UNIPDE-BENCH: EVALUATING RECOVERABILITY AND SELECTIVE PREDICTION IN PARTIALLY OBSERVED PDE SYSTEMS

Anonymous authors
Paper under double-blind review

ABSTRACT

Aggregate reconstruction error measures how well a learned estimator predicts a hidden PDE state on a given distribution, but not how strongly the observations constrain its predictions. We introduce UniPDE-Bench, a direction-wise evaluation protocol that relates empirical recovery and confidence-based selection to local observation geometry. The protocol whitens the joint observation Jacobian by the noise covariance and normalizes it by a state metric. A complete right singular basis captures joint state variations, which a relative sensitivity threshold separates into retained and below-threshold directions. In this common basis, we evaluate recovery and agreement between prediction claims and the geometric partition, then trace recovery yield and abstention across claim coverages. Controlled inverse problems and PDE simulations show that low prediction error can coexist with a nontrivial below-threshold subspace. Moreover, confidence rankings that outperform chance overall can still produce below-random abstention on below-threshold directions at high claim coverages. UniPDE-Bench separates recovery capability from direction-selection quality, providing an observation-based reference for evaluating partially observed state estimation beyond aggregate error.

1 INTRODUCTION

A learned estimator reconstructing a PDE state from partial observations draws on two sources: information in the measurements and statistical correlations in its training data. Aggregate reconstruction error evaluates their combined predictive performance, but does not distinguish how strongly the measurements constrain different state variations. Two estimators with similar errors may recover different joint variations of the state, or assign confidence to directions with very different observational support. Evaluating their predictions therefore requires more than measuring accuracy on a test distribution.

The distinction already appears in $y = c_1 + c_2$. Changing either coefficient alone changes the observation, yet equal and opposite changes leave it unchanged. Coordinate-wise sensitivity misses this joint ambiguity. Strong correlations in the data can nevertheless make both coefficients predictable. Prior-supported predictability and observation-based distinguishability are different properties, and should be assessed in a common directional representation.

Classical inverse-problem analysis characterizes observational constraints through spectral geometry (Backus & Gilbert, 1968; Tarantola, 2005); null-space diagnostics examine prior-induced reconstruction components (Bhadra et al., 2021). Scientific machine learning benchmarks compare predictions across tasks and distributions (Takamoto et al., 2022; Ohana et al., 2024), while selective and conformal prediction study acceptance decisions and statistical coverage (Geifman & El-Yaniv, 2017; Angelopoulos & Bates, 2023). We connect these perspectives through a specific evaluation question: *which directions does a model recover, and which directions does its confidence prioritize, relative to local observation sensitivity?*

UniPDE-Bench uses the joint observation Jacobian as an estimator-independent reference. The observation map includes the inputs actually used by the estimator, including histories and auxiliary channels. Noise whitening and state-metric normalization define the comparison scales. A complete right singular basis retains all state directions, including the right null space, and a relative thresh-

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.

old separates retained from below-threshold directions. This is a *local sensitivity partition*, not a classification of globally recoverable and unrecoverable states.

Predictions and ground truth are projected into this basis to determine which retained directions meet a recovery tolerance. Separately, confidence determines which predictions are claimed and which are left unclaimed. We call the claimed fraction *claim coverage*, distinct from statistical interval coverage. Sweeping this fraction yields recovery-yield and below-threshold-abstention curves. Their areas summarize ranking quality, but do not replace the curves: an above-chance ranking can still perform below the random reference at particular budgets.

Our contributions are threefold. First, we specify a local geometric reference that matches the estimator's actual inputs and supports direction-wise comparison. Second, we separate recovery, geometric agreement, and confidence-based selection, and derive the relationships between coverage-curve areas and AUROC. Third, controlled inverse problems and multi-model PDE experiments demonstrate differences between prediction accuracy and direction selection, and between overall ranking and budget-specific abstention. These findings motivate joint reporting of observation geometry, empirical recovery, and selection behavior.

2 Related Work

Inverse-problem geometry. Resolution analysis and spectral regularization characterize directional observational constraints (Backus & Gilbert, 1968; Tarantola, 2005). Likelihood-informed subspaces and low-rank Bayesian posterior approximations additionally account for prior geometry (Cui et al., 2014; Spantini et al., 2015); their interpretation is not interchangeable with sensitivity under an arbitrary state metric. UniPDE-Bench uses established spectral tools to construct a local evaluation reference, rather than proposing a new inverse solver.

Learned reconstruction and null-space diagnostics. Bhadra et al. (2021) use measurement-space and null-space decompositions to diagnose prior-induced structures in tomography. Gottschling et al. (2025) analyze kernel-induced limitations and accuracy–stability trade-offs in inverse problems. We apply this perspective to prediction claims, comparing recovery and confidence rankings in the same basis across claim budgets. Our below-threshold set includes both exact null directions and nonzero low-sensitivity directions; prior-supported predictions are not treated as intrinsically erroneous.

Scientific machine learning benchmarks. PDEBench and The Well compare models across PDE tasks and distributions (Takamoto et al., 2022; Ohana et al., 2024), while the seismic wavefield common task framework evaluates sparse wavefield reconstruction (Yermakov et al., 2025). These task-level evaluations complement our direction-wise analysis. UniPDE-Bench adds an observation-geometry reference that must include all estimator inputs; it neither replaces existing task suites nor substitutes geometric agreement for prediction accuracy.

Selective prediction and uncertainty. Selective prediction studies the trade-off between accepting predictions and their errors (El-Yaniv & Wiener, 2010; Geifman & El-Yaniv, 2017). Conformal prediction provides statistical coverage under exchangeability and suitable calibration (Angelopoulos & Bates, 2023); deep ensembles provide uncertainty signals from prediction disagreement (Lakshminarayanan et al., 2017). Exchangeability is a property of samples, whereas injectivity is a property of the observation map. We evaluate joint state directions as selection units, separating recovery from preference for retained directions. Interval width and ensemble spread are empirical ranking signals here, not tests of conformal coverage guarantees.

3 UniPDE-Bench: Constructing Observation Subspaces

Figure 1 summarizes the protocol. Observation geometry supplies a shared directional basis and sensitivity labels. Predictions and test ground truth determine recovery outcomes; independently obtained confidence scores determine the claim order. The labels are used for evaluation, not to automatically suppress below-threshold predictions.

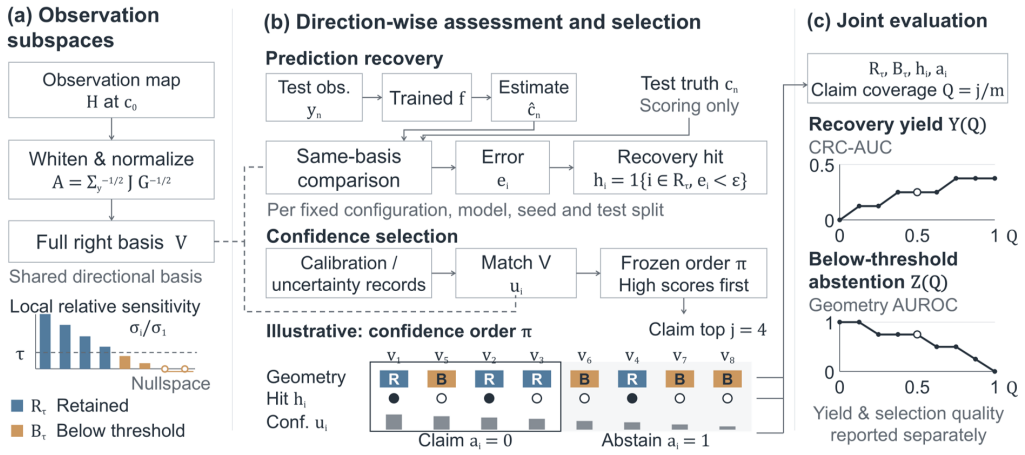

Figure 1: UniPDE-Bench workflow. Geometry defines directions and sensitivity labels; prediction errors define recovery hits. Confidence and a claim budget determine accepted directions, which are then scored against the geometric reference. Test ground truth is used for scoring only, not for ranking.

3.1 Problem Setup and Observation Specification

Let $c \in \mathcal{C} \subseteq \mathbb{R}^m$ be the interface parameters and $y \in \mathbb{R}^p$ the observations actually used by the estimator. For a differentiable observation map $\mathcal{H}$ and working point c_0,

$$y = \mathcal{H}(c) + \eta, \qquad J = D\mathcal{H}(c_0), \qquad \mathrm{Cov}(\eta) = \Sigma_y \succ 0. \tag{1}$$

A positive-definite shape metric G defines $\|d\|_G^2 = d^\top G d$ and local coordinates $z = G^{1/2}(c - c_0)$. The observation specification fixes the PDE, sensor functions, auxiliary channels, history, c_0, G, Σ_y, and numerical settings. A separate scoring specification fixes the sensitivity threshold, recovery tolerance, confidence rule, and aggregation procedure.

3.2 Joint Sensitivity and Subspace Partition

Whitening observations and normalizing the state metric give

$$A = \Sigma_y^{-1/2} J G^{-1/2}, \qquad A^\top A = V \,\mathrm{diag}(\sigma_1^2, \dots, \sigma_m^2) V^\top, \tag{2}$$

where $V = [v_1, \dots, v_m]$ is orthogonal and $\sigma_1 \geq \cdots \geq \sigma_m \geq 0$. Computation uses the *complete* right singular basis: when $p < m$, an economy-size SVD alone omits right-null-space directions. For $0 < \tau \leq 1$ and $\sigma_1 > 0$, define

$$\begin{aligned} \mathcal{R}_\tau &= \{i : \sigma_i/\sigma_1 \geq \tau\}, \quad \mathcal{B}_\tau = \{i : \sigma_i/\sigma_1 < \tau\}, \\ P_\tau &= \sum_{i \in \mathcal{R}_\tau} v_i v_i^\top, \qquad r_\tau = |\mathcal{R}_\tau|, \quad b_\tau = m - r_\tau. \end{aligned} \tag{3}$$

P_τ projects onto the retained subspace in z coordinates. If $\sigma_1 = 0$, set $\mathcal{R}_\tau = \varnothing$, $\mathcal{B}_\tau = \{1, \dots, m\}$, and $P_\tau = 0$. Unlike the exact null space, $\mathcal{B}_\tau$ can include nonzero low-sensitivity directions. We record τ separately from numerical rank tolerance.

3.3 What the Geometric Labels Mean

Proposition 1 (Joint ambiguity). *Nonzero Jacobian columns do not imply joint injectivity. For* $J = [1\ 1]$, $(1, -1)^\top \in \ker J$*; in general,* $m > p$ *implies* $\dim \ker J \geq m - p$.

Proposition 2 (Indistinguishable states). *If* $c_+, c_- \in \mathcal{C}$ *satisfy* $\mathcal{H}(c_+) = \mathcal{H}(c_-)$*, any deterministic estimate at that observation incurs error at least* $\frac{1}{2}\|c_+ - c_-\|_G$ *for one of the two states.*

The local partition in Equation (3) does not establish the premise of Proposition 2. For example, $\mathcal{H}(c) = c^3$ is injective despite a zero derivative at $c_0 = 0$. Uniformly scaling noise standard deviations leaves the relative partition unchanged; absolute resolution also depends on noise level and perturbation amplitude. Basis choices within repeated-singular-value subspaces must be recorded, because rotations can change directional errors and rankings. Finite perturbations $c_\pm = c_0 \pm \delta G^{-1/2} v_i$ must remain in $\mathcal{C}$; small nonzero observation differences support only local diagnostics. Appendix A gives proofs and finite-perturbation bounds, and Appendix C examines numerical sensitivity.

4 Direction-Wise Selective Evaluation

4.1 Directional Predictions and Successful Recovery

For a fixed configuration, estimator, seed, and test split, project ground truth and predictions into the same basis:

$$q_n = V^\top G^{1/2}(c_n - c_0), \qquad \widehat{q}_n = V^\top G^{1/2}(\widehat{c}_n - c_0).$$

With N test samples, define signal scale, normalized error, and recovery hits by

$$\begin{aligned} s_i^2 &= \frac{1}{N}\sum_{n=1}^{N} q_{ni}^2, \qquad e_i = \frac{\sqrt{N^{-1}\sum_{n=1}^{N}(\widehat{q}_{ni} - q_{ni})^2}}{s_i}, \\ h_i &= \mathbf{1}\{i \in \mathcal{R}_\tau,\ e_i < \varepsilon\}. \end{aligned} \tag{4}$$

If $s_i = 0$, set $e_i = +\infty$, $h_i = 0$, and report the zero-signal count. A hit requires both retention and adequate prediction; geometry alone does not determine recovery. Setting $\varepsilon = 1$ means outperforming the zero-perturbation predictor, not obtaining a worst-case error guarantee.

4.2 Claims, Abstention, and Fixed-Budget Scores

Let $a_i = 1$ mean abstention and $a_i = 0$ a prediction claim. For $r_\tau, b_\tau > 0$, report

$$\begin{aligned} Q &= \frac{1}{m}\sum_i (1 - a_i), \qquad R = \frac{1}{r_\tau}\sum_i (1 - a_i) h_i, \\ A_B &= \frac{1}{b_\tau}\sum_{i \in \mathcal{B}_\tau} a_i, \qquad H_B = \sum_{i \in \mathcal{B}_\tau} (1 - a_i). \end{aligned} \tag{5}$$

These are, respectively, claim coverage, the successfully recovered fraction of retained directions, the abstained fraction of below-threshold directions, and the number of below-threshold claims. Acceptance is not an identifiability certificate. The experiments also report the soft composite score

$$\mathrm{sCRS} = \max\left(0, 1 - \lambda \frac{H_B}{b_\tau}\right) \frac{2RA_B}{R + A_B}, \qquad \lambda = 2. \tag{6}$$

Set the score to zero if $R + A_B = 0$; use $\mathrm{sCRS} = R$ if $b_\tau = 0$, and zero if $r_\tau = 0$. A positive score does not require $H_B = 0$, so its components must also be reported. *Oracle gating* sets $a_i = \mathbf{1}\{i \in \mathcal{B}_\tau\}$, providing a geometric reference for a given predictor rather than a learned selection rule.

4.3 Coverage Curves and Confidence Rankings

A confidence score u_i gives higher-priority directions larger values. Scores must match the prediction basis and must not be obtained from test errors e_i or hit labels h_i. Calibration error, interval width, and ensemble spread are converted to confidence by prespecified rules. Calibration-error and interval-width rankings are fixed per configuration, rather than updated for each test observation.

Let π sort confidence in descending order. At budget $j = 0, \ldots, m$, claim the first j directions and abstain from the rest:

$$Q_j = \frac{j}{m}, \quad Y_j = \frac{1}{m}\sum_{\ell=1}^{j} h_{\pi(\ell)}, \quad Z_j = \frac{1}{b_\tau}\sum_{\ell=j+1}^{m} \mathbf{1}\{\pi(\ell) \in \mathcal{B}_\tau\}. \tag{7}$$

Recovery yield $Y_j = (r_\tau/m)R_j$ and below-threshold abstention $Z_j = A_{B,j}$ separate recovery quantity from selection behavior. For $k = \sum_i h_i$, random rankings have expected curves $Y(Q) = (k/m)Q$ and $Z(Q) = 1 - Q$. We integrate piecewise-linearly interpolated curves and average uniformly over orderings within tied groups.

For $0 < k < m$, normalized recovery-coverage area is

$$\text{CRC-AUC} = \frac{\int_0^1 Y(Q)\, dQ - k/(2m)}{k(m-k)/(2m^2)} = 2\,\text{AUROC}_h - 1, \tag{8}$$

where AUROC_h takes recovery hits as positives. CRC-AUC measures selection quality conditional on the prediction outcomes, not recovery quantity; it must accompany k/m and the recovery curve. It is undefined for $k = 0$ or $k = m$, although the curves and counts remain informative.

For geometric agreement alone, take retained directions as positives and below-threshold directions as negatives. Using the standard pairwise interpretation of AUROC (Hanley & McNeil, 1982), define, for $r_\tau, b_\tau > 0$,

$$U_B = \Pr(u_I > u_J) + \tfrac{1}{2}\Pr(u_I = u_J), \quad I \sim \text{Unif}(\mathcal{R}_\tau),\ J \sim \text{Unif}(\mathcal{B}_\tau), \tag{9}$$

with independent draws. The excess abstention area over the random reference satisfies

$$\begin{aligned} D &= \int_0^1 [Z(Q) - (1-Q)]\, dQ = \frac{r_\tau}{m}\left(U_B - \frac{1}{2}\right), \\ \frac{D}{D_{\text{geom}}} &= 2U_B - 1, \qquad D_{\text{geom}} = \frac{r_\tau}{2m}. \end{aligned} \tag{10}$$

Appendix B derives both area identities. Each area and its corresponding AUROC express the same ranking information. Crucially, $U_B > 1/2$ does not imply $Z(Q) > 1 - Q$ at every budget. Singular-value ranking gives $U_B = 1$ as a geometric reference, not as evidence of learned uncertainty quality.

4.4 Evaluation Records

Algorithm 1 in Appendix E evaluates each configuration, model, seed, split, and selection rule separately before aggregation. Records contain observation/scoring identifiers, the complete spectrum and basis, directional errors, hits and confidence, curves, and reasons for undefined metrics. Test truth is used only for scoring. Reusing a confidence vector does not constitute additional geometric-ranking evidence; interval-width results are empirical ranking analyses only.

5 Experiments

We ask which joint state directions observations constrain, whether accurate predictions accompany appropriate direction selection, and how selection changes with claim coverage.

5.1 Experimental Setup

Tasks and observations. We use three interface-sensing simulations: a diffuse-domain battery-potential model (Battery), low-Péclet-number advection–diffusion (Melt-pool), and high-Péclet-number transport (Flooding). A real Fourier expansion with $K = 8$ and a constant term gives $m = 17$ interface coefficients. Reference diagnostics use a 64×64 grid, $\tau = 0.05$, and channel-normalized observation noise of 1% with an absolute floor. The reference Battery configuration includes two auxiliary channels. Learning experiments use five separately specified configurations: Battery M0/M2/M8, Melt-pool M12, and Flooding M12. Reference diagnostics and learning configurations are not interchangeable: their geometry must match all inputs, including histories and auxiliary channels.

Estimators and metrics. Controlled local inverse problems use ridge regression and TSVD, with 1,600/500/800 training/validation/test samples and five seeds. Neural comparisons use UniPDE, UniPDE-noalign, GRU, static MLP, FNO (Li et al., 2021), and DeepONet (Lu et al., 2021): six

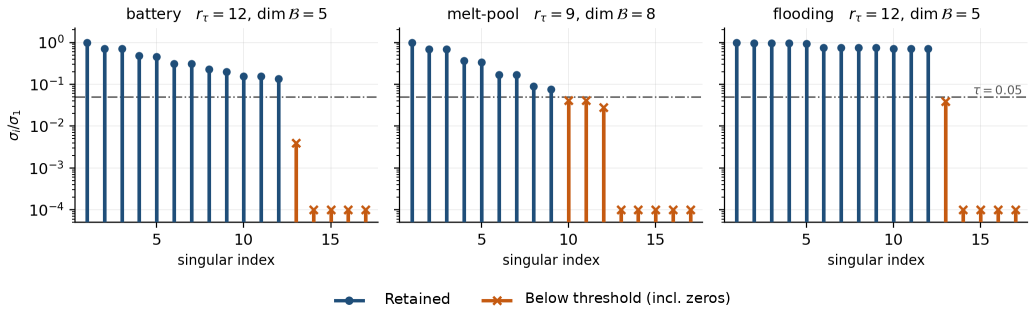

Figure 2: Normalized joint spectra for the reference configurations. The dashed threshold $\tau = 0.05$ gives retained dimensions 12, 9, and 12. Padded zeros and values below the display range are plotted at the lower limit; these display positions do not determine exact rank.

Table 1: Controlled local inverse problems. Coefficient NRMSE is mean $\pm$ standard deviation over five seeds. The last four columns are sCRS; Oracle uses the geometric partition.

Operator	Ridge OOD NRMSE	No abstention	Validation gate	Oracle gate	TSVD + Oracle
Battery	0.669 ± 0.007	0.00	0.00	1.00	0.40
Melt-pool	0.140 ± 0.004	0.00	0.00	1.00	0.875
Flooding	0.054 ± 0.001	0.00	0.00	1.00	1.00

families, five configurations, and five seeds give 150 training runs. We report ID and OOD NRMSE, fixed-point sCRS, CRC-AUC, geometric AUROC, and coverage curves. Controlled and Poseidon-T experiments measure *coefficient* NRMSE; the six-model comparison measures *field* NRMSE. Neural macro-averages first average seeds within each family, then average the six families. Poseidon-T adaptation and prediction averaging are reported separately in Appendix D.4.

5.2 Joint Sensitivity Reveals Coordinate-Wise Ambiguities

For the reference configurations, the Battery Jacobian is 14×17 and the other two are 13×17 (Appendix C). Coordinate-wise analysis reports no blind modes for Flooding, although its Jacobian cannot have full column rank. Joint analysis instead yields retained dimensions of 12, 9, and 12 for Battery, Melt-pool, and Flooding, respectively (Figure 2). In Flooding, at least four directions lie in the exact null space; the thirteenth normalized singular value, 0.039, lies below $\tau = 0.05$, adding a fifth below-threshold direction. Coordinate-wise responses therefore do not rule out joint ambiguity.

On a 48×48 grid, effective ranks are 12, 7, and 12; refining to 64×64 changes only Melt-pool, to 9. Near-threshold directions require numerical sensitivity checks: small changes in a matrix norm alone do not establish subspace convergence. Appendix C provides spectra and finite-difference diagnostics.

5.3 Low Error Does Not Imply Geometrically Aligned Selection

Controlled inverse problems. We sample coefficients from a smooth correlated prior with a small full-rank residual, generate noisy observations through the local Jacobian, and fit ridge regression. The map retains joint ambiguity while the estimator can exploit prior correlations. Flooding achieves the lowest OOD coefficient NRMSE, 0.054 ± 0.001, despite five below-threshold directions (Table 1). Both no abstention and validation-error gating have zero sCRS, whereas Oracle gating reaches 1.00. TSVD with the same Oracle rule reaches 0.40, 0.875, and 1.00 across the three tasks. Thus geometry specifies retained directions, but the estimator still determines which of them are recovered.

Neural estimators under ID and OOD conditions. The three Battery configurations have similar ID field NRMSEs (0.048, 0.048, 0.047), but validation-gated sCRS varies from 0.465 to 0.049 (Table 2). Field accuracy and directional selection are not interchangeable. Under shift, macro-

Table 2: Six-model macro-averages, after averaging five seeds per model. Field NRMSE is lower-better. Val. and Oracle denote the corresponding gated sCRS.

	ID			OOD		
Configuration	NRMSE	Val.	Oracle	NRMSE	Val.	Oracle
Battery M0	0.048	0.465	1.000	0.168	0.438	0.900
Battery M2	0.048	0.345	0.800	0.170	0.345	0.800
Battery M8	0.047	0.049	0.228	0.180	0.050	0.244
Melt-pool M12	0.072	0.559	0.668	0.530	0.000	0.000
Flooding M12	0.071	0.281	0.767	0.483	0.000	0.000
Macro-average	0.057	0.340	0.692	0.306	0.166	0.389

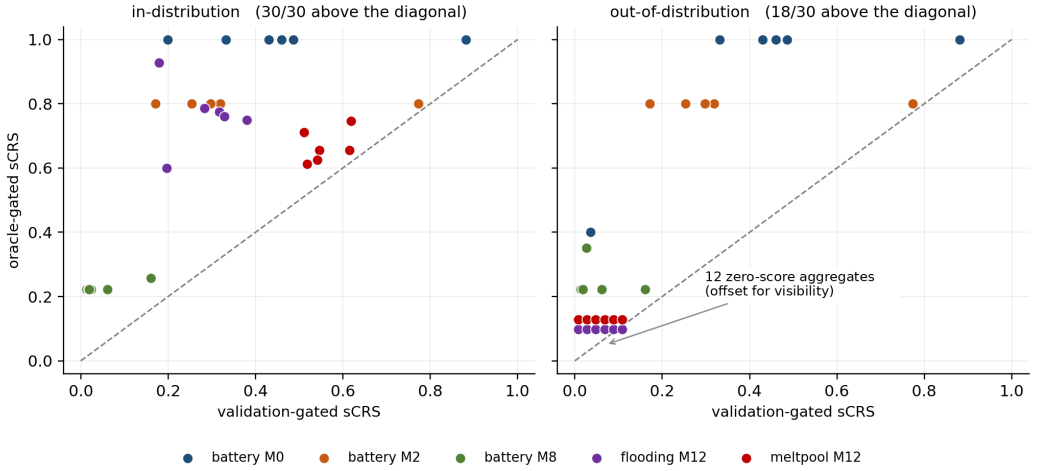

Figure 3: Validation versus Oracle gating. Each point is a five-seed mean for one model–configuration pair. Dashed lines indicate equal sCRS. The 12 OOD zero-score points are offset for visibility only; all have actual coordinates $(0, 0)$.

averaged NRMSE rises from 0.057 to 0.306. Melt-pool and Flooding account for much of this increase and have no recovery hits in the reported OOD records, giving zero scores under both gates. Changing selection cannot compensate for missing recovery.

Oracle gating exceeds validation gating in all 30 model–configuration combinations under ID and 18 under OOD; the remaining 12 OOD combinations have zero scores (Figure 3). FNO on Flooding M12 illustrates the distinction: ID field NRMSE is 0.039 ± 0.002, yet validation-gated sCRS is 0.179 ± 0.183, compared with 0.928 ± 0.026 under Oracle gating. Separately, five-model Poseidon-T averaging lowers OOD *coefficient* NRMSE from 0.164 to 0.139; it does not evaluate ensemble-spread-based selection or provide a direct comparison with field errors (Appendix D.4).

5.4 Confidence Rankings and Geometry Measure Different Outcomes

We compare calibration error, conformal interval width, and ensemble spread, prioritizing smaller values. No abstention, geometric gating, and random ranking are references. Calibration errors and interval widths are computed on calibration data and induce fixed per-configuration direction rankings. We assess their empirical ranking behavior, not conformal statistical coverage.

Table 3 summarizes fixed operating points for 60 model–configuration–split combinations. Passing requires $H_B = 0$ and $\mathrm{sCRS} > 0$. Calibration-error selection passes only 9 combinations and claims below-threshold directions in 46, despite a mean score of 0.375. Ensemble spread and conformal width never pass at their fixed operating points, although their rankings exceed chance. A soft score, a fixed operating point, and a full ranking describe distinct properties.

Table 3: Fixed-point scores and ranking summaries. Counts are out of 60 combinations; a pass requires $H_B = 0$ and $\mathrm{sCRS} > 0$. Ranking entries are reported means and dispersion, not confidence intervals. CRC-AUC and geometric AUROC have different positive labels.

Rule	sCRS	Passes	$H_B > 0$	CRC-AUC	U_B
No abstention	0.000	0/60	60/60	0.000 ± 0.000	0.500 ± 0.000
Calibration error	0.375	9/60	46/60	0.821 ± 0.176	0.649 ± 0.139
Ensemble spread	0.000	0/60	0/60	0.758 ± 0.191	0.640 ± 0.184
Conformal width	0.000	0/60	0/60	0.753 ± 0.196	0.742 ± 0.161
Geometric partition	0.545	47/60	0/60	0.618 ± 0.095	1.000 ± 0.000
Random ranking	—	—	—	0.000	0.500

Table 4: ID excess abstention areas. Dimensions correspond to the learning configurations, not the reference operators. Hits count recovered retained directions. Ratio is conformal D divided by geometric D.

			Excess area D				
Configuration	r_τ/b_τ	Hits	Conf.	Ens.	Cal.	Geom.	Ratio
Battery M0	1/16	1	0.029	0.029	0.029	0.029	100%
Battery M2	3/14	2	0.072	0.074	0.053	0.088	82%
Battery M8	8/9	1	0.126	0.121	0.050	0.235	54%
Melt-pool M12	9/8	4–6	0.195	0.178	0.141	0.265	74%
Flooding M12	12/5	7–11	0.063	0.047	0.053	0.353	18%

Within the reported aggregation scopes, calibration error gives higher CRC-AUC than conformal width (0.821 versus 0.753), but lower geometric AUROC (0.649 versus 0.742). CRC-AUC ranks *recovery hits*; geometric AUROC ranks *retained directions*. Their values are not directly comparable. Conformal-width geometric AUROC also varies by configuration: 0.909 for Battery M2, 0.769 for Battery M8, 0.868 for Melt-pool M12, and 0.589 for Flooding M12 (Appendix D).

5.5 Overall Ranking Can Conceal Budget-Specific Abstention

We sweep all $m + 1$ claim budgets in each of the five configurations, recording recovery yield $Y(Q)$ and below-threshold abstention $Z(Q)$. Figure 4 shows all abstention curves; complete recovery-yield curves appear in Appendix D.2. Curves average six model families; shading represents one standard deviation *across families*, not a seed-level confidence interval.

Conformal width yields positive excess area in every configuration, but relative to the geometric reference ranges from 100% in Battery M0 to 18% in Flooding (Table 4). Battery M0 remains a valid ranking comparison between one retained and 16 below-threshold directions. The 54% and 74% ratios for Battery M8 and Melt-pool show that the retained/below-threshold proportion alone does not explain performance.

Flooding M12 makes the budget dependence explicit (Figure 5). Although $U_B = 0.589 > 0.5$, mean conformal-width abstention falls below $1 - Q$ at all seven tested budgets $Q = 10/17, \ldots, 16/17$. At $Q = 14/17 \approx 0.824$, $Z = 0.067$, versus a random expectation of $3/17 \approx 0.176$. At $Q = 1$, both are zero. An overall AUROC above chance therefore does not ensure better-than-random abstention at every budget.

Prior correlations are one possible explanation: some low-sensitivity directions may be predictable in distribution and receive narrow intervals. Controlled experiments establish that such predictability can coexist with observational ambiguity, but do not establish this mechanism for all neural ranking differences. The evidence supports joint reporting of recovery, geometric ranking, and budget-dependent abstention; it does not show that conformal coverage guarantees fail.

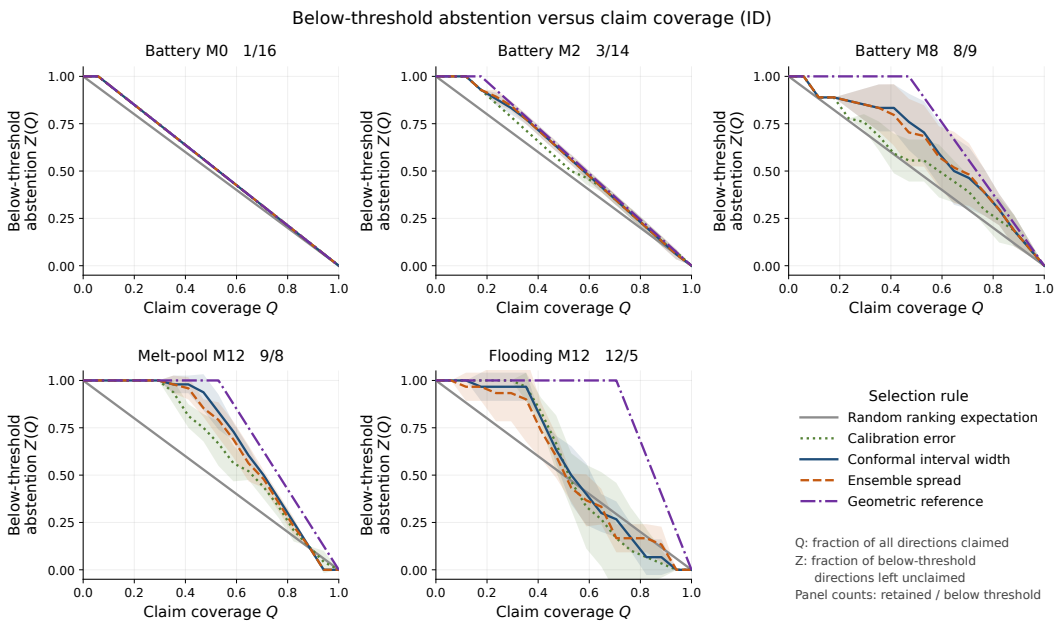

Figure 4: Below-threshold abstention across claim coverages (ID). Curves average six model families; shading shows one between-family standard deviation. The gray line is the random reference $1 - Q$. Panel titles give retained/below-threshold dimensions.

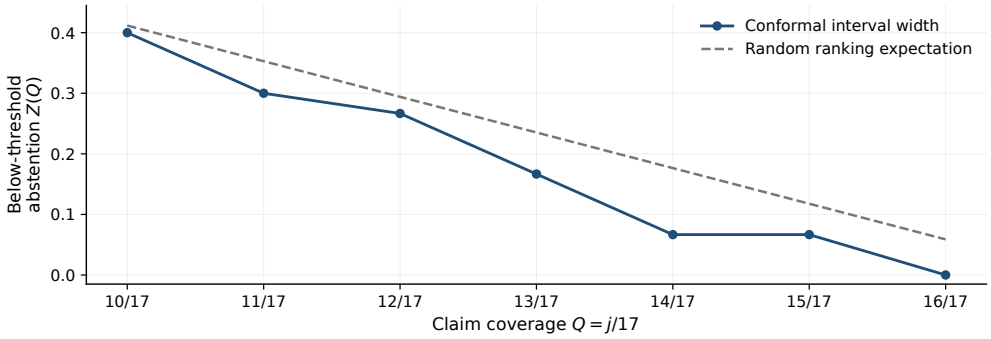

Figure 5: Flooding M12 at high claim coverages. Points average six model families; the dashed line is $1 - Q$. An above-chance overall U_B does not prevent below-random abstention at these budgets.

6 Conclusion

UniPDE-Bench evaluates local observation sensitivity, empirical recovery, and confidence-based selection in a shared directional basis. Its area identities clarify what ranking summaries contain, while its coverage curves expose behavior hidden by aggregate error and AUROC. Across three simulated operators and five learning configurations, low error coexists with below-threshold directions, and above-chance overall ranking can accompany below-random abstention at specific budgets. Partially observed PDE estimation should therefore be assessed through observation geometry, recovery yield, and budget-dependent selection together. The reference remains local, metric- and threshold-dependent, and sensitive to discretization near the threshold; the reported results concern the studied simulations, not global identifiability or statistical coverage certification.

AI USE STATEMENT

Generative AI assisted Chinese-to-English translation, language editing, and LaTeX typesetting. This manuscript-preparation assistance did not generate data or run the reported experiments.

REPRODUCIBILITY STATEMENT

Appendices A–E provide proofs, metric derivations, numerical diagnostics, supplementary results, and the complete evaluation procedure.

REFERENCES

Anastasios N. Angelopoulos and Stephen Bates. Conformal prediction: A gentle introduction. *Foundations and Trends in Machine Learning*, 16(4):494–591, 2023. doi: 10.1561/2200000101.

G. Backus and F. Gilbert. The resolving power of gross earth data. *Geophysical Journal International*, 16(2):169–205, 1968.

S. Bhadra, V. A. Kelkar, F. J. Brooks, and M. A. Anastasio. On hallucinations in tomographic image reconstruction. *IEEE Transactions on Medical Imaging*, 40(11):3249–3260, 2021.

T. Cui, J. Martin, Y. M. Marzouk, A. Solonen, and A. Spantini. Likelihood-informed dimension reduction for nonlinear inverse problems. *Inverse Problems*, 30(11):114015, 2014.

R. El-Yaniv and Y. Wiener. On the foundations of noise-free selective classification. *Journal of Machine Learning Research*, 11:1605–1641, 2010.

Y. Geifman and R. El-Yaniv. Selective classification for deep neural networks. In *Advances in Neural Information Processing Systems*, 2017.

N. M. Gottschling, V. Antun, A. C. Hansen, and B. Adcock. The troublesome kernel: On hallucinations, no free lunches, and the accuracy–stability tradeoff in inverse problems. *SIAM Review*, 67 (1):73–104, 2025. doi: 10.1137/23M1568739.

J. A. Hanley and B. J. McNeil. The meaning and use of the area under a receiver operating characteristic (ROC) curve. *Radiology*, 143(1):29–36, 1982. doi: 10.1148/radiology.143.1.7063747.

M. Herde, B. Raonić, T. Rohner, R. Käppeli, R. Molinaro, E. de Bézenac, and S. Mishra. Poseidon: Efficient foundation models for PDEs. In *Advances in Neural Information Processing Systems*, 2024.

B. Lakshminarayanan, A. Pritzel, and C. Blundell. Simple and scalable predictive uncertainty estimation using deep ensembles. In *Advances in Neural Information Processing Systems*, 2017.

Zongyi Li, Nikola Kovachki, Kamyar Azizzadenesheli, Burigede Liu, Kaushik Bhattacharya, Andrew Stuart, and Anima Anandkumar. Fourier neural operator for parametric partial differential equations. In *International Conference on Learning Representations*, 2021.

L. Lu, P. Jin, G. Pang, Z. Zhang, and G. E. Karniadakis. Learning nonlinear operators via DeepONet based on the universal approximation theorem of operators. *Nature Machine Intelligence*, 3: 218–229, 2021.

R. Ohana et al. The well: A large-scale collection of diverse physics simulations for machine learning. In *Advances in Neural Information Processing Systems, Datasets and Benchmarks Track*, 2024.

A. Spantini, A. Solonen, T. Cui, J. Martin, L. Tenorio, and Y. Marzouk. Optimal low-rank approximations of Bayesian linear inverse problems. *SIAM Journal on Scientific Computing*, 37(6): A2451–A2487, 2015.

M. Takamoto et al. PDEBench: An extensive benchmark for scientific machine learning. In *Advances in Neural Information Processing Systems, Datasets and Benchmarks Track*, 2022.

A. Tarantola. *Inverse Problem Theory and Methods for Model Parameter Estimation*. SIAM, 2005.

A. Yermakov, Y. Zhao, M. Denolle, Y. Ni, P. M. Wyder, J. Goldfeder, S. Riva, J. Williams, D. Zoro, A. S. Rude, M. Tomasetto, J. Germany, J. Bakarji, G. Maierhofer, M. Cranmer, and J. N. Kutz. The seismic wavefield common task framework. *arXiv preprint arXiv:2512.19927*, 2025.

A Additional Theory for Local Observation Geometry

A.1 Joint Ambiguity and Indistinguishability

Proof of Proposition 1. The two columns of $J = [1\ 1]$ are nonzero, yet $(1, -1)^\top$ lies in its null space. More generally, the rank–nullity theorem gives $\dim \ker J = m - \operatorname{rank}(J) \geq m - p$. Nonzero coordinate-wise responses therefore do not guarantee that the joint linear map is injective.

Proof of Proposition 2. When two states produce the same observation, any deterministic estimator returns the same estimate $\widehat{c}$. The triangle inequality gives

$$\|c_+ - c_-\|_G \leq \|c_+ - \widehat{c}\|_G + \|\widehat{c} - c_-\|_G.$$

At least one error is therefore no smaller than half the distance between the states. The result requires identical observations under the observation map itself, not merely a nontrivial Jacobian null space.

Proposition 3 (Indistinguishability under bounded noise)**.** *Suppose admissible noise satisfies* $\|\Sigma_y^{-1/2}\eta\|_2 \leq \beta$. *If* $c_+, c_- \in \mathcal{C}$ *and*

$$\|\Sigma_y^{-1/2}(\mathcal{H}(c_+) - \mathcal{H}(c_-))\|_2 \leq 2\beta,$$

then there are admissible noises associated with the two states that make their observations identical. The worst-case error of any deterministic estimator over these state–noise combinations is at least $\frac{1}{2}\|c_+ - c_-\|_G$.

Proof. Choose the common observation $y_* = (\mathcal{H}(c_+) + \mathcal{H}(c_-))/2$. The whitened norm of the noise required for either state is at most β, so the triangle-inequality argument in Proposition 2 applies directly. A noise covariance does not itself imply bounded noise; resolution under unbounded noise, such as Gaussian noise, requires a probabilistic criterion. □

A.2 Finite Perturbations and Local Linearization

Suppose the Jacobian of the whitened observation map

$$\widetilde{\mathcal{H}}(z) = \Sigma_y^{-1/2}\mathcal{H}(c_0 + G^{-1/2}z)$$

is L-Lipschitz on the relevant line segments. The linearization remainder then satisfies

$$\|\widetilde{\mathcal{H}}(z) - \widetilde{\mathcal{H}}(0) - Az\|_2 \leq \frac{L}{2}\|z\|_2^2.$$

Taking $z_\pm = \pm\delta v_i$ along a right singular direction gives

$$\|\widetilde{\mathcal{H}}(z_+) - \widetilde{\mathcal{H}}(z_-)\|_2 \leq 2\delta\sigma_i + L\delta^2.$$

This bound requires admissible perturbed states and control of the Lipschitz constant and perturbation amplitude. We do not estimate these bounds here; the finite-perturbation results are used only as local sensitivity diagnostics.

A.3 Observation Histories and Statistical Coverage

Observation histories. With an unchanged state parameterization, if O_{T+1} is obtained by appending observation rows to O_T, then $\ker O_{T+1} \subseteq \ker O_T$, so the exact rank cannot decrease. Relative effective rank, however, need not be monotone. At $\tau = 0.05$, $A_0 = \operatorname{diag}(1, 0.1)$ has effective rank 2; appending the row $(100, 0)$ changes its singular values to $\sqrt{10001}$ and 0.1, reducing the effective rank to 1. This change results from normalization by the largest singular value, not from a loss of observation information.

Exchangeability. For independently and identically distributed state samples (c_1, c_2), let $y = c_1$. The second coordinate is not identifiable from the observation over a two-dimensional admissible set, while the observation–target samples remain exchangeable. Marginal coverage guarantees therefore do not require the observation map to be injective, although distribution shift can invalidate these guarantees.

B Coverage-Curve Areas and AUROC

The derivations below use the classical equivalence between AUROC and the probability of correctly ordering a positive–negative pair, counting a tie as one half (Hanley & McNeil, 1982). They establish its relationship to the recovery-yield and below-threshold-abstention areas defined in this paper.

B.1 Recovery-Coverage Area

Let $t_1, \ldots, t_k$ be the ranks of the k successfully recovered directions in descending confidence order, starting at 1. Each included hit increases recovery yield by $1/m$. Integrating the piecewise-linear curve in Equation (7) gives

$$\int_0^1 Y(Q)\, dQ = \frac{1}{m^2} \sum_{a=1}^{k} \left(m - t_a + \frac{1}{2} \right).$$

The expected area under random ranking is $k/(2m)$. When all hits rank first, the area is $k/m - k^2/(2m^2)$. Let L_h count direction pairs in which a non-hit precedes a hit. For $0 < k < m$,

$$\mathrm{AUROC}_h = 1 - \frac{L_h}{k(m-k)}, \qquad \int_0^1 Y(Q)\, dQ = \frac{k}{m} - \frac{k^2}{2m^2} - \frac{L_h}{m^2}.$$

Substituting into Equation (8) yields the linear relationship between normalized recovery area and the AUROC of hit labels. This score describes ranking quality rather than recovery quantity.

B.2 Below-Threshold-Abstention Area

Write $r = r_\tau$ and $b = b_\tau$, with both classes nonempty. A below-threshold direction at rank t contributes $(t - \frac{1}{2})/(mb)$ to the area under the abstention curve. Summing its pairwise ordering relative to retained directions gives

$$\int_0^1 Z(Q)\, dQ = \frac{b}{2m} + \frac{r}{m} U_B.$$

Subtracting the random-reference area $1/2$ yields $D = (r/m)(U_B - 1/2)$. The geometric reference has $U_B = 1$, so $D_{\mathrm{geom}} = r/(2m)$, proving Equation (10). These identities compare areas over the full curve; even when $U_B > 1/2$, some budgets may satisfy $Z(Q) < 1 - Q$.

B.3 Ties and Boundary Cases

Taking the expectation over uniform random orderings within each tied group preserves the area identities above. The normalization denominator for recovery area is zero only when $k = 0$ or $k = m$; one or two hits remain scorable within the nondegenerate range. Geometric AUROC requires both direction classes to be nonempty. Whenever a score is undefined, the curves, direction counts, and reason are still reported.

C Reference Spectra and Numerical Sensitivity

C.1 Singular-Value Spectra of Reference Configurations

Table 6 lists the normalized spectra underlying Figure 2, rounded to three decimal places. At $\tau = 0.05$, the retained dimensions of Battery, Melt-pool, and Flooding are 12, 9, and 12. Determining exact rank requires the unrounded spectra and numerical rank tolerance.

C.2 Grid Resolution and Finite-Difference Step Size

Refining the grid from 48×48 to 64×64 leaves the effective ranks of Battery and Flooding at 12, while increasing Melt-pool from 7 to 9. This change indicates sensitivity of near-threshold singular directions to discretization; effective rank alone is insufficient to establish subspace convergence.

Table 5: Joint-Jacobian diagnostics for the three reference configurations ($m = 17$, $\tau = 0.05$, 64×64 grid). The below-threshold dimension is the state dimension minus the effective rank, not the exact null-space dimension. Coordinate-wise conclusions concern independent perturbations of individual coefficients.

Operator	J size	Structural rank[†]	Effective rank	Below threshold	Coordinate-wise finding
Battery	14×17	13	12	5	High-frequency gain decay
Melt-pool	13×17	13	9	8	High-frequency gain decay
Flooding	13×17	13	12	5	"No blind modes"

[†]Structural-rank values are reported as in the reference diagnostics. Exact numerical rank requires the unrounded spectrum and a separately specified numerical rank tolerance.

Table 6: Normalized singular-value spectra for the three reference configurations. Displayed zeros may result from padding or rounding and do not establish exact null-space dimensions.

Index	Battery	Melt-pool	Flooding
1	1.000	1.000	1.000
2	0.713	0.682	0.961
3	0.713	0.682	0.960
4	0.488	0.362	0.960
5	0.452	0.335	0.927
6	0.311	0.170	0.757
7	0.311	0.170	0.740
8	0.226	0.089	0.739
9	0.200	0.077	0.739
10	0.154	0.041	0.703
11	0.154	0.041	0.701
12	0.135	0.028	0.701
13	0.004	0.000	0.039
14–17	0.000	0.000	0.000

As the finite-difference step varies from 10^{-2} to 3×10^{-4}, the largest change in the Battery Jacobian norm between adjacent step sizes is 5.0×10^{-6}. Along the least sensitive direction, the ratios of observation change to shape change are 2.0×10^{-7} for Battery, 9.2×10^{-14} for Melt-pool, and 4.1×10^{-13} for Flooding.

These norms and ratios support only local diagnostics. Establishing subspace convergence additionally requires comparing $\|J_h - J_{h/2}\|$, near-threshold singular values, and projection-matrix distances, and reporting perturbation amplitudes and absolute whitened observation differences.

D Supplementary Experimental Results

D.1 Controlled Prediction and Geometric Ranking

Figure 6 provides the graphical counterpart of Table 1. Table 7 reports configuration-wise geometric AUROC for conformal-width rankings; the normalized excess area contains the same ranking information.

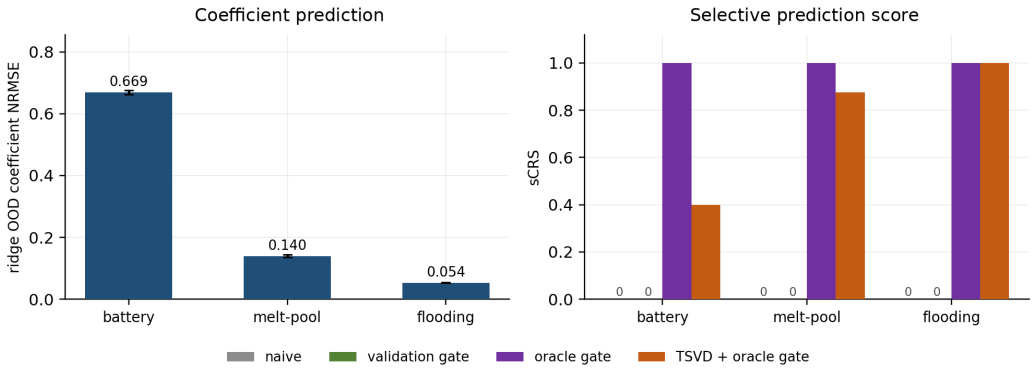

Figure 6: Prediction accuracy and selective scores in the controlled inverse problems. Left: ridge-regression OOD coefficient NRMSE, with error bars showing standard deviations over five seeds. Right: sCRS for no abstention, validation-error gating, Oracle gating, and TSVD + Oracle.

Table 7: Geometric AUROC for rankings by conformal interval width. Retained directions are positives; larger AUROC means they tend to receive higher confidence. The last column is the normalized excess area from Equation (10), which contains the same ranking information as AUROC.

Configuration	r_τ	b_τ	U_B	$2U_B - 1$
Battery M2	3	14	0.909	0.818
Battery M8	8	9	0.769	0.538
Melt-pool M12	9	8	0.868	0.736
Flooding M12	12	5	0.589	0.178

D.2 Complete Recovery and Abstention Curves

Figure 7 shows recovery yield for all five learning configurations; Figure 4 in the main text shows the corresponding abstention curves. Model-family means and between-family standard deviations describe the reported curves; they are not confidence intervals over seeds. The recovery and abstention curves are separate views of the same confidence order, and should be read together with hit counts and the geometric dimensions.

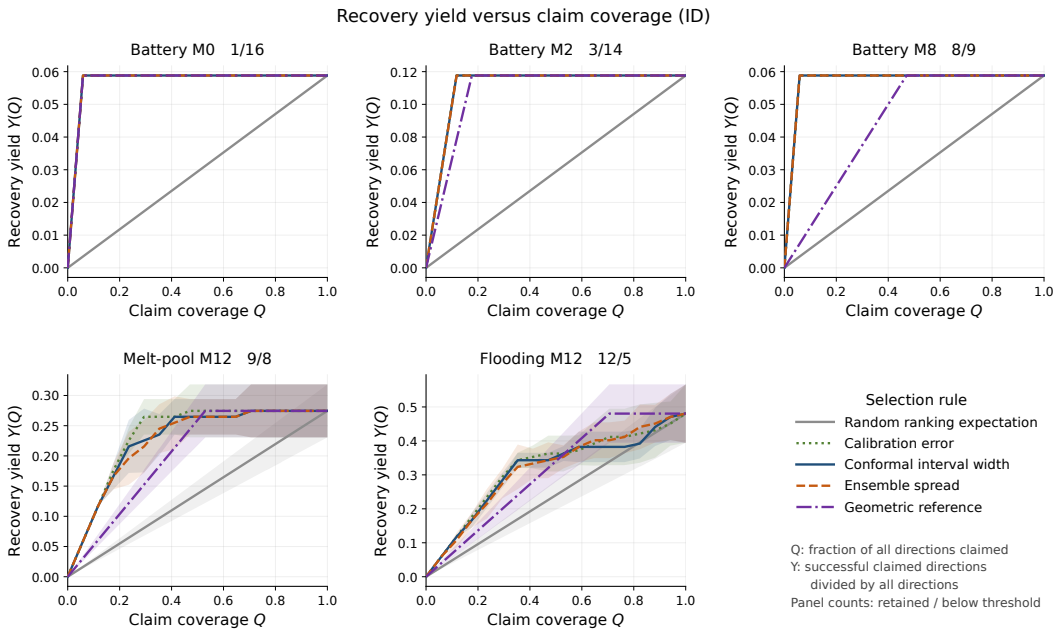

Figure 7: Recovery yield $Y(Q)$ versus claim coverage Q under ID conditions. The horizontal axis is the fraction of all directions claimed; the vertical axis is the number of successfully recovered claimed directions divided by all directions. Curves are means over six model families. Shading shows one standard deviation across model families, not a random-seed confidence interval. The gray straight line is the expected random-ranking curve under constant confidence, $Y(Q) = (k/m)Q$. Panel titles give retained/below-threshold dimensions; vertical scales differ across panels.

D.3 High Claim Budgets in Flooding

Table 8 supplements the Flooding M12 ID results in Section 5.5, covering claim budgets $j = 10, \ldots, 16$. The random reference uses the exact coverage $Q = j/17$. Abstention is below the random reference at all seven budgets, consistent with an overall geometric AUROC above chance (Appendix B.2).

Table 8: High-claim-budget results for Flooding M12, averaged over six model families.

j	$Q = j/17$	Conformal Z	Random $1 - Q$	Difference
10	0.588	0.400	0.412	−0.012
11	0.647	0.300	0.353	−0.053
12	0.706	0.267	0.294	−0.027
13	0.765	0.167	0.235	−0.069
14	0.824	0.067	0.176	−0.110
15	0.882	0.067	0.118	−0.051
16	0.941	0.000	0.059	−0.059

D.4 Poseidon-T Adaptation and Prediction Averaging

Experimental setup. The adapted Poseidon-T (Herde et al., 2024) has 20.77M parameters, takes rasterized sensor histories as input, and predicts 17 interface coefficients. The dataset contains 576 trajectories. For each configuration, five random seeds are trained for a budget of 200 epochs with validation-based early stopping. We compare OOD coefficient NRMSE for individual models and the average prediction of five models. This experiment assesses point estimation, not backbone rankings under equal computational budgets.

Table 9: Individual-model predictions and five-model prediction averaging for Poseidon-T. Single-model OOD coefficient NRMSE is the mean ± standard deviation over five seeds.

Configuration	Single-model OOD NRMSE	Averaged OOD NRMSE	Relative change	No abstention sCRS	Oracle sCRS
Battery M0	0.149 ± 0.007	0.148	-0.7%	0.000	1.000
Battery M2	0.159 ± 0.026	0.156	-1.8%	0.000	0.840
Battery M8	0.148 ± 0.026	0.144	-2.5%	0.000	0.857
Melt-pool M12	0.193 ± 0.040	0.121	-37.1%	0.000	0.562
Flooding M12	0.171 ± 0.056	0.126	-26.4%	0.000	0.836
Macro-average	0.164	0.139	—	—	—

Results. Prediction averaging reduces OOD coefficient error in all five configurations, lowering macro-averaged NRMSE from 0.164 to 0.139. The reductions are larger for Melt-pool M12, from 0.193 to 0.121, and Flooding M12, from 0.171 to 0.126 (Figure 8).

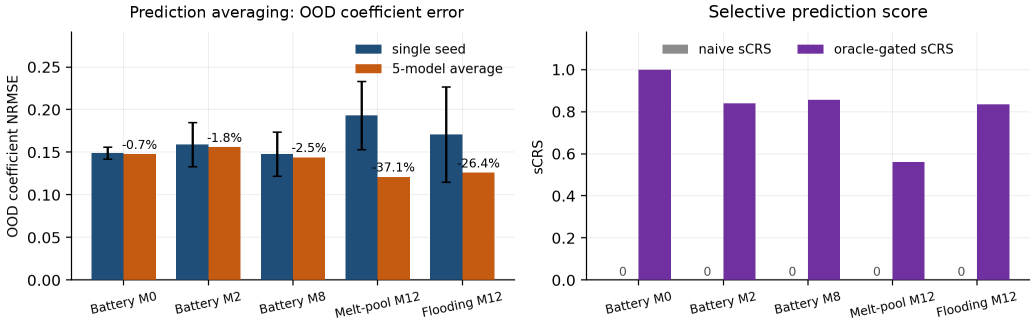

Figure 8: Prediction averaging and selective scores for Poseidon-T. Left: single-model and five-model-averaged OOD coefficient NRMSE; error bars show the five-seed standard deviations of individual models. Right: sCRS for no abstention and Oracle gating.

All 25 point-estimation models have zero no-abstention sCRS, reflecting their no-abstention behavior in configurations with below-threshold directions; this alone does not determine their recovery capability. Prediction averaging is also not equivalent to effective confidence-based selection. The latter requires a separate evaluation of ensemble spread using observation geometry that includes the same history inputs.

E Evaluation Algorithm and Reporting Conventions

Algorithm 1 makes the separation between scoring and selection explicit. Geometry is constructed once per observation specification, while recovery outcomes depend on the estimator and evaluation split. Confidence records must identify the same directional basis. All metrics are computed before aggregation across models or seeds.

Algorithm 1 Direction-wise evaluation with UniPDE-Bench

Require: Observation specification $(\mathcal{H}, c_0, G, \Sigma_y)$; scoring specification (τ, ε); numerical, tie, and aggregation conventions; trained estimator f; test set $\{(y_n, c_n)\}_{n=1}^{N}$; basis-identified confidence record u.
Ensure: Directional records, fixed-budget metrics, coverage curves, and ranking scores.
1: Check that $\mathcal{H}$ includes all history and auxiliary inputs used by f.
2: Compute $J = D\mathcal{H}(c_0)$ and $A = \Sigma_y^{-1/2} J G^{-1/2}$.
3: Compute the full right singular basis V, keeping all m directions and singular values.
4: Form $\mathcal{R}_\tau, \mathcal{B}_\tau, P_\tau$ using Equation (3); if $\sigma_1 = 0$, all directions are below threshold.
5: Compute $\widehat{c}_n = f(y_n)$ and project $c_n, \widehat{c}_n$ to $q_n, \widehat{q}_n$.
6: Compute s_i, e_i, h_i by Equation (4); set $e_i = +\infty, h_i = 0$ if $s_i = 0$.
7: Read the u_i matched to V and freeze ranking π; do not update it with test errors.
8: **for** $j = 0, \ldots, m$ **do**
9: Set $a_{\pi(\ell)} = 0$ for $\ell \le j$ and $a_{\pi(\ell)} = 1$ otherwise.
10: Compute $Q_j, R_j, A_{B,j}, H_{B,j}, Y_j, Z_j$ by Equations (5) and (7).
11: **end for**
12: Average curve contributions uniformly over tied orderings; integrate piecewise linearly.
13: Set $k = \sum_i h_i$; compute CRC-AUC if $0 < k < m$, otherwise record it as undefined.
14: Compute $U_B, D, D/D_{\mathrm{geom}}$ if $r_\tau, b_\tau > 0$; otherwise record geometry AUROC as undefined.
15: Save specification identifiers, $V, \sigma, e, h, u, k, r_\tau, b_\tau$, curves, scores, and undefined-value reasons.

Boundary and tie conventions. Set $R_j = 0$ when $r_\tau = 0$. When $b_\tau = 0$, Z_j and $A_{B,j}$ are not applicable. Curves and direction counts are reported even when an AUROC normalization is undefined. Tied groups use expected curve contributions; hard decision criteria are evaluated on actual claim decisions rather than on an expected fractional mask. Soft and strict composite criteria are recorded separately.

Records and aggregation. Each evaluation record contains the observation and scoring specifications; numerical rank and relative-threshold conventions; the full spectrum and basis; directional errors, hit labels and confidence scores; the hit count; all fixed-budget scores and coverage curves; and reasons for undefined metrics. Results are first computed for individual configurations, models, seeds, splits, and selection rules, and only then aggregated according to the declared rule. Reuse of a confidence vector does not supply an independent geometric-ranking observation.