PROTECTING SMALL-COEFFICIENT TERMS IN PDE DISCOVERY WITH A PHYSICAL-CATEGORY MAGNITUDE-DISENTANGLED TRANSFORMER

Anonymous authors
Paper under double-blind review

ABSTRACT

PDE discovery recovers the governing equation from observed data. An equation has terms and coefficients. Recovering it has two steps: choosing terms, and fitting coefficients. Terms often have coefficients of very different sizes. Existing methods compare all terms together in one step. Large-coefficient terms then dominate, and small-coefficient terms are dropped when choosing terms. When a small term is dropped, coefficient fitting gives biased values to compensate, so the recovered equation is incomplete. We introduce the **Physical-Category Magnitude-Disentangled Transformer (PC-MDT)** with two mechanisms. **Masked local category competition** compares each term only with terms of the same physical category, time, space, or nonlinear, and a learned gate exchanges information across categories only when the equation needs it. **Soft-gated global assembly** elects at most one term per category or none and fits coefficients only after the structure is fixed, so size plays no role in the choice. On four benchmark equations, PC-MDT recovers all exactly, while every baseline fails on at least one. A model trained on heat transfers to Burgers with two of three true terms already in place.

1 INTRODUCTION

Data-driven discovery of partial differential equations (PDEs) extracts explicit governing equations from observational data (Makke & Chawla, 2024; Brunton & Kutz, 2022).These equations support prediction, control, and theoretical analysis (Rao et al., 2023), and have been applied in fluid dynamics, climate modeling, and materials science (Makke & Chawla, 2024).The standard pipeline for PDE discovery constructs a fixed candidate library whose columns are the candidate terms (Brunton et al., 2016; Rudy et al., 2017).Given this library, sparse regression separates term selection, which decides which candidates enter the equation, and coefficient estimation, which fits their coefficients to minimize the equation residual (Brunton et al., 2016; Rudy et al., 2018).Rudy et al. (2017) classified the candidate library by dominant physical operations into temporal evolution, spatial transport and diffusion, and nonlinear mode coupling. Stephany & Earls (2024) further noted that these operations are separable in physical systems.We follow this classification and denote the three categories as Temporal (T), Spatial (S), and Nonlinear (N).

Recent Transformer-based PDE discovery methods, such as EqGPT (Xu et al., 2025), LLM-SR (Shojaee et al., 2025), and Verifier-Guided (Faraji & Belardinelli, 2026), place candidates from different physical categories into one sequence and share a single normalization across all candidates (Vaswani et al., 2017).Terms in a PDE candidate library have coefficients that differ by orders of magnitude (Dang et al., 2026).In the Burgers equation, the viscosity coefficient can be orders of magnitude smaller than the convection coefficient (Rudy et al., 2017; Dang et al., 2026).Large-coefficient terms dominate the equation residual and, under the shared normalization, the attention weights; small-coefficient terms are suppressed (Dang et al., 2026).Once a small-coefficient term is missed, the fitted coefficients become biased to compensate for the missing physical mechanism (Du et al., 2024a).BG-SINDy named this phenomenon **coefficient magnitude imbalance** (Dang et al., 2026).This imbalance arises from a single normalization shared by all candidates, which lets coefficient magnitude affect selection (Dang et al., 2026; Vaswani et al., 2017).

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.

Existing remedies on **coefficient magnitude imbalance** follow sparse regression (Dang et al., 2026), derivative estimation (Stephany & Earls, 2024), or search strategy (Petersen et al., 2021; Du et al., 2024b; Brugger et al., 2025).The closest work is BG-SINDy, which uses balance-guided pruning to retain small-coefficient terms, but its fixed threshold acts on a globally mixed library and is hard to tune across equations (Dang et al., 2026). PDE-LEARN replaces finite differences with neural derivative estimates, but its equation identification still relies on STRidge (Stephany & Earls, 2024), which applies a single global threshold to all coefficients (Rudy et al., 2017).The three Transformer methods discussed above also have no mechanism for coefficient magnitude imbalance (Xu et al., 2025; Shojaee et al., 2025; Faraji & Belardinelli, 2026).Each remedy adjusts one stage, but all candidates share a single normalization, so small-coefficient terms remain suppressed (Vaswani et al., 2017; Dang et al., 2026).

To address this, we introduce the Physical-Category Magnitude-Disentangled Transformer (PC-MDT), which removes the shared normalization that causes the suppression. Our contributions are summarized as follows:

1. **PC-MDT.** A Transformer that divides term selection into **masked local category competition** and **soft-gated global assembly**. Local competition protects small-coefficient terms inside each category and exchanges information across categories only when the equation requires it. Global assembly combines the category selections into a complete equation and fits coefficients only after the structure is fixed.
2. **Masked Local Category Competition.** A category mask restricts attention to one category, so each term competes only with terms of the same operation type. A soft gate then exchanges information across categories only when the equation structure requires it. The competition protects small-coefficient terms and produces the representations that global assembly consumes.
3. **Soft-Gated Global Assembly.** Each category elects at most one term from its own candidates, and the elected terms are assembled into a complete equation. Coefficients are fitted only after the structure is fixed, so magnitude differences never enter selection. The verification statistics of the assembled equation start the next round.

2 Related Work

2.1 Sparse Regression for PDE Discovery

SINDy (Brunton et al., 2016) and PDE-FIND (Rudy et al., 2017) established the paradigm of discovering equations from a candidate library with sparse regression. Given a fixed library, sparse regression solves two subproblems: term selection chooses which terms enter the equation, and coefficient estimation fits their coefficients (Rudy et al., 2018; Du et al., 2024b). When coefficients span orders of magnitude, global sparse regression retains large-coefficient terms and removes small-coefficient terms before coefficient estimation (Dang et al., 2026). This difficulty comes from one normalization shared by all candidates, not from threshold selection or library construction.

2.2 Coefficient Magnitude Imbalance and Existing Remedies

Remedies operate at three levels. At the regression level, BG-SINDy uses balance-guided pruning to retain small-coefficient terms (Dang et al., 2026). At the derivative level, PDE-LEARN replaces finite differences with neural derivative estimates (Stephany & Earls, 2024). At the search level, DSR uses reinforcement learning with a sparsity reward (Petersen et al., 2021), and DISCOVER adds dimensional constraints (Du et al., 2024b). Each method adjusts one stage of the three levels, but all candidates are still compared together under one shared normalization (Vaswani et al., 2017). BG-SINDy prunes on a globally mixed library with a fixed threshold and does not change the shared normalization (Dang et al., 2026).

2.3 Transformer-based PDE Discovery

Recent Transformer methods use attention or language models for candidate selection and equation generation (Xu et al., 2025; Shojaee et al., 2025; Faraji & Belardinelli, 2026). Their attention shares

one normalization across all candidates (Vaswani et al., 2017), and T-SHRED and MDLFormer do not decouple by category (Yermakov et al., 2025; Yu et al., 2025). Iterative generation and verification loops exist: LLM-SR feeds ranked equations back (Shojaee et al., 2025), Verifier-Guided filters with a verifier (Faraji & Belardinelli, 2026), and PG-SR uses a reflection pool (Xiao et al., 2026). These loops use language, learned, or set-filtering feedback. In contrast, PC-MDT intervenes on the normalization itself and conditions on fixed verification statistics with no learned component.

3 PRELIMINARIES

The classical approach to equation discovery is sparse regression (Brunton et al., 2016; Rudy et al., 2017). Starting from an observed field $u(x,t)$, we first estimate the time derivative $\hat{u}_t$, then construct a candidate library Θ, each of whose columns ϕ_k is one candidate term. Sparse regression fits

$$\hat{u}_t = \frac{\partial u}{\partial t}, \quad \Theta = [\phi_1 \mid \cdots \mid \phi_K], \quad \Theta\xi \approx \hat{u}_t, \xi^* = \arg\min_{\xi} \|\Theta\xi - \hat{u}_t\|_2^2 \quad \text{s.t.} \quad \|\xi\|_0 \le s. \tag{1}$$

Here ξ is the coefficient vector whose nonzero entries are the selected terms (Rudy et al., 2018).

Transformer-based methods represent each candidate as a token and replace sparse regression with attention (Xu et al., 2025; Shojaee et al., 2025). Standard attention normalizes over all candidates once (Vaswani et al., 2017):

$$q_i = W_Q h_i, \quad k_j = W_K h_j, \quad v_j = W_V h_j, \quad \alpha_{ij} = \frac{\exp(q_i \cdot k_j/\sqrt{d})}{\sum_{l=1}^{K} \exp(q_i \cdot k_l/\sqrt{d})}. \tag{2}$$

Here q_i, k_j, v_j are query, key, and value vectors, h_i, h_j are token representations, d is the dimension, and α_{ij} is the attention weight that candidate j receives from query i. The weights α_{ij} sum to 1 over j, so each row of the attention matrix is a probability distribution over all candidates; the candidates sharing this normalization form the normalization range.

When candidate coefficients span orders of magnitude (Dang et al., 2026), the score gap between large- and small-coefficient terms widens; in equation 2 the weights are exponential in the scores, so the weight of a small-coefficient term approaches zero exponentially. This paper replaces the global normalization with category-wise normalization: the token side in Section 4.2, and the readout side in Section 4.3.

4 METHOD

4.1 PC-MDT STRUCTURE

Figure 1 shows the overall structure. Two mechanisms each remove one global normalization, and the verification step sends the selection result back to the model as the next input.

A standard Transformer uses positional encoding to distinguish sequence elements (Vaswani et al., 2017). Candidates form an unordered set, so positional encoding does not apply. We use type encoding to mark the physical category of each candidate (Lee et al., 2019). Each term ϕ_k is a token, and its initial representation is

$$h_k = e_\phi(\phi_k) + e_c(c_k). \tag{3}$$

Here $e_\phi : \phi_k \mapsto \mathbb{R}^d$ is the term embedding, $e_c : c_k \mapsto \mathbb{R}^d$ is the category embedding, and $c_k \in \{T, S, N\}$.The search proceeds by iterative decoding. The input at step t is the verification statistics R_t for PC-MDT, computed from the current candidate equation and the observed data:

$$R_t = [r_t; g_t; a_t; m_t; c_t; h_t]. \tag{4}$$

r_t holds the residual mean, variance, minimum, and maximum, the statistical criteria used for model selection in sparse regression (Meissner & Glasner, 2024); g_t the residual gradient norms, following derivative-based identification (Pal et al., 2024); a_t the number of active terms and the ℓ_1 norm of their coefficients; m_t the field mean, variance, gradient norm, and spectral energy ratio (Prokop & Gelens, 2026); c_t the active-term count of each category; and h_t the recent residual history, the state feedback of neural-guided search (Brugger et al., 2025). Definitions of the components are in

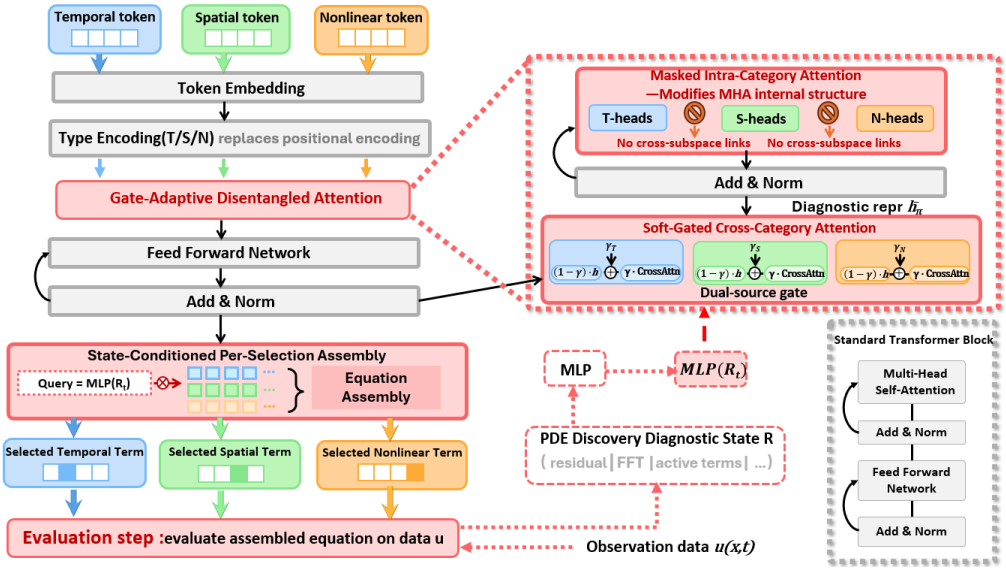


Figure 1: PC-MDT structure. Each candidate term is a token with type encoding marking its physical category. Masked local category competition restricts attention within a category and exchanges information across categories through a soft gate. Soft-gated global assembly elects at most one term per category, assembles them into a complete equation, and fits coefficients after the structure is fixed. The verification step recomputes the statistics and starts the next round. Red boxes mark modifications from a standard Transformer; red dashed arrows mark the verification feedback loop.

Appendix B. We use fixed statistics rather than a learnable state (Petersen et al., 2021), because the model cannot rewrite its own verification result. The gate and the readout both read R_t, because both decisions must answer the current mismatch. The readout selects, from the candidate library, the elected term for each category:

$$z_t = \mathrm{Proj}(R_t), \quad \hat{y}_t = \mathrm{Dec}_\theta(\{h_k\}, z_t). \tag{5}$$

Here $\mathrm{Proj} : R_t \mapsto \mathbb{R}^d$ is a latent projection of the statistics vector (Koulali et al., 2026), and Dec_θ is the readout, a cross-attention query over the candidate representations (Carion et al., 2020). $\hat{y}_t$ is the structural selection at step t: the set of terms the model elects this round.

The suppression of a candidate by global normalization has an upper bound. Let ℓ_j be the pre-normalization score of candidate j, and p_j its post-normalization probability. By equation 2, a candidate with score gap Δ has probability at most

$$\begin{aligned} \ell_j = q \cdot k_j/\sqrt{d}, \quad p_j = \frac{\exp(\ell_j)}{\sum_{l=1}^{K} \exp(\ell_l)}, \quad \ell_{\max} = \max_l \ell_l, \\ p_j = \frac{\exp(\ell_j)}{\exp(\ell_{\max}) + \sum_{l \neq \max} \exp(\ell_l)} \leq \frac{\exp(\ell_j)}{\exp(\ell_{\max})} = e^{-\Delta}, \quad \Delta = \ell_{\max} - \ell_j. \end{aligned} \tag{6}$$

equation 6 explains why the suppression is structural: a shared normalization turns a score gap $\Delta = \ell_{\max} - \ell_j$ into a weight $p_j \leq e^{-\Delta}$. Both interventions therefore remove cross-category scores from the exponent: the category mask equation 7 keeps attention scores inside each category, and the category-wise readout equation 11 normalizes each category separately.

4.2 Masked Local Category Competition

The suppression in equation 6 first occurs in the attention layer. Candidates sharing one normalization compete for the same probability mass, so a large-coefficient term of another category takes the mass a small-coefficient term needs. The mask changes who competes with whom.

Masked Intra-category competition. Terms of different operation types no longer compete. We set cross-category scores to $-\infty$: the attention matrix becomes block-diagonal, the normalization

range is exactly the category, and the suppression channel from other categories is cut:

$$Q_c = W_Q H_c, \quad K_c = W_K H_c, \quad V_c = W_V H_c,$$
$$\Lambda_c[i,j] = \begin{cases} 0 & c_j = c \\ -\infty & c_j \neq c \end{cases}, \quad A_c = \mathrm{softmax}\big((Q_c K_c^\top + \Lambda_c)/\sqrt{d}\big)V_c. \tag{7}$$

Here H_c is the token representation of category c, Q_c, K_c, V_c are projections of H_c, and Λ_c is the category mask. A_c is the attention output within the category, the representation each category keeps for itself, used in equation 8. Λ_c is partitioned by physical category, not by sequence.

Inter-category interaction. Intra-category competition alone isolates every category, but some equations need terms of different categories to appear together. The interaction strength cannot be fixed by hand: the wave equation fails without interaction, and the Burgers equation fails with full interaction (ablation in Section 5.5). Cross-attention reopens the channel cut by equation 7 in a controlled form: each subspace uses itself as query and the tokens of the other subspaces as keys and values (Vaswani et al., 2017):

$$q_k = W_Q h_k, \quad k_{c'} = W_K h_{c'}, \quad v_{c'} = W_V h_{c'}, \quad \hat{h}_k = \sum_{c' \neq c_k} \mathrm{softmax}(q_k \cdot k_{c'}/\sqrt{d}) v_{c'}. \tag{8}$$

$\hat{h}_k$ summarizes what the other categories currently hold: the softmax in equation 8 normalizes only over the other categories, so outside magnitudes cannot alter the within-category shares.

Soft gating. How much cross-category context to take cannot be fixed in advance, because the equation structure decides it. A scalar gate γ_c makes the decision. It is soft: it ranges over the full interval between isolation ($\gamma_c = 0$) and replacement by the cross-category context ($\gamma_c = 1$), so one model serves both coupled and decoupled equations without hand-tuning. The gate is computed from the verification statistics and the subspace mean representations:

$$\bar{h}_c = \frac{1}{|c|} \sum_{k \in c} h_k, \quad \gamma_c = \sigma\big(w_c^\top [z_t; \bar{h}_T; \bar{h}_S; \bar{h}_N]\big). \tag{9}$$

Here z_t comes from equation 5. The three means average the token representations h_k of equation 3 within each category, as shown in the first line of equation 9. z_t reports how far the current equation is from the data, and $\bar{h}_c$ summarizes what category c encodes. The gate softly interpolates the two representations, so a candidate keeps its own representation unless the equation requires outside information:

$$h'_k = (1 - \gamma_{c_k}) h_k + \gamma_{c_k} \hat{h}_k. \tag{10}$$

Here γ_{c_k} is the gate value for the category of h_k. The gate is initialized to favor decoupling (Appendix O). The readout in Section 4.3 uses the resulting h' as keys and values equation 11.

T, S, and N characterize operation types, the mask and gate parameters are shared across equations, and γ is recomputed from the verification statistics of the target equation, so the transferred model starts from the correct structure (Section 5.4).

4.3 Soft-Gated Global Assembly

The suppression in equation 6 also occurs on the readout side. Standard methods normalize over all candidates once; here each category has its own scores ℓ_c, and equation 12 normalizes only inside the category. The selection of a small-coefficient term is decided by its own category alone, and the large terms of other categories cannot enter. Global assembly works on the representations produced by the local competition: the readout consumes the gated representations h'_k of equation 10, not the raw ones. Each h'_k mixes its own representation with the cross-category context, and the soft gate γ_c of equation 9 sets the mix ratio.

Gated Category-wise readout. Each category votes among its own candidates. The readout uses z_t as query and h' as keys and values for cross-attention:

$$q = W_Q z_t, \quad k_k = W_K h'_k, \quad v_k = W_V h'_k, \quad \ell_c = W_c \sum_k \mathrm{softmax}(q \cdot k_k/\sqrt{d}) v_k. \tag{11}$$

Here z_t comes from equation 5, h'_k from equation 10, and W_c is the output projection for category c. Let N_c be the number of candidates in category c. Each category performs a softmax over its own candidates plus one extra entry:

$$p_{c,k} = \frac{\exp(\ell_{c,k})}{\sum_{k'=1}^{N_c+1} \exp(\ell_{c,k'})}, \quad \hat{y}_c = \arg\max_k p_{c,k}. \tag{12}$$

$\ell_c \in \mathbb{R}^{N_c+1}$ is the candidate score vector for category c. $p_{c,k}$ is normalized only over the candidates of category c plus the extra entry, which elects no term, borrowing the no-object class of set prediction (Carion et al., 2020). Without it, every category must elect a term, so the wave equation, which has no nonlinear term, cannot be represented. $\hat{y}_c$ is the selection of category c; $\hat{y}_t = [\hat{y}_T; \hat{y}_S; \hat{y}_N]$ is the complete structural selection.

Structure before coefficients. The structure is fixed before any coefficient is fitted, so magnitude differences only affect coefficients and never enter structural selection. The coefficients are fitted by least squares on the selected columns, and this refitting repeats at every step:

$$\hat{\xi} = \arg\min_{\xi} \|(\hat{y}_t \odot \Theta)\xi - \hat{u}_t\|_2^2. \tag{13}$$

Here $\odot$ denotes column-wise selection. The least-squares refitting is the same as in DSR (Petersen et al., 2021). The selections of the three categories are assembled into a complete equation:

$$\hat{u}_t \approx \sum_c \sum_k \hat{y}_{c,k} \hat{\xi}_{c,k} \phi_{c,k}. \tag{14}$$

Here $\hat{y}_{c,k}$ comes from equation 12, $\hat{\xi}_{c,k}$ from equation 13, and $\phi_{c,k}$ from equation 1.

Verification feedback. The verification step closes the loop: it measures the assembled equation against the observed data, recomputes the statistics for the next round, and scores the selections of this round.

$$R_{t+1} = \mathrm{Stats}(\hat{y}_t, u). \tag{15}$$

Stats is the statistics function defined in equation 4. The iteration starts from an initial structure (Appendix C) and stops when the structure converges or reaches the step limit. After one decoding round, the sequence of selections is scored by the least-squares residual of its final equation:

$$\begin{gathered} r(\hat{y}) = \min_{\xi} \|(\hat{y} \odot \Theta)\xi - \hat{u}_t\|_2^2, \quad \rho(\hat{y}) = -\log\max(r(\hat{y}), \delta), \\ L(\theta) = -\sum_t \log \mathrm{Dec}_\theta(\hat{y}_t^* \mid R_t). \end{gathered} \tag{16}$$

Here $r(\hat{y})$ is the least-squares residual, $\rho(\hat{y})$ is the score, and δ is a clipping constant. The score depends only on the selected structure, and the coefficients are optimized out. The sequence with the highest score $\rho(\hat{y})$ in the round is the target sequence $\hat{y}_t^*$, and the model is updated by maximum likelihood (Anthony et al., 2017). When a category elects a spurious term instead of the no-term entry, the residual increases and the objective corrects the entry. The final equation is the converged $\hat{y}_T$ combined with $\hat{\xi}$.The mask protects the local competition, the soft gate controls cross-category exchange, the category-wise readout protects selection, and the verification step closes the loop.

5 Experiments

5.1 Experimental Setup

Data and library. We evaluate on the four standard benchmark systems of PDE discovery (Rudy et al., 2017; Stephany & Earls, 2024), simulated on 256×256 spatiotemporal grids: Burgers ($u_t + u \cdot u_x = 0.1\, u_{xx}$), heat ($u_t = 0.5\, u_{xx}$), wave ($u_{tt} = u_{xx}$), and Korteweg–de Vries ($u_t + 6u \cdot u_x + u_{xxx} = 0$). All methods share the library of Section 3, fixed across equations so the search space leaks no ground-truth information. KdV is the only system that activates all three categories at once and forces competition inside the spatial category (u_{xx} versus u_{xxx}).

Table 1: Exact recovery on four equations (each cell: SR↑/TPR↑-FPR↓/RE↓; mean over 5 seeds; * = published value under the original protocol, for context only). Bold marks the proposed method.

	Burgers			Heat			Wave			KdV		
Method	SR	TPR-FPR	RE	SR	TPR-FPR	RE	SR	TPR-FPR	RE	SR	TPR-FPR	RE
PC-MDT	**1.00**	**1.00-0.00**	**1.5e-4**	**1.00**	**1.00-0.00**	**2.0e-4**	**1.00**	**1.00-0.00**	**2.7e-5**	**1.00**	**1.00-0.00**	**1.3e-5**
DSR	0.00	1.00-0.57	16.0	0.00	1.00-0.50	9.31	0.00	1.00-0.50	0.00	0.00	1.00-0.57	95.1
T-SHRED†	1.00	1.00-0.00	2.0e-4	1.00	1.00-0.00	2.0e-4	0.00	0.00-1.00	1.00	1.00	1.00-0.00	0.00
Vanilla	0.00	0.67-0.00	5.5e-1	1.00	1.00-0.00	2.0e-4	1.00	1.00-0.00	2.7e-5	1.00	1.00-0.00	1.3e-5
PDE-LEARN*	1.00	1.00-0.00	1.1e-2	0.00	1.00-0.33	3.1e-1	0.00	1.00-0.55	4.7e-1	0.00	1.00-0.55	4.2e-1
BG-SINDy	1.00	1.00-0.00	1.5e-4	0.00	0.50-0.50	4.5e-1	1.00	1.00-0.00	0.00	1.00	1.00-0.00	0.00

Baselines. Table 1 compares six methods. Four baselines are re-run under our protocol on the same library, data, and success criterion: DSR (Petersen et al., 2021; Dang et al., 2026), T-SHRED† († = transplanted onto the shared fixed library), Vanilla (PC-MDT with the category masks and gated coupling removed, leaving global attention; it is also the vanilla ablation of Section 5.5), and BG-SINDy (Dang et al., 2026) (re-implemented; Appendix G). PDE-LEARN (Stephany & Earls, 2024) is listed at its published values under its original protocol, for context only; EqGPT, DISCOVER, and MDLFormer do not fit a shared protocol.

Metrics. A run succeeds if the selected term set equals the ground truth and every coefficient satisfies $|\xi - \xi^*| \le 0.05|\xi^*| + 10^{-6}$, DSR's recovery-rate criterion (Petersen et al., 2021). We report success rate SR↑ (fraction of successful seeds), term recall TPR↑$= \mathrm{TP}/(\mathrm{TP}+\mathrm{FN})$, false-positive rate FPR↓$= \mathrm{FP}/(\mathrm{FP}+\mathrm{TP})$, and coefficient error RE↓$= \|\xi_{\mathrm{disc}} - \xi^*\|/\|\xi^*\|$ over truth terms. A method that keeps the truth plus spurious terms reads RE 0.00 (DSR on Wave in Table 1); the TPR–FPR pair carries that information.

Protocol. For PC-MDT one training iteration is one decoding round followed by one maximum-likelihood update (Section 4.3); every method in Table 1 runs 500 iterations under its own loop.

5.2 Main Results: Exact Recovery Across Four Equations

This experiment tests the central claim of Section 1: exact recovery is structural, not an artifact of easy data. PC-MDT recovers all four equations exactly: $\mathrm{SR} = 1.00$ with $\mathrm{TPR} = 1.00$ and $\mathrm{FPR} = 0.00$, and coefficient error below 2.0×10^{-4} on every system. DSR over-selects on every equation (FPR 0.50–0.57) and never succeeds. Vanilla misses the viscosity term on Burgers (TPR $= 0.67$): the only term whose coefficient is an order of magnitude smaller than the rest is the one dropped, the suppression predicted in Section 1. BG-SINDy passes Burgers, Wave, and KdV but prunes u_{xx} on Heat (TPR $= 0.50$): one shared threshold misfires on the term it was designed to protect. T-SHRED† recovers Burgers, Heat, and KdV but fails Wave entirely (TPR $= 0.00$), because its regression is anchored to u_t, which cannot express a u_{tt}-governed system. Since some method passes every equation, the failures track mechanisms, not the data.

5.3 Small-Coefficient Recovery

This experiment measures the magnitude-imbalance suppression of Section 1 across term categories. We sweep the small coefficient of four equation families over 30 parameter points; at each point every method runs the full search on a fresh dataset with 5 seeds and is scored as in Table 1, with per-point time windows (Appendix D). Figure 2 reports $\mathrm{TPR} \times \mathrm{SR}$ against the swept coefficient. Only the PC-MDT line never leaves 1.00; the per-family failure details follow.PC-MDT is the only method that recovers all 30 points: $\mathrm{SR} = 1.00$ with $\mathrm{TPR} = 1.00$ everywhere, and $\mathrm{RE} \le 4.9 \times 10^{-4}$ at every point; its KdV passes use the coordinate-descent revisit (Appendix E). On Burgers, Vanilla drops u_{xx} at every ν ($\mathrm{TPR} = 0.67$), T-SHRED† from $\nu \le 0.05$, and PDE-LEARN from $\nu \le 0.005$. On KdV, Vanilla drops u_{xxx} from $\beta \le 0.1$, T-SHRED† from $\beta \le 0.05$, and PDE-LEARN from $\beta \le 0.1$. On the weak-reaction family, Vanilla and T-SHRED† drop u^2 from $\varepsilon \le 0.1$, where the coefficient falls at or below their thresholds 0.15 and 0.1; PDE-LEARN drops it from $\varepsilon \le 0.02$. On weak advection the roles flip at $\varepsilon = 0.5$, where diffusion becomes the small coefficient: Vanilla keeps $u \cdot u_x$ but drops u_{xx}, and from $\varepsilon \le 0.1$ both Vanilla and T-SHRED† drop both terms ($\mathrm{TPR} = 0.33$); PDE-LEARN drops $u \cdot u_x$ from $\varepsilon \le 0.05$. DSR over-selects on all 30 points; on weak advection at $\varepsilon \le 0.01$ it also drops $u \cdot u_x$. The failures track the threshold, not the term or the category; PC-MDT's recovery comes from category disentanglement, not a better-tuned global rule.

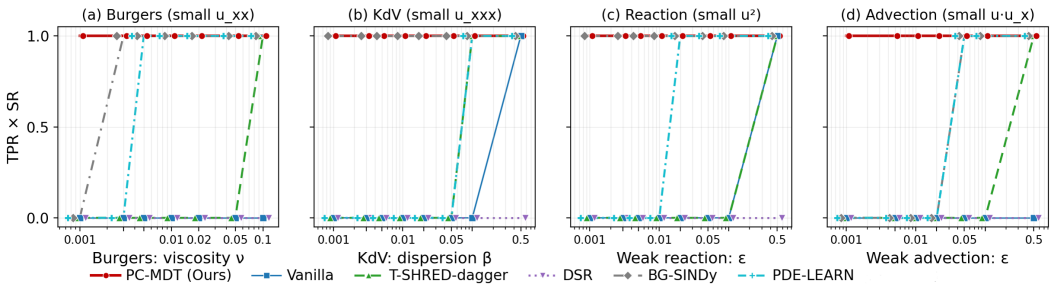


Figure 2: Four-family small-coefficient sweep, one panel per family, logarithmic x-axis, six methods per panel. The y-axis is $\mathrm{TPR} \times \mathrm{SR}$ per run: the term-set recall times the exact-recovery indicator (1 on success, 0 otherwise), averaged over the 5 seeds of each point. Panels (a)–(d): Burgers (ν on u_{xx}), KdV (β on u_{xxx}), weak reaction (ε on u^2), weak advection (ε on $u \cdot u_x$). PC-MDT is the only line that stays at 1.00 on all 30 points.

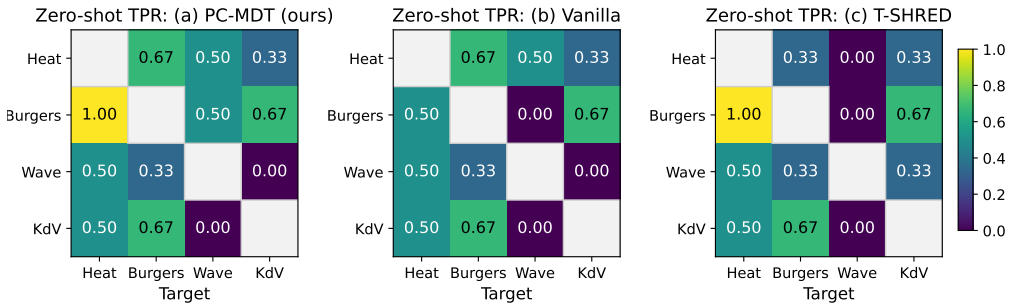


Figure 3: Zero-shot term-set recall (TPR) over the source-to-target matrix (mean over 5 seeds): (a) PC-MDT, (b) Vanilla, (c) T-SHRED†. PC-MDT matches or beats Vanilla on every pair, strictly better on Burgers→Heat and Burgers→Wave.

5.4 Transfer Across Equations

This experiment tests the second structural claim: a model trained on one equation should start close to the correct equation on another, because the T/S/N categories describe operations, not particular terms. We pre-train on each equation (500 iterations), then continue training on each other equation, reusing the category masks and gate parameters (Section 4.2); training from scratch with identical seeds is the control. Figure 3 reports the zero-shot term-set recall of three methods over the source-to-target matrix (12 directed pairs; mean over 5 seeds). PC-MDT's off-diagonal values track structural overlap: mean zero-shot TPR 0.75 on pairs sharing two terms and 0.44 on pairs sharing one, against 0.40 from random initialization on average. DSR is omitted from this figure: its zero-shot TPR is 1.00 on all 12 pairs, but only by keeping two to five spurious terms per pair, so $\mathrm{SR} = 0$ throughout.In Figure 4, the same two-phase protocol targets Burgers at $\nu = 0.001$, the point at which every baseline fails (Section 5.3). The transferred PC-MDT reaches $\mathrm{TPR} \times \mathrm{SR} = 1.00$ in all 15 runs (3 sources $\times$ 5 seeds), with final coefficient error 5.9×10^{-8}; the from-scratch control also reaches 1.00, but starts from 3.6 wrong terms on average, against one to three after transfer. Every baseline stays at 0 ($\mathrm{SR} = 0$ throughout). The bottom row counts wrong terms, $\mathrm{FP} + \mathrm{FN}$, in the proposed term set at each iteration: Vanilla, T-SHRED†, and PDE-LEARN stay at one, the missing u_{xx}; BG-SINDy stays at one, a ghost u_x at coefficient 5.1×10^{-12}; DSR stays at four. At this target a vector RE cannot separate BG-SINDy from PC-MDT: both end at 5.9×10^{-8}, because a coefficient of 10^{-12} is invisible to it. Only the term-set count, the success criterion of Section 5.1, separates the methods, and only PC-MDT reaches zero. What transfers is the category-consistent selection, not the data: both attention architectures carry the source equation's selection regularities to the target, but only PC-MDT turns them into exact recovery.

5.5 Component Ablation

This experiment shows what each component contributes: removing one should fail exactly on the equation whose mechanism depends on it.We test four variants against the full model, all with identical post-processing. no_intra removes the category masks (Section 4.2) but keeps the per-

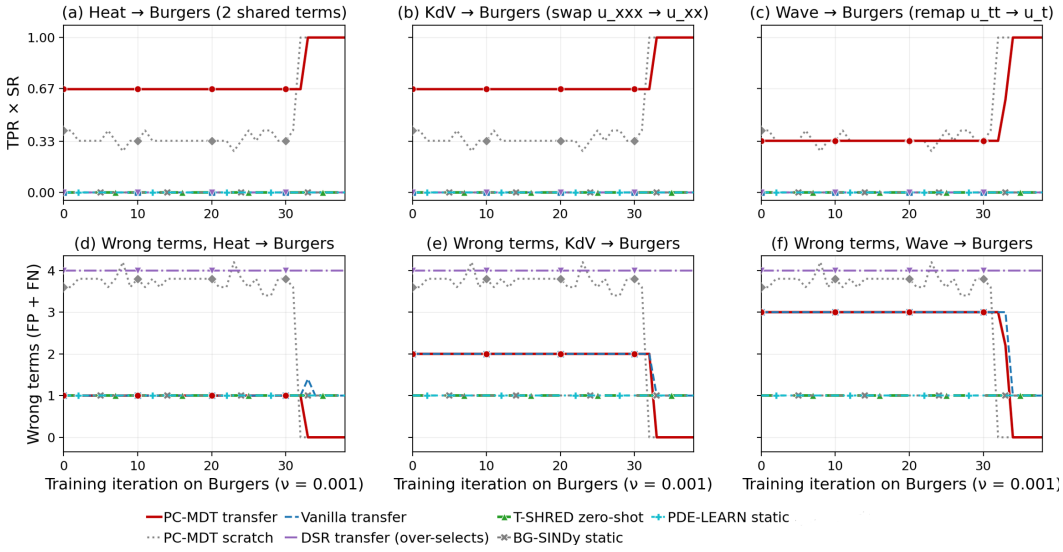


Figure 4: Transfer to Burgers at $\nu = 0.001$. (a)–(c) TPR $\times$ SR per run (term-set recall times the exact-recovery indicator, mean over 5 seeds): transferred PC-MDT (red) and the from-scratch control (dotted grey) reach 1.00; Vanilla, DSR, T-SHRED†, BG-SINDy, and PDE-LEARN stay at 0. (d)–(f) Wrong-term count FP + FN of the proposed term set: PC-MDT removes its one to three inherited wrong terms and ends at 0; the from-scratch control starts at 3.6 wrong terms on average and also ends at 0; Vanilla, T-SHRED†, and PDE-LEARN stay at 1 (missing u_{xx}), BG-SINDy at 1 (ghost u_x at coefficient 5.1×10^{-12}), and DSR at 4. T-SHRED†, BG-SINDy, and PDE-LEARN have no fine-tuning loop and appear as flat levels.

category selection heads, so we predict no failure: the category-wise readout already selects, and the masked layer's role is representational. no_cross fixes the gate at $\gamma = 0$, cutting cross-category communication, so Wave should fail.full_cross fixes the gate at $\gamma = 1$, which reduces to global mixing, and vanilla replaces the attention stack by global attention; both should suppress the small viscosity term and fail on Burgers.The failure pattern (Table 2) matches the prediction. no_cross fails only on Wave, where five seeds collapse to the equation anchored at u_t. full_cross and vanilla fail only on Burgers (TPR$= 0.67$). full and no_intra pass all four equations.

Table 2: Ablation (each cell: SR↑/TPR↑-FPR↓/RE↓; mean over 5 seeds). Bold marks the complete model (full). Variant definitions: no_intra removes the category masks; no_cross fixes the gate at $\gamma = 0$ (categories decoupled); full_cross fixes the gate at $\gamma = 1$ (global mixing); vanilla replaces the masked stack with global attention.

	Burgers			Heat			Wave			KdV		
Variant	SR	TPR-FPR	RE	SR	TPR-FPR	RE	SR	TPR-FPR	RE	SR	TPR-FPR	RE
full	**1.00**	**1.00-0.00**	**1.5e-4**	**1.00**	**1.00-0.00**	**2.0e-4**	**1.00**	**1.00-0.00**	**2.7e-5**	**1.00**	**1.00-0.00**	**1.3e-5**
no_intra	1.00	1.00-0.00	1.5e-4	1.00	1.00-0.00	2.0e-4	1.00	1.00-0.00	2.7e-5	1.00	1.00-0.00	1.3e-5
no_cross	1.00	1.00-0.00	1.5e-4	1.00	1.00-0.00	2.0e-4	0.00	0.00-1.00	1.00	1.00	1.00-0.00	1.3e-5
full_cross	0.00	0.67-0.00	5.5e-1	1.00	1.00-0.00	2.0e-4	1.00	1.00-0.00	2.7e-5	1.00	1.00-0.00	1.3e-5
vanilla	0.00	0.67-0.00	5.5e-1	1.00	1.00-0.00	2.0e-4	1.00	1.00-0.00	2.7e-5	1.00	1.00-0.00	1.3e-5

6 Conclusion

We address coefficient magnitude imbalance in PDE discovery: candidates sharing one global normalization let large terms suppress small ones. PC-MDT removes this suppression structurally. Candidates are normalized only within their own physical category, and a gate computed from the verification statistics restores the cross-category coupling that some equations require. Each category selects terms with a learned empty option, and coefficients are fitted only after selection. On four benchmark equations, PC-MDT recovers each exactly, whereas every baseline fails on at least one; the small viscosity term survives down to $\nu = 0.001$; a model trained on heat transfers to the Burgers equation with two of three true terms in place; and each ablation variant fails exactly where expected. Limitations are the fixed candidate library and noise-free data; noisy observations and higher-dimensional fields are left to future work.

Reproducibility Statement

We describe the PC-MDT architecture and training procedure in Section 4, and the data, baselines, success criterion, metrics, and protocol in Section 5.1. The verification-statistics components, library composition, search initialization, sweep data and windows, coordinate-descent revisit, transfer protocol, and the BG-SINDy and PDE-LEARN re-implementations are in the appendices. All datasets are simulated in-house; no external dataset is required. Each equation is evaluated with five random seeds and the same success criterion for all methods. An anonymized version of the code, simulation scripts, and evaluation scripts is included in the supplementary material; the code will be released publicly upon acceptance.

AI Use Statement

In this work, we used generative AI tools for literature search and grammar checking. We did not use generative AI tools for generating research ideas, designing experiments, analyzing data, or generating code. We did not use generative AI for any other purpose. We have reviewed all AI-assisted work, checked generated text for factual accuracy against primary sources, and verified that all citations correspond to real publications. We take responsibility for the final content of this work, including text, claims, or artifacts produced with the aid of generative AI.

Ethics Statement

This work focuses on automated discovery of physical equations from numerically simulated benchmark data. It does not involve human subjects, personal data, or IRB approval. We do not foresee direct harmful applications. Recovered equations are intended for scientific analysis and should be validated before deployment in physical systems. The authors declare no conflicts of interest.

References

Thomas Anthony, Zheng Tian, and David Barber. Thinking fast and slow with deep learning and tree search. In *Advances in Neural Information Processing Systems 30 (NeurIPS)*, pp. 3389–3399, 2017.

Jonas Brugger, Mattia Cerrato, David Richter, Cedric Derstroff, Daniel Maninger, Mira Mezini, and Stefan Kramer. Neural-guided equation discovery. *arXiv preprint arXiv:2503.16953*, 2025.

Steven L. Brunton and J. Nathan Kutz. *Data-Driven Science and Engineering: Machine Learning, Dynamical Systems, and Control*. Cambridge University Press, 2nd edition, 2022.

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.

Nicolas Carion, Francisco Massa, Gabriel Synnaeve, Nicolas Usunier, Alexander Kirillov, and Sergey Zagoruyko. End-to-end object detection with transformers. In *European Conference on Computer Vision (ECCV)*, pp. 213–229. Springer, 2020.

Zhenhua Dang, Lei Zhang, Long Wang, and Guowei He. Balance-guided sparse identification of multiscale nonlinear PDEs with small-coefficient terms. *arXiv preprint arXiv:2604.18414*, 2026.

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

Mengge Du, Yuntian Chen, and Dongxiao Zhang. DISCOVER: Deep identification of symbolically concise open-form partial differential equations via enhanced reinforcement learning. *Physical Review Research*, 6(1):013182, 2024b.

Farbod Faraji and Francesco Belardinelli. Verifier-guided model discovery for physical dynamical systems with pretrained symbolic transformers. *arXiv preprint arXiv:2608.02662*, 2026.

Imane Koulali, Erhan Turan, and M. Taner Eskil. LaSIPDE: Latent-space identification of partial differential equations from indirect, high-dimensional measurements. *Frontiers in Applied Mathematics and Statistics*, 2026. doi: 10.3389/fams.2026.1807939.

Juho Lee, Yoonho Lee, Jungtaek Kim, Adam R. Kosiorek, Seungjin Choi, and Yee Whye Teh. Set transformer: A framework for attention-based permutation-invariant neural networks. In *Proceedings of the 36th International Conference on Machine Learning (ICML)*, volume 97 of *Proceedings of Machine Learning Research*, pp. 3744–3753. PMLR, 2019.

Nour Makke and Sanjay Chawla. Interpretable scientific discovery with symbolic regression: a review. *Artificial Intelligence Review*, 57(1):2, 2024.

Teddy Meissner and Karl Glasner. Discovery of differential equations using sparse state and parameter regression. *arXiv preprint arXiv:2406.06707*, 2024.

Ashish Pal, Sutanu Bhowmick, and Satish Nagarajaiah. Physics-informed AI and ML-based sparse system identification algorithm for discovery of PDEs representing nonlinear dynamic systems. *arXiv preprint arXiv:2410.10023*, 2024.

Brenden K. Petersen, Mikel Landajuela, T. Nathan Mundhenk, Claudio P. Santiago, Soo K. Kim, and Joanne T. Kim. Deep symbolic regression: Recovering mathematical expressions from data via risk-seeking policy gradients. In *The Ninth International Conference on Learning Representations (ICLR)*, 2021.

Bartosz Prokop and Lendert Gelens. Data-driven discovery of dynamical models in biology. *Nature Reviews Physics*, 2026. doi: 10.1038/s42254-026-00955-4.

Chengping Rao, Pu Ren, Qi Wang, Oral Buyukozturk, Hao Sun, and Yang Liu. Encoding physics to learn reaction–diffusion processes. *Nature Machine Intelligence*, 5(7):765–779, 2023.

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.

Samuel H. Rudy, Anil Alla, Steven L. Brunton, and J. Nathan Kutz. Data-driven identification of parametric partial differential equations. *arXiv preprint arXiv:1806.00732*, 2018.

Parshin Shojaee, Kazem Meidani, Chandan K. Reddy, and Amir Barati Farimani. LLM-SR: Scientific equation discovery with large language models. In *The Thirteenth International Conference on Learning Representations (ICLR)*, 2025.

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

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N. Gomez, Łukasz Kaiser, and Illia Polosukhin. Attention is all you need. In *Advances in Neural Information Processing Systems 30 (NeurIPS)*, pp. 5998–6008, 2017.

Yifan Xiao et al. PG-SR: Policy-guided symbolic regression with a reflection pool. *arXiv preprint arXiv:2602.13021*, 2026.

Hao Xu, Yuntian Chen, Rui Cao, Tianning Tang, Mengge Du, Jian Li, Adrian H. Callaghan, and Dongxiao Zhang. Generative discovery of partial differential equations by learning from math handbooks. *Nature Communications*, 16(1):10255, 2025.

Alexey Yermakov, David Zoro, Mars Liyao Gao, and J. Nathan Kutz. T-SHRED: Symbolic regression for regularization and model discovery with transformer shallow recurrent decoders. *arXiv preprint arXiv:2506.15881*, 2025.

Zihan Yu, Jingtao Ding, Yong Li, and Depeng Jin. Symbolic regression via MDLformer-guided search: From minimizing prediction error to minimizing description length. In *The Thirteenth International Conference on Learning Representations (ICLR)*, 2025.

A Candidate Library Composition

All methods share the same library of Section 3 on every equation. The ground-truth term sets are Burgers $\{u_t, u \cdot u_x, u_{xx}\}$, heat $\{u_t, u_{xx}\}$, wave $\{u_{tt}, u_{xx}\}$, and KdV $\{u_t, u \cdot u_x, u_{xxx}\}$.

B Verification Statistics

The verification statistics $R_t = [r_t; g_t; a_t; m_t; c_t; h_t] \in \mathbb{R}^{19}$ are defined as follows: $r_t \in \mathbb{R}^4$ holds residual statistics (mean, variance, min, max), $g_t \in \mathbb{R}^2$ the residual gradient norms $\|\partial R/\partial x\|$ and $\|\partial R/\partial t\|$, $a_t \in \mathbb{R}^2$ the number of active terms and the ℓ_1 norm of their coefficients, $m_t \in \mathbb{R}^4$ field statistics (mean, variance, gradient norm, FFT energy ratio), $c_t \in \mathbb{R}^3$ the number of active terms in each category, and $h_t \in \mathbb{R}^4$ the history of recent residuals (trend, volatility, recent min, recent max). The clipping constant is $\delta = 10^{-10}$.

C Search Initialization and Step Limit

The pivot equation keeps only the best single temporal term: each temporal candidate is fitted alone, the one with the lowest residual is kept, and the search starts from it; R_0 holds its verification statistics. The search stops when the assembled equation converges or after a fixed step limit of 40, whichever comes first.

D Sweep Data and Windows

Each parameter point of Section 5.3 uses its own 256×256 dataset, generated with the same finite-difference solvers, grid, and boundary treatment as the main benchmark of Table 1; only the swept coefficient and the time window change. The window is calibrated per point so that the 90th percentile of the periodic finite-difference residual of the truth equation stays at the known-good level of the calibrated setups (the ceilings calibrated at Burgers $\nu = 0.1$ and KdV $\beta = 1$ are 5.7×10^{-6} and 1.2×10^{-3}): smaller Burgers viscosities and KdV dispersions sharpen their solutions and use shorter windows, ending before the under-resolved shock (KdV in the $\beta \to 0$ limit develops a shock at $t \approx 1/6$), while the weak-reaction and weak-advection families are diffusion-dominated and keep the full window. Table 3 lists the window per point. The weak-reaction family uses a non-negative Gaussian initial condition: under the damping sign $u_t = 0.5u_{xx} - \varepsilon u^2$ the maximum principle confines the solution to $[0, \max u_0]$, so the equation is globally well-posed, and the reaction data are indistinguishable in kind from the diffusion-dominated benchmark data. A small-coefficient time term cannot enter this sweep: the fixed library has no second temporal term to pair with u_t (and u_{tt} is reserved for the wave target), so a time category competition would structurally misalign with the library; the Wave row of Table 1, where the T-SHRED† anchor fails, exhibits the corresponding mechanism.

Family	Parameter grid	Time window per point
Burgers	$\nu \in \{0.1, 0.05, 0.02, 0.01, 0.005, 0.003, 0.001\}$	0.1: $[0, 10]$; 0.05: $[0, 3]$; 0.02: $[0, 2.5]$; 0.01: $[0, 2]$; 0.005–0.001: $[0, 1]$
KdV	$\beta \in \{0.5, 0.1, 0.05, 0.02, 0.01, 0.005, 0.003, 0.001\}$	0.5: $[0, 1.5]$; 0.1: $[0, 0.2]$; 0.05–0.001: $[0, 0.1]$
Weak reaction	$\varepsilon \in \{0.5, 0.1, 0.05, 0.02, 0.01, 0.005, 0.003, 0.001\}$	$[0, 10]$ for every point
Weak advection	$\varepsilon \in \{0.5, 0.1, 0.05, 0.02, 0.01, 0.005, 0.001\}$	$[0, 10]$ for every point

Table 3: Parameter grids and per-point time windows for the four sweep families.

E Coordinate-Descent Revisit

After the category-by-category pass of Section 4.3, one coordinate-descent revisit per subspace re-examines the assembled equation on the least-squares-only chain: for each subspace in turn, every

candidate term (including the currently selected one) is tried in the slot while all other slots stay fixed, and the slot is updated only when the replacement improves the residual by a strict relative margin (above 10^{-5}) and the refit coefficient is non-negligible (above 10^{-3} in magnitude). The revisit adds about seven least-squares fits per decoding round (Burgers and KdV: 3,507 to 7,014 least-squares evaluations per run), and it is what lets KdV converge at $\beta \leq 0.1$: in the pass-1 search the spatial slot locks u_{xx} as a proxy for the advection term before u_{xxx} competes, and only the post-pass revisit, which evaluates each candidate against the coefficients calibrated so far, replaces the proxy. The revisit is applied uniformly to every equation and every sweep point, not selectively to the KdV family.

F TRANSFER PROTOCOL

Phase A trains the source equation for 500 iterations and exports the weights; phase B continues training on the target from those weights, with the decoder's proposed term set recorded before every update. The zero-shot TPR of Figure 3 is the proposal recall at phase-B iteration 0; from-scratch controls use identical seeds. The subspace masks and the gate parameters are reused across phases, and the gate value is recomputed from the target's verification statistics at every step. The DSR variant is transferred under the same two-phase protocol (its ablation removes only the decoupling); T-SHRED† is evaluated zero-shot only, as it has no learned decoder to fine-tune, and contributes its flat level to panels (a)–(f) of Figure 4. BG-SINDy and PDE-LEARN have no learned decoder to transfer and are omitted from the matrix; BG-SINDy appears as a flat static level in Figure 4 at its $\nu = 0.001$ sweep level (Appendix H);PDE-LEARN appears as a flat static level there at its $\nu = 0.001$ sweep level with one wrong term, the missing u_{xx} (Appendix J).

G BG-SINDY RE-IMPLEMENTATION

BG-SINDy (Dang et al., 2026) has no public implementation; we transcribe the algorithm of its method section (Eqs. 4–11) faithfully: (i) form the candidate library Φ and the target vector (in the pivot convention shared by all methods in this paper, the target is $-u_t$, or $-u_{tt}$ for the wave equation, and the discovered equation reads $\text{target} + \sum_j \xi_j \varphi_j = 0$); (ii) extract a maximal linearly independent subset by QR factorization with column pivoting (relative tolerance 10^{-10}); (iii) fit the active-set least-squares coefficients ξ; (iv) compute the per-sample local relative importance $w_{ij} = |\Phi_{ij}\xi_j|/(\max_{l\in A}|\Phi_{il}\xi_l| + \epsilon)$ with $\epsilon = 10^{-12}$, averaged to a global importance per term; (v) remove the least important term, refit, one term per iteration; (vi) stop when the residual after a removal exceeds $\tau = 3$ times the residual before it, and keep the previous iteration as the selected model. The runs use the identical cached 256×256 benchmark fields, interior mask, library, and success criterion as every other method in this paper (the term set must equal the truth and each coefficient must satisfy $|\xi - \xi^*| \leq 0.05|\xi^*| + 10^{-6}$). As a self-check against the original paper's claim, the implementation recovers $\{u_t, u \cdot u_x, u_{xx}\}$ on Burgers with $\nu = 0.1$ (coefficient error 1.5×10^{-4}, all five seeds). The algorithm is deterministic; the seeds are recorded only for protocol uniformity.

H BG-SINDY PRUNING BEHAVIOR

The pruning trajectories explain where BG-SINDy's balance-guided ranking fails. On Heat, the initial condition is a single mode $\sin(2\pi x)$, so the observed field keeps u_{tt} and u_{xx} exactly collinear (measured correlation -1.000, $u_{tt} \approx -9.9u_{xx}$); the full-library least-squares fit is then degenerate and redistributes the diffusion coefficient between the two columns, fitting $\xi_{xx} = -0.005$ against $\xi_{tt} = 0.050$ at a residual of 10^{-26}. The relative-importance ranking therefore places u_{tt} above u_{xx}, and the pruning removes u^2, $u \cdot u_x$, u_x, u_{xxx}, and finally u_{xx} in that order, each at an essentially unchanged residual: u_{tt} substitutes exactly for the removed term, so the $\tau = 3$ residual-ratio criterion never trips, and the selected model is $\{u_t, u_{tt}\}$ (TPR 0.50, coefficient error 0.45). The same data supports identification: Vanilla, T-SHRED†, and PC-MDT all recover $u_t = 0.5u_{xx}$ on it (Table 1), so the failure lies in the importance ranking under column degeneracy, not in the data. The collinearity is exact on this field ($u_{tt} = -9.900\,u_{xx}$, measured correlation -1.000), and Appendix M M.2 derives the observed redistribution ($\xi_{xx} = -0.005$, $\xi_{tt} = 0.050$) as the minimum-norm least-squares

solution. On the weak-advection sweep, the initial full-library fit assigns the smallest coefficient mass to the true small term $u \cdot u_x$, which is consequently first in the removal queue after the zero-coefficient u^2; the residual-ratio check refuses that removal, and the search freezes with the dust columns u_{tt}, u_x, u_{xxx} still active ($\varepsilon \leq 0.02$; it over-selects to protect the small term). On Burgers at $\nu = 0.001$ the search freezes one step earlier, keeping u_x: on the steepening front the residual-ratio check refuses the next removal as well. On KdV, no column proxies the dispersion term u_{xxx}, the ranking never inverts, and all five points pass. These trajectories show a deterministic per-problem re-optimization with no learned structure: each new equation re-runs the full algorithm from scratch, and the protection of the small term is purchased where the data are least informative.

I Extended-Library Boundary Test

The main benchmark fixes the seven-term library of Section 3. To check that the category mechanism does not depend on that particular size, we extend the library to eleven terms by adding u^3 and $u \cdot u_t$ to the nonlinear subspace and u_x^2 and u_{xx}^2 to the spatial subspace ($u \cdot u_t$ is classified nonlinear by the product-first rule, since products always carry the nonlinear role regardless of the factor categories). With everything else unchanged (256×256 fields, 500 iterations, five seeds), PC-MDT recovers Burgers, Heat, and Wave exactly: $\mathrm{SR} = 1.00$ with $\mathrm{TPR} = 1.00$ and $\mathrm{FPR} = 0.00$ on all three, and coefficient error at most 2.7×10^{-3} (Burgers) with 1.1×10^{-4} (Wave) and 2.0×10^{-4} (Heat) elsewhere. The subspace competition absorbs the four additional candidates without any change to the architecture.

J PDE-LEARN Re-implementation

The main text lists PDE-LEARN (Stephany & Earls, 2024) at its published values under its original protocol. We additionally ran its three-stage pipeline (neural network derivative estimation followed by thresholded sparse regression) on our benchmark fields under the shared protocol. It recovers Burgers exactly ($\mathrm{SR} = 1.00$, coefficient error 1.1×10^{-2}) but fails Heat, Wave, and KdV: on Heat it over-selects (FPR 0.33), on Wave it admits spurious terms (FPR 0.55), and on KdV it drops the true terms (TPR 0.33). The re-implementation is reported here rather than in Table 1 because its derivative-estimation stage uses its own protocol choices; the qualitative failure pattern, threshold misfires on the small-coefficient term, is consistent with the published picture.On the sweep of Section 5.3 and the $\nu = 0.001$ target of Section 5.4, the pipeline runs unchanged on the same datasets and windows (Appendix D); it keeps the small term while its masked estimate stays above the threshold and drops it below, missing u_{xx} at $\nu = 0.001$ ($\mathrm{SR} = 0$, one wrong term).

K Datasets and Hyperparameters

Table 4 summarizes the four benchmark fields (all 256×256, periodic boundary conditions, finite-difference solvers, and the interior mask $5{:}-5$ for all fits). PC-MDT uses the hyperparameters of Section 4: model width 64, two decoder layers, 500 training iterations per run, five seeds, and the verification statistics of Appendix B. The baseline constants are: Vanilla's STRidge threshold 0.15; T-SHRED†'s single global threshold 0.1; BG-SINDy's $\tau = 3$, $\epsilon = 10^{-12}$, and QR tolerance 10^{-10} (Appendix G); DSR trains the Vanilla architecture (global attention, no decoupling) with its risk-seeking RL loop. The sweep datasets of Section 5.3 share the solver settings and differ only in the swept coefficient and the window of Table 3.

Table 5 reports wall time per run (mean $\pm$ std over seeds, single CPU machine). All methods are evaluated under the same protocol and the same cached fields; wall time is machine-dependent and reported for budget transparency only. The coordinate-descent revisit of Appendix E raises PC-MDT wall times by 5–50% depending on the equation.

L Pseudocode

Algorithms L–L formalize the method of Section 4; Algorithm L transcribes the BG-SINDy re-implementation of Appendix G. Notation follows the main text: Θ is the candidate library, R_t the diagnostic state, $\hat{\boldsymbol{y}}_t$ the per-category selection vector, δ the clipping constant.

Equation	Truth (pivot convention)	Time window	Initial condition
Burgers	$u_t = -u \cdot u_x + 0.1u_{xx}$	$[0, 10]$	sin
Heat	$u_t = 0.5u_{xx}$	$[0, 1]$	$\sin(2\pi x)$
Wave	$u_{tt} = u_{xx}$	$[0, 2]$	$\sin(2\pi x)$
KdV	$u_t = -6u \cdot u_x - u_{xxx}$	$[0, 2]$	sin
Weak reaction	$u_t = 0.5u_{xx} - \varepsilon u^2$	Table 3	non-negative Gaussian
Weak advection	$u_t = -\varepsilon u \cdot u_x + 0.1u_{xx}$	Table 3	sin

Table 4: Benchmark fields. All datasets are 256×256 on periodic domains with finite-difference solvers; the per-point sweep windows are in Table 3.

Method	Burgers	Heat	Wave	KdV
PC-MDT	39.8 ± 1.0	43.1 ± 2.0	66.8 ± 5.4	51.5 ± 1.4
Vanilla	37.8 ± 0.4	36.6 ± 0.1	59.9 ± 0.2	39.1 ± 0.3
DSR	28.3 ± 0.7	38.1 ± 1.0	251.7 ± 47.9	42.2 ± 4.9
T-SHRED†	0.2 ± 0.1	0.2 ± 0.0	0.5 ± 0.1	0.2 ± 0.0
BG-SINDy	0.06	0.05	0.22	0.05
PDE-LEARN	870.6 ± 30.7	924.4 ± 35.3	945.9 ± 27.1	706.4 ± 63.9

Table 5: Wall time in seconds per run (one seed, 500 iterations), mean $\pm$ std over 5 seeds on the recorded CPU runs; BG-SINDy is deterministic and its timing varies with machine load.

Algorithm 1: PC-MDT training (one equation, one seed)

Input: library Θ, data $u(x,t)$, iterations $N = 500$, step limit $H = 40$, clipping $\delta = 10^{-10}$; truth equations and success criterion as in Section 5.1.
Init: pivot selection $\hat{\boldsymbol{y}}_0$ = best single temporal term (Appendix C); decoder parameters θ; verification statistics $R_0 = \mathrm{Stats}(\hat{\boldsymbol{y}}_0, u)$.
for $n = 1, \ldots, N$ **do**
  run one decoding round: **for** $t = 1, \ldots, H$ (stop if converged)
    token embeddings $h_k = e_\varphi(\varphi_k) + e_c(c_k)$ (Eq. 3)
    gated representations h'_c (Algorithm L)
    per-category readout with empty option $\hat{y}_c = \arg\max p_c$ (Eq. 12)
    coefficients $\hat{\boldsymbol{\xi}} = \arg\min_\xi \|(\hat{\boldsymbol{y}}_t \odot \Theta)\boldsymbol{\xi} - \hat{u}_t\|_2^2$ (Eq. 13)
    statistics update $R_{t+1} = \mathrm{Stats}(\hat{\boldsymbol{y}}_t, u)$ (Eq. 15)
  coordinate-descent revisit over the round's best selection (Algorithm L)
  target sequence $\hat{y}^* = \arg\max_{\hat{y} \in \text{round}} \rho(\hat{y})$ (Eq. 16)
  update θ to increase $\sum_t \log \mathrm{Dec}_\theta(\hat{y}^*_t \mid R_t)$
return final selection $\hat{\boldsymbol{y}}$ and its least-squares coefficients.

Algorithm 2: Gated forward pass (Section 4.2)

Input: token representations h_k with category labels $c_k \in \{T, S, N\}$; verification statistics R_t.
Subspace-Masked Self-Attention: for each category c, apply attention restricted to tokens of c only: $A_c = \mathrm{softmax}((Q_c K_c^\top \odot M_c)/\sqrt{d})V_c$ with $M_c[i,j] = \mathbb{1}[c_j = c]$ (Eq. 7).
Gated Cross-Subspace Attention: each subspace queries the other two subspaces, $\hat{h}_c = \mathrm{CrossAttn}(q = h_c, kv = \{h_{c'} : c' \neq c\})$ (Eq. 8); compute gate $\gamma_c = \sigma(w_c^\top[\mathrm{MLP}(R_t); \bar{h}_T; \bar{h}_S; \bar{h}_N])$ (Eq. 9); interpolate $h'_c = (1 - \gamma_c)h_c + \gamma_c \hat{h}_c$ (Eq. 10).
Output: gated representations h'_c; the gate value γ_c is recomputed from R_t at every step.

Algorithm 3: Coordinate-descent revisit (Appendix E)

Input: assembled equation $\hat{\boldsymbol{y}}$ from the category-by-category pass, with coefficients refit by least squares.
for each subspace $c \in \{T, S, N\}$ **do**
  for each candidate k of subspace c (including the currently selected one) **do**
    swap k into the c-slot, refit all coefficients by least squares, record residual
  replace the slot only if the new residual improves on the old one by a strict relative margin $> 10^{-5}$ **and** the refit coefficient of the replacement satisfies $|\hat{\xi}| > 10^{-3}$.
return the revisited equation (adds ≈ 7 least-squares fits per decoding round; applied uniformly to every equation and sweep point).

Algorithm 4: BG-SINDy (transcription of Dang et al., 2026, Eqs. 4–11) (Appendix G)

Input: library Φ, target vector (pivot convention, $-u_t$ or $-u_{tt}$ for Wave), residual-ratio threshold $\tau = 3$, $\epsilon = 10^{-12}$, QR tolerance 10^{-10}.
1. form Φ and the target vector;
2. maximal linearly independent subset by QR factorization with column pivoting (relative tolerance 10^{-10});
3. active-set least-squares fit $\boldsymbol{\xi}$;
4. per-sample local importance $w_{ij} = |\Phi_{ij}\xi_j|/(\max_{l\in A}|\Phi_{il}\xi_l| + \epsilon)$, averaged to one global importance per term;
5. remove the least important term, refit; repeat one term per iteration;
6. stop when the residual after a removal exceeds τ times the residual before it; keep the previous iteration.

M WHY THRESHOLD RULES DROP SMALL TERMS

This section collects the analytical statements behind the failure patterns of Sections 5.3 and Appendix H.

M.1 A hard threshold cannot recover a term whose coefficient is below it. Let the selection rule keep candidate j only if its fitted coefficient satisfies $|\hat{\xi}_j| \geq \lambda$ (Vanilla's STRidge uses $\lambda = 0.15$; T-SHRED†'s single global threshold is $\lambda = 0.1$). On noise-free data the least-squares estimate of a truth coefficient is consistent, $\hat{\xi}_j \to \xi_j^*$, so a truth term with $|\xi_j^*| < \lambda$ is zeroed out for any finite residual floor, however informative its column: the rule cannot tell a small true coefficient from a dust term. The sweep of Section 5.3 realizes this bound pointwise: Vanilla drops u_{xx} on Burgers at every swept ν (all $\nu \leq 0.1 < 0.15$), T-SHRED† drops it from $\nu \leq 0.05$ on, and both drop u^2 on the weak-reaction family from $\varepsilon \leq 0.1$—each failure sits exactly at $\xi^* < \lambda$. PC-MDT has no threshold or global normalization to intersect, so no coefficient value can trigger removal.

M.2 Collinear columns redistribute the fitted coefficients. On the Heat benchmark field the observed solution is a single mode $\sin(2\pi x)$, so the library columns u_{tt} and u_{xx} are exactly collinear, $u_{tt} = -c\,u_{xx}$ with measured $c = 9.900$ and correlation -1.000. The full-library least-squares problem for the pivot target $-u_t = 0.5\,u_{xx}$ is then $\min_{\xi_{tt},\xi_{xx}} \|(\xi_{tt}u_{tt} + \xi_{xx}u_{xx} + 0.5u_{xx})\|_2^2 = \min \| - c\,\xi_{tt} + \xi_{xx} + 0.5 \|_2 \|u_{xx}\|_2$, degenerate along the line $\xi_{xx} = c\,\xi_{tt} - 0.5$ (residual exactly 10^{-26}). The minimum-norm solution splits the diffusion coefficient between the two columns,

$$\xi_{tt} = \frac{0.5c}{1+c^2} = 0.0500, \qquad \xi_{xx} = -\frac{0.5}{1+c^2} = -0.0050,$$

which is the fit observed in Appendix H. The BG-SINDy importance ratio then becomes $\frac{w_{tt}}{w_{xx}} = \frac{|\varphi_{tt}\xi_{tt}|}{|\varphi_{xx}\xi_{xx}|} = \frac{c|u_{xx}|\cdot 0.0500}{|u_{xx}|\cdot 0.0050} \approx 99$, so the balance-guided ranking places u_{tt} an order of magnitude above the true diffusion column and the pruning removes u_{xx} while u_{tt} substitutes for it at an essentially unchanged residual—the $\tau = 3$ residual-ratio criterion never trips. PC-MDT is exposed to the same columns, but its selection probabilities are normalized per category and trained against the structure-only score of Eq. 16; no coefficient importance ranking enters the decision.

M.3 The score is a monotone function of the structure only. For $\rho(\hat{\boldsymbol{y}}) = -\log\max(r(\hat{\boldsymbol{y}}), \delta)$: if $r \geq \delta$ then $\partial\rho/\partial r = -1/r < 0$, and if $r < \delta$ then $\rho = -\log\delta$ is flat, so ρ is strictly decreasing in the residual on its informative range and never diverges. The coefficients are optimized out, $r(\hat{\boldsymbol{y}}) = \min_{\boldsymbol{\xi}} \|(\hat{\boldsymbol{y}} \odot \Theta)\boldsymbol{\xi} - \hat{u}_t\|_2^2$, so the score is a function of the selected structure alone, which is what the maximum-likelihood objective of Eq. 16 optimizes.

M.4 The per-category assembly shrinks the search space and bounds the suppression. A global selection over the seven candidates spans $2^7 = 128$ subsets. The per-category heads with one learned empty option per category span $(2{+}1)(3{+}1)(2{+}1) = 36$ assemblies, and the empty option keeps sparse truths representable: without it every category would contribute at least one term, and the two-term Wave equation $\{u_{tt}, u_{xx}\}$ could never be assembled. Each head normalizes over at most $K_c + 1 \leq 4$ entries, so the logit-gap bound of Eq. 6 holds only within a category: the largest possible gap between two candidates of the same category is bounded by the head's own entries, and no cross-category gap exists at all.

M.5 GADA overhead. The masked layer scores $\sum_c K_c^2$ token pairs instead of $(\sum_c K_c)^2$ dense pairs (17 versus 49 for the seven-term library); the cross-subspace layer and the gate add one $O(K^2)$ pass and one MLP evaluation per step, so the total cost stays in the same asymptotic class as global attention with a small constant factor. Measured wall time per run is reported in Table 5 (Appendix K).

N SEED-LEVEL RESULTS

Table 6 reports the per-seed success counts and coefficient errors behind Table 1: SR is the number of successful seeds out of five, RE is the mean over the five seeds of the same relative coefficient error as Table 1. Because the least-squares coefficients are deterministic given the selected term set, RE varies across seeds only where different seeds select different term sets (the largest such standard deviation over all cells is 0.11, on PDE-LEARN Wave). Table 7 lists, per method and equation, the union of the term sets the five seeds selected; the ground-truth sets are given below the table.

Table 6: Per-seed success counts (n/5) and mean RE over seeds, per method and equation (shared success criterion of Section 5.1). PDE-LEARN is our re-implementation of Appendix J; the starred row of Table 1 carries its published values.

	Burgers		Heat		Wave		KdV	
Method	SR	RE	SR	RE	SR	RE	SR	RE
PC-MDT	5/5	1.5e-4	5/5	2.0e-4	5/5	2.7e-5	5/5	1.3e-5
Vanilla	0/5	5.5e-1	5/5	2.0e-4	5/5	2.7e-5	5/5	1.3e-5
T-SHRED†	5/5	1.5e-4	5/5	2.0e-4	0/5	1.0e+0	5/5	1.3e-5
DSR	0/5	1.6e+1	0/5	9.3e+0	0/5	0.00	0/5	9.5e+1
BG-SINDy	5/5	1.5e-4	0/5	4.5e-1	5/5	0.00	5/5	1.3e-5
PDE-LEARN	5/5	1.1e-2	0/5	3.1e-1	0/5	4.7e-1	0/5	9.9e-1

Table 7: Recovered equations: union of the term sets selected by the five seeds, per method and equation. Ground truth: Burgers $\{u_t, u \cdot u_x, u_{xx}\}$; Heat $\{u_t, u_{xx}\}$; Wave $\{u_{tt}, u_{xx}\}$; KdV $\{u_t, u \cdot u_x, u_{xxx}\}$.

Method	Burgers	Heat	Wave	KdV
PC-MDT	$\{u_t, u_{xx}, u \cdot u_x\}$	$\{u_t, u_{xx}\}$	$\{u_{tt}, u_{xx}\}$	$\{u_t, u_{xxx}, u \cdot u_x\}$
Vanilla	$\{u_t, u \cdot u_x\}$	$\{u_t, u_{xx}\}$	$\{u_{tt}, u_{xx}\}$	$\{u_t, u_{xxx}, u \cdot u_x\}$
T-SHRED†	$\{u_t, u_{xx}, u \cdot u_x\}$	$\{u_t, u_{xx}\}$	$\{u_t\}$	$\{u_t, u_{xxx}, u \cdot u_x\}$
DSR	$\{u_t, u_{tt}, u_x, u_{xx}, u_{xxx}, u^2, u \cdot u_x\}$	$\{u_t, u_{tt}, u_{xx}, u \cdot u_x\}$	$\{u_t, u_{tt}, u_{xx}, u_{xxx}\}$	$\{u_t, u_{tt}, u_x, u_{xx}, u_{xxx}, u^2, u \cdot u_x\}$
BG-SINDy	$\{u_t, u_{xx}, u \cdot u_x\}$	$\{u_t, u_{tt}\}$	$\{u_{tt}, u_{xx}\}$	$\{u_t, u_{xxx}, u \cdot u_x\}$
PDE-LEARN	$\{u_t, u_{xx}, u \cdot u_x\}$	$\{u_t, u_{tt}, u_{xx}, u_{xxx}\}$	$\{u_t, u_{tt}, u_x, u_{xx}, u_{xxx}, u \cdot u_x\}$	$\{u_t, u_x\}$

The failing cells isolate the mechanism behind each failure: Vanilla drops u_{xx} on Burgers in all five seeds; T-SHRED† collapses Wave to its u_t anchor in all five seeds; DSR over-selects on every equation and on Burgers and KdV selects the entire library; BG-SINDy keeps $\{u_t, u_{tt}\}$ on Heat (the pruning trajectory of Appendix H); PDE-LEARN drops $u \cdot u_x$ and u_{xxx} on KdV.

O GATE DYNAMICS

Figure 5 shows the gate values $\gamma_T, \gamma_S, \gamma_N$ over the 500 training iterations (per-seed trajectories in faint lines, means in bold). The learned gate settles at the coupling each equation needs: on Heat all three gates collapse to zero (pure diffusion needs no cross-category coupling); on Wave the temporal and spatial gates rise to 0.99 while the nonlinear gate stays at zero, forming exactly the u_{tt}–u_{xx} coupling the wave equation requires; on Burgers the spatial and nonlinear gates settle at 0.72, partial coupling, strictly between the $\gamma = 1$ of full_cross and the $\gamma = 0$ of no_cross ablation of Table 2, both of which fail; on KdV all gates settle near 0.1. The gate values are recomputed from the diagnostic state at every step, so the coupling strength is read from the residual evidence of the equation being searched, not set by hand.

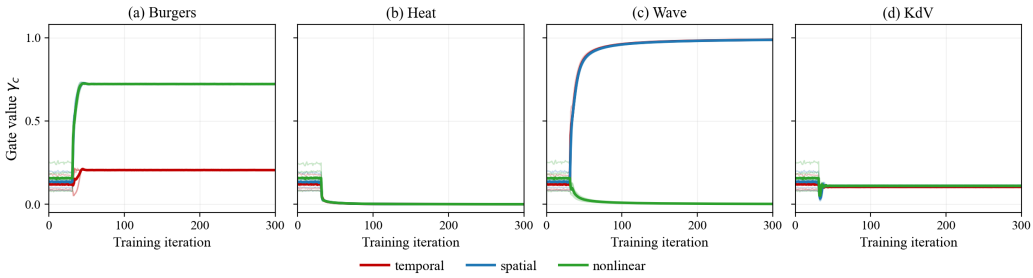


Figure 5: Gate evolution γ_c over the 500 training iterations for PC-MDT, one panel per equation, mean over 5 seeds (bold) with per-seed trajectories (faint). The gate settles at the coupling each equation needs: Heat ≈ 0 everywhere; Wave 0.99 on the temporal–spatial pair and 0 on the nonlinear subspace; Burgers 0.72 on the spatial–nonlinear pair; KdV ≈ 0.1.

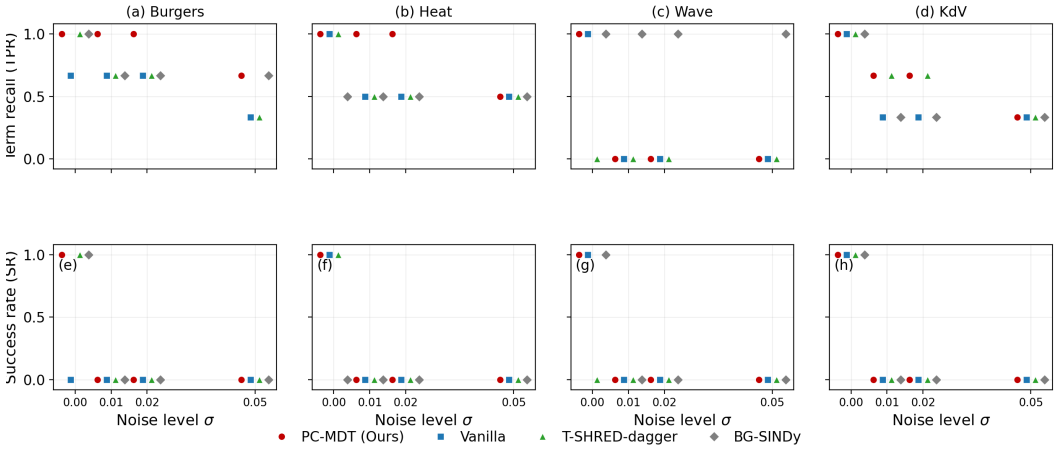


Figure 6: Structure recall (TPR, top row) and success rate (SR, bottom row) under additive Gaussian noise of level σ (relative to the field standard deviation), one column per benchmark equation, 5 seeds per point. The $\sigma = 0.00$ column is Table 1; markers only, no connecting lines, since many configurations share identical values. All methods collapse to SR $= 0.00$ at every $\sigma > 0$; TPR is the only quantity that separates the methods (e.g. PC-MDT keeps the Burgers/Heat structure at $\sigma \leq 0.02$ but loses Wave, while BG-SINDy keeps the Wave structure but fails its coefficients).

P Noise Robustness

Noise injected at the field level is amplified by the finite-difference derivative stencils ($1/dx^2 \approx 6.5 \times 10^4$ on the 256×256 grid): the derivative columns become noise-dominated and the least-squares coefficients shrink systematically (on Heat the fitted u_{xx} drops to -9.3×10^{-5} against -0.5; on Burgers $u \cdot u_x$ drops to 0.23 against 1.0). The success rate therefore falls to zero for every method at every tested $\sigma > 0$; the methods differ only in structure recall, shown in Figure 6 and Table 8 (the $\sigma = 0.00$ row is Table 1). We report this as a shared limitation of field-level symbolic discovery rather than a comparative advantage of any method.

Table 8: Term recall (TPR) under additive Gaussian noise of level σ relative to the field standard deviation (256×256, 5 seeds, shared protocol of Section 5.1). For $\sigma > 0$, $\mathrm{SR} = 0.00$ in all $4 \times 4 \times 3 = 48$ configurations for every method, so only TPR is tabulated.

Equation	σ	PC-MDT	Vanilla	T-SHRED†	BG-SINDy
Burgers	0.00	1.00	0.67	1.00	1.00
	0.01	1.00	0.67	0.67	0.67
	0.02	1.00	0.67	0.67	0.67
	0.05	0.67	0.33	0.33	0.67
Heat	0.00	1.00	1.00	1.00	0.50
	0.01	1.00	0.50	0.50	0.50
	0.02	1.00	0.50	0.50	0.50
	0.05	0.50	0.50	0.50	0.50
Wave	0.00	1.00	1.00	0.00	1.00
	0.01	0.00	0.00	0.00	1.00
	0.02	0.00	0.00	0.00	1.00
	0.05	0.00	0.00	0.00	1.00
KdV	0.00	1.00	1.00	1.00	1.00
	0.01	0.67	0.33	0.67	0.33
	0.02	0.67	0.33	0.67	0.33
	0.05	0.33	0.33	0.33	0.33