NEARESTPDE: STRUCTURED BACKWARD ERROR FOR AMBIGUITY-AWARE PDE DIAGNOSIS

Anonymous authors
Paper under double-blind review

ABSTRACT

We introduce NEARESTPDE, a post-hoc framework for diagnosing computed partial differential equation (PDE) fields through structured backward error. It ranks admissible coefficient, source, boundary, and geometry changes using physically scaled residual sensitivities, retaining indistinguishable explanations in a set. We establish conditions for unique recovery and limits of point diagnosis under observational overlap. Split-conformal calibration provides finite-sample marginal coverage under exchangeability within the declared dictionary. A disjoint residual grid selects a corrective re-solve without test-time reference solutions. Controlled tests recover all 1,980 identifiable cases and flag all 240 collinear controls. On nonlinear finite-difference fields, top-1 accuracy averages 100% across three noise levels, versus 78% for matched orthogonal matching pursuit (OMP). Contextual screening recovers three-cause supports with up to 1,023 localized atoms; OMP remains stronger in several transfer settings. Independent re-solves reduce forward error in 1,820/1,824 one-dimensional transfer cases. The framework separates ranked explanations, calibrated uncertainty, and testable model corrections within the declared perturbation dictionary.

1 INTRODUCTION

A numerical field rarely satisfies its stated partial differential equation (PDE) exactly. Standard diagnostics quantify this mismatch with a residual or an error estimator. A reference solution also permits a forward-error comparison. Those quantities answer *how far* the computation is from an idealized target. They generally do not answer a different question that matters during scientific debugging: *what nearby physical problem is the computed field consistent with?* A large residual may arise from coefficient errors, omitted transport, or incorrect forcing. Boundary errors, misplaced interfaces, and approximation error can also produce large residuals. These causes call for different actions even when their residual norms coincide.

Backward error asks which nearby problem the computed field solves. Structure is essential: an unrestricted pointwise forcing can cancel any discrete residual and yields no diagnosis. This issue is well understood for structured linear systems (Higham & Higham, 1992; Corless & Fillion, 2026), but a practical solver-introspection procedure must also decide which physical perturbations to permit, how to compare their units, and how to report equivalent causes.

The diagnostic requires one computed field, its stated PDE, held-out residual points, and a user-declared perturbation dictionary. We use the random-feature method (RFM) as the primary implementation because its fixed nonlinear features and linear output solve provide analytic spatial jets (Chen et al., 2022). The same post-hoc construction applies when derivatives are supplied by finite differences, finite elements, or a trained tanh field. From the field alone it forms residual sensitivities to diffusion, advection, reaction, forcing, boundary values, local material properties, and a differentiable interface position. Perturbation labels and reference solutions are never test-time inputs. Labels are used only on disjoint calibration and evaluation instances.

Each physical cause spans a range of possible residual changes. Diagnosis is possible only when these ranges are sufficiently separated: a small residual or a sparse coefficient vector cannot distinguish causes whose ranges overlap. We fit every admissible support by scaled least squares, penalize perturbation size and model complexity, and return all near-optimal explanations. Principal angles

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.

expose unstable point labels. For a single sine mode, for example, constant diffusion and reaction have collinear sensitivities and must remain competing explanations.

We make three contributions.

1. **Structured diagnosis.** We formulate nearest-PDE inference over coefficient, source, boundary, local-material, and geometry perturbations. Physical scaling and admissibility turn residual fitting into a comparison of meaningful nearby problems.
2. **Identifiability and reliability.** A subspace-margin theorem gives a sufficient condition for point recovery, while an overlap construction shows when no point selector can be uniformly correct. Calibrated explanation sets then provide finite-sample marginal coverage under exchangeability.
3. **Actionable evaluation.** Matched baselines, exact ambiguity controls, solver transfer, and dictionary shift identify where point recovery succeeds and fails. Independent re-solves test whether the selected physical change improves the solution.

The central modeling choice is to represent each candidate cause by the range it can generate in residual space. This gives coefficients, sources, boundary conditions, and geometry a common representation without pretending that their raw parameters are directly comparable. It also ties the form of the answer to the evidence: separated ranges support a point label, overlapping ranges call for a set, and residual left outside every candidate range indicates a missing mechanism. Because a fitted cause specifies a concrete nearby PDE, the diagnosis can be tested by re-solving that PDE. This chain from residual geometry to uncertainty and then to action is the main distinction from ordinary sparse selection over a residual library.

The experiments cover manufactured linear elliptic and one-dimensional nonlinear parabolic problems. A new equation family requires its own perturbation dictionary and calibration data. Within that scope, the method returns either a testable physical diagnosis, a set of indistinguishable causes, or an explicit rejection.

2 Related work

Structured backward error. Classical backward analysis asks for the smallest perturbation under which a computed object becomes exact. Higham & Higham (1992) show that linearly parameterized matrix structure converts componentwise backward error to an underdetermined system. Corless & Fillion (2026) develop the broader perturbative viewpoint for differential equations. Equation-error methods also fit PDE coefficients by minimizing a governing-equation defect (Acar, 1993; Hanke & Scherzer, 1999). Modern finite-element analyses give convergence rates for related coefficient-identification objectives (Jin & Zhou, 2021). These methods reconstruct a preselected distributed coefficient from external observations. We diagnose one computed field over a finite union of competing coefficient, source, boundary, and geometry models, and return an equivalence-aware set when the cause is not identifiable. Modified-equation analyses derive a mesh-dependent formal equation from a known scheme (McLachlan & Offen, 2022; Zhu et al., 2024). Our field-level residual construction treats the forward solver as a black box, enforces physical admissibility, and tests the selected change by re-solving. The new object is a set-valued backward diagnosis across heterogeneous PDE perturbations.

Random features for forward and inverse PDEs. RFM collocation combines fixed nonlinear features with a linear output solve (Chen et al., 2022). Random-feature maps have also been used as data-driven surrogates between function spaces (Nelsen & Stuart, 2021). Interface-aware RFMs use separate feature representations and interface conditions to handle low regularity and complex geometry (Chi et al., 2024). Inverse random-weight solvers jointly estimate declared PDE parameters and a solution from measurements (Dong & Wang, 2023). Our inference uses only the solver output and the known nominal problem. It selects the nearest member of a competing structured family without external measurements.

Equation discovery and model discrepancy. Sparse equation-discovery methods infer governing terms from spatiotemporal measurements, using strong or weak libraries (Rudy et al., 2017; Messenger et al., 2022). Recent group-sparse weak-form methods also recover spatially varying coefficients

from noisy trajectories (Tang et al., 2026). Neural approaches can learn an unknown operator before symbolic compression (Stephany & Earls, 2022; Podina et al., 2023), while universal differential equations embed learnable components in mechanistic models (Rackauckas et al., 2020). Numerical differentiation followed by subspace pursuit provides a particularly relevant sparse baseline for noisy time-dependent observations (He et al., 2022). These methods discover an unknown equation from external spatiotemporal data. Nearest-PDE inference starts from a known nominal equation and asks which small change explains one computed field. Matched OMP/lasso baselines isolate support selection under the same residual and dictionary. The distinction is also related to calibration with model discrepancy: a flexible discrepancy absorbs mismatch but can be confounded with calibrated parameters (Kennedy & O'Hagan, 2001). Our fixed groups make the confounding visible through range angles and explanation sets.

Recent scientific machine-learning benchmarks clarify the neighboring tasks. PDEBench evaluates forward surrogate accuracy and efficiency across a broad PDE collection (Takamoto et al., 2022). PHYSPDE benchmarks physical-hypothesis selection for PDE discovery from data (Feng et al., 2025), and Fan et al. (2026) use inference-time defect correction to improve a PDE solution. Our task diagnoses one returned field, represents non-identifiability by a calibrated set, and tests the selected nearby PDE through a new solve. We report support recovery, set coverage, abstention, and corrective error separately.

Residual diagnostics and certification. Residual-based stability estimates can connect PDE defects to solution error under operator-dependent assumptions. Analogous analyses exist for physics-informed neural approximations (De Ryck et al., 2022). Functional a posteriori estimators provide sharper scalar certificates (Fanaskov et al., 2024), and empirical work documents optimization and loss pathologies in physics-informed models (Basir & Senocak, 2022). Our goal is complementary: identify a plausible physical cause. Residual-spectrum diagnostics decompose neural-solver defects by spatial frequency (Baheri, 2026). Our projection instead asks which physically parameterized perturbation ranges explain the defect. Selective prediction trades coverage for controlled risk (Geifman & El-Yaniv, 2017). Our coherent singleton gate adopts that evaluation view, while the calibrated explanation set retains several causes when point prediction is not justified. The methodological contribution is their combination: a post-hoc union of heterogeneous nearby-PDE models, explicit observational equivalence, calibrated set output, and label-free selection of a corrective re-solve.

3 STRUCTURED NEAREST-PDE INFERENCE

3.1 PROBLEM STATEMENT

Let $\mathcal{F}_\theta(u) = 0$ denote a nominal PDE with its boundary conditions and let u_M be any differentiable computed field. We evaluate its weighted residual on inference points that were not used by the forward solve,

$$r = W\mathcal{F}_\theta(u_M) \in \mathbb{R}^n, \tag{1}$$

where W balances interior and boundary observations. A perturbation group g specifies physical coordinates p_g, a diagonal matrix D_g of physical scales, admissibility constraints, and a residual sensitivity matrix A_g. We write $p_g = D_g z_g$, so z_g is dimensionless and comparable across coefficient, forcing, and boundary units. For operator and data parameters that enter linearly, $A_g p_g$ is exact for fixed u_M. The interface-position group uses a first-order tangent. The resulting diagnostic is exact in residual-linear coordinates and local in tangent coordinates.

For a support S containing at most K groups, concatenate A_g and D_g and solve

$$\widehat{z}_S \in \arg\min_z \|r + A_S D_S z\|_2^2 + \gamma \|z\|_2^2, \qquad \widehat{p}_S = D_S \widehat{z}_S. \tag{2}$$

We solve Eq. (2) by singular value decomposition (SVD) and assign each admissible support the score

$$q_S = \frac{\|r + A_S D_S \widehat{z}_S\|_2}{\max(\|r\|_2, \eta)} + \lambda_p \|\widehat{z}_S\|_2 + \lambda_c \frac{d_S \log n}{n}, \tag{3}$$

where η is a residual noise floor and d_S is the number of scalar coordinates. The perturbation term therefore measures change in units of the chosen physical scales instead of mixing raw coefficients with different units. Diffusion candidates must remain positive on a check grid, and the interface

center must remain inside the domain. The null support is selected when $\|r\|_2 \le \tau\eta$. Otherwise, the minimum-score support is top-1.

The three score terms answer different questions. The normalized remainder asks whether a support explains the observed defect; the scaled coefficient norm asks how far the fitted PDE is from the nominal one; and the complexity term asks whether extra mechanisms are warranted. The remainder alone favors flexible, high-dimensional supports, while an unscaled coefficient norm makes the ranking depend on units. Their combination implements the backward-error preference for the closest admissible explanation that fits the residual.

Top-1 is accompanied by an explanation set. If $q_\star$ is the best admissible score, write the excess score as $\Delta_S(r) = q_S(r) - q_\star(r)$. For a threshold $\kappa \ge 0$, the explanation set is

$$\mathcal{E}_\kappa(r) = \{S : \Delta_S(r) \le \kappa\}. \tag{4}$$

Exploratory summaries use the fixed threshold $\kappa = \max(\epsilon_{\text{abs}}, \epsilon_{\text{rel}} \max(|q_\star|, \eta))$. Coverage claims instead use the disjoint-calibration quantile $\widehat{q}$ in Theorem 3; only the latter carries a coverage guarantee. We write $\mathcal{E}(r)$ for the set used by the stated protocol. Every member includes coefficients, units, remaining residual, effective rank, and condition number. Appendix A summarizes the complete procedure.

For large localized dictionaries, let $\mathcal{G}_0$ and $\mathcal{G}_{\text{loc}}$ be the global and localized groups. We screen a local group after conditioning on up to $K-1$ global causes,

$$\widetilde{q}_g^{(K)} = \min_{T \subseteq \mathcal{G}_0,\, |T| \le K-1} q_{T \cup \{g\}}, \qquad g \in \mathcal{G}_{\text{loc}}. \tag{5}$$

We then run exact search on the best L atoms. The cost is $O(|\mathcal{G}_{\text{loc}}|\,|\mathcal{G}_0|^{K-1})$ small fits. Unlike singleton screening, this score can expose a cause masked by the others. Theorem 6 states a sufficient retention margin.

The support label has two levels. An *atom* identifies a particular dictionary element, such as a Gaussian center, while its *family* forgets that coordinate and retains "localized diffusion." Atom coverage implies family coverage under this map, but not conversely. If a physical center is off-grid, there is no true atom in the dictionary, so we report family coverage and localization error rather than calling family inclusion exact-atom recovery.

A three-way split separates threshold tuning, risk certification, and final evaluation. The tuning split freezes the set threshold, remainder gate, and a finite confidence-threshold list. A disjoint certificate split selects among those thresholds. Offline calibration and evaluation use hidden labels, but test-time construction of r, ranking, rejection, and correction is label-free. The full algorithm and split rule are in Appendix B.

3.2 Why structure and set-valued output are necessary

An unrestricted alternative is non-informative. If the candidate dictionary contains one forcing value per interior observation, its sensitivity includes $-I_n$ and can cancel any interior residual. We use this case only as a negative control. Meaningful backward error requires a specified perturbation class.

For a fixed support and $\gamma = 0$, the following elementary result makes the identifiability boundary explicit.

Proposition 1 (Discrete characterization)**.** *For every finite support S, a minimizer of Eq.* (2) *exists. Writing $X_S = A_S D_S$, the minimum-norm solution is $-X_S^+ r$, and the minimum unexplained residual is $\left\|(I - X_S X_S^+) r\right\|_2$. The normalized coefficients are unique if and only if X_S has full column rank. Two supports are observationally equivalent when their scaled sensitivity ranges coincide.*

Proposition 2 (Invariance to parameter units)**.** *Let a change of units be $p'_S = U_S p_S$ for a positive diagonal U_S, and set $A'_S = A_S U_S^{-1}$ and $D'_S = U_S D_S$. Then $A'_S D'_S = A_S D_S$, so Eq.* (2)*, every score q_S, the selected support, and $\mathcal{E}(r)$ are unchanged (up to numerical roundoff).*

For $X_S = A_S D_S$, write $P_S = X_S X_S^+$ and $\pi_S = \lambda_p \|\widehat{z}_S\|_2 + \lambda_c d_S \log(n)/n$.

Theorem 1 (Subspace-margin support recovery)**.** *For $\gamma = 0$, suppose $r = -X_{S_\star} z_\star + e$ with admissible $z_\star$ and $\|e\|_2 \le \varepsilon$, and set $m_\star = \min_{T \ne S_\star} \|(I - P_T) X_{S_\star} z_\star\|_2$ and $d = \max(\|r\|_2, \eta)$. Residual-only selection uniquely recovers $S_\star$ if $m_\star > 2\varepsilon$. The scored rule in Eq.* (3) *also recovers it when $\max_{T \ne S_\star}(\pi_{S_\star} - \pi_T) < (m_\star - 2\varepsilon)/d$.*

Theorem 2 (Overlapping-support impossibility). *If $v_S \in \mathrm{range}(X_S)$ and $v_T \in \mathrm{range}(X_T)$ satisfy $\|v_S - v_T\|_2 \le 2\varepsilon$, one residual observation is compatible with either support under error at most ε. No deterministic point selector is uniformly correct. Under an equal prior its error is at least $1/2$. For equal-amplitude one-dimensional signals at angle ϑ, overlap occurs when $a \sin(\vartheta/2) \le \varepsilon$.*

The recovery and impossibility results mark the boundary between justified point output and a necessary set. The proofs and angle specialization are in Appendix A.

Theorem 3 (Calibrated explanation-set coverage). *Fix the dictionary, score, and fitting rule independently of m exchangeable calibration pairs (r_i, T_i) and a test pair (r_{m+1}, T_{m+1}), with every T_i in the declared support family. Let $\Delta_i = q_{T_i}(r_i) - \min_S q_S(r_i)$, $k = \lceil (m+1)(1-\alpha) \rceil$, and let $\widehat{q}$ be the kth calibration order statistic, with $\widehat{q} = +\infty$ if $k = m+1$. Then $\mathcal{E}_{\mathrm{cal}}(r) = \{S : q_S(r) - \min_H q_H(r) \le \widehat{q}\}$ satisfies $\Pr\{T_{m+1} \in \mathcal{E}_{\mathrm{cal}}(r_{m+1})\} \ge 1 - \alpha$.*

The split-conformal rank argument (Lei et al., 2018) gives only marginal coverage under exchangeability, not protection against an undeclared mechanism or conditional coverage for every PDE.

Point selection is modular. We can pair any selector h—including OMP with the Bayesian information criterion (BIC)—with $\mathcal{E}_{\mathrm{cal}}$, then accept a point diagnosis only when the singleton set agrees with h and passes remainder and confidence gates. This leaves both the point label and structured set unchanged (Appendix A). For a non-singleton set, a second collocation grid selects the candidate with smallest validation defect. Theorem 7 bounds its regret relative to the best set member, and instance-randomized binomial bounds control selective risk.

We report effective rank and condition number within supports and principal angles between groups. Small angles trigger an identifiability warning for the observed field, grid, and perturbation dictionary (Appendix A).

Appendix A bounds simultaneous perturbations of r and $X_S = A_S D_S$, separating derivative error, ridge conditioning, and score margin. Its RFM consequence preserves ranking when a weighted C^2 jet error is smaller than the observed margin.

3.3 Perturbation groups

In one dimension the nominal operator is

$$L_\theta u = -(a u_x)_x + b u_x + c u, \qquad u(0) = g_0, \quad u(1) = g_1. \tag{6}$$

For a fixed u_M, scalar diffusion and advection produce columns $-u_{M,xx}$ and $u_{M,x}$. Reaction and rank-one forcing produce u_M and $-\psi_f$, respectively. Boundary changes occupy separate endpoint rows. A Gaussian local diffusion element q_j contributes $-(q_j u_{M,xx} + q_j' u_{M,x})$. The 2D construction applies the same divergence- form differentiation to scalar and localized diffusion, two advection directions, reaction, rank-one forcing, and Neumann data.

For the smooth vertical material interface

$$a(x; c) = a_0 + J \tanh((x - c)/\varepsilon), \tag{7}$$

we differentiate $-\nabla \cdot (a \nabla u_M)$ analytically with respect to c. Thus a finite shift obeys $L_{c+\delta c} u_M - L_c u_M = \delta c\, \partial_c L_c u_M + O(\delta c^2)$. The reported displacement is the corresponding local tangent estimate.

3.4 Correction and conditional forward error

After inference, we construct the selected nearby PDE and rerun the forward solver. An independent reference for the perturbed problem evaluates the result but is never used to rank or select explanations.

Let $u_\star$ solve the reported perturbed problem and $\rho = L_{\theta+\widehat{\delta}} u_M - f_{\theta+\widehat{\delta}}$. If the operator is invertible with $\left\|L^{-1}\right\|_{V' \to U} \le C_{\mathrm{cond}}$, then

$$\|u_\star - u_M\|_U \le C_{\mathrm{cond}} \|\rho\|_{V'}. \tag{8}$$

The bound controls the remaining residual for the inferred problem. Comparing $u_\star$ with the nominal solution additionally requires an operator- and data-perturbation estimate; the perturbation norm in Eq. (3) alone is therefore insufficient to bound forward error.

4 EXPERIMENTAL DESIGN

We evaluate point accuracy against matched selectors, calibrated set coverage, reliable abstention, the quality of a corrective re-solve, and transfer across solvers and derivative estimates.

Forward solvers and split integrity. RFM fields use fixed features and a linear SVD solve. The 1D field enforces Dirichlet data by construction. The 2D field uses boundary-adapted spectral features and a Neumann-flux lift. The first two spatial derivatives are analytic. Fit and inference points are disjoint. The full representations and grids are given in Appendix B.

One 2D study includes the manufactured solution modes in every feature draw; three additional PDE instances omit these anchors to test representation sensitivity.

All hidden perturbation types and amplitudes are read only after inference for metric calculation. Forward discrepancy is evaluated against a closed-form manufactured solution where available and otherwise against an independent sparse finite-difference solve on a 65×65 grid. The fit grid, inference grid, and reference solver therefore serve distinct roles.

Controlled splits. The 1D suite contrasts an identifiable two-mode field with a collinear single-mode control. A disjoint OOD split tests omitted local centers and two-cause combinations. Mixed-boundary 2D and smooth-interface suites test localized coefficients and geometry perturbations. Appendix B specifies the equations, amplitudes, randomization design, and sample counts.

Ranking and baselines. Every suite fixes hyperparameters before evaluation. Controlled and cross-solver experiments use $\lambda_p = \lambda_c = 0.02$; paired RFM experiments use $\lambda_p = \lambda_c = 0.01$. An independent penalty study evaluates $\lambda_c = 0.04$. Ridge is 10^{-10}. Relative and absolute explanation gaps are 0.10 and 0.03. The principal-angle warning threshold is 7.5°. Noise floors are 5×10^{-3} in 1D and 10^{-4} in 2D.

All selectors receive the same held-out residual observable, candidate dictionary, physical scaling, and support budget. Besides raw residual, we compare against the simplest dense simultaneous inverse and against group OMP+BIC (Tropp, 2004), group lasso+BIC, and sparse-group lasso+BIC (Yuan & Lin, 2006). Evaluation labels are used only to compute metrics. Diagnostic metrics are top-1 support accuracy, explanation-set coverage and size, and rejection performance. Action metrics are correction win rate and corrected-to-nominal forward-error ratio. We also report computational cost and paired intervals that resample independently generated PDE instances. Solver-feature seeds are treated as repeated measurements rather than independent PDE instances. Controlled runs also include all-one parameter scales to check unit invariance.

Cross-solver tests add finite-difference method (FDM) and finite-element method (FEM) fields, continuous localization, sharper interfaces, and off-dictionary shapes under fixed physical scales. Full protocols and machine-generated result tables are in Appendix B.

Statistical unit and uncertainty. Confirmatory uncertainty is computed at the independently seeded PDE-instance level, never at the row level. Accuracy differences use paired seed-cluster bootstrap intervals and exact paired discordance tests. For selective-risk claims, a disjoint tuning split fixes every acceptance rule. Each certificate cell then draws one task from each of 120 independent PDEs. Exact upper binomial error bounds are union-bounded over all method–threshold candidates. Each certificate draw contributes one task from one PDE; repeated tasks from the same PDE are not counted as independent observations.

5 RESULTS

On nonlinear finite-difference fields, structured top-1 accuracy is 1.000 versus 0.780 for matched OMP. Table 1 also shows regimes where OMP leads and reports set coverage where comparable. Uncertainty uses independent PDE instances; solver-seed sweeps remain descriptive. The following subsections examine identification, reliability, ambiguity, transfer, and action. Figure 1 compares off-dictionary detection and point-versus-set identification.

Table 1: Held-out identification results. Set coverage measures inclusion of the correct support, while top-1 measures point selection. Coherent singleton coverage for literal faults is reported separately in Section 5.

Split	Rows	Structured top-1	Set coverage	OMP+BIC top-1	Replication unit
Controlled 1D identifiable	1,980	1.000	1.000	–	descriptive seeds
Cross-solver 1D declared	1,944	0.967	0.970	**0.981**	6 seeds/cell
Nonlinear FDM, three noise levels	1,728	**1.000**	–	0.780	72 PDEs
Nonlinear RFM, three noise levels	1,728	0.786	0.889	**0.832**	72 PDEs
Literal assembly faults	3,744	**1.000**	–	**1.000**	72 PDEs
Hard 2D, no forced anchors	1,632	0.965	0.971	**0.972**	96 PDEs
Collinear ambiguity controls	240	–	1.000	–	descriptive seeds

Nonlinear rows average the three noise cells; solver-specific strata appear in Appendix Table 10. Dashes indicate that set inclusion is not compared in those rows; the literal-fault coherent singleton coverage is 0.923. Uncertainty is computed at the PDE-instance level. Selective-risk bounds use 120 independent PDEs with one randomized task each (Appendix Table 11).

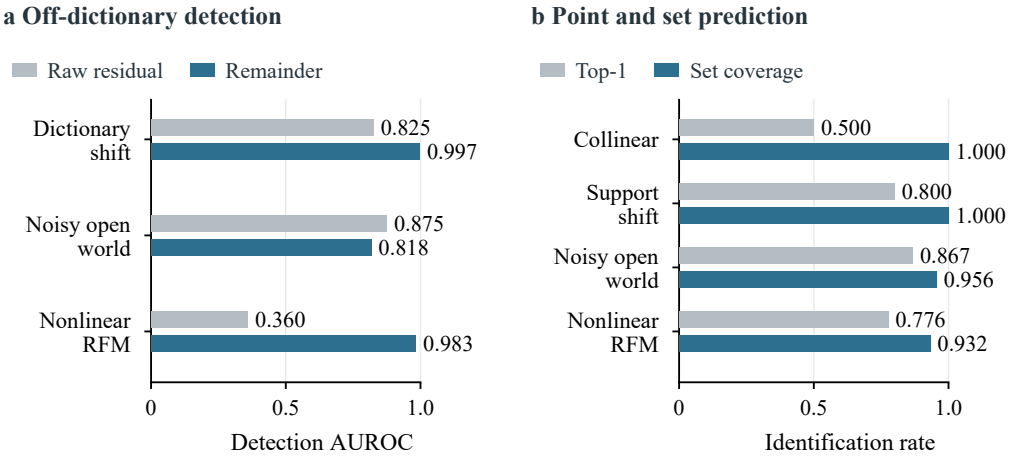

Figure 1: Reliability across evaluation settings. (a) OOD detection measured by area under the receiver operating characteristic curve (AUROC). Remainder improves on raw residual in two settings but trails it under noisy open-world shift. (b) Explanation-set coverage exceeds top-1 accuracy in all four settings. Nonlinear RFM results use the noiseless case.

5.1 Q1: Where does structure help under matched baselines?

On 1,944 declared-support 1D rows, NearestPDE reaches 0.9666 top-1 and 0.9697 set coverage. It exceeds dense inverse by 0.206 (seed-cluster 95% confidence interval (CI) $[0.165, 0.243]$), group lasso by 0.443 $[0.425, 0.471]$, and sparse-group lasso by 0.417 $[0.401, 0.438]$. Group OMP+BIC is stronger at 0.9815 (structured-minus-OMP -0.0149 $[-0.0220, -0.0078]$). On 352 scaled 2D and sharp-interface rows, structured and OMP both obtain $349/352$, versus 0.8182 for dense and 0.6307 for both lasso variants. Changing parameter units by factors from 10^{-3} to 10^{3} preserves every selected support and changes scores by at most 2.22×10^{-16}, confirming the unit invariance in Proposition 2.

The added structure has a measurable computational cost. In the cross-solver suite, median structured inference takes 0.0467 seconds, compared with 0.0019 seconds for OMP. With 1,023 dictionary atoms, contextual screening recovers the correct cause on all tested tasks in 5.60 seconds per task; OMP takes 0.224 seconds but does not recover it. The current implementation therefore trades speed for a more reliable search in this regime.

5.2 Q2: What reliability is gained beyond point ranking?

On 1,312 declared-support evaluation rows, structured and matched OMP are top-1 correct and both reject 384/384 off-dictionary rows. Structured remainder has AUROC/average precision (AP) $0.997/0.986$ versus raw residual's $0.825/0.617$. A 3,744-row fault evaluation gives both selectors top-1 1.000 and coherent coverage 0.923 at zero observed error. Requiring set–point agreement

prevents acceptance when the singleton set conflicts with the point label. Across 48 independently generated unanchored 2D PDEs, structured top-1/set coverage is $0.9596/0.9596$ and OMP top-1 is 0.9602. Their paired difference -0.0006 has instance-bootstrap 95% interval $[-0.0043, 0.0025]$. Here the gain comes from calibrated uncertainty and rejection rather than point accuracy.

An instance-randomized certificate tests the same distinction. On 120 independent PDEs per solver–noise cell, structured FDM accepts 105–112 cases with simultaneous exact risk upper bounds 0.075–0.080. Every RFM rule abstains, so the corresponding risk is undefined; coverage and risk must be reported together.

5.3 Q3: When is a unique cause identifiable?

At zero principal angle, both selectors have 0.500 top-1 but 1.000 set coverage (Figure 1b). From 20° both are 1.000. The controlled benchmark likewise recovers all 1,980 separated cases, flags all 240 exactly collinear controls, and covers all 2,220 supports. Under open-world shift, however, structured identification top-1/coverage are $0.867/0.956$ versus OMP's $0.953/0.966$, and its OOD remainder AUROC/AP $0.818/0.521$ trail raw residual's $0.875/0.798$. The subspace results predict this transition from point recovery to set-valued output. Under misspecification, neither score is uniformly reliable.

5.4 Q4: Where is the structured selector strongest?

Nonlinear results depend on the solver. On FDM fields, structured top-1 is 1.000 averaged across the three noise levels, versus 0.780 for OMP. Single-RFM gives 0.785–0.786 versus OMP's 0.832–0.833, and every RFM gate abstains. Only the paired width-384/mean-five profile beats the baselines (0.938/0.938/0.958 across noise versus OMP's 0.667/0.677/0.771). Its paired width gains are 0.063 $[0.000, 0.125]$, 0.063 $[0.010, 0.125]$, and 0.094 $[0.031, 0.156]$. Ensemble and one-member width intervals contain zero.

The complexity penalty matters in this family: setting $\lambda_c = 0$ drops pooled top-1 from 0.944 to 0.260. On an independent confirmation set, $\lambda_c = 0.04$ is correct on 191/192 PDEs at each noise level versus 0.943–0.953 for $\lambda_c = 0.01$. Paired gains are 0.042–0.052 and all 95% intervals exclude zero. Across 40 seeds, contextual top-16/32/64 screening retains all 600 size–shortlist–seed cases and recovers every three-group support through 1,023 atoms. Singleton top-64 retains only the 40 smallest-dictionary cases. In the controlled scaling suite, contextual screening remains correct from 255 through 1,023 atoms while matched OMP drops to zero accuracy. The independent confirmation set reproduces both effects.

Two boundary tests delimit this result. Finite differences with $h \leq 10^{-2}$ preserve all rankings, but $h = 2 \times 10^{-2}$ lowers both selectors to 0.750. On 96 high-frequency unanchored 2D PDEs, structured top-1/set coverage is 0.9645/0.9706 versus OMP's 0.9724 (paired difference -0.0080 $[-0.0147, -0.0025]$). At noise 10^{-4}, structured/OMP top-1 is 0.9396/0.9583 for RFM and 0.8896/0.9250 for FDM/FEM. OMP remains stronger in these transfer cells.

5.5 Q5: Does diagnosis produce a useful action?

The label-free set action improves 84/84 perturbed fields at each noise level (exact two-sided 95% interval $[0.957, 1.000]$). The median error ratio is 2.72–2.73×10^{-3}. Matched OMP, subset-BIC, and oracle actions also win 84/84 with nearly identical ratios, showing that the action layer works similarly across these selectors. Independent re-solving also improves 1,820/1,824 declared 1D fields, every 2D field, all 512 unanchored-2D corrections, and all 1,008 nonlinear corrections. These re-solves show that the diagnosis suggests a useful intervention within the perturbation dictionary; the conditional error analysis is given in Appendix C.

Figure 2 visualizes this model correction, not numerical refinement of the observed input field. Appendix C provides 2D maps and reconstruction details.

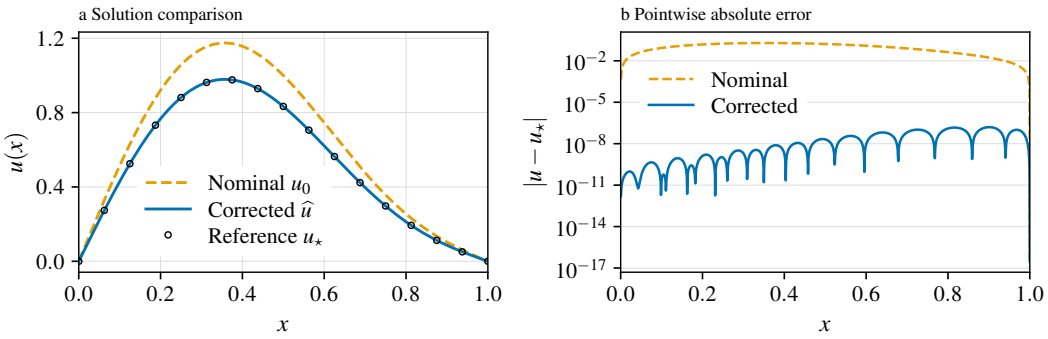

Figure 2: Solution-level model correction: a fixed diffusion perturbation $\delta a = +0.20$, $M = 128$. (a) Nominal and inferred-PDE re-solves use the same basis. The independent reference is used only for evaluation. (b) Relative L^2 error falls from 0.200 to 8.79×10^{-8}. Pointwise errors are shown on a log scale.

6 Discussion and Limitations

What the method adds. Separated sensitivity ranges permit point diagnosis; overlapping ranges call for a set. Physical scaling gives unit invariance, calibration controls marginal set coverage, and a disjoint grid turns an explanation into a candidate action. The structured selector is particularly effective for nonlinear finite-difference fields and large localized dictionaries.

Comparative limits. OMP is more accurate in several transfer settings, raw residual is better for noisy open-world detection, and matched alternatives produce similar corrections. Because point selection is modular, OMP can be paired with the same explanation-set and validation layers where it performs better.

Scope and assumptions. The experiments use manufactured problems; adaptive meshes, neural operators, and real solver incidents remain open. A small remainder means *explained by this dictionary*, which is weaker than causal identification. Point recovery requires range separation, coverage requires exchangeability, and forward-error control requires operator stability.

Failure signals. Small principal angles or a large within-support condition number mean that the field cannot separate the proposed causes. A large unexplained remainder points instead to derivative error or a missing dictionary element, while a failed independent re-solve rejects the fitted local explanation. Reporting these signals separately distinguishes the need for better data, a richer model, or no action.

Using the diagnosis. The dictionary should contain faults that a practitioner can inspect or change, and its scales should encode meaningful physical tolerances. A well-separated singleton supports checking that mechanism first. A non-singleton explanation set instead identifies the small group of mechanisms that requires another measurement or intervention; silently choosing one member would discard information. If every candidate leaves a large remainder, the model should be expanded rather than forced into a familiar label. A disjoint residual grid selects the action, which is then checked by an independent re-solve; reference-based forward error is an evaluation metric. This workflow remains useful even when another point selector is more accurate on a particular benchmark.

7 Conclusion

NearestPDE turns a PDE residual into ranked physical perturbations, returns a calibrated set when the data do not identify one cause, and tests the selected change through an independent re-solve. Its structured selector is strongest for nonlinear finite-difference fields and large multi-cause dictionaries; other point selectors can be used when they are more accurate. This separation between ranking, uncertainty, and action is the main practical contribution.

ARTIFICIAL INTELLIGENCE (AI) USE STATEMENT

OpenAI Codex was used for coding assistance and language editing. The authors reviewed all AI-assisted content and take full responsibility for the manuscript, implementation, and reported results.

REPRODUCIBILITY STATEMENT

Appendix B reports split definitions, randomization design, and hyperparameters. It also specifies solvers and grids, baselines, and record-validation rules. Detailed resource records accompany the reproduction files. Appendix A states the assumptions and full proofs, and the artifact manifest identifies every machine-generated table. Independent PDE instances, rather than multiple tasks from one PDE, are the replication unit for statistical claims.

ETHICS STATEMENT

The study uses synthetic PDE instances and no human participants, personal data, or learned generative datasets. Its main foreseeable risk is over-interpreting a diagnosis as causal. Section 6 and the reported abstention rules delimit that use.

REFERENCES

Robert Acar. Identification of the coefficient in elliptic equations. *SIAM Journal on Control and Optimization*, 31(5):1221–1244, 1993. doi: 10.1137/0331058.

Ali Baheri. What does a neural PDE solver really learn? a residual-spectrum diagnostic. In *ICLR Workshop on AI and Partial Differential Equations*, 2026. URL https://openreview.net/forum?id=e49Aef9YtL.

Shamsulhaq Basir and Inanc Senocak. Critical investigation of failure modes in physics-informed neural networks. *arXiv preprint arXiv:2206.09961*, 2022. URL https://arxiv.org/abs/2206.09961.

Jingrun Chen, Xurong Chi, Weinan E, and Zhouwang Yang. Bridging traditional and machine learning-based algorithms for solving PDEs: The random feature method. *Journal of Machine Learning*, 1:268–298, 2022.

Xurong Chi, Jingrun Chen, and Zhouwang Yang. The random feature method for solving interface problems. *Computer Methods in Applied Mechanics and Engineering*, 420:116719, 2024. doi: 10.1016/j.cma.2023.116719.

Robert M. Corless and Nicolas Fillion. *Perturbation Methods Using Backward Error*. Society for Industrial and Applied Mathematics, 2026. doi: 10.1137/1.9781611978865.

Tim De Ryck, Ameya D. Jagtap, and Siddhartha Mishra. Error estimates for physics-informed neural networks approximating the navier–stokes equations. *arXiv preprint arXiv:2203.09346*, 2022. URL https://arxiv.org/abs/2203.09346.

Suchuan Dong and Yiran Wang. A method for computing inverse parametric PDE problems with random-weight neural networks. *Journal of Computational Physics*, 489:112263, 2023. doi: 10.1016/j.jcp.2023.112263.

Zexi Fan, Yan Sun, Shihao Yang, and Yiping Lu. Physics-informed inference time scaling for solving high-dimensional partial differential equations via defect correction. In *International Conference on Learning Representations*, 2026. Conference proceedings.

Vladimir Fanaskov, Alexander Rudikov, and Ivan Oseledets. Neural functional a posteriori error estimates. *arXiv preprint arXiv:2402.05585*, 2024. URL https://arxiv.org/abs/2402.05585.

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*, 2025. Conference proceedings.

Yonatan Geifman and Ran El-Yaniv. Selective classification for deep neural networks. In *Advances in Neural Information Processing Systems*, volume 30, 2017. Conference proceedings.

Martin Hanke and Otmar Scherzer. Error analysis of an equation error method for the identification of the diffusion coefficient in a quasi-linear parabolic differential equation. *SIAM Journal on Applied Mathematics*, 59(3):1012–1027, 1999. doi: 10.1137/S0036139997331628.

Yuchen He, Sung-Ha Kang, Wenjing Liao, Hao Liu, and Yingjie Liu. Robust identification of differential equations by numerical techniques from a single set of noisy observation. *SIAM Journal on Scientific Computing*, 44(3):A1145–A1175, 2022. doi: 10.1137/20M134513X.

Desmond J. Higham and Nicholas J. Higham. Backward error and condition of structured linear systems. *SIAM Journal on Matrix Analysis and Applications*, 13(1):162–175, 1992. doi: 10.1137/0613014.

Bangti Jin and Zhi Zhou. Error analysis of finite element approximations of diffusion coefficient identification for elliptic and parabolic problems. *SIAM Journal on Numerical Analysis*, 59(1):119–142, 2021. doi: 10.1137/20M134383X.

Marc C. Kennedy and Anthony O'Hagan. Bayesian calibration of computer models. *Journal of the Royal Statistical Society: Series B*, 63(3):425–464, 2001. doi: 10.1111/1467-9868.00294.

Jing Lei, Max G'Sell, Alessandro Rinaldo, Ryan J. Tibshirani, and Larry Wasserman. Distribution-free predictive inference for regression. *Journal of the American Statistical Association*, 113(523):1094–1111, 2018. doi: 10.1080/01621459.2017.1307116.

Robert I. McLachlan and Christian Offen. Backward error analysis for variational discretisations of PDEs. *Journal of Geometric Mechanics*, 14(3):447–471, 2022. doi: 10.3934/jgm.2022014.

Daniel A. Messenger, Emiliano Dall'Anese, and David M. Bortz. Online weak-form sparse identification of partial differential equations. In *Proceedings of Mathematical and Scientific Machine Learning*, volume 190 of *Proceedings of Machine Learning Research*, pp. 241–256. PMLR, 2022. URL https://proceedings.mlr.press/v190/a-messenger22a.html.

Nicholas H. Nelsen and Andrew M. Stuart. The random feature model for input-output maps between Banach spaces. *SIAM Journal on Scientific Computing*, 43(5):A3212–A3243, 2021. doi: 10.1137/20M133957X.

Lena Podina, Brydon Eastman, and Mohammad Kohandel. Universal physics-informed neural networks: Symbolic differential operator discovery with sparse data. In *Proceedings of the 40th International Conference on Machine Learning*, volume 202 of *Proceedings of Machine Learning Research*, pp. 27948–27956. PMLR, 2023. URL https://proceedings.mlr.press/v202/podina23a.html.

Christopher Rackauckas, Yingbo Ma, Julius Martensen, Collin Warner, Kirill Zubov, Rohit Supekar, Dominic Skinner, Ali Ramadhan, and Alan Edelman. Universal differential equations for scientific machine learning. *arXiv preprint arXiv:2001.04385*, 2020. URL https://arxiv.org/abs/2001.04385.

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. doi: 10.1126/sciadv.1602614.

Robert Stephany and Christopher Earls. PDE-READ: Human-readable partial differential equation discovery using deep learning. *arXiv preprint arXiv:2111.00998*, 2022. URL https://arxiv.org/abs/2111.00998.

Makoto Takamoto, Timothy Praditia, Raphael Leiteritz, Daniel MacKinlay, Francesco Alesiani, Dirk Pflüger, and Mathias Niepert. PDEBench: An extensive benchmark for scientific machine learning. In *Advances in Neural Information Processing Systems: Datasets and Benchmarks Track*, volume 35, 2022. Conference proceedings.

Cheng Tang, Roy Y. He, and Hao Liu. WG-IDENT: Weak group identification of PDEs with varying coefficients. *Journal of Computational Physics*, 545:114454, 2026. doi: 10.1016/j.jcp.2025.114454. URL https://doi.org/10.1016/j.jcp.2025.114454.

Joel A. Tropp. Greed is good: Algorithmic results for sparse approximation. *IEEE Transactions on Information Theory*, 50(10):2231–2242, 2004. doi: 10.1109/TIT.2004.834793.

Ming Yuan and Yi Lin. Model selection and estimation in regression with grouped variables. *Journal of the Royal Statistical Society: Series B*, 68(1):49–67, 2006. doi: 10.1111/j.1467-9868.2005.00532.x.

Aiqing Zhu, Sidi Wu, and Yifa Tang. Error analysis based on inverse modified differential equations for discovery of dynamics using linear multistep methods and deep learning. *SIAM Journal on Numerical Analysis*, 62(5):2087–2120, 2024. doi: 10.1137/22M152373X.

A Theoretical Guarantees and Algorithm

Proof of the discrete characterization. Fix a support S and write $A = X_S = A_S D_S$. Because the range of a finite matrix is a closed subspace of $\mathbb{R}^n$, the orthogonal projection of $-r$ onto $\mathrm{range}(A)$ exists. Every least-squares minimizer maps to that projection. The Moore–Penrose solution $-A^+ r$ is the unique minimizer orthogonal to $\mathrm{null}(A)$ and therefore has minimum Euclidean norm. The residual is

$$r - AA^+ r = (I - AA^+)r, \tag{9}$$

up to the sign convention in Eq. (2). Hence exact explanation is possible if and only if $-r \in \mathrm{range}(A)$. Coefficients are unique exactly when $\mathrm{null}(A) = \{0\}$, equivalently when A has full column rank. If two supports have the same range, they attain the same unpenalized residual for every r. No statistic derived only from this residual observable can distinguish them. With ridge $\gamma > 0$, the coefficient vector is unique, but ridge does not restore physical identifiability between nearly coincident ranges.

Proof of the subspace-margin theorem. Let $v = X_{S_\star} z_\star$. The true support leaves residual $(I - P_{S_\star})e$, whose norm is at most ε. For any competing support T, the reverse triangle inequality gives

$$\|(I - P_T)r\|_2 = \|-(I - P_T)v + (I - P_T)e\|_2 \geq \|(I - P_T)v\|_2 - \varepsilon \geq m_\star - \varepsilon. \tag{10}$$

Thus $m_\star > 2\varepsilon$ makes every competing residual strictly larger than the true-support residual for unregularized fitting. The admissibility of $z_\star$ keeps the true-support fitted residual at most ε. After division by the common denominator d, the residual advantage is at least $(m_\star - 2\varepsilon)/d$. The stated penalty condition prevents any competitor from cancelling that advantage in Eq. (3). If both ranges are one-dimensional and v spans the true range, elementary orthogonal projection gives $\|(I - P_T)v\|_2 = \|v\|_2 \sin \vartheta_{S_\star, T}$. This also shows why the conclusion is conditional on the realized field and observation: the same dictionary can have a large margin for one field and zero margin for another.

Proof of the overlapping-support impossibility theorem. Let $r = -(v_S + v_T)/2$. Then

$$r = -v_S + e_S = -v_T + e_T, \qquad e_S = (v_S - v_T)/2, \quad e_T = (v_T - v_S)/2, \tag{11}$$

and both error norms are at most ε. A deterministic selector has one value at r. If it returns S it fails on the T-generated experiment, and otherwise it fails on the S-generated experiment. Thus its maximum error over the pair is one, and its average error under an equal prior is at least $1/2$. For unit vectors spanning two one-dimensional ranges, the chord length is $2\sin(\vartheta/2)$. Multiplying both vectors by amplitude a proves the last condition in Theorem 2. This lower bound and Theorem 1 are complementary: the former rules out uniform point recovery under overlap, whereas the latter certifies recovery outside a score-adjusted separation margin.

Conditional residual-to-forward bound. Let $L : U \to V'$ be the operator for the reported perturbed problem, assume that L^{-1} exists, and let $Lu_\star = f$. For any approximate field u_M, define $\rho = Lu_M - f$. Then

$$u_M - u_\star = L^{-1}\rho, \qquad \|u_M - u_\star\|_U \le \left\|L^{-1}\right\|_{V'\to U} \|\rho\|_{V'}. \tag{12}$$

This is Eq. (8). To compare $u_\star$ with the solution u_0 of $L_0 u_0 = f_0$, one must additionally control $L^{-1}[(f - f_0) - (L - L_0)u_0]$. Neither the support score nor $\|\widehat{z}\|$ contains this condition factor in general.

Proof of calibrated explanation-set coverage. If $k = m + 1$, the infinite threshold includes every declared support and the claim is immediate. Otherwise, condition on the pre-fixed dictionary and scoring rule. Exchangeability makes the rank of Δ_{m+1} among $\Delta_1, \ldots, \Delta_m, \Delta_{m+1}$ uniform when there are no ties and stochastically no larger under the conservative order-statistic convention when ties occur. Therefore $\Pr\{\Delta_{m+1} \le \widehat{q}\} \ge k/(m+1) \ge 1 - \alpha$. The event $\Delta_{m+1} \le \widehat{q}$ is exactly $T_{m+1} \in \mathcal{E}_{\mathrm{cal}}(r_{m+1})$. Labels enter only through the calibration values $\Delta_1, \ldots, \Delta_m$. Evaluating the test set requires only its ranked scores. If the test pair is not exchangeable with calibration, or its true mechanism is outside the dictionary, the theorem makes no coverage claim.

Selective-risk proof and hybrid modularity. For completeness, the reported hybrid output is

$$\big(h(r), \mathcal{E}_{\mathrm{cal}}(r), A_{\widehat{j}}(r)\big), \qquad A_j(r) = \mathbf{1}\{|\mathcal{E}_{\mathrm{cal}}(r)| = 1,\ h(r) \in \mathcal{E}_{\mathrm{cal}}(r),\ R(r) \le t_R,\ c(r) \ge t_j\}, \tag{13}$$

where R is unexplained remainder and c is the point selector's confidence.

Theorem 4 (Instance-randomized selective-risk control)**.** *Let a certified example draw a PDE instance and then one task. A disjoint tuning split fixes $\{A_{hj}\}_{j=1}^{J_h}$ for method–cells $h = 1, \ldots, H$. Each cell then receives one task from each of m independent certificate PDEs. Let U_{hj} be the exact upper binomial error bound at level $1 - \delta/(\sum_h J_h)$, with $U_{hj} = 1$ after zero acceptances. Selecting the greatest-coverage rule with $U_{hj} \le \rho$ (or abstaining) yields, with probability at least $1 - \delta$, selective risk at most ρ for every selected rule simultaneously.*

Proof of Theorem 4. Condition on the tuning split. For method–cell h and candidate j, each $A_{h,j}$ is now a fixed measurable acceptance rule. A certificate observation first draws an independent PDE instance and then one task by the specified randomization. Conditional on $N_{h,j} = n > 0$, the accepted errors are Bernoulli draws with parameter $R_{h,j}$, so the one-sided exact binomial bound satisfies $\Pr\{R_{h,j} > U_{h,j}\} \le \delta/(\sum_h J_h)$. The convention $U_{h,j} = 1$ covers $n = 0$. A union bound over every method–cell and candidate gives simultaneous validity. Thus every selected rule with $U_{h,j} \le \rho$ has $R_{h,j} \le \rho$, even when selection maximizes certificate coverage. If none passes, universal abstention makes no risk claim. Integrating over the independent tuning split proves the result. Multiple tasks evaluated on the same PDE are dependent cluster members and cannot replace independent certificate observations in this argument. □

Proposition 3 (Hybrid modularity)**.** *For any point selector h and any structured set/rejection rule, the first coordinate of Eq.* (13) *equals $h(r)$ on every input. Consequently the hybrid's unconditional top-1 accuracy is exactly that of h, and its explanation-set coverage is exactly that of $\mathcal{E}_{\mathrm{cal}}$. Only selective coverage and selective risk depend on the agreement and remainder gates.*

The statement follows immediately by coordinate projection. It prevents an ambiguous comparison: adding structured diagnostics cannot be credited with OMP's point accuracy, and OMP cannot be credited with structured set coverage.

Unit invariance and joint perturbation stability. For the unit-invariance proposition, write $X_S = A_S D_S$. Under the stated coordinate change,

$$A'_S D'_S = (A_S U_S^{-1})(U_S D_S) = X_S. \tag{14}$$

The optimization variable z_S, its ridge problem, fitted residual, dimensionless norm, and complexity are therefore identical. Admissibility is a condition on the same physical perturbation $p_S = D_S z_S = U_S^{-1} D'_S z_S$, so the admissible support set is unchanged as well. This proves invariance of the complete ranked output, not merely of its least-squares term.

The following result allows derivative error to perturb both the residual and the sensitivity matrix. It is conditional on a fixed support family and on not crossing an admissibility boundary.

Proposition 4 (Joint residual–sensitivity stability). *For each support let $X = X_S$, $\widetilde{X} = X + E$, $\|X\|_2 \le M$, $\|E\|_2 \le \delta$, and $\gamma > 0$. Set*

$$B = (X^\top X + \gamma I)^{-1} X^\top, \quad \widetilde{B} = (\widetilde{X}^\top \widetilde{X} + \gamma I)^{-1} \widetilde{X}^\top, \tag{15}$$

$h = 2M\delta + \delta^2$, $C_B = \delta/\gamma + Mh/\gamma^2$, and $C_P = \delta(M+\delta)/\gamma + MC_B$. If $\widetilde{r} = r + e$ with $\|e\|_2 \le \varepsilon$, then

$$\left\| \widetilde{B}\widetilde{r} - Br \right\|_2 \le C_B \|r\|_2 + (M+\delta)\varepsilon/\gamma, \tag{16}$$

$$\left\| (I - \widetilde{X}\widetilde{B})\widetilde{r} - (I - XB)r \right\|_2 \le C_P \|r\|_2 + \varepsilon. \tag{17}$$

Consequently the non-complexity part of Eq. (3) *changes by at most*

$$C_P + 2\varepsilon/\eta + \lambda_p\{C_B \|r\|_2 + (M+\delta)\varepsilon/\gamma\}. \tag{18}$$

If this bound is L_S on support S, a clean winner T with margin m remains top-1 whenever $m > L_T + \max_{S \ne T} L_S$. The analogous excess-score bound yields membership in the calibrated explanation set.

For the proof, both ridge Gram matrices dominate γI. The resolvent identity and $\left\| \widetilde{X}^\top \widetilde{X} - X^\top X \right\|_2 \le h$ give $\left\| \widetilde{B} - B \right\|_2 \le C_B$. Writing $P = I - XB$ and $\widetilde{P} = I - \widetilde{X}\widetilde{B}$ then gives $\left\| \widetilde{P} - P \right\|_2 \le C_P$. Moreover $\|P\|_2, \left\| \widetilde{P} \right\|_2 \le 1$. The displayed coefficient and fitted- residual bounds follow by adding and subtracting $\widetilde{B}r$ and $\widetilde{P}r$. Finally, $\|Pr\|_2 / \max(\|r\|_2, \eta)$ is $2/\eta$-Lipschitz in r and changes by at most C_P when P changes. Taking supportwise score differences proves the margin statements. When $E = 0$, the result reduces to the earlier residual-only bound. When $\gamma \downarrow 0$, fixed-rank pseudoinverse stability requires a separate singular-gap assumption.

RFM jet and screening certificates. Write $J_2 v = (v, \nabla v, \nabla^2 v)$ on the weighted inference grid and let $\|J_2 v\|_{W,2}$ denote its discrete Euclidean norm. For the groups used here, the residual and dictionary columns are affine combinations of entries of $J_2 v$, with bounded PDE coefficients and basis functions on the compact domain. They are therefore Lipschitz in this norm.

Theorem 5 (RFM jet-to-ranking stability). *Fix a finite support family, ridge $\gamma > 0$, and an admissible reference field u. Suppose*

$$\|J_2 u_M - J_2 u\|_{W,2} \le \delta_M, \quad \|r(u_M) - r(u)\|_2 \le L_r \delta_M, \quad \|X_S(u_M) - X_S(u)\|_2 \le L_{X,S}\delta_M \tag{19}$$

for every support S. Substitute $\varepsilon = L_r\delta_M$ and $\delta = L_{X,S}\delta_M$ in the joint residual–sensitivity bound and call the resulting score bound $L_S(\delta_M)$. If the reference winner T has margin m with

$$m > L_T(\delta_M) + \max_{S \ne T} L_S(\delta_M), \tag{20}$$

then inference from the analytic RFM jet selects the same support T. Explanation-set membership is stable for every support whose reference excess score is separated from the calibrated threshold by more than its two supportwise score bounds.

The first two perturbation inequalities follow from the jet bound with constants determined by the PDE coefficients, weights, and declared basis. The conclusion follows from the preceding proposition. Thus any deterministic or high-probability RFM C^2 estimate transfers to a ranking certificate with the same probability. The theorem does not assume a universal finite-width rate.

Theorem 6 (Top-L screening retention). *Let $s_g^\star$ be an ideal screening score for each localized group and let $\widehat{s}_g$ satisfy $|\widehat{s}_g - s_g^\star| \le \delta_q$ uniformly. The score may be the singleton value $q_{\{g\}}$ or the contextual value in Eq.* (5)*. A required group g is retained by the L smallest computed scores if fewer than L groups h satisfy $s_h^\star \le s_g^\star + 2\delta_q$. If the global pool is always retained and this condition holds for every localized group in a full-search minimizer $S_\star$, exhaustive search on the reduced pool attains the same minimum score as the full search.*

Indeed, $\widehat{s}_h \le \widehat{s}_g$ implies $s_h^\star \le s_g^\star + 2\delta_q$. If fewer than L groups meet the latter condition, g cannot be ranked below position L. Applying the argument to each group in $S_\star$ proves that $S_\star$ remains feasible in the reduced search. The reduced feasible set is contained in the full one, so their minimum scores coincide. This is not an accuracy guarantee when score separation fails, so the experiment records retention directly.

Table 2: Experimental configurations used in the paper.

	1D confirmation	1D OOD	2D variable	2D interface
Solver initializations	20	10	10	10
Widths	32,64,128	32,64,128	64,128	64,128
Fit points	512	512	34^2	34^2
Inference points	384	384	27^2	27^2
Boundary points/side	endpoints	endpoints	48	48
Reference grid	1D adaptive/closed	1D adaptive/closed	65^2	65^2
Maximum groups	1	2	1	1
Rows	2220	150	340	100

Theorem 7 (Set-to-action validation regret)**.** *Let $\mathcal{E}$ be a finite explanation set, let $L(S)$ be the population loss of the corrective action induced by S, and let $\widehat{L}_V(S)$ be its loss on a disjoint validation grid. If*

$$\sup_{S\in\mathcal{E}} |\widehat{L}_V(S) - L(S)| \le \varepsilon_V, \qquad \widehat{S} \in \arg\min_{S\in\mathcal{E}} \widehat{L}_V(S), \tag{21}$$

then

$$L(\widehat{S}) \le \min_{S\in\mathcal{E}} L(S) + 2\varepsilon_V. \tag{22}$$

Consequently, on the joint event that the true support T lies in $\mathcal{E}$ and the uniform bound holds, $L(\widehat{S}) \le L(T) + 2\varepsilon_V$.

Let $S^\star \in \arg\min_{S\in\mathcal{E}} L(S)$. Uniform control and empirical optimality give $L(\widehat{S}) \le \widehat{L}_V(\widehat{S}) + \varepsilon_V \le \widehat{L}_V(S^\star) + \varepsilon_V \le L(S^\star) + 2\varepsilon_V$. The final claim follows because $T \in \mathcal{E}$ makes $S^\star$ no worse than T. The theorem is conditional on a uniform validation bound. It neither assumes nor implies that the explanation set contains the true support under arbitrary distribution shift.

Interface-center derivative. For Eq. (7), with $s = (x - c)/\varepsilon$,

$$\partial_c a = -\frac{J}{\varepsilon}\operatorname{sech}^2(s), \qquad \partial_c a_x = \frac{2J}{\varepsilon^2}\operatorname{sech}^2(s)\tanh(s). \tag{23}$$

For a fixed field, the strong-form tangent column is $-[\partial_c a\,(u_{xx} + u_{yy}) + \partial_c a_x\, u_x]$. Taylor's theorem gives a second-order remainder for finite δc when the coefficient family and field are sufficiently smooth. This explains the shrinkage visible for $|\delta c| = 0.04$ without implying that all interface errors are second order in nonsmooth settings.

Inference and action procedure. Given a computed field, nominal PDE, inference grid, and perturbation dictionary, NEARESTPDE proceeds as follows.

1. Evaluate the weighted residual r and each scaled sensitivity $X_g = A_g D_g$ on the inference grid.
2. Enumerate admissible supports up to size K, using contextual screening in Eq. (5) for large local dictionaries.
3. Solve Eq. (2), reject physically inadmissible fits, and rank the remainder by Eq. (3).
4. Return the best support together with coefficients, units, residual remainder, rank, condition number, and principal-angle warnings.
5. Form the calibrated explanation set from a disjoint calibration split. A point diagnosis is accepted only when the set is a singleton that agrees with the point selector and passes the remainder and confidence gates.
6. If the set contains several candidates, select an action using residual error on a second, disjoint collocation grid, then re-solve the chosen PDE.

B EXPERIMENTAL SPECIFICATION AND REPRODUCIBILITY

Table 2 records the data splits and numerical resolutions. The paragraphs below specify the corresponding equations, amplitudes, and candidate dictionaries.

Controlled split design. The 1D RFM representation is

$$u_M(x) = g_0(1-x) + g_1 x + x(1-x)\sum_{j=1}^{M} \alpha_j \tanh(w_j x + \beta_j). \tag{24}$$

It satisfies Dirichlet data exactly and has analytic first and second derivatives. Output coefficients use column-scaled, ridge-augmented SVD least squares on 512 midpoint collocation points. Inference uses 384 shifted points without overlap. The 2D field uses $\sin(k_x \pi x)\cos(k_y \pi y)$ features and a Neumann-flux lift. Before the lift, it vanishes at $x = 0, 1$ and has zero normal derivative at $y = 0, 1$.

The identifiable 1D probe is $u(x) = \sin(\pi x) + 0.35\sin(2\pi x)$. Its dictionary contains scalar diffusion, advection, reaction, rank-one forcing, left and right boundary data, and five Gaussian diffusion elements. Twenty solver seeds and widths $M \in \{32, 64, 128\}$ produce 1,980 non-ambiguous rows, including 60 nominal controls. Another 240 rows use $u(x) = \sin(\pi x)$, whose diffusion and reaction sensitivities are collinear. The OOD split has ten disjoint seeds, the same widths, and five two-group perturbations (150 rows). Local centers 0.45 and 0.75 lie between the dictionary centers. Reaction plus a local change at 0.45 therefore tests dictionary misspecification. Its maximum support size is two.

The mixed-boundary 2D suite crosses two widths and ten seeds with 17 null, global, boundary-flux, forcing, and localized-diffusion tasks (340 rows). The interface suite has 20 controls and 80 shifts with the same widths and new seeds, testing both a geometry coordinate and finite-shift tangent bias. Its residual observable uses 48 points per boundary segment, with 34^2 fit and 27^2 inference points.

The smooth 2D benchmark on $[0, 1]^2$ uses

$$a(x, y) = 1 + 0.2\sin(2\pi x)\cos(\pi y), \quad b = (0.12, -0.08), \quad c = 0.15, \tag{25}$$

with manufactured field $u(x, y) = \sin(\pi x)\cos(\pi y) + 0.35\sin(2\pi x)\cos(2\pi y)$, Dirichlet conditions at $x = 0, 1$, and Neumann conditions at $y = 0, 1$. The interface benchmark replaces a by $1.1 + 0.25\tanh((x - c)/0.06)$ at nominal center $c = 0.5$ and uses shifts $\delta c \in \{-0.04, -0.02, 0.02, 0.04\}$.

For the 1D confirmation split, scalar amplitudes have both signs and at least two magnitudes. Diffusion and boundary changes use $\{\pm 0.08, \pm 0.20\}$. Reaction and advection use $\{\pm 0.08, \pm 0.25\}$, while rank-one forcing uses $\{\pm 0.04, \pm 0.10\}$. Local diffusion uses the diffusion amplitudes at centers 0.45 and 0.80. The local dictionary is centered at $\{0.20, 0.35, 0.50, 0.65, 0.80\}$ with width 0.12.

The 2D scalar amplitudes are ± 0.12 for diffusion and bottom flux and ± 0.15 for reaction and both advection components. Rank-one forcing uses ± 0.05. Local diffusion uses ± 0.15 at $(0.3, 0.7)$ and $(0.7, 0.3)$ while the candidate dictionary additionally includes $(0.3, 0.3)$ and $(0.7, 0.7)$. All configurations were fixed before evaluation.

Cross-solver protocol. The disjoint cross-solver split adds 3,088 tasks with fresh solver seeds. The 1D core crosses RFM and centered finite-difference fields with two solver resolutions, three field-noise levels, two signed amplitudes, and seven physical groups. The misspecification split holds out a super-Gaussian coefficient shape, fifth-harmonic forcing, and coarse-grid discretization error. Tight and loose score/set thresholds are evaluated without tuning on these labels. The 2D extension repeats the sparse baselines under fixed physical scales and adds a sharper interface of thickness 0.025.

Every comparison consumes the same held-out residual and candidate matrices. Baselines are dense inverse regression, group OMP+BIC, group lasso+BIC, and sparse-group lasso+BIC. An oracle-support residual is evaluation-only. Continuous localization profiles a Gaussian center by bounded scalar optimization. Correction is re-evaluated with both RFM and an independent finite-difference solve. Explanation-set coverage is reported with its mean, median, maximum, and full size distribution. Table 3 (Panels A–C) separates matched-baseline accuracy, robustness and correction, and two-dimensional transfer. Each panel is computed from the complete validated raw records.

Calibration and representation sensitivity. The independent calibration split contains 2,240 one-dimensional rows (544 calibration and 1,696 evaluation), including 128 transfers from independently seeded, fully trainable one-hidden-layer tanh collocation solvers. It further contains 288 anchor-ablation rows, 432 unanchored off-spectrum 2D rows across three PDE instances, six sharp-interface grid-convergence cases, and 72 scaling measurements. All six experimental arms use fixed configurations, and aggregation rejects failed records.

Table 3: (A) Cross-solver comparison under a common held-out residual observable. Differences use a seed-cluster bootstrap. No hidden label selects hyperparameters. (B) Robustness, explanation-set calibration, and independent-solver correction in the cross-solver suite. (C) Two-dimensional and sharp-interface transfer under matched observables and physical scales.

Panel A

Method	Top-1 accuracy	Scaled minus method (95% CI)
NEARESTPDE (physical scales)	0.967	–
NEARESTPDE (all-one scales)	0.969	-0.002 [-0.006,0.000]
Dense inverse	0.760	0.206 [0.165,0.243]
Group OMP+BIC	0.981	-0.015 [-0.022,-0.008]
Group lasso+BIC	0.523	0.443 [0.425,0.471]
Sparse-group lasso+BIC	0.550	0.417 [0.401,0.438]

Panel B

Diagnostic	Result
Explanation-set coverage / mean size	0.970 / 1.28
Unit-rescaled support agreement	1.000
Max unit-rescaled score difference	2.22e-16
Independent FDM correction median ratio / win rate	3.51e-04 / 0.998
Continuous-center median absolute error	2.41e-05

Panel C

Method	Top-1 accuracy	Structured minus method (95% CI)
NEARESTPDE	0.991	–
Dense inverse	0.818	0.173 [0.065,0.367]
Group OMP+BIC	0.991	0.000 [0.000,0.000]
Group lasso+BIC	0.631	0.361 [0.246,0.426]
Sparse-group lasso+BIC	0.631	0.361 [0.246,0.426]
Set coverage / mean size	0.994 / 1.86	–
Independent correction median ratio / win rate	0.010 / 1.000	–

We construct the explanation set as in Theorem 3 with $\alpha = 0.10$ and separately calibrate the 95th percentile of the selected support's unexplained residual fraction. A point diagnosis is accepted only when its explanation set is a singleton and its remainder is below this threshold.

Tables 4, 5, and 6 report one-dimensional calibration, anchor sensitivity, and numerical convergence and cost. The 1D metrics include group coverage, set size, off-dictionary rejection, coefficient error, and localization error. AUROC/AP compares unexplained remainder with raw residual and ensemble disagreement. Wall-clock time records computational cost. For the trainable solver, we also report terminal collocation residual and optimizer convergence. The 2D comparison reports anchor policy, exact solution-mode overlap, top-1, set coverage, and a hierarchy that resamples PDE instances before seeds. The reference study reports successive-grid differences and observed order. The scaling study reports candidate supports, exact design storage, and inference time.

Table 4: Calibration, transfer, and off-dictionary diagnostics on disjoint calibration and evaluation seeds.

Metric	NearestPDE	Group OMP+BIC
1D declared top-1	1.000	1.000
Trainable tanh solver transfer ($n = 128$)	1.000	1.000
OOD rejection	1.000	1.000
Conformal coverage / size	0.965 / 1.00	–
Solver-fault top-1	1.000	–
Median scalar relative error	4.79×10^{-6}	–
Median center absolute error	3.07×10^{-6}	–
OOD remainder AUROC / AP	0.997 / 0.986	–
OOD raw residual AUROC / AP	0.825 / 0.617	–

All 1,312 declared 1D evaluation rows are top-1 correct for both structured selection and admissible OMP, and both reject 384/384 held-out off-dictionary rows. The calibrated group set has coverage

Table 5: Sensitivity to deterministic feature anchors on four 2D PDEs. For all unanchored rows, the structured-minus-OMP difference is -0.017 with hierarchical 95% interval [-0.043, 0.002].

Setting	n	NearestPDE	OMP	Set coverage
PDE 0; anchored	144	1.000	1.000	1.000
PDE 0; unanchored	144	0.979	0.979	0.986
PDE A; unanchored	144	0.854	0.854	0.861
PDE B; unanchored	144	0.882	0.903	0.889
PDE C; unanchored	144	0.715	0.764	0.722

Table 6: Reference convergence and computational scaling. Timings are wall-clock measurements.

Quantity	Result
Maximum $129 \to 257$ reference difference	6.06×10^{-5}
Median observed grid order	1.94
Maximum groups / enumerated supports	37 / 703
Maximum design storage (MiB)	0.109
Median structured / OMP time (s)	0.0467 / 0.0019

0.965 and mean size 1.00. Median scalar relative error is 4.79×10^{-6} and median continuous-center error is 3.07×10^{-6}.

The unanchored condition does not force exact solution modes, although the random basis may still sample nearby modes. Across the three additional PDEs, structured top-1 ranges from 0.715 to 0.882, and explanation-set coverage from 0.722 to 0.889. The drop quantifies sensitivity to the feature representation.

Fault, correction, and computational scaling. Table 7 combines four experiments: literal 1D assembly faults, matched independent-FDM correction, independent 2D PDE instances, and dictionary scaling. It contributes 3,413 records with no failed row. The fault arm uses 416 calibration and 832 evaluation rows across two finite-difference implementations and two resolutions. The 2D arm uses eight calibration and 16 evaluation PDE instances (544 and 1,088 rows), width 192, and no forced exact-mode anchors. The correction arm compares structured, OMP, and oracle supports on the same 512 fields. The scaling arm reaches 1,029 groups, 8,911 admissible supports, and 1.523 MiB of exact design storage.

Selective acceptance requires agreement between the singleton explanation set and the point label. This rule prevents 64 null controls from being accepted with conflicting diffusion labels; the resulting fault coverage is 0.923 with zero observed selective error, equal to OMP.

Table 7: Held-out fault, transfer, correction, and scaling results.

Experiment	Rows	NEARESTPDE	OMP+BIC	Oracle support
Literal assembly faults: top-1	832	1.000	1.000	–
Literal faults: coherent selective coverage	832	0.923	0.923	–
Independent 2D instances: top-1	1088	1.000	1.000	–
Independent 2D instances: set coverage	1088	1.000	–	–
Correction: median FDM ratio	512	0.00019	0.00019	0.00019

Fault selective risks are 0.000 for both methods. Coherent selection requires the singleton set to equal the point prediction. The scaling experiment reaches 1029 groups. For $K = 1$, both methods have accuracy 1.000, while median times at the largest dictionary are 85.9 s and 0.061 s.

Subspace separation and open-world behavior. Table 8 reports identifiability, open-world, and large-dictionary outcomes. The subspace experiment has 8 angles, 4 noise levels, 200 calibration and 400 evaluation vectors per cell (19,200 records). It alternates the two labels at zero angle, so 0.5 point accuracy is the identifiable limit. The scaling arm has six seeds and 63, 127, 255, 511, or 1,023 localized atoms. Exact search is run only through 127, while three shortlist sizes and OMP cover every size. The open-world arm has eight calibration and 16 evaluation seeds, RFM and finite-difference observations, three noise levels, four declared and two off-dictionary mechanisms. Calibration is separate within each solver–noise block.

Table 8: Identifiability, open-world, and large-dictionary results on disjoint calibration and evaluation seeds. Scaling rows use a 16-local shortlist.

Experiment	NEARESTPDE	Matched OMP	Raw residual
Angle $\leq 5^\circ$: top-1 / set coverage	0.860 / 0.989	0.860 / 0.989	–
Angle $\geq 20^\circ$: top-1 / set coverage	1.000 / 1.000	1.000 / 1.000	–
Open-world identification: top-1 / set coverage	0.867 / 0.956	0.953 / 0.966	–
Open-world AUROC / AP	0.818 / 0.521	–	0.875 / 0.798
127 locals: exact / screened median s	2.203 / 0.070	0.018	–
1023 locals: exact-grid accuracy / median s	1.000 / 0.142	0.000 / 0.143	–
1023 locals: median center error	0.00e+00	7.83e-04	–

At zero principal angle, both selectors have top-1 accuracy 0.500 but calibrated set coverage 1.000 with mean size 2.00. Open-world OOD rejection is 0.333 for the remainder and 0.500 for the raw residual. Neither threshold transfers across nonzero noise. At 1,023 atoms, all six OMP exact-grid misses select the immediately adjacent local atom. They are localization offsets, not wrong physical mechanisms.

Table 9: Nonlinear dynamics, derivative stability, and exact three-cause search on disjoint calibration and evaluation PDEs.

Experiment	structured statistic	comparator	diagnostic
Dynamics (RFM), top-1	0.776	0.849 (OMP)	set 0.932/size 5.78
Dynamics (FDM), top-1	1.000	0.875 (group lasso)	set 1.000/size 1.00
Dynamics (RFM), OOD AUROC	0.983	0.360 (raw)	reject 1.000
Analytic-jet check, top-1	1.000	1.000 (OMP)	cert. violations 0
$K = 3$ exact (63 locals), acc.	1.000	1.000 (OMP)	time 4.922/0.014 s

Nonlinear dynamics, derivatives, and multiple causes. The nonlinear protocol fixes 8 calibration and 24 evaluation instances of $u_t - au_{xx} + buu_x + cu = f$ before execution. Each instance randomizes nominal coefficients, initial data, and forcing without consulting task labels. Ten tasks include null, four single causes, two double causes, one triple operator cause, and two off-dictionary forcing shifts. Analytic-jet RFM fields and scheme-consistent finite-difference trajectories are evaluated at three noise levels. Structured enumeration, admissible group OMP+BIC, exhaustive subset BIC, group lasso+BIC, and sparse-group lasso+BIC consume the same observable and dictionary. Calibration is separate within each solver–noise block, and PDE instance is the resampling unit.

The correction experiment compares no change, structured, OMP, subset-BIC, and oracle re-solves on every noiseless in-dictionary evaluation field. It reports relative and absolute L^2/L^∞ error, median, 90th percentile, maximum, win rate, and cost. The RFM-jet arm crosses five widths, three inference counts, and analytic derivatives with three finite-difference steps. The scaling arm uses exact search through 63 local atoms and shortlist sizes 8, 16, and 32 through 1,023 atoms for $K = 3$, recording true-atom singleton rank and screening retention.

All 4,986 rows completed. Structured selection is strongest on finite-difference fields; OMP has higher RFM top-1 accuracy, while the structured remainder has higher RFM OOD AUROC. Every perturbed re-solve improves over no change. No empirical jet-margin certificate is violated, and exact three-support recovery holds through 63 local atoms. Contextual screening at larger scales is reported in Section B. Table 9 summarizes these nonlinear, derivative, correction, and scaling results.

Independent-instance reliability and solver transfer. The reliability experiments contain 5,760 nonlinear rows, 4,992 literal assembly-fault rows, 2,176 generated-2D rows, and 320 ensemble summaries. The nonlinear split uses 24 calibration PDEs and 72 disjoint evaluation PDEs; eight calibration instances tune the set, remainder, and confidence thresholds. The hybrid uses OMP+BIC for its point label and structured inference for set agreement and remainder checks.

Table 10 reports solver-specific accuracy, set coverage, and OOD detection. Gate coverage based on eight tasks per PDE is descriptive. Formal selective-risk certification instead uses 120 independent PDEs per solver–noise cell and one keyed-randomized task per PDE, with familywise $\delta = 0.05$ across methods, cells, and threshold candidates.

The solver-transfer study adds an independently assembled piecewise-linear Galerkin FEM at two resolutions and two noise levels. Table 11 reports the certificate and FEM results. All 720 certificate rows and 3,744 FEM rows are finite and unique.

Table 10: Reliability across independent PDE instances. Nonlinear ranges span root-mean-square (RMS) noise 0, 10^{-4}, and 5×10^{-4}. Gate coverage is descriptive because eight tasks share each of 16 PDEs.

Experiment	structured	comparator	reliability/OOD statistic
Nonlinear FDM top-1	1.000–1.000	0.365–0.988 (OMP)	descriptive gate cov. 0.896–0.924; 0 row errors
Nonlinear RFM top-1	0.785–0.786	0.832–0.833 (OMP)	gate cov. 0; OOD AUROC ≥ 0.993 (raw ≤ 0.380)
Literal assembly faults	1.000	1.000 (OMP)	coherent cov./risk 0.923/0.000
Independent 2D PDEs	0.9596	0.9602 (OMP)	$\Delta = -0.0006$; 48-instance CI contains 0
Matched RFM OOD AUROC	0.9995	0.9957 (ensemble)	raw 0.3719

Table 11: Selective-risk certification and finite-element fault transfer. Each Panel A entry lists top-1 accuracy, accepted certificate instances (of 120), and exact risk upper bound U; a dash indicates no accepted instances. Bounds are simultaneous over four methods, six solver–noise cells, and 13 thresholds per cell. Panel B uses 72 solver seeds per condition with piecewise-linear (P_1) FEM.

Panel A: nonlinear instance certificate (top-1; accepted; U)

Solver	noise RMS	Structured	OMP+BIC	Subset+BIC	OMP–structured hybrid
RFM	0	0.785; 0; –	0.833; 0; –	0.828; 0; –	0.833; 0; –
RFM	10^{-4}	0.786; 0; –	0.832; 0; –	0.826; 0; –	0.832; 0; –
RFM	5×10^{-4}	0.786; 0; –	0.832; 0; –	0.825; 0; –	0.832; 0; –
FDM	0	1.000; 112; 0.075	0.365; 0; –	0.363; 0; –	0.365; 0; –
FDM	10^{-4}	1.000; 108; 0.078	0.988; 107; 0.078	0.988; 106; 0.079	0.988; 102; 0.082
FDM	5×10^{-4}	1.000; 105; 0.080	0.988; 91; 0.092	0.988; 91; 0.092	0.988; 102; 0.082

Panel B: independent P_1 FEM literal faults (top-1 accuracy)

resolution	noise	Structured	OMP+BIC	Dense	group / sparse-group lasso
129	0	1.000	1.000	1.000	0.692 / 0.692
129	10^{-5}	0.956	0.999	0.912	0.708 / 0.696
257	0	1.000	1.000	1.000	0.692 / 0.692
257	10^{-5}	0.944	0.998	0.916	0.706 / 0.698

Structured − OMP at 129, 10^{-5}	-0.043 [-0.051,-0.034]
Structured − OMP at 257, 10^{-5}	-0.053 [-0.061,-0.045]

RFM fidelity, action selection, and contextual screening. The factorial experiment pairs width $192/384$ and one/five RFM members on identical PDE instances and tasks. It reports 17 score, ridge, cardinality, and explanation-gap ablations. No variant is selected using evaluation labels. The action arm uses a third collocation grid, disjoint from fit and inference grids, to choose among the explanation-set candidates before an independent forward re-solve.

The transfer arm evaluates RFM, a fully trainable tanh network, two independently assembled finite-difference schemes, and a piecewise-linear Galerkin field. Every inverse method receives the same residual and dictionary for a row. The scaling arm crosses 63–1,023 local atoms, eight seeds, top-16/32/64 shortlists, and both singleton and Eq. (5) scores. Its primary endpoint is true-local retention. Exact-support accuracy and elapsed time are secondary. The four factorial profiles contribute 432 rows each, solver transfer contributes 576, and scaling contributes 288; all 2,592 rows are complete and finite.

Tables 12 and 13 (Panel A) report the factorial comparison with paired intervals and complexity sensitivity. Tables 15 (Panel A) and 16 (Panel A) cover solver transfer and contextual screening. Table 18 reports set-to-action outcomes. Because the profiles differ in width, fit density, and jet averaging, their contrast measures the combined fidelity change.

Independent stress tests. Five additional configurations test the complexity penalty, harder unanchored 2D PDEs, analytic versus finite-difference jets, five-solver transfer, and contextual $K = 3$ screening. Tables 13 (Panel B), 14, 15 (Panel B), 17, and 16 (Panel B) report the results. Inadmissible candidates have score $+\infty$; all selected fits are finite.

Reproducibility and compute. Reproduction files contain experiment configurations, raw records, and analysis scripts, including all random seeds. Validation checks cover record completeness and configuration consistency.

All runs used central processing unit (CPU) resources without a graphics processing unit (GPU). Detailed resource records remain in the reproduction files.

Table 12: (A) Paired 2×2 RFM ablation. All profiles use identical PDE instances, tasks, fit grids, inference grids, and tuning/evaluation splits. Entries are top-1 / calibrated coverage / mean set size. (B) Paired factorial effects with 95% instance-bootstrap intervals (10,000 resamples). Each cell is an accuracy difference [interval] over the same 96 evaluation PDEs.

Panel A

Profile	noise	Structured	OMP+BIC	Subset+BIC
w192_m1	0	0.844/0.979/6.28	**0.885**	0.875
w192_m1	0.0001	0.854/0.979/6.28	**0.896**	0.885
w192_m1	0.0005	0.854/0.979/6.31	**0.906**	0.896
w192_m5	0	0.875/0.969/6.23	**0.896**	0.885
w192_m5	0.0001	0.875/0.979/6.24	**0.896**	0.885
w192_m5	0.0005	0.865/0.979/6.25	**0.906**	0.896
w384_m1	0	**0.906**/0.958/6.09	0.771	0.760
w384_m1	0.0001	**0.906**/0.958/6.14	0.771	0.760
w384_m1	0.0005	**0.906**/0.969/6.09	0.812	0.812
w384_m5	0	**0.938**/0.906/5.53	0.667	0.667
w384_m5	0.0001	**0.938**/0.906/5.48	0.677	0.677
w384_m5	0.0005	**0.958**/0.906/5.43	0.771	0.760

Panel B

Effect	noise 0	noise 10^{-4}	noise 5×10^{-4}
ensemble at width 192	0.031 [-0.031, 0.094]	0.021 [-0.042, 0.083]	0.010 [-0.052, 0.073]
ensemble at width 384	0.031 [-0.042, 0.104]	0.031 [-0.042, 0.104]	0.052 [-0.021, 0.125]
width at one member	0.062 [-0.031, 0.156]	0.052 [-0.042, 0.146]	0.052 [-0.042, 0.146]
width at five members	0.062 [0.000, 0.125]	0.062 [0.010, 0.125]	0.094 [0.031, 0.156]

Table 13: (A) Component and hyperparameter ablations on the width-384, five-member profile, pooled over three paired noise levels. Coverage is recalibrated on the tuning split for each variant. (B) Complexity-penalty evaluation on 192 independent PDE instances per noise cell. The pre-specified rule uses $\lambda_c = 0.04$; $\lambda_c = 0.01$ and the baselines are shown for comparison.

Panel A

Variant	top-1	calibrated coverage	mean set size
full	0.944	0.906	5.48
lambda_c_0	0.260	0.899	5.59
lambda_c_0025	0.715	0.899	5.56
lambda_c_04	1.000	0.917	1.02
lambda_c_08	1.000	0.969	1.00
lambda_p_0	0.955	0.906	5.49
lambda_p_0025	0.955	0.906	5.48
lambda_p_04	0.899	0.910	5.07
lambda_p_08	0.889	0.917	5.66
max_groups_1	0.688	–	–
max_groups_2	0.816	–	–
residual_only	0.260	0.899	5.58
ridge_0	0.944	0.906	5.48
ridge_1e12	0.944	0.906	5.48
ridge_1e6	0.951	0.906	5.48
ridge_1e8	0.944	0.906	5.48
top1_only	0.944	0.906	5.48
wide_set	0.944	0.906	5.48

Panel B

noise	$\lambda_c = .04$	$\lambda_c = .01$	OMP	subset	group lasso	set cov.
0	0.995	0.943	0.635	0.625	0.490	1.000
0.0001	0.995	0.953	0.667	0.651	0.479	1.000
0.0005	0.995	0.948	0.786	0.786	0.562	1.000

Table 14: High-frequency unanchored 2D transfer. Metrics use independently generated PDE instances. No solution mode is forced into the RFM basis.

partition	structured	set cov.	OMP	dense	group lasso	sparse group
calibration	0.995	1.000	1.000	0.779	0.632	0.630
evaluation	0.964	0.971	0.972	0.754	0.627	0.626

Table 15: (A) Independent solver-transfer comparison. Every method consumes the same residual observable and dictionary within a row. Bold marks the best top-1 value in each solver/noise cell. (B) Five-solver weak/strong-amplitude transfer. Every row uses matched observables and dictionaries.

Panel A

Solver	noise	Structured	Dense	OMP	Group lasso	Sparse-group
fdm_centered	0	**1.000**	**1.000**	**1.000**	0.625	0.625
fdm_centered	0.0001	0.875	0.375	**0.922**	0.625	0.594
fdm_flux	0	**1.000**	**1.000**	**1.000**	0.625	0.625
fdm_flux	0.0001	0.875	0.375	**0.922**	0.625	0.594
fem_linear	0	**1.000**	**1.000**	**1.000**	0.625	0.625
fem_linear	0.0001	0.891	0.375	**0.922**	0.625	0.594
rfm	0	**1.000**	**1.000**	**1.000**	0.625	0.625
rfm	0.0001	0.891	0.469	**0.938**	0.594	0.594
trainable_mlp	0	0.984	0.984	**1.000**	0.625	0.625

Panel B

solver/noise	structured	OMP	dense	group lasso	sparse group
fdm_centered—0	1.000	1.000	1.000	0.600	0.600
fdm_centered—0.0001	0.890	0.925	0.200	0.469	0.444
fdm_flux—0	1.000	1.000	1.000	0.600	0.600
fdm_flux—0.0001	0.890	0.925	0.202	0.469	0.444
fem_linear—0	1.000	1.000	1.000	0.600	0.600
fem_linear—0.0001	0.890	0.925	0.204	0.471	0.435
rfm—0	1.000	1.000	1.000	0.600	0.600
rfm—0.0001	0.940	0.958	0.340	0.519	0.506
trainable_mlp—0	0.975	0.975	0.960	0.585	0.585

Table 16: (A) Three-cause large-dictionary screening. Contextual screening ranks each local atom after conditioning on up to two global causes; singleton screening is the unconditioned baseline. (B) Contextual $K = 3$ screening across 40 seeds. Retention is evaluated before exact reduced-dictionary search.

Panel A

Atoms	method	accuracy	truth retention	median seconds
63	structured singleton 64	1.000	1.000	5.371
63	structured contextual 64	1.000	1.000	5.362
63	group omp bic	1.000	–	0.015
127	structured singleton 64	0.000	0.000	5.599
127	structured contextual 64	1.000	1.000	5.610
127	group omp bic	1.000	–	0.031
255	structured singleton 64	0.000	0.000	5.590
255	structured contextual 64	1.000	1.000	5.579
255	group omp bic	0.000	–	0.058
511	structured singleton 64	0.000	0.000	5.571
511	structured contextual 64	1.000	1.000	5.576
511	group omp bic	0.000	–	0.111
1023	structured singleton 64	0.000	0.000	5.566
1023	structured contextual 64	1.000	1.000	5.604
1023	group omp bic	0.000	–	0.224

Panel B

atoms	singleton-64 ret.	contextual-16 ret.	contextual-32 ret.	contextual-64 ret.	OMP acc.
63	1.000	1.000	1.000	1.000	1.000
127	0.000	1.000	1.000	1.000	1.000
255	0.000	1.000	1.000	1.000	0.000
511	0.000	1.000	1.000	1.000	0.000
1023	0.000	1.000	1.000	1.000	0.000

Table 17: Derivative-evaluation ablation across widths, seeds, tasks, and inference grids.

derivative	structured	OMP	certificate	median score deviation
analytic	1.000	1.000	0.750	1.11e-04
finite_difference_h1e-02	1.000	1.000	0.750	4.42e-03
finite_difference_h1e-03	1.000	1.000	0.750	1.28e-04
finite_difference_h2e-02	0.750	0.750	0.750	1.70e-02
finite_difference_h5e-03	1.000	1.000	0.750	1.32e-03

Table 18: Action utility on the width-384, mean-five profile. Win denotes lower forward error than leaving the nominal problem unchanged. The set action selects a candidate by residual on a disjoint collocation grid and never uses the true support.

noise	method	corrected cases	win rate	median error ratio
0	structured	84	1.000	2.735e-03
0	structured set action	84	1.000	2.730e-03
0	omp	84	1.000	2.728e-03
0	subset bic	84	1.000	2.728e-03
0	oracle	84	1.000	2.678e-03
0.0001	structured	84	1.000	2.733e-03
0.0001	structured set action	84	1.000	2.727e-03
0.0001	omp	84	1.000	2.721e-03
0.0001	subset bic	84	1.000	2.721e-03
0.0001	oracle	84	1.000	2.678e-03
0.0005	structured	84	1.000	2.720e-03
0.0005	structured set action	84	1.000	2.720e-03
0.0005	omp	84	1.000	2.726e-03
0.0005	subset bic	84	1.000	2.726e-03
0.0005	oracle	84	1.000	2.678e-03

C Additional Results and Solution Visualizations

Figures 3–6 break down controlled identifiability, off-dictionary combinations, interface displacement, and the mixed-boundary 2D problem. Table 19 gives per-class values.

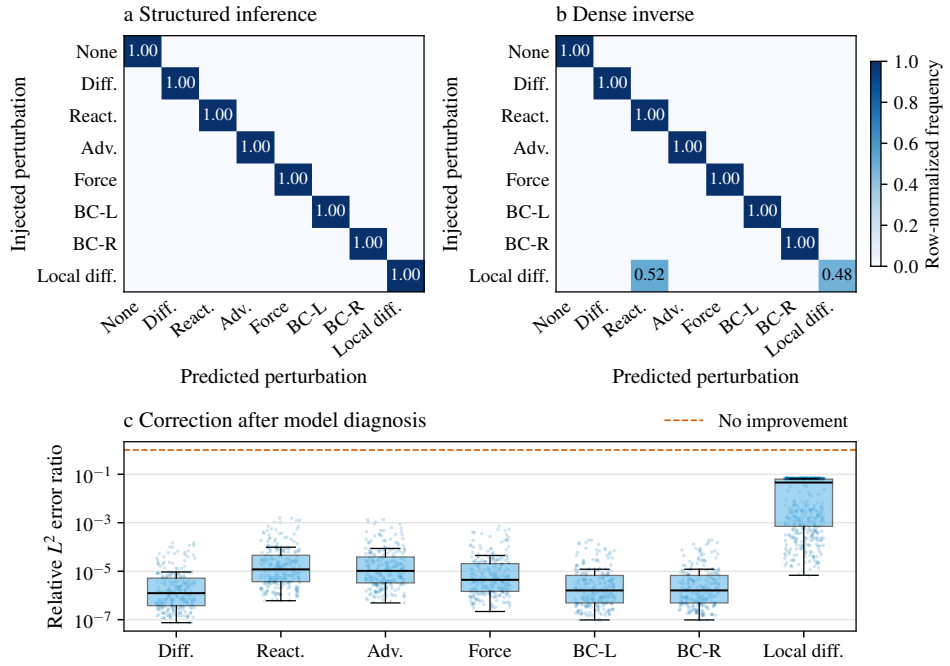

Figure 3: Controlled 1D identifiability and correction. (a,b) Row-normalized confusion matrices for structured inference and the dense simultaneous inverse on 1,980 identifiable rows. (c) Corrected-to-nominal relative L^2 error. The dashed line is no improvement.

Figure 3 shows exact identification by structured inference, while the dense inverse confuses some localized-diffusion perturbations with reaction. The inferred correction remains beneficial across all declared groups. No ambiguity flags occur in the 1,980 identifiable cases, whereas all 240 collinear controls are flagged.

Top-1 identifies all runs in four combinations. All 30 reaction-plus-local diffusion runs are instead assigned to forcing plus local diffusion. The true support remains in every explanation set across the 150 runs. Figure 4 shows that correction improves every run, but both local-diffusion combinations retain larger errors. The pooled Spearman correlation between backward-error norm and forward discrepancy is $\rho = -0.10$, reported descriptively for this fixed suite.

Table 19: Per-class results for the principal failure regimes.

Class	n	Structured top-1	Set coverage	Median correction ratio
1D OOD: reaction + local	30	0.000	1.000	0.111
1D OOD: advection + local	30	1.000	1.000	0.0626
2D: x-advection	40	0.900	0.950	0.148
2D: localized diffusion	80	1.000	1.000	0.0816
Interface shifts	80	0.950	1.000	0.373

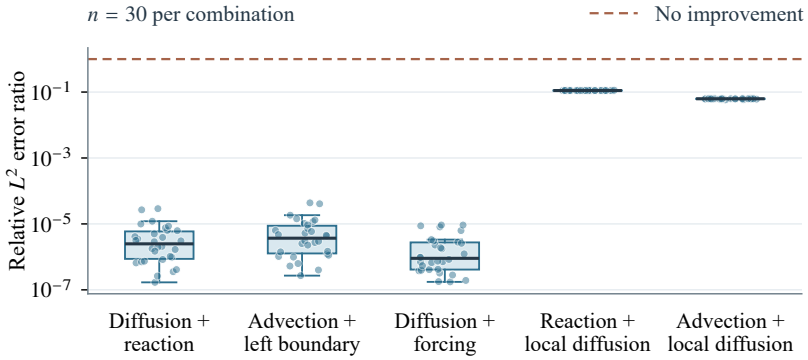

Figure 4: Correction under OOD combinations. Corrected-to-nominal relative L^2 error is shown on a logarithmic scale. Boxes show medians and interquartile ranges, with points for all 30 runs per combination. The dashed line marks no improvement.

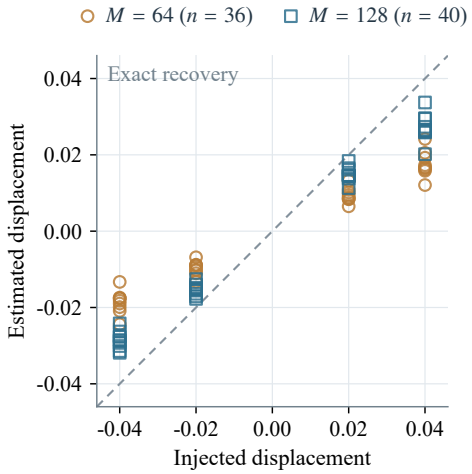

Figure 5: Estimated versus injected material-interface displacement for the 76 correctly selected interface cases. The dashed line is exact recovery. The four wrong-support cases are excluded from position recovery.

On the 100 interface tasks, structured top-1 is correct on 46/50 and 50/50 at widths 64 and 128, compared with 10/50 and 17/50 for the dense inverse. Explanation sets cover all 100 tasks. All four errors occur in shifted width-64 cases: one is assigned to global diffusion and three to local diffusion. Figure 5 shows underestimation of displacement magnitude. Median absolute position errors are 0.01455 and 0.00779 for the correctly selected cases at widths 64 and 128, respectively. All 80 shifted-case corrective solves improve the solution. The two widths show the direction of the effect but are insufficient for a convergence-rate estimate.

On the 340 mixed-boundary tasks, the four top-1 errors assign x-advection to forcing, all at width 64. At that width, top-1 is correct on 166/170 tasks and explanation sets cover 168/170. Both are

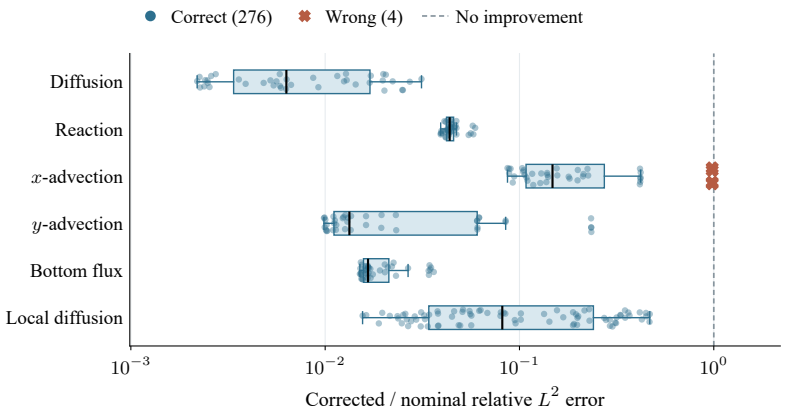

Figure 6: Variable-coefficient 2D problem with Dirichlet x boundaries and Neumann y boundaries. Boxes show medians and interquartile ranges of corrected-to-nominal relative L^2 error. Points show all 280 non-forcing cases, with crosses marking the four wrong top-1 labels. The dashed line marks no improvement.

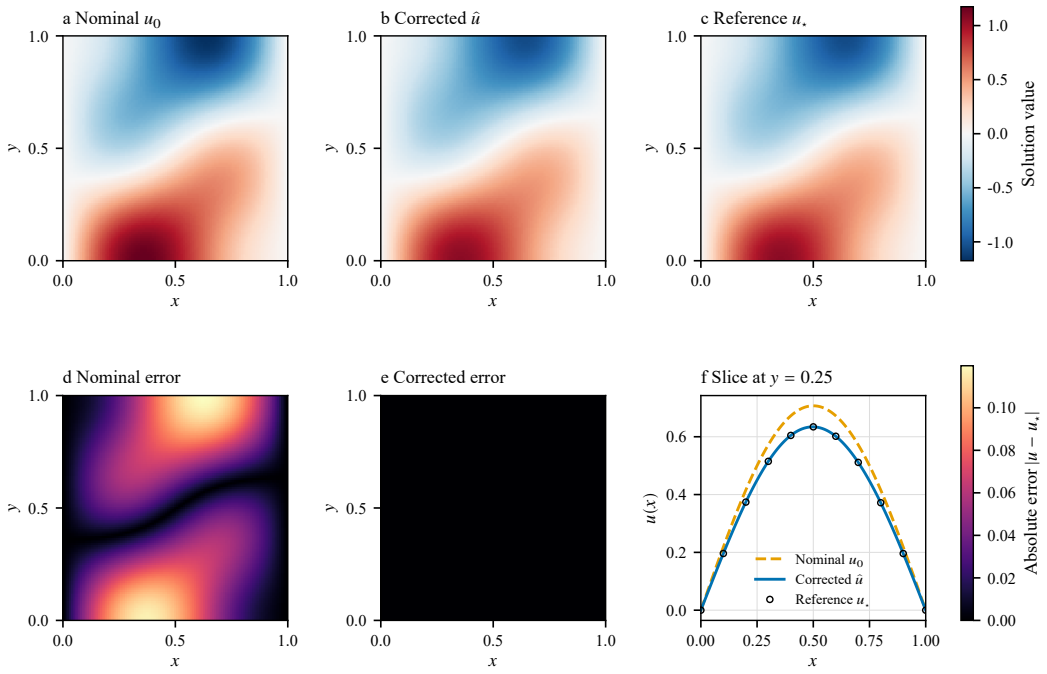

Figure 7: Reconstructed 2D diffusion case ($\delta a = +0.12$, $M = 128$). (a–c) Nominal, inferred-PDE, and independent reference solutions share one color scale. (d,e) Absolute errors share a linear scale. Relative L^2 error decreases from 0.1154 to 2.75×10^{-4}. (f) A fixed slice at $y = 0.25$ shows the corrected and reference curves nearly coinciding.

170/170 at width 128. All 320 perturbed cases improve. The four misidentified cases in Figure 6 have error ratios of approximately 0.981, so the corresponding correction is only marginal. The 40 forcing cases (not plotted) have error ratios near 1.9×10^{-11}.

Solution-field reconstruction. Figures 2 and 7 reconstruct one 1D and one 2D diffusion case. We choose the first seed at the largest width and largest positive amplitude, independently of correctness or correction gain. The 1D reference is closed form; the 2D reference is an independent 65×65

finite-difference solution. Nominal and corrected re-solves use the same basis, resolution, and ridge setting, changing only the inferred coefficient.

The reconstruction script reads the selected fits and checks the relative errors and their ratio against the stored records. Reproduction files include the field arrays and settings. Figure 7 shows that correction removes the broad amplitude discrepancy in the 2D field, leaving a small residual. Both error panels use the same scale.

Solution geometry and resolved correction error. Figures 8 and 9 give complementary views of the same frozen 2D case in Figure 7, without selecting additional runs or changing the inferred coefficient. The surfaces and profiles use the stored 101×101 display grid. The reference is evaluated from the independent 65×65 finite-difference solve. These are visualizations of the reconstructed case, not additional benchmark trials.

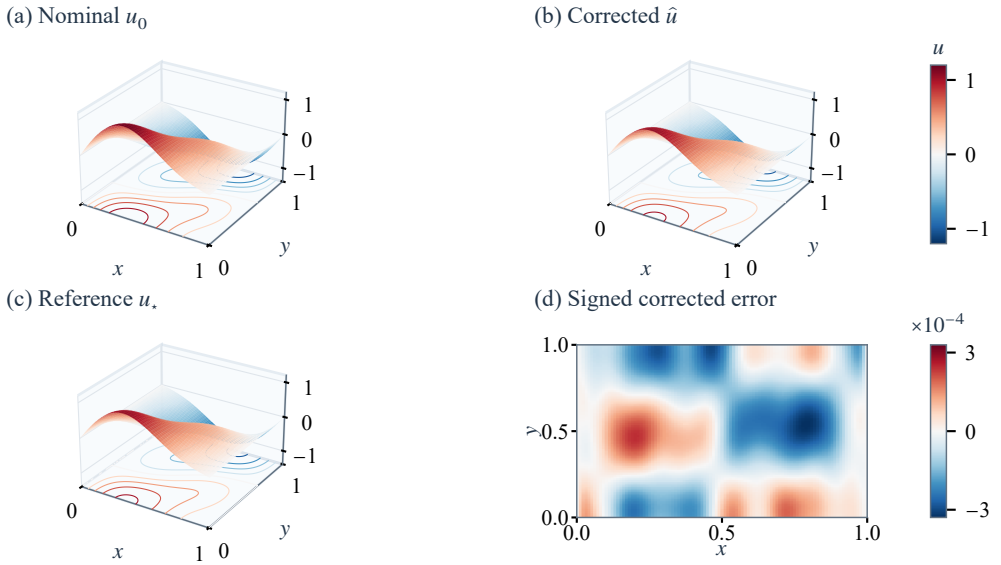

Figure 8: **Solution geometry before and after model correction.** (a–c) Nominal, inferred-PDE, and independent reference fields share the same height and color scales; projected contours reveal their spatial structure. (d) The signed discrepancy $\widehat{u} - u_\star$ uses a separate, magnified scale to resolve the small remaining error. Unlike the common absolute-error scale in Figure 7, this panel emphasizes residual structure rather than the size of the improvement.

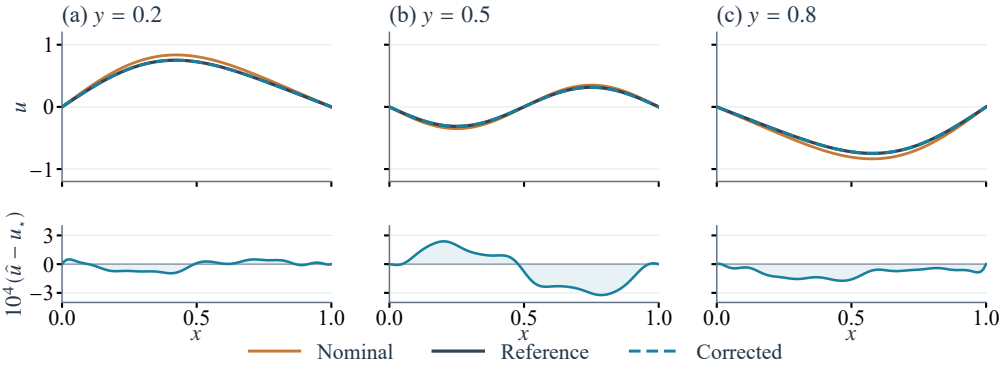

Figure 9: **Fixed spatial sections of the reconstructed solution.** Top: nominal, reference, and corrected fields at $y = 0.2$, 0.5, and 0.8. Bottom: signed corrected errors, multiplied by 10^4, on a common scale. These sections are fixed geometrically, not selected by error. All curves use the existing display-grid samples; no smoothing or new fitting is applied.