-dictionary Wave configuration the neural-only field reaches only 0.0719 ± 0.0042 (consistent with the neural-only ablation of Appendix J, 0.077 ± 0.007). Projecting it onto the dictionary improves the core to 0.0284 ± 0.0122, yet that core remains a factor of 20 worse than the symbolic-first core, 0.0014 ± 0.0006 at either operating point ($p = 0.01$), and the delivered total, being the teacher itself, is worse by factors of 44 and 29 against the accuracy-refined and interpretability-first points ($p < 10^{-4}$ for both).

On Allen–Cahn and Burgers, where the neural-only field is accurate, no core comparison is significant at either operating point. On Allen–Cahn the post-hoc core, 0.0691 ± 0.0009, lies between the interpretability-first core, 0.0681 ± 0.0001 ($p = 0.08$), and the accuracy-refined core, 0.0860 ± 0.0137 ($p = 0.07$); its numerical margin over the latter is the Stage III–IV core degradation of Sec. 5.4. On Burgers the post-hoc core, 0.2771 ± 0.0000, is numerically slightly better than both, 0.2785 ± 0.0011 and 0.2789 ± 0.0013 ($p = 0.07$ and $p = 0.06$). Total reconstruction separates differently: on Allen–Cahn the accuracy-refined total is significantly better than the post-hoc total (0.0073 ± 0.0018 vs. 0.0140 ± 0.0039, $p = 0.02$), whereas the interpretability-first total is not (0.0110 ± 0.0031, $p = 0.27$); on Burgers the totals are indistinguishable at both points ($p = 0.66$ and $p = 0.57$).

Across the twelve comparisons (three benchmarks, two metrics, two operating points) the curriculum is significantly better in five and never significantly worse. By metric, the core is significantly better only on minimal Wave, at both operating points, and no core difference is significant on Allen–Cahn or Burgers. The total is significantly better on minimal Wave at both points and on Allen–Cahn at the accuracy-refined point only, and indistinguishable on Burgers.

Complexity, stability, and what is actually delivered. The two routes are indistinguishable on the other two axes the comparison fixes. Active-term counts match (2.0 vs. 2.0 on minimal Wave, 18.0 vs. 18.0 on Allen–Cahn, 9.6 vs. 9.8 on Burgers), and so does cross-seed support agreement (1.00 on the first two benchmarks for both; 0.82 vs. 0.91 on Burgers). Neither procedure is the more

parsimonious or the more structurally stable one. The difference that remains on minimal Wave is in the delivered object itself: the curriculum drives the corrector to $\rho_{\mathrm{NN}} \approx 3 \times 10^{-6}$, so what it returns is a closed-form expression, whereas the post-hoc route leaves a branch share $\rho_{\mathrm{NN}} = 0.018$, more than three orders of magnitude larger, and therefore returns a core plus a non-negligible corrector. On the benchmark where a closed form exists, only one of the two procedures actually delivers it.

This is a structural difference rather than a tuning artifact. The post-hoc core is a projection of the teacher field, so the part of the teacher's error that lies in the dictionary's span enters the core, and the total inherits the teacher's error exactly; a core fitted directly to the observations in Stage I depends on no teacher. The comparison also locates where the difference lives. The regime in which forming the decomposition during training pays is the one in which the dictionary is adequate and the analytic core can carry the solution – the configuration whose delivered object is a genuine closed form ($\rho_{\mathrm{NN}} \approx 3 \times 10^{-6}$, Sec. 5.2). Where the dictionary cannot express the field at all, as on Burgers, both routes return the same inadequate core. We therefore do not claim that training-time decomposition dominates a post-hoc fit everywhere: its core advantage is significant where the analytic core can carry the solution, the cores are statistically indistinguishable elsewhere, and outside minimal Wave only the accuracy-refined total separates from the post-hoc total (Allen–Cahn).

Two readings are excluded by this comparison and one is not. It is not enough to say that PS-PINN retains a neural fallback, since this baseline does too; and it is not enough to compare total reconstruction error, since the total here is f by construction. What the comparison does isolate is the quality of the analytic core each procedure leaves behind, at equal dictionary, equal sparsity rule and equal compute.

L Solution-Library Sensitivity

The Wave results benefit from a dictionary whose wavenumbers sit near the true mode, and this appendix quantifies what happens as that prior degrades. The protocol was registered before any run, all runs use the deployed 1D Wave configuration at $\sigma = 0.10$ with $n = 3$ seeds.

L.1 Frequency mismatch. The two-element traveling-wave dictionary is rebuilt at $k = k^\star(1+\delta)$ with $\delta \in \{0, 1, 2, 5, 10\}\%$, so for $\delta > 0$ no candidate sits at the true mode. The analytic-first hypothesis predicts graceful degradation. As the library worsens, r_θ should gradually absorb the unexplained share while hybrid accuracy degrades slowly, rather than the method failing outright the moment the basis stops being exact. Figure 5 and Table 17 confirm this. ρ_{NN} rises continuously by over four orders of magnitude ($7 \times 10^{-6} \to 0.29$), the symbolic-only error grows steeply ($0.0015 \to 0.249$, approaching the field amplitude), while the hybrid degrades an order of magnitude more slowly ($0.0019 \to 0.028$, which is $9\times$ better than the core alone at $\delta = 10\%$). The representation hands the unexplained share to the corrector instead of failing.

L.2 Library size. The dictionary is grown with the four-form family $\{\sin(kx)\cos(kt), \sin(kt)\cos(kx), \sin(k(t\pm x))\}$ instantiated at $1/2/4/7$ wavenumbers around the true mode ($K \in \{4, 8, 16, 28\}$; true mode always included, distractors within $\pm 10\%$). Hybrid accuracy is robust to dictionary growth (0.0019 at $K{=}4$ to 0.0037–0.0042 at $K{\geq}16$) and the corrector share stays small throughout ($\rho_{\mathrm{NN}} \leq 1.2 \times 10^{-3}$). One caveat is verified rather than assumed: the delivered expression is not compact under heavy redundancy, at either level. At the coefficient level the support does not stay minimal, with $|\mathrm{supp}(\boldsymbol{\alpha})| \approx K$. Equivalence reduction does not rescue it either. The within-frequency standing/traveling redundancy is an exact identity, $\sin(kx)\cos(kt) \equiv \frac{1}{2}[\sin(k(x{+}t)) + \sin(k(x{-}t))]$, and collapses 28 terms to 14 traveling-wave coefficients, but the cross-frequency directions are near-degenerate rather than identical, and after reduction the energy remains spread over all seven candidate wavenumbers. At $K{=}28$, seed 0, the true mode holds only 24% of the reduced-coefficient energy and its $\pm 2\%$ neighbours hold 21–22% each. Nearby sinusoids are strongly correlated on the bounded domain, so the ℓ_1 pressure cannot isolate a single frequency. Compactness of the delivered expression is therefore not an automatic consequence of sparsity. Near-degenerate dictionaries impede compact selection, and among the deployed configurations only the minimal Wave dictionaries yielded compact cores, with two active terms throughout L.1; the deployed Burgers and noisy Allen–Cahn cores keep most candidates active ($10/11$ and $18/19$; Appendices G, C). Good conditioning alone does not guarantee that the target admits a sparse representation. We do not analyze the degeneracy structure itself here.

Table 17: Library sensitivity on 1D Wave ($\sigma = 0.10$, mean±s.d., n=3).

	Config	Hybrid RMSE	Symbolic RMSE	ρ_{NN}	$\lvert\mathrm{supp}(\boldsymbol{\alpha})\rvert$
L.1 (δ)	0%	0.0019 ± 0.0011	0.0015 ± 0.0007	7.3×10^{-6}	2.0
	1%	0.0050 ± 0.0014	0.0282 ± 0.0005	3.2×10^{-3}	2.0
	2%	0.0067 ± 0.0003	0.0554 ± 0.0004	1.3×10^{-2}	2.0
	5%	0.0160 ± 0.0008	0.1331 ± 0.0012	7.8×10^{-2}	2.0
	10%	0.0280 ± 0.0029	0.2489 ± 0.0005	2.9×10^{-1}	2.0
L.2 (K)	4	0.0019 ± 0.0012	0.0015 ± 0.0007	8.5×10^{-6}	3.3
	8	0.0034 ± 0.0005	0.0177 ± 0.0019	1.2×10^{-3}	7.3
	16	0.0042 ± 0.0004	0.0080 ± 0.0021	2.0×10^{-4}	15.3
	28	0.0037 ± 0.0005	0.0062 ± 0.0006	1.3×10^{-4}	27.3

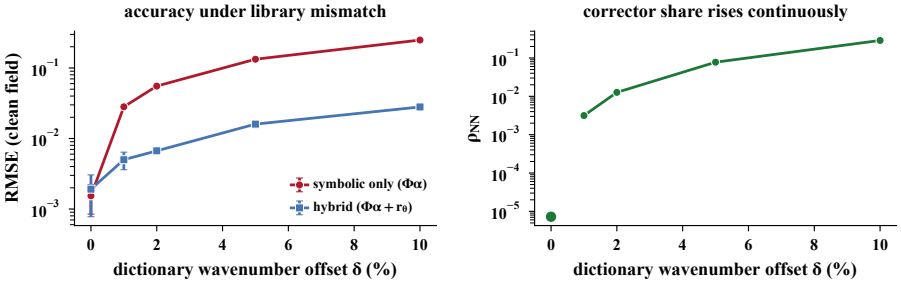


Figure 5: **Graceful degradation under library mismatch** (L.1). Left: as the dictionary wavenumber offset δ grows, the symbolic core alone degrades steeply while the hybrid degrades an order of magnitude more slowly. Right: the corrector share ρ_{NN} rises continuously by over four orders of magnitude, so the unexplained share migrates to the corrector instead of breaking the method.

M Structural Extrapolation Details

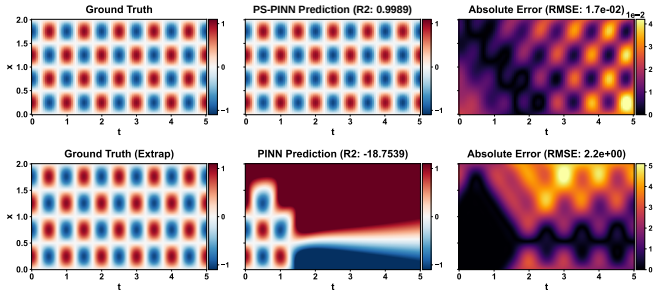


Figure 6: **Structural extrapolation on the Wave equation**, trained on $[0,1]^2$ and evaluated on $[0,5] \times [0,2]$. Top: PS-PINN, whose selected structure generalizes. Bottom: a standard PINN collapses outside the training box. Archived single-seed fields from the original pipeline, retained as an illustration; the quantitative claims are those of Table 6, which repeats the comparison at n=5 against an oracle PINN.

Wave. Training uses the noisy box $\mathcal{D}_{\mathrm{train}} = [0,1]^2$ and evaluation runs on $t \in [0,5]$, $x \in [0,2]$: a $5\times$ temporal horizon and $2\times$ spatial range, against the analytic solution $\sin(2\pi x)\cos(2\pi t)$. All arms share data, noise realizations, initialization scheme, budget and the coupled analytic reference operator; the only difference is the symbolic branch, which the *oracle PINN* arm gates to zero exactly as in the component ablation (Appendix J). That arm is therefore handed the true operator, making it a stronger baseline than an ordinary data-fit network. We use n=5 seeds at $\sigma = 0.10$.

Three things follow. First, the failure of an unconstrained network outside its training box is not marginal and not a single-seed artifact: the oracle PINN fits the training box well ($R^2 = 0.975$) and then collapses on the extended box in every seed, from -0.68 to -20.5. Second, the selected analytic structure transfers essentially exactly, and does so under the standard K=14 dictionary as well as the minimal one, even though the standard dictionary contains three near-identical wavenumbers whose small frequency errors have a $5\times$ horizon over which to compound. Third, it is specifically the

core that transfers: the $K{=}14$ hybrid loses accuracy outside the box (0.9929) while its core does not (0.99998), because the neural corrector is fitted only where data exist. The transferable object the paper claims to deliver is therefore the analytic one, and this is a property of the representation rather than of the training-box fit.

As stated in the main text, the dictionary encodes a periodicity prior near the true wavenumber, so this measures whether a *selected* structure extrapolates, not whether structure is discovered without prior.

The quantitative Wave extrapolation study above has been rerun with $n{=}5$; Figure 6 retains the original single-seed fields only as an archived illustration. The Burgers generalization study below remains a single-seed legacy result and is used only qualitatively.

Burgers generalization (single-seed original-pipeline protocol): "Discovered PDE" numerically integrates the recovered equation; "closed-form fit" is the PySR expression.

Table 18: Generalization on Burgers (single-seed protocol).

Task	Method	RMSE ↓	MAE ↓	R^2 ↑
Interpolation, $t \in [0, 1]$	Discovered PDE	0.0519	0.0147	0.9927
	Closed-form fit	0.2343	0.1802	0.8511
	PS-PINN	**0.0372**	**0.0135**	**0.9962**
Time extrap., $t \in [1, 2]$	Discovered PDE	**0.0599**	**0.0205**	**0.9646**
	Closed-form fit	0.3437	0.3207	-0.1651
	PS-PINN	0.1329	0.0646	0.8257
Space extrap., $x \in [-2, 2]$	Discovered PDE	**0.0357**	**0.0107**	**0.9965**
	Closed-form fit	0.5705	0.4528	0.1030
	PS-PINN	0.3166	0.1661	0.7238

The pattern mirrors the mechanism/realization distinction of Sec. 1. In interpolation the hybrid realization is best. Outside the data window the equation, integrated numerically, dominates, and the pure closed-form fit turns negative. The law carries the transferable physics; the fitted surface does not.

N Additional PDE Benchmarks

The purpose of these benchmarks is to show the representation is not tuned to two systems. Each reports configuration (Appendix A), reconstruction, and the delivered decomposition, and none carries any operator-reliability narrative here. All three rows use the four-stage curriculum with $n{=}5$ seeds. **1D Heat.** RMSE 0.0034 ± 0.0010 at $\sigma = 0.10$. The core selects the correct separable mode $\sin(2\pi x)e^{-\alpha(2\pi)^2 t}$ from the nuisance-padded dictionary. **2D Wave.** RMSE 0.0035 ± 0.0006 at $\sigma = 0.10$ with the product-mode dictionary. **Allen–Cahn.** RMSE 0.0402 ± 0.0064 at $\sigma = 0.10$ with physics off, improving to 0.0075 ± 0.0016 with the analytic reference operator coupled. This is the one benchmark where the optional operator coupling of Equation (6) is decisive ($5\times$), included to document that the coupling exists and helps when the operator merits enforcement. The core is a combination of $\tanh$-front profiles, and the corrector concentrates at the interfaces.

O Conditional Optimization Analysis

O.1 Probabilistic Derivation of the Objective

We first justify the data and sparsity terms of Equation (6) as MAP estimation under stated assumptions. Treating the recovery of the symbolic coefficients $\boldsymbol{\alpha}$ and the neural parameters θ as probabilistic inference with independent priors, Bayes' theorem gives the posterior

$$P(\boldsymbol{\alpha}, \theta \mid y) \propto P(y \mid \boldsymbol{\alpha}, \theta)\, P(\boldsymbol{\alpha})\, P(\theta), \tag{7}$$

and the MAP estimate minimizes the negative log-posterior

$$\mathcal{J}_{\text{MAP}} = -\log P(y \mid \boldsymbol{\alpha}, \theta) - \log P(\boldsymbol{\alpha}) - \log P(\theta). \tag{8}$$

Gaussian likelihood → data fidelity. With independent additive Gaussian noise $\epsilon \sim \mathcal{N}(0, \sigma^2)$,

$$P(y \mid \boldsymbol{\alpha}, \theta) = \prod_{i=1}^{N} \frac{1}{\sqrt{2\pi}\sigma} \exp\Big(- \frac{(y_i - \hat{u}_i)^2}{2\sigma^2} \Big) \implies -\log P(y \mid \boldsymbol{\alpha}, \theta) \propto \|y - \hat{u}_{\boldsymbol{\alpha},\theta}\|_2^2, \quad (9)$$

so minimizing the mean squared error is equivalent to maximizing the likelihood under Gaussian noise (Bishop, 2006).

Laplacian prior → sparsity. Imposing $P(\boldsymbol{\alpha}) = \prod_{j=1}^{K} \frac{\lambda}{2} \exp(-\lambda|\alpha_j|)$ gives $-\log P(\boldsymbol{\alpha}) \propto \lambda\|\boldsymbol{\alpha}\|_1$. As established in the statistical literature (Park & Casella, 2008; Tibshirani, 1996), ℓ_1-penalized regression is MAP estimation with a Laplacian prior; unlike a Gaussian prior (which yields ℓ_2 ridge regularization), the Laplacian density's non-differentiable peak at zero drives small coefficients exactly to zero, promoting exact sparsity.

Assembly. Substituting into Equation (8) under a flat (improper) prior on θ, and adding the operator-coupling term as a regularizer (an informal prior on physically valid functions, not part of the probabilistic model), recovers Equation (6). Thus the data and sparsity terms admit a Bayesian MAP interpretation; the physics term does not.

O.2 Optimization Analysis for a Fixed, Known Linear Operator

Remark (scope of the assumptions). This subsection states a single optimization result, for a *fixed, known, continuous linear* operator $\mathcal{N}$ and full-rank Φ. The operator condition holds for 2D Wave, 1D Heat, and minimal-library 1D Wave, but not for Burgers (uu_x) or Allen–Cahn ($u - u^3$), for which the analysis is heuristic guidance only. Even on the linear benchmarks the result does not cover the implemented dictionaries as they stand: the 1D Heat and 2D Wave dictionaries contain a constant candidate, which mean-standardization turns into an all-zero column and so violates the linear independence required in (A1). That column receives zero gradient, so its coefficient stays at zero from the zero initialization, and the result applies to the dictionary with it removed. The exact block solves of (A2) and the asymptotic regularity of (A3) are further not verified for the finite-budget implementation (Appendix E). All function spaces and norms below are taken on the space–time domain $Q = \Omega \times (0, T]$ of Sec. 3, since $\mathcal{N}$ involves time derivatives. The analysis is stated for $\Phi\alpha + r_\theta$ without the trainable offset b_0 of Sec. 4.2: for a fixed b_0 and an operator that annihilates constants, as the Wave and Heat operators do, the statement is unchanged after replacing y by $y - b_0$, whereas the trainable, unpenalized b_0 of the implementation is not covered. All conditions used in the proof appear explicitly in the theorem statement. Below, λ denotes the sparsity weight (λ_{sym} of Equation (6)).

Theorem 1 (Monotone descent and block-stationary accumulation points)**.** *Let*

$$\mathcal{J}(\alpha, \theta) = \|y - (\Phi\alpha + r_\theta)\|_{L^2(Q)}^2 + \lambda\|\alpha\|_1 + \lambda_{\mathrm{phys}}\|\mathcal{N}[\Phi\alpha + r_\theta]\|_{L^2(Q)}^2,$$

and assume:

- *(A1)* $\Phi : Q \to \mathbb{R}^K$ *is a feature map with components in* $H^2(Q)$*, linearly independent in* $L^2(Q)$*; the parameter* θ *ranges over a* compact *set* $\Theta \subset \mathbb{R}^p$ *(e.g. norm-bounded weights), and the map* $\theta \mapsto r_\theta$ *is* continuously differentiable as an $H^2(Q)$-valued map *on an open neighborhood of* Θ*, meaning* $\theta \mapsto \nabla_\theta r_\theta$ *is continuous into* $H^2(Q)^p$*, which justifies differentiating the quadratic terms of* $\mathcal{J}$ *(including the PDE-residual term) under the integral and makes* $\theta \mapsto \nabla_\theta \mathcal{J}$ *continuous on* Θ*, as used in Step 5 (hence* $\{r_\theta : \theta \in \Theta\}$ *is bounded in* $H^2(Q)$*);* $\mathcal{N} : H^2(Q) \to L^2(Q)$ *is a* fixed *continuous linear differential operator, independent of the training sample.*
- *(A2)* (Block solves.) *At every outer step* t*, the convex* α*-subproblem* $\min_\alpha \mathcal{J}(\alpha, \theta^{(t)})$ *is solved exactly (or by proximal gradient run to its unique minimizer), and the* θ*-update does not increase its subproblem objective and returns a point whose* ambient-space *gradient satisfies* $\|\nabla_\theta \mathcal{J}(\alpha, \cdot)\| \le \epsilon_t$ *with* $\epsilon_t \to 0$*. For limit points on the boundary of* Θ *this ambient condition is stronger than tangent-cone (KKT) stationarity, and we assume it in this ambient form. No claim is made that generic gradient descent finds such points; their availability is assumed.*
- *(A3)* (Asymptotic regularity.) $\|(\alpha^{(t+1)}, \theta^{(t+1)}) - (\alpha^{(t)}, \theta^{(t)})\| \to 0$ *as* $t \to \infty$.

Then the energy values $\{\mathcal{J}(\alpha^{(t)}, \theta^{(t)})\}$ are monotonically non-increasing and converge to a finite limit $\mathcal{J}^* \geq 0$, and $\{(\alpha^{(t)}, \theta^{(t)})\}$ is bounded in $\mathbb{R}^K \times \Theta$. Hence there is a subsequence with $(\alpha^{(t_k)}, \theta^{(t_k)}) \to (\alpha^*, \theta^*)$ in the parameter space*; every such limit point is a block-coordinate stationary point of $\mathcal{J}$, i.e.*

$$0 \in \partial_\alpha \mathcal{J}(\alpha^*, \theta^*), \qquad \nabla_\theta \mathcal{J}(\alpha^*, \theta^*) = 0$$

(the θ-gradient vanishing in the ambient sense assumed in (A2), which implies stationarity along every feasible direction, including at boundary points of Θ), and, by continuity of $\theta \mapsto r_\theta$ and of $\mathcal{N}$, the fields and the PDE residual converge strongly *along the same subsequence:*

$$r_{\theta^{(t_k)}} \to r_{\theta^*} \text{ in } H^2(Q), \qquad \mathcal{N}[\Phi\alpha^{(t_k)} + r_{\theta^{(t_k)}}] \to \mathcal{N}[\Phi\alpha^* + r_{\theta^*}] \quad \text{in } L^2(Q).$$

The theorem does not *assert that (α^*, θ^*) is a global minimizer of the nonconvex $\mathcal{J}$, and it says nothing about the discovery step that produces $\mathcal{N}$ in the full pipeline.*

Proof. **Step 1 (nonnegativity, coercivity, continuity).** Since $\|y - (\Phi\alpha + r_\theta)\|^2_{L^2(Q)} \geq 0$, $\|\alpha\|_1 \geq 0$, and $\|\mathcal{N}[\Phi\alpha + r_\theta]\|^2_{L^2(Q)} \geq 0$, we have $\mathcal{J}(\alpha, \theta) \geq 0$. The ℓ_1 term gives coercivity in α; θ ranges over the compact Θ by (A1). Since $\theta \mapsto r_\theta$ is continuous into $H^2(Q)$ and $\mathcal{N}$ is continuous linear, $(\alpha, \theta) \mapsto \mathcal{J}(\alpha, \theta)$ is continuous on $\mathbb{R}^K \times \Theta$ (cf. (Ekeland & Temam, 1999) for the lower-semicontinuity of the quadratic terms).

Step 2 (α-block subproblem). For fixed $\theta^{(t)}$, define

$$\mathcal{J}_\alpha(\alpha) = \|\Phi\alpha - (y - r_{\theta^{(t)}})\|^2_{L^2(Q)} + \lambda_{\text{phys}}\|\mathcal{N}[\Phi\alpha + r_{\theta^{(t)}}]\|^2_{L^2(Q)} + \lambda\|\alpha\|_1 = \mathcal{J}(\alpha, \theta^{(t)}).$$

This functional is convex, proper, and lower semicontinuous. It admits a unique minimizer $\alpha^{(t+1)}$ characterized by the subgradient condition

$$0 \in 2\Phi^\top\big(\Phi\alpha^{(t+1)} - (y - r_{\theta^{(t)}})\big) + 2\lambda_{\text{phys}}(\mathcal{N}\Phi)^\top\big(\mathcal{N}\Phi\,\alpha^{(t+1)} + \mathcal{N}[r_{\theta^{(t)}}]\big) + \lambda\,\partial\|\alpha^{(t+1)}\|_1,$$

where $\mathcal{N}\Phi$ applies the (linear) operator $\mathcal{N}$ to each basis function. If the subproblem is solved by proximal gradient (Beck & Teboulle, 2009) with step size $\eta < 1/L$, $L = 2\|\Phi^\top\Phi + \lambda_{\text{phys}}(\mathcal{N}\Phi)^\top(\mathcal{N}\Phi)\|_2$, the inner iterates converge to this unique minimizer (the subproblem is convex); otherwise, by (A2), the subproblem is solved exactly. In either case,

$$\mathcal{J}(\alpha^{(t+1)}, \theta^{(t)}) \leq \mathcal{J}(\alpha^{(t)}, \theta^{(t)}).$$

Step 3 (θ-block subproblem). For fixed $\alpha^{(t+1)}$, consider

$$\mathcal{J}_\theta(\theta) = \|y - (\Phi\alpha^{(t+1)} + r_\theta)\|^2_{L^2(Q)} + \lambda_{\text{phys}}\|\mathcal{N}[\Phi\alpha^{(t+1)} + r_\theta]\|^2_{L^2(Q)}.$$

By (A1), r_θ is continuously differentiable in θ, so $\mathcal{J}_\theta$ is continuously differentiable. By (A2), the block update returns $\theta^{(t+1)}$ with $\mathcal{J}_\theta(\theta^{(t+1)}) \leq \mathcal{J}_\theta(\theta^{(t)})$ and $\|\nabla_\theta \mathcal{J}_\theta(\theta^{(t+1)})\| \leq \epsilon_t$, $\epsilon_t \to 0$, the gradient taken in the ambient space. We emphasize that (A2) is an assumption on the block solver, not a consequence of running plain gradient descent on the nonconvex θ-subproblem; the availability of such near-stationary points is assumed, not derived. Consequently,

$$\mathcal{J}(\alpha^{(t+1)}, \theta^{(t+1)}) \leq \mathcal{J}(\alpha^{(t+1)}, \theta^{(t)}).$$

Step 4 (global monotonicity). Combining Steps 2–3,

$$\mathcal{J}(\alpha^{(t+1)}, \theta^{(t+1)}) \leq \mathcal{J}(\alpha^{(t+1)}, \theta^{(t)}) \leq \mathcal{J}(\alpha^{(t)}, \theta^{(t)}),$$

so $\{\mathcal{J}^{(t)}\}$ is monotonically non-increasing and bounded below, implying the finite limit $\mathcal{J}^* = \lim_{t\to\infty} \mathcal{J}(\alpha^{(t)}, \theta^{(t)})$.

Step 5 (block-stationarity of parameter-space limit points). The α-iterates are bounded by coercivity (Step 1) and $\theta^{(t)} \in \Theta$ with Θ compact, so by Bolzano–Weierstrass there is a subsequence with $(\alpha^{(t_k)}, \theta^{(t_k)}) \to (\alpha^*, \theta^*)$ *in* $\mathbb{R}^K \times \Theta$. We emphasize that the convergence used here is convergence of the *parameters*, not of the fields: stationarity is a statement about $\nabla_\theta \mathcal{J}$, and weak convergence

of $r_{\theta^{(t_k)}}$ alone would not justify passing to the limit in it. With parameter convergence in hand, continuity of $\mathcal{J}$ (Step 1) gives $\mathcal{J}(\alpha^*, \theta^*) = \mathcal{J}^*$, and passing to the limit in the first-order conditions of Steps 2 and 3 (upper-hemicontinuity of $\partial\|\cdot\|_1$, continuity of $\theta \mapsto \nabla_\theta \mathcal{J}$ on the compact Θ, the tolerance $\epsilon_t \to 0$ of (A2), and the asymptotic regularity (A3)) yields

$$0 \in \partial_\alpha \mathcal{J}(\alpha^*, \theta^*), \qquad \nabla_\theta \mathcal{J}(\alpha^*, \theta^*) = 0,$$

i.e., (α^*, θ^*) is a block-coordinate stationary point of $\mathcal{J}$, the θ-gradient vanishing in the ambient sense of (A2). We do *not* claim global minimality: the limit value $\mathcal{J}^*$ is the value attained along the alternating iterates and is not in general the global infimum of the nonconvex $\mathcal{J}$.

Step 6 (strong convergence of the fields and the PDE residual). By (A1), $\theta^{(t_k)} \to \theta^*$ implies $r_{\theta^{(t_k)}} \to r_{\theta^*}$ strongly in $H^2(Q)$, and $\alpha^{(t_k)} \to \alpha^*$ gives $\Phi\alpha^{(t_k)} \to \Phi\alpha^*$ in $H^2(Q)$. Continuity of $\mathcal{N} : H^2(Q) \to L^2(Q)$ then yields

$$\mathcal{N}[\Phi\alpha^{(t_k)} + r_{\theta^{(t_k)}}] \to \mathcal{N}[\Phi\alpha^* + r_{\theta^*}] \quad \text{strongly in } L^2(Q).$$

Field convergence is thus a *consequence* of parameter convergence, not an ingredient of the stationarity argument.

Conclusion. The energy sequence is monotonically non-increasing with finite limit $\mathcal{J}^*$; every parameter-space limit point of the iterates (which exist by compactness) is block-coordinate stationary, and the fields and PDE residual converge strongly along the corresponding subsequences. The result is conditional on the stated assumptions on $\Phi, \Theta, \mathcal{N}$ and does not constitute a global-convergence guarantee for the full nonconvex training procedure. □

Remark (what we do not claim). Reallocation invariance of the prediction terms (why the objective does not identify the allocation). The data and physics terms of the objective depend on $(\boldsymbol{\alpha}, \theta)$ only through the sum $\hat{u} = \Phi\boldsymbol{\alpha} + r_\theta$, while the ℓ_1 penalty acts on $\boldsymbol{\alpha}$ alone. Hence if at the current iterate there exists θ' with $r_{\theta'} = r_\theta + \Phi\boldsymbol{\alpha}$ (a *representable reallocation*), the prediction terms are unchanged and

$$\mathcal{J}(0, \theta') = \mathcal{J}(\boldsymbol{\alpha}, \theta) - \lambda_{\text{sym}}\|\boldsymbol{\alpha}\|_1 \;\le\; \mathcal{J}(\boldsymbol{\alpha}, \theta) :$$

with the decrease strict whenever $\lambda_{\text{sym}} > 0$ and $\boldsymbol{\alpha} \neq 0$. Along representable reallocation directions the objective therefore does not merely fail to reward an analytic-first allocation but prefers emptying the symbolic core. This is a conditional statement, since an over-parameterized MLP can typically approximate such a θ' though exact representability is not guaranteed, and its scope must be stated precisely. The identity compares objective values. It does not establish that the reallocated point $(\mathbf{0}, \theta')$ is itself stationary, and we make no such claim. What it does establish is that the stationarity result above provides no guarantee that the symbolic core is retained. Along representable reallocations that preserve the total prediction, the objective does not favor retaining the symbolic component, so the finite-budget allocation actually reached is decided by the optimization path (initialization, freezing, staging) and must be evaluated empirically. This is the formal sense in which prediction accuracy does not identify the division of explanatory labor. It is consistent with, but does not predict, what the budget-matched controls of the main text observe, where joint training matches hybrid RMSE while collapsing the core.

We make no statistical claims about the estimator. In particular, we do not claim coefficient consistency or support recovery for $\boldsymbol{\alpha}$. Such results would additionally require a minimum-signal (beta-min) condition, a non-circular treatment of the physics term, and control of the coupling between the noise and the data-dependent residual $r_{\hat{\theta}}$, none of which we verify; Stage-I support recovery holds only under classical incoherence conditions on the dictionary (Tibshirani, 1996; Wainwright, 2009). We do not claim that the PDE residual vanishes asymptotically. A finite physics weight does not by itself guarantee a vanishing residual, since the implemented schedule raises the weight only to $\lambda_{\text{phys}} = 1$, though the residual can still vanish when data and physics admit a common minimizer. We do not claim a generalization bound for the full pipeline. Stability-based bounds require the enforced operator to be independent of the training sample, whereas a discovered operator is learned from the same noisy data, and we do not analyze the reliability of that coupling here. The empirical multi-seed experiments of the main text, not asymptotic theory, are the evidence base of this paper.

P COMPUTE AND REPRODUCIBILITY

Implementation scope. All benchmarks and controls use the same staged training algorithm and explicit seed control. The one benchmark-specific algorithmic difference is the Burgers objective (shock-weighted data loss and no Stage-IV ℓ_1 term; Appendix E). The released reproducibility materials include the configurations, per-run loss traces, recovered $\boldsymbol{\alpha}$, corrector amplitudes, and reported metrics.

Compute. All experiments run on a single NVIDIA GeForce RTX 3090 Ti (24 GB) per job. Representative per-run wall-clock at full budget is ≈ 6 minutes for Allen–Cahn and ≈ 10 minutes for Burgers; the complete multi-seed evaluation comprises several hundred runs. No method is given more compute than its comparisons: budgets are matched per benchmark for every training-schedule control except the truncated two-stage arm, a deliberate stage ablation (Appendix E), and the constant-weight joint control matches the curriculum's wall-clock within 15%. The neural corrector is a 4-layer, 128-unit MLP throughout ($\approx 5 \times 10^4$ parameters); the symbolic core adds K coefficients, with K set per benchmark (Appendix B): 2 for the minimal Wave diagnostic, 14 for the standard Wave dictionary, 11 for Burgers, 19 for Allen–Cahn, 43 for 1D Heat, 9 for 2D Wave, and up to $K = 28$ in the Appendix L dictionary-growth stress test.

Seeds and reporting. Every reported cell fixes its random seeds, and mean$\pm$s.d. denotes the population standard deviation over seeds throughout (Welch p-values use the sample variance, as required by the test). Software versions and the pinned environment are included in the reproducibility materials. Dictionary size K denotes the number of candidates, whereas the delivered active-term counts $|\mathrm{supp}(\boldsymbol{\alpha})|$ are reported with each expression (Appendices G and L).