EVALBENCH: BENCHMARKING REFERENCE-FREE PDE SCORES FROM RANKING TO SEARCH

Anonymous authors
Paper under double-blind review

ABSTRACT

PDE program search relies on reference-free scores, yet accurate ranking of a fixed candidate bank need not translate into accurate solutions when a score guides search. We introduce EVALBENCH, a benchmark that separates target identifiability, static score–error rank agreement (M1), and the error of the program returned by active search (M2). Core comparisons hold the proposer and fitter fixed; parameter, grammar, and proposer shifts test robustness. Across seven identifiable one-dimensional targets, five combined or resampled scores achieve median Spearman correlations of 0.813–0.817, compared with 0.417 for raw residual. Nevertheless, held-out scoring gives a median paired $\log_{10}$ error ratio of -0.0036 against baseline (95% interval $[-0.032, 0.003]$), showing no clear active-search gain. A Keller–Segel scale-family control demonstrates why ranking cannot resolve an incomplete public specification. SAAS, an operator-aware rank aggregator, fails its prespecified success criterion; a separate ten-operator development study localizes gains and persistent failures. We provide conditional transfer bounds under distribution coverage and simultaneous risk certificates for fixed selectors using independent calibration seeds from the deployment distribution. EVALBENCH makes specification ambiguity, ranking failure, and search limitations separately measurable, while reporting per-operator errors, catastrophic outcomes, coverage, and evaluation cost.

1 INTRODUCTION

Program-search systems construct mathematical objects by proposing code, executing it, and retaining promising candidates (Romera-Paredes et al., 2024; Novikov et al., 2025; Yu & Yang, 2026). For a PDE, the desired solution is usually unavailable during search. Candidate programs must therefore be evaluated through the equation, initial and boundary data, and numerical consistency checks. The practical question is whether selecting for these scores produces more accurate solutions.

Three obstacles arise. The public problem may admit several solutions while the evaluator chooses one as its target. Even when that target is identified, a score can rank candidate programs incorrectly. Finally, a score can rank a fixed bank well yet lose that relationship as repeated selection changes the candidate distribution. This last obstacle resembles the error–regret gap in learned rewards (Gao et al., 2023; Fluri et al., 2025): predictive agreement on observed candidates does not determine the outcome of optimizing a proxy.

Existing PDE benchmarks mainly compare solvers against reference fields (Takamoto et al., 2022; Hao et al., 2023). Equation-discovery benchmarks measure recovery of governing laws from observations (Brunton et al., 2016; Udrescu & Tegmark, 2020; Shojaee et al., 2025b), while a-posteriori analysis provides error estimates under assumptions on the operator and approximation (Becker & Rannacher, 2001; Hillebrecht & Unger, 2022; Liu et al., 2023). Here the object of evaluation is the score used inside solution search. We compare scores with a common proposer and fitter, then test how those comparisons change when the fitting loss or proposal distribution changes.

EVALBENCH makes three contributions. First, it separates target identifiability, static rank agreement (M1), and the error of the publicly selected return (M2; Table 1). Second, paired experiments across local, radial, free-boundary, and nonlocal problems show where improved rank agreement fails to translate into search gains. Parameter, grammar, and proposer shifts test the stability of this finding. Third, theoretical results state the conditions needed for identification, error transfer, and

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.

risk calibration. We also test SAAS, which combines operator-compatible signals and penalizes disagreement. Its mixed results help distinguish failures of the scoring rule from limitations of the candidates available to it.

2 PROTOCOL

Public problem and withheld reference. For case c, let $P_c = (\mathcal{N}_c, \mathcal{B}_c, \mathcal{I}_c)$ denote the equation and boundary/initial constraints on the declared domain. A score S uses this specification, the candidate program u, public collocation points, and operator metadata. Scores are oriented so that smaller values are preferred. The benchmark runner holds the reference $u_c^\star$ and evaluates relative error

$$E_c(u) = \frac{\|u - u_c^\star\|_2}{\|u_c^\star\|_2}. \tag{1}$$

Reference errors are attached after search and never passed to the selector.

Stage 0: specification gate. Define the public zero set

$$\mathcal{Z}_c = \{u : \mathcal{N}_c[u] = 0,\ \mathcal{B}_c[u] = 0,\ \mathcal{I}_c[u] = 0\}. \tag{2}$$

If the specification leaves several exact solutions admissible, it does not tell the search which one the hidden reference selects. We test this ambiguity constructively. Radial Keller–Segel Liouville profiles form a scale family at critical mass 8π (Dávila et al., 2024). Matched cases with and without the center-value anchor $u(0) = 8a^2$ isolate the missing scale. The unanchored case tests the specification and is excluded from evaluator leaderboards.

The executable audit checks declared scale, translation, and sign symmetries. Each must have a finite public anchor, with scale-dependent anchors consistent with the public parameter. The checks detect removed, zeroed, non-finite, and inconsistent anchors independently of case names. They audit declared symmetries; establishing PDE uniqueness requires a separate argument.

Theorem 1 (Reference-free non-identifiability)**.** *Let a public specification P admit two solutions $u_1, u_2 \in \mathcal{Z}(P)$ with $\|u_1 - u_2\|_2 > 2\varepsilon$. For any possibly randomized algorithm A whose inputs are measurable functions of P alone, at least one choice of hidden target $u^\star \in \{u_1, u_2\}$ satisfies*

$$\Pr(\|A(P) - u^\star\|_2 \le \varepsilon) \le \tfrac{1}{2}. \tag{3}$$

Thus no reference-free score or search rule can uniformly recover an arbitrarily chosen member of an unresolved public equivalence class.

The proof in Appendix A.1 uses two disjoint error balls. Stage 0 therefore precedes any comparison of empirical scores.

Stage 1: static evaluability (M1). For a fixed bank $\mathcal{C}_c$, we report

$$\mathrm{M1}_{c,S} = \rho_{\mathrm{Spearman}}\big(S(u; P_c), E_c(u)\big), \qquad u \in \mathcal{C}_c, \tag{4}$$

where positive correlation means that low scores tend to identify low-error candidates. Each 300-program bank is balanced across eight baseline search seeds. The 95% interval resamples whole seeds to retain dependence among related programs. Top-decile precision is a secondary measure because ties near the selection threshold can change it substantially.

Stage 2: active behavior (M2). For selector S, seed s, and case c, let $\hat{u}_{c,S,s}$ be the rank-0 candidate returned at the final iteration. We compare it with the baseline four-component physics score S_0 on the same seed:

$$G_{c,S,s} = \log_{10} \frac{E_c(\hat{u}_{c,S,s})}{E_c(\hat{u}_{c,S_0,s})}. \tag{5}$$

Negative G is improvement, zero is parity, and positive G is degradation. We report the paired median, seed-bootstrap interval, and win rate. Equation 5 uses the returned program, even when another survivor has lower reference error. The best survivor is recorded separately to diagnose pool quality. Aggregate intervals resample cases as clusters; per-case results show which operators drive an apparent gain or loss.

Stage	Question	Failure interpretation
0	Does the public problem identify the target?	Hidden-label ranking is not a valid task.
M1	Does a fixed-bank score agree with withheld error?	Evaluator or discretization mismatch.
M2	What error does score-driven search return?	Selection, fitting, or proposal failure.
Shift	Does the result survive parameter, grammar, and proposer changes?	Benchmark-distribution dependence.

Table 1: The stages answer different questions and must be applied in order.

Target	Public structure	Role
Poisson	2-D elliptic, Dirichlet	dimension control
Heat	diffusion, initial/boundary	bounded control
Burgers	viscous traveling front	sharp layer
Allen–Cahn	steady reaction–diffusion	nonlinear control
NLS	steady homoclinic profile	anchor control
Keller–Segel	radial log diffusion	scale control
Porous medium	Barenblatt free boundary	nonsmooth support
gCLM	nonlocal spectral dynamics	nonlocal stress

Table 2: Eight identifiable targets. Keller–Segel is additionally run without its scale anchor as a Stage-0 specification control, not as a leaderboard case.

M1 measures order, not error calibration. Theorem 3 connects bank-level calibration to active error when the selected distribution is covered by the bank distribution with a bounded density ratio. Selection outside that support is unconstrained by bank calibration (Appendix A.1).

Theorem 2 (Simultaneous calibration-seed risk certificate)**.** *Fix K selectors before calibration and let $Z_{ij} = \mathbf{1}\{E(\hat{u}_{ij}) > \tau\}$ be i.i.d. Bernoulli losses across n calibration seeds for each selector j. Let U_j be its one-sided Clopper–Pearson upper limit at level $1 - \delta/K$ (Clopper & Pearson, 1934). Then, with probability at least $1 - \delta$, $R_j = \mathbb{E}Z_{ij} \le U_j$ simultaneously for every j. Consequently any data-dependent choice from $\{j : U_j \le \alpha\}$ has catastrophic risk at most α on a future seed from the same distribution, on the same event.*

Simultaneous coverage allows the calibration data to choose among the fixed selectors. An interval for only the empirically best selector would omit this selection step. Appendix A.3 gives the proof.

Independent validation protocol. Selectors are frozen on seeds 0–7; seeds 8–23 are opened only after case-wise choice and certification. The executable Stage 0 contract audits the public specification, and a frozen public operator router is evaluated on two held-out families, 2-D Helmholtz and the 1-D wave equation. We use $\tau = 0.5$, family-wise $\delta = 0.05$, and $\alpha = 0.5$; the system abstains if no selector satisfies $U_j \le \alpha$. Coverage, catastrophic/all, solved/all, and conditional median error are reported together.

3 Benchmark Construction

Cases. Table 2 lists the operators; Figure 1 shows representative targets. The two-dimensional Poisson case uses both spatial coordinates in the candidate and residual. The porous-medium target is a first-kind Barenblatt profile with a moving support boundary. The nonlocal Constantin–Lax–Majda (gCLM) residual is evaluated by a spectral Hilbert transform.

Programs and proposers. Candidates are sums of sinusoidal, front, localized, self-similar, Gaussian, and polynomial atoms. High-frequency and narrow-spike atoms test errors between collocation points. In the core comparison, L-BFGS fits continuous parameters using the four-component loss; only the outer selection score changes. The proposer retains eight programs from each pool of 24 over 12 iterations. The no-decoy grammar removes the two adversarial atoms. A second proposer keeps

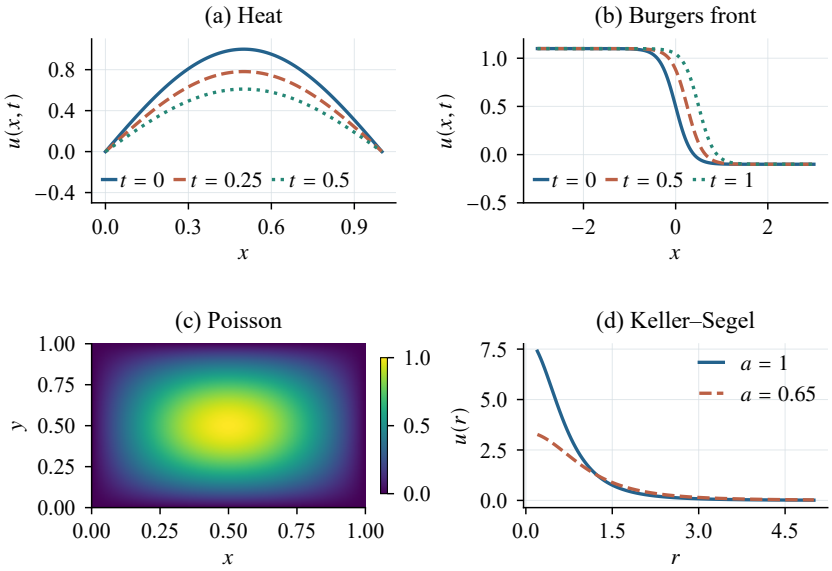

Figure 1: Representative target solutions. The Keller–Segel scale family motivates the specification gate.

elites but samples each new structure independently, using the same grammar, fitter, and pool size. This random-restart control tests whether score performance depends on evolutionary parent–child relationships.

Signals. The scalar library contains residual, initial, boundary, compatibility, their weighted four-component sum, held-out collocation, a four-draw ensemble, multi-slice consistency, higher-order stencil consistency, curvature, parameter count, and a variation-weighted residual. The latter uses $|u_x| + |u_t| + 10^{-3}$, replacing u_t by u_y in the 2-D case. It weights the residual by local variation and requires no adjoint solve. Active comparisons also include rank aggregation and a development-selected operator router. Each score sees the same public problem and its permitted sampling rule.

Specification-aware agreement selection (SAAS). SAAS combines public signals whose numerical assumptions match the operator. For a pool of m programs, $r_k(u) \in [0, 1]$ is the zero-based rank under signal k, divided by $\max(m - 1, 1)$; the best rank is zero. Public operator tags determine the signal set $K(P)$. Free-boundary cases use the baseline, held-out, ensemble, multi-slice, and variation-weighted scores. Nonlocal cases omit the baseline from that set. Radial cases replace multi-slice by higher-order consistency. Other cases use baseline, held-out, ensemble, multi-slice, and higher-order scores. The selector minimizes

$$A_P(u) = 0.45 \operatorname*{mean}_{k \in K(P)} r_k(u) + 0.45 \max_{k \in K(P)} r_k(u) + 0.10 \operatorname*{sd}_{k \in K(P)} r_k(u). \tag{6}$$

The mean favors candidates that score well overall. The maximum prevents a poor signal rank from being hidden by averaging, and the standard deviation adds a disagreement penalty. Ranks allow signals with different units to be combined. SAAS retains the lowest-scoring programs and returns the final rank-zero candidate. Each component uses $\max(16, \lfloor n/4 \rfloor)$ nominal points for a requested grid size n; actual prediction counts also depend on the component and operator.

SAAS-mean keeps only the average rank. SAAS-no-metadata uses all six signals for every operator. SAAS-no-holdout removes held-out and ensemble scores. These ablations test aggregation, operator-specific signal choice, and independent sampling. Proposition 5 characterizes the fixed-pool rank aggregator; the active comparison measures its effect on subsequent pools and returned error.

Experimental matrix. The core and no-decoy suites each contain $8 \times 10 \times 8 = 640$ searches; ten parameter variants contribute 800 searches; the random-restart suite adds 640; and the 2-D control

adds 80. Every search logs all 288 pool events, including lineage, score, acceptance, and offline error. We use deterministic grids for M1, distinct public seeds for resampled signals, and no hidden-label tuning within the static-evaluation suite. Extreme non-finite candidate evaluations receive the existing worst-score guard and remain visible in the logs.

The independent selection study contains 1,440 baseline-fitter searches on 16 new seeds and 576 held-out-fitter searches. A matched 2×4 factorial crosses two equal-cost continuous objectives (baseline and a fixed held-out draw) with baseline, held-out, ensemble, and higher-order outer selectors. The same nine cases, seeds, grammar, population, L-BFGS limit, and outer budget are used in every cell. This tests how the fitting objective interacts with the outer selector.

The registered selection study contains 5,200 searches. Seeds 24–47 cover ten identifiable operators and 12 selectors: eight controls, SAAS, and its three ablations. Seeds 48–55 cover no-decoy, random-restart, and generic-only proposal regimes; seeds 56–63 cover ten public parameter shifts. Its compound criterion requires a 50% catastrophe reduction, no new catastrophes, solved-rate loss at most two percentage points, a nonpositive upper interval for median G, and a paired catastrophe-test $p < 0.05$.

The ten-operator diagnostic uses finite-domain manufactured data and seeded gCLM time slices with 12 selectors on seeds 24–31 (960 searches). These seeds are development data, and the numerical specifications differ from those of the registered study. We analyze the studies separately. Each diagnostic run records time and actual prediction points for fitting and scoring, allowing error differences to be assessed alongside the work required to obtain them.

Bank sampling. The baseline log can contain the same elite in several consecutive iterations. Before M1 sampling we deduplicate exact program JSON within each seed, then allocate the 300 slots as evenly as possible across seeds. Rediscovery in a different seed is retained as an independent cluster. Sorting precedes seeded sampling, so parallel completion order cannot change a bank. Offline errors are attached only after search and do not affect bank sampling.

4 Results

The core suite shows stronger fixed-bank ranking without a clear aggregate gain in returned error. The results locate this ranking-to-search gap.

4.1 An identifiable target is a prerequisite for evaluating scores

Figure 2 isolates the effect of the Keller–Segel scale anchor. With critical mass but no scale anchor, residual M1 is -0.233 (95% CI $[-0.372, -0.057]$); with the center-value anchor it becomes 0.111 ($[-0.072, 0.283]$). Variation weighting rises from 0.349 to 0.637 after anchoring. A high-order score nevertheless reaches 0.868 on the unanchored bank. This ranking can reflect a stencil's preference for a particular scale: every exact member of the family still satisfies the unanchored continuum problem. A high observed correlation therefore cannot supply the missing public constraint.

4.2 Static ranking improves, with operator-dependent failure modes

Table 3 summarizes the per-operator rankings in Figure 3. Raw residual M1 has median 0.417 and is useful on several local cases, but reverses on the nonlocal gCLM target (-0.569). Variation weighting, the four-component score, and three resampled scores have median M1 0.813–0.817 with positive cluster intervals on every case. Higher-order consistency is negative on the porous-medium profile (-0.484), where the true free-boundary curvature is not a numerical artifact. No signal is uniformly best.

The 2-D Poisson control supplies a useful positive case. Residual M1 is 0.773 with seed-cluster 95% CI $[0.722, 0.822]$, close to held-out M1 0.772. Residual ranking can therefore be informative in two dimensions. Together with the radial and nonlocal cases, this supports reporting score quality by operator and specification rather than treating dimension as its explanation.

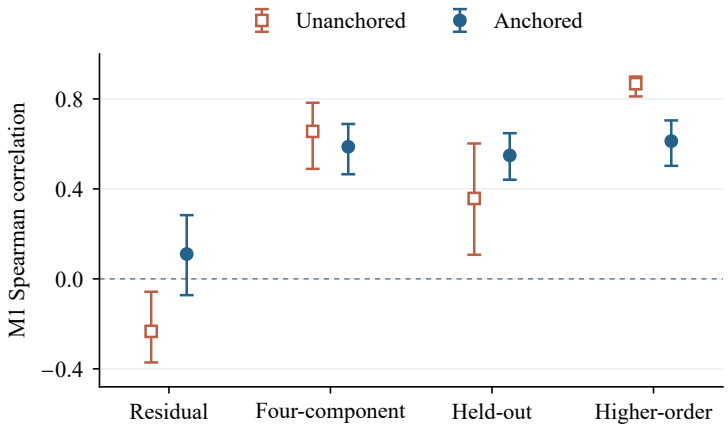

Figure 2: Matched Keller–Segel specifications. Static correlation cannot replace the analytic identifiability check; an apparently high M1 may exploit a numerical preference among continuum-equivalent solutions.

Signal	Median M1	Worst M1	CI > 0
Variation-weighted	0.817	0.350	7/7
Four-component	0.817	0.367	7/7
Ensemble	0.816	0.368	7/7
Multi-slice	0.815	0.368	7/7
Held-out	0.813	0.366	7/7
Higher-order	0.658	-0.484	6/7
Residual	0.417	-0.569	5/7
Curvature	-0.051	-0.761	2/7

Table 3: Five combined or resampled signals have positive M1 intervals on all seven identifiable 1-D targets; residual and higher-order scores reverse on particular operators. The unidentified control is excluded. "CI > 0" counts seed-cluster intervals with a positive lower endpoint.

4.3 Better static ranking does not ensure better active search

Figure 4 summarizes active selection. Across all seven 1-D targets, held-out scoring has median paired $G = -0.0036$ with case-cluster 95% CI $[-0.032, 0.003]$ and wins 54% of seeds: no clear gain despite strong static M1. Ensemble and variation weighting are at median parity. Higher-order consistency has $G = +0.074$ in the aggregate, but this hides severe operator failures: $G = 2.13$ on Burgers and 4.57 on porous medium. Smoothness has aggregate $G = 0.180$ ($[0.013, 1.826]$) and wins only 16% of seeds.

The 2-D control quantifies the cost of optimizing a score that M1 flags as unreliable. The baseline returns median error 8.24×10^{-7}. Higher-order and held-out objectives remain within sampling variation of that value, but the smoothness objective returns median error 2.08 ($G = +6.41$). Curvature has negative static M1 (-0.487) here. This example supports the value of static screening; the strong-M1 selectors above show why active testing is additionally needed.

4.4 Mitigation tests and the limits of selector improvements

In the independent selection test, the frozen case-wise best-M1 policy has zero catastrophic returns and solves 86 of 112 case–seed decisions, compared with two catastrophes and 88 solved for baseline, at full coverage. Risk-based abstention reduces solved/all by 8.9 percentage points. The matched fitting factorial yields negative median G on six of nine operators and positive values on three (Appendix B.1). Thus typical error, tail failures, and coverage respond differently to evaluator choice.

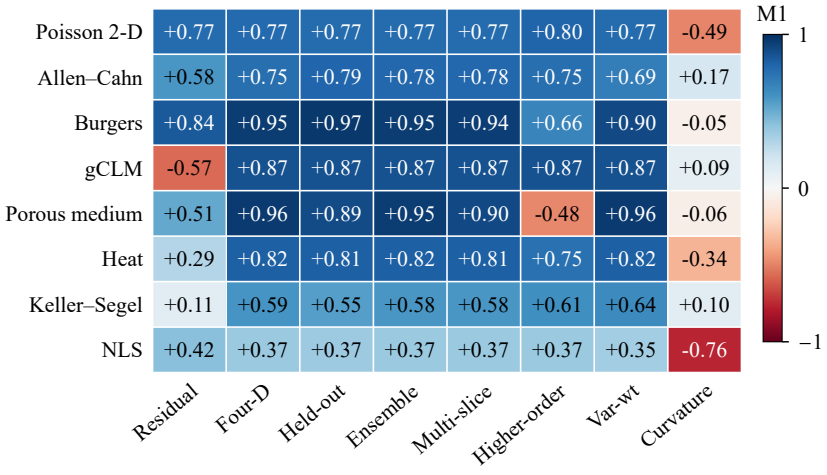

Figure 3: Static score–error agreement varies by operator, including reversals for residual on gCLM and higher-order consistency on porous medium. Programs are deduplicated within each seed before M1 is computed.

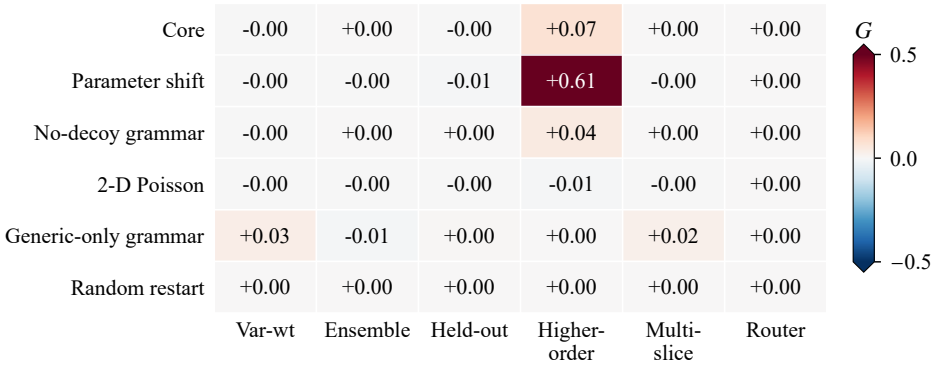

Figure 4: Better static ranking does not consistently reduce returned error. Paired log-error ratios compare each active selector with its same-seed baseline; negative values improve and positive values degrade it.

In the registered SAAS comparison, baseline solves 203 of 240 main searches with two catastrophes; SAAS solves 209 with none. The paired exact catastrophe test gives $p = 0.5$, and the interval for median log-error crosses zero.

The prespecified compound criterion is not met. With only two discordant catastrophic outcomes, the observed decrease from two failures to none gives too little evidence to replace the baseline with SAAS on these data.

The ten-operator diagnostic separates well-resolved cases from persistent failures (Table 4). Baseline solves all 72 runs outside gCLM, including $8/8$ Keller–Segel runs with median error 2.37×10^{-5}. It solves $0/8$ gCLM runs, whose median error is 0.982, with six catastrophes. Under the finite-domain specification, the radial task is already solved by baseline. The nonlocal task accounts for all its unsolved runs.

SAAS and baseline both solve 72 runs with six catastrophes. Variation weighting reduces gCLM catastrophes from six to five, fixing two seeds while introducing one failure. This mixed paired outcome explains why the count advantage alone gives little evidence of a consistent improvement.

Selector	Solved	Cat.	Med. G	Pred. (M)
Baseline	72	6	0.0000	115.4
Held-out	72	7	-0.0004	117.5
Ensemble	72	7	0.0005	119.2
Multi-slice	72	7	0.0005	115.4
Higher-order	65	7	0.6591	157.7
Variation-weighted	72	5	-0.0000	115.0
Rank combination	72	6	0.0020	117.8
Operator router	72	6	0.0000	115.4
SAAS	72	6	-0.0002	122.6
SAAS-mean	72	6	-0.0004	125.1
SAAS-no-metadata	72	6	0.0006	114.4
SAAS-no-holdout	72	6	0.0017	114.1

Table 4: All 12 selectors in the ten-operator diagnostic study (80 runs each). Solved and catastrophic counts use $E < 0.05$ and $E > 0.5$; G is paired with baseline. "Pred." is the median number of candidate-prediction points and reveals unequal actual cost. Seeds 24–31 are development data.

Higher-order selection loses seven solved runs and uses 37% more prediction points. A localized development effect is large: on Keller–Segel, SAAS-no-holdout reduces median error from 2.37×10^{-5} to 9.35×10^{-7} (median $G = -1.364$, seed-bootstrap 95% interval $[-1.541, -0.905]$), with $8/8$ solved in both arms. The gain concerns accuracy within the solved regime, rather than additional solved runs.

Appendix B.2 reports all 120 operator–selector cells for error, success, and cost. The solution atlas in Appendix C shows where errors occur in the returned fields. These views distinguish small improvements across solved cases from concentrated failures that are obscured by a single aggregate.

Finally, a privileged fitting diagnostic refits the public top-four gCLM structures without adding atoms. Neither public nor exact-label fitting solves any of 64 structure–budget pairs; even at 400 iterations the best exact-label error is 0.0858. Even direct access to the reference during this diagnostic does not bring the tested structures below the solved threshold. Their error therefore cannot be explained solely by which one is selected. This motivates testing a broader proposal distribution alongside the fitting objective.

4.5 Distribution shifts delimit the benchmark's scope

The shift suites assess how much these findings depend on the candidate distribution. Case–signal M1 agrees with the core suite at $r = 0.956$ under physical-parameter changes, 0.922 without decoys, 0.900 under random restart, and 0.793 with a generic-only grammar; median absolute shifts are 0.039, 0.105, 0.108, and 0.097. The only active aggregate whose cluster interval excludes zero is the small no-decoy rank-combination effect, $G = -0.012$ ($[-0.034, -0.002]$), which does not transfer to the other regimes. Static rankings remain similar under the tested shifts, but active gains depend on the operator and search distribution.

5 Discussion: Implications for Agentic Search

Treat the public specification as part of the evaluator. A scoring rule has meaning only relative to a fully stated target. The case object should include the domain, equation, initial and boundary conditions, conservation laws, and normalization. The same object can generate symmetry and consistency tests. A failed anchor check calls for completing the specification before changing the score. This separates ambiguity in the task from numerical errors in its evaluation.

Separate ranking quality from search behavior. A common candidate bank makes scores directly comparable, but active search changes which structures are proposed and fitted. Two selectors can agree on most bank candidates yet retain different parents near the cutoff, leading to different later pools. M2 captures this interaction. Keeping the proposer and fitter fixed isolates the choice of selection rule; the fitting factorial then tests whether its effect depends on the continuous loss. Logging whole pools makes both comparisons reproducible.

Use agreement as a diagnostic signal. Rank aggregation prevents raw signal scales from dominating selection, but several signals can agree on the same erroneous candidate. Operator tags remove clearly mismatched checks, such as smoothness at a free boundary; they do not calibrate the remaining ranks to solution error. Disagreement can guide extra collocation or an operator-specific

estimate. Agreement still needs to be checked against active outcomes, as the SAAS comparisons show.

Choose interventions from the pattern of errors. For already solved operators, continuous error reveals gains hidden by binary success counts. When all tested candidates remain inaccurate even after reference-based fitting, changing the final ranking is unlikely to be enough. The next test should vary proposal structure or fitting capacity. In both settings, report the added prediction work: a lower error obtained with more evaluation is a different result from better selection at the same cost.

Risk–coverage accounting. Let A denote acceptance, $c = \Pr(A)$, and C a catastrophic return. For $c > 0$, $\Pr(A \cap C) = c \Pr(C \mid A)$. A policy can lower accepted-set risk by rejecting difficult searches, so we report coverage, catastrophic/all, and solved/all: $(c, \Pr(A \cap C), \Pr(A \cap \{E < 0.05\}))$. Conditional median error describes accepted returns; it does not account for rejected runs. Comparing all three rates reveals whether a reduction in failures comes with a loss of usable solutions.

6 Related Work

Physics-informed learning and PDE benchmarks. PINNs optimize residuals and public constraints (Raissi et al., 2019), with failures even for expressive networks (Krishnapriyan et al., 2021). PDEBench and PINNacle compare solvers across equations and architectures (Takamoto et al., 2022; Hao et al., 2023). EVALBENCH evaluates the objective used to select structural solutions.

A-posteriori estimation. Classical goal-oriented DWR solves an adjoint problem to weight residuals by the quantity of interest (Becker & Rannacher, 2001); residual-based analyses provide bounds for particular operator classes (Hillebrecht & Unger, 2022; Liu et al., 2023). These estimates depend on operator stability and approximation assumptions. EVALBENCH addresses the empirical question of whether a public score orders structural candidates correctly and guides their search. Where a calibrated estimator is available, it can be tested as a selector within the same protocol.

Scientific program search and proxy optimization. Equation discovery recovers governing expressions from observations (Brunton et al., 2016; Udrescu & Tegmark, 2020; Cranmer, 2023; Shojaee et al., 2025a). ASYS instead searches candidate PDE solutions under physical feedback (Yu & Yang, 2026), creating the evaluation setting studied here. Reward-gaming work formalizes how optimizing a proxy can move behavior away from the intended objective (Manheim & Garrabrant, 2018; Skalse et al., 2022; Laidlaw et al., 2025). We instantiate that concern for reference-free physics scores and add the prior question of whether the public specification identifies the offline target at all.

Selective prediction and risk control. Selective prediction studies error–coverage trade-offs (Geifman & El-Yaniv, 2017); conformal risk control treats bounded monotone losses (Angelopoulos et al., 2024). Our exact-binomial certificate covers a fixed selector set on independent search seeds from the calibration distribution.

7 Limitations and Conclusion

The suite uses analytic references and a restricted atom library. The symmetry audit covers declared ambiguities, and oracle separation assumes trusted benchmark modules. The ten-operator diagnostic uses development seeds with unequal realized costs; the gCLM refit covers four structures per seed. Independent replication and fitting-budget studies are specified in Appendix D and remain pending.

EVALBENCH evaluates reference-free scores at the point where they are used: selecting programs during search. Its experiments show that target ambiguity, incorrect ranking, and poor active returns require different tests. Strong static correlations often bring little active gain, while failures remain concentrated in particular operators. Checking the public specification, fixed-bank ranking, and selected return in sequence identifies whether to repair the task, score, or candidate search. High M1 warrants an active test before baseline replacement; returned error, failures, coverage, and evaluation cost determine the decision. The same protocol can assess new proposers with the evaluator fixed.

Reproducibility Statement

Section 3 specifies the candidate banks, search settings, and seed splits. Appendix A gives proofs and assumptions; Appendix B.2 reports per-operator outcomes and costs. The accompanying artifact provides selector definitions, candidate pools, seed assignments, and the numerical tables and solution atlas.

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.

References

Anastasios N. Angelopoulos, Stephen Bates, Adam Fisch, Lihua Lei, and Tal Schuster. Conformal risk control. In *The Twelfth International Conference on Learning Representations*, 2024. URL https://openreview.net/forum?id=33XGfHLtZg.

Roland Becker and Rolf Rannacher. An optimal control approach to a posteriori error estimation in finite element methods. *Acta Numerica*, 10:1–102, 2001. doi: 10.1017/S0962492901000010.

Steven L. Brunton, Joshua L. Proctor, and J. Nathan Kutz. Discovering governing equations from data by sparse identification of nonlinear dynamical systems. *Proceedings of the National Academy of Sciences*, 113(15):3932–3937, 2016. doi: 10.1073/pnas.1517384113.

C. J. Clopper and E. S. Pearson. The use of confidence or fiducial limits illustrated in the case of the binomial. *Biometrika*, 26(4):404–413, 1934. doi: 10.1093/biomet/26.4.404.

Miles Cranmer. Interpretable machine learning for science with PySR and SymbolicRegression.jl. *arXiv preprint arXiv:2305.01582*, 2023. URL https://arxiv.org/abs/2305.01582.

Juan Dávila, Manuel del Pino, Jean Dolbeault, Monica Musso, and Juncheng Wei. Existence and stability of infinite time blow-up in the Keller–Segel system. *Archive for Rational Mechanics and Analysis*, 248(4):61, 2024. doi: 10.1007/s00205-024-02006-7.

Lukas Fluri, Leon Lang, Alessandro Abate, Patrick Forré, David Krueger, and Joar Max Viktor Skalse. The perils of optimizing learned reward functions: Low training error does not guarantee low regret. In *Proceedings of the 42nd International Conference on Machine Learning*, volume 267 of *Proceedings of Machine Learning Research*, pp. 17306–17377. PMLR, 2025. URL https://proceedings.mlr.press/v267/fluri25a.html.

Leo Gao, John Schulman, and Jacob Hilton. Scaling laws for reward model overoptimization. In *Proceedings of the 40th International Conference on Machine Learning*, volume 202 of *Proceedings of Machine Learning Research*, pp. 10835–10866. PMLR, 2023. URL https://proceedings.mlr.press/v202/gao23h.html.

Yonatan Geifman and Ran El-Yaniv. Selective classification for deep neural networks. In *Advances in Neural Information Processing Systems*, volume 30, pp. 4878–4887, 2017. URL https://proceedings.neurips.cc/paper/2017/hash/4a8423d5e91fda00bb7e46540e2b0cf1-Abstract.html.

Zhongkai Hao, Jiachen Yao, Chang Su, Hang Su, Ziao Wang, Fanzhi Lu, Zeyu Xia, Yichi Zhang, Songming Liu, Lu Lu, and Jun Zhu. PINNacle: A comprehensive benchmark of physics-informed neural networks for solving PDEs. *arXiv preprint arXiv:2306.08827*, 2023. URL https://arxiv.org/abs/2306.08827.

Birgit Hillebrecht and Benjamin Unger. Certified machine learning: A posteriori error estimation for physics-informed neural networks. *arXiv preprint arXiv:2203.17055*, 2022. URL https://arxiv.org/abs/2203.17055.

Wassily Hoeffding. Probability inequalities for sums of bounded random variables. *Journal of the American Statistical Association*, 58(301):13–30, 1963. URL https://doi.org/10.1080/01621459.1963.10500830.

Aditi S. Krishnapriyan, Amir Gholami, Shandian Zhe, Robert M. Kirby, and Michael W. Mahoney. Characterizing possible failure modes in physics-informed neural networks. In *Advances in Neural Information Processing Systems*, volume 34, pp. 26548–26560, 2021. URL https://proceedings.neurips.cc/paper_files/paper/2021/hash/df438e5206f31600e6ae4af72f2725f1-Abstract.html.

Cassidy Laidlaw, Shivam Singhal, and Anca Dragan. Correlated proxies: A new definition and improved mitigation for reward hacking. In *International Conference on Learning Representations*, 2025. URL https://openreview.net/forum?id=msEr27EejF.

Shuheng Liu, Xiyue Huang, and Pavlos Protopapas. Residual-based error bound for physics-informed neural networks. *arXiv preprint arXiv:2306.03786*, 2023. URL https://arxiv.org/abs/2306.03786.

David Manheim and Scott Garrabrant. Categorizing variants of Goodhart's law. *arXiv preprint arXiv:1803.04585*, 2018. URL https://arxiv.org/abs/1803.04585.

Alexander Novikov, Ngân Vũ, Marvin Eisenberger, Emilien Dupont, Po-Sen Huang, Adam Zsolt Wagner, Sergey Shirobokov, Borislav Kozlovskii, Francisco J. R. Ruiz, Abbas Mehrabian, M. Pawan Kumar, Abigail See, Swarat Chaudhuri, George Holland, Alex Davies, Sebastian Nowozin, Pushmeet Kohli, and Matej Balog. Alphaevolve: A coding agent for scientific and algorithmic discovery. *arXiv preprint arXiv:2506.13131*, 2025. URL https://arxiv.org/abs/2506.13131.

Maziar Raissi, Paris Perdikaris, and George E. Karniadakis. Physics-informed neural networks: A deep learning framework for solving forward and inverse problems involving nonlinear partial differential equations. *Journal of Computational Physics*, 378:686–707, 2019. doi: 10.1016/j.jcp.2018.10.045.

Bernardino Romera-Paredes, Mohammadamin Barekatain, Alexander Novikov, Matej Balog, M. Pawan Kumar, Emilien Dupont, Francisco J. R. Ruiz, Jordan S. Ellenberg, Pengming Wang, Omar Fawzi, Pushmeet Kohli, and Alhussein Fawzi. Mathematical discoveries from program search with large language models. *Nature*, 625(7995):468–475, 2024. doi: 10.1038/s41586-023-06924-6.

Parshin Shojaee, Kazem Meidani, Shashank Gupta, Amir Barati Farimani, and Chandan K. Reddy. LLM-SR: Scientific equation discovery via programming with large language models. In *International Conference on Learning Representations*, 2025a. URL https://openreview.net/forum?id=m2nmp8P5in.

Parshin Shojaee, Ngoc-Hieu Nguyen, Kazem Meidani, Amir Barati Farimani, Khoa D. Doan, and Chandan K. Reddy. LLM-SRBench: A new benchmark for scientific equation discovery with large language models. *arXiv preprint arXiv:2504.10415*, 2025b. URL https://arxiv.org/abs/2504.10415.

Joar Skalse, Nikolaus H. R. Howe, Dmitrii Krasheninnikov, and David Krueger. Defining and characterizing reward gaming. In *Advances in Neural Information Processing Systems*, volume 35, pp. 9460–9471, 2022. URL https://proceedings.neurips.cc/paper_files/paper/2022/hash/3d719fee332caa23d5038b8a90e81796-Abstract-Conference.html.

Makoto Takamoto, Timothy Praditia, Raphael Leiteritz, Dan MacKinlay, Francesco Alesiani, Dirk Pflüger, and Mathias Niepert. PDEBench: An extensive benchmark for scientific machine learning. *arXiv preprint arXiv:2210.07182*, 2022. URL https://arxiv.org/abs/2210.07182.

Silviu-Marian Udrescu and Max Tegmark. AI feynman: A physics-inspired method for symbolic regression. *Science Advances*, 6(16):eaay2631, 2020. doi: 10.1126/sciadv.aay2631.

Zongmin Yu and Liu Yang. Agentic symbolic search: Characterizing PDEs beyond hand-crafted expressions, meshes, and neural networks. *arXiv preprint arXiv:2606.20467*, 2026. URL https://arxiv.org/abs/2606.20467.

A Theoretical Foundations and Proofs

These results address identifiability and distribution transfer, validation and rank aggregation, and the scope and sample requirements of risk certification.

A.1 Identifiability and distribution transfer

Proof of Theorem 1. Let $B_i = \{u : \|u - u_i\|_2 \le \varepsilon\}$. The strict separation assumption and the triangle inequality imply $B_1 \cap B_2 = \varnothing$. Because A observes the same public specification under either hidden target, its output law is identical in both worlds. Hence $\Pr(A(P) \in B_1) + \Pr(A(P) \in B_2) \le 1$, so at least one success probability is at most $1/2$.

Candidate-distribution transfer.

Theorem 3 (Candidate-distribution transfer bound)**.** *Let P be the observational-bank distribution and Q the distribution induced by active selection. Suppose $Q \ll P$, $dQ/dP \le \kappa$, and a calibrated map g obeys $\mathbb{E}_P[(E - g(S))^2] \le \epsilon^2$. Then*

$$\mathbb{E}_Q|E - g(S)| \le \sqrt{\kappa}\,\epsilon. \tag{7}$$

If active selection places mass outside the support of P, no such guarantee follows from bank calibration alone.

Write $w = dQ/dP$ and $e = E - g(S)$. Change of measure and Cauchy–Schwarz give

$$\mathbb{E}_Q|e| = \mathbb{E}_P[w|e|] \le \sqrt{\mathbb{E}_P[w^2]}\sqrt{\mathbb{E}_P[e^2]}. \tag{8}$$

In equation 8, since $0 \le w \le \kappa$ and $\mathbb{E}_P w = 1$, $\mathbb{E}_P[w^2] \le \kappa$. Substitution yields equation 7. Absolute continuity is essential: if Q assigns mass to a P-null set, E and S may be extended arbitrarily there without changing any bank statistic.

A.2 Validation and rank aggregation

Fresh validation of an adaptively generated pool.

Theorem 4 (Finite-pool validation and conditional solution control)**.** *At round $t \le T$, condition on the complete search history before drawing a fresh validation batch. Let the resulting candidate pool $\mathcal{U}_t$ have at most M elements, with positive integers M, T, n. Let the public loss $\ell(u, Z) \in [0, B]$, for $B > 0$, and draw n validation observations independently from a fixed law, independently of that history. Define $L(u) = \mathbb{E}\ell(u, Z)$ and the batch mean $\widehat{L}_t(u)$. For $\delta \in (0, 1)$ set*

$$\epsilon_n = B\sqrt{\frac{\log(2MT/\delta)}{2n}}. \tag{9}$$

With probability at least $1 - \delta$, simultaneously at all rounds, the empirical minimizer $\widehat{u}_t$ satisfies

$$L(\widehat{u}_t) \le \min_{u \in \mathcal{U}_t} L(u) + 2\epsilon_n. \tag{10}$$

If an independently established PDE stability estimate gives $E(u) \le C\sqrt{L(u)} + a$ for every candidate under the complete public specification, with fixed $C, a \ge 0$, then on the same event

$$E(\widehat{u}_t) \le C\sqrt{\min_{u \in \mathcal{U}_t} L(u) + 2\epsilon_n} + a. \tag{11}$$

Here E is the chosen solution-error norm, and a can account for a known approximation or discretization allowance in the assumed stability estimate.

Proof. Conditional on the history, each candidate is fixed before the fresh batch. The bounded-variable exponential inequality gives $\Pr(|\widehat{L}_t(u) - L(u)| > e \mid \text{history}) \le 2e^{-2ne^2/B^2}$ (Hoeffding, 1963). A union bound over at most M candidates makes the conditional failure probability at most δ/T when $e = \epsilon_n$. Taking expectations removes the conditioning; a second union bound covers all T

rounds. Neither union bound requires independent candidates or independent search histories. On this simultaneous event, let $u_t^\star$ minimize L on the finite pool. Then

$$\begin{aligned} L(\widehat{u}_t) &\le \widehat{L}_t(\widehat{u}_t) + \epsilon_n \le \widehat{L}_t(u_t^\star) + \epsilon_n \\ &\le L(u_t^\star) + 2\epsilon_n. \end{aligned} \tag{12}$$

This proves equation 10. Apply the assumed stability estimate and monotonicity of the square root to obtain equation 11. □

Application to collocation-based selection. Fresh samples keep validation independent of the current pool, even though earlier samples influenced how that pool was generated. Reusing a queried grid breaks this condition. Boundedness must hold for the evaluated loss; clipping a residual changes the target of the estimate. The PDE stability assumption connects that loss to solution error. SAAS instead minimizes a rank aggregate, whose fixed-pool properties are analyzed next.

Stability of the SAAS rank aggregate. For a fixed pool and public-metadata route, use the aggregator A from equation 6 with K signal ranks $r \in [0,1]^K$. The standard deviation is the population quantity, $\mathrm{sd}(r)^2 = K^{-1}\sum_k (r_k - \bar{r})^2$.

Proposition 5 (Envelope and perturbation bounds)**.** *Let $\widehat{u}$ minimize $A(r(u))$ over a fixed finite pool. If some u_0 has $\max_k r_k(u_0) \le \eta$, for $\eta \in [0,1]$, then*

$$\max_k r_k(\widehat{u}) \le \min\{1, 19\eta/9\}. \tag{13}$$

If rank vectors are perturbed to $r'(u)$ with $\max_u \|r'(u) - r(u)\|_\infty \le \gamma$, then $|A(r'(u)) - A(r(u))| \le \gamma$ for every u. A unique minimizer with a score gap greater than 2γ is unchanged by this perturbation.

Proof. For ranks in $[0,\eta]$, the variance is at most $\eta^2/4$: $K^{-1}\sum r_k^2 - \bar{r}^2 \le \eta\bar{r} - \bar{r}^2 \le \eta^2/4$. Hence $A(r(u_0)) \le 0.95\eta$. Nonnegativity gives $0.45 \max_k r_k(\widehat{u}) \le A(r(\widehat{u})) \le A(r(u_0))$, proving equation 13. The mean and maximum are each 1-Lipschitz in the infinity norm. With $H = I - \mathbf{1}\mathbf{1}^\top/K$, the reverse triangle inequality and the Euclidean contraction of H give

$$|\,\mathrm{sd}(r') - \mathrm{sd}(r)| \le K^{-1/2}\|H(r' - r)\|_2 \le \|r' - r\|_\infty. \tag{14}$$

The nonnegative aggregation weights sum to one, proving the score bound. Each competitor's difference from the minimizer can decrease by at most 2γ, which proves preservation under the strict gap condition. □

These bounds concern a fixed signal set. Near-tied raw scores can produce large rank changes, so the perturbation size must be measured in rank space. An error guarantee would additionally require a relation between signal ranks and physical error.

A.3 Risk certification and finite-seed limits

Proof and scope of Theorem 2. For a fixed selector j, inversion of the binomial test gives $\Pr(R_j > U_j) \le \delta/K$. A union bound over the K selectors yields $\Pr(\exists j : R_j > U_j) \le \delta$. On the complementary event every member of $\{j : U_j \le \alpha\}$ has $R_j \le \alpha$, so a selector chosen from that set after observing calibration data inherits the bound. The argument requires the calibration seeds to be i.i.d. for each fixed selector and the future seed to follow the same distribution. It does not cover a new PDE family, adaptive addition of selectors, or repeated split appearances of the same case–seed pair.

A finite-seed limit for safety certification.

Proposition 6 (Zero-failure sample requirement)**.** *Consider the simultaneous one-sided binomial certificate in Theorem 2 for K fixed selectors and family-wise failure probability δ. If a selector has zero catastrophic outcomes in n independent calibration seeds, its exact upper bound is*

$$U_0 = 1 - (\delta/K)^{1/n}. \tag{15}$$

For a target risk $\alpha \in (0,1)$, even this best observed outcome can certify $U_0 \le \alpha$ only if

$$n \ge \left\lceil \frac{\log(\delta/K)}{\log(1-\alpha)} \right\rceil. \tag{16}$$

Proof. For a Bernoulli failure probability p, the chance of zero failures is $(1-p)^n$. Inverting this probability at level δ/K gives equation 15. Requiring $(\delta/K)^{1/n} \geq 1-\alpha$, taking logarithms, and multiplying by positive n yields equation 16, with reversal when dividing by the negative number $\log(1-\alpha)$. □

With $K = 12$ and $\delta = 0.05$, zero observed failures require at least 8, 53 and 107 calibration seeds to certify risks of at most 50%, 10% and 5%, respectively. Sixteen zero-failure seeds give $U_0 \approx 0.290$. Thus a 16-seed study is too small to certify a 5% per-operator risk with this procedure, even with no observed failures. Pooling operators would certify a different target distribution.

B Supplementary Experimental Results

B.1 Matched fitting and selection

Table 5 gives the matched factorial cell in which both the fitter and outer selector use held-out scoring. Its baseline uses the four-component loss for both. Win rate counts paired seeds with lower returned error; median G measures the size and direction of the change.

Case	Median G	Win rate	Catastrophic
Heat	0.001	43.8%	0.0%
Burgers	−0.039	62.5%	0.0%
Allen–Cahn	−0.005	50.0%	0.0%
NLS	0.013	43.8%	0.0%
Keller–Segel	0.151	37.5%	0.0%
Porous medium	−0.073	62.5%	0.0%
gCLM	−0.015	56.3%	0.0%
Helmholtz 2-D	−0.001	68.8%	0.0%
Wave	−0.116	62.5%	0.0%

Table 5: Per-case outcomes for the held-out fitter with held-out outer selector. Ratios are paired against the original baseline-fitter, baseline-selector run.

The executable specification audit has precision and recall one on 12 mutations: it accepts six valid declarations and rejects six removed, zeroed, or inconsistent anchors. On the two unseen families, the frozen router and baseline return the same candidate for every seed (maximum paired absolute error difference zero).

B.2 Full operator-by-selector comparisons

Tables 6, 7, and 8 give relative error, solved counts, and prediction work for each operator–selector pair. Each cell summarizes eight paired development seeds (24–31) using the final publicly selected program. Errors and prediction counts are medians over runs; solved counts use the threshold in Table 4. Prediction points include fitting and scoring. Table 9 reports the corresponding timing distributions. Source-log hashes in the artifact identify the runs contributing to each entry.

C Returned-Solution Atlas

All plates use final rank-zero programs from the completed ten-operator diagnostic study. Seed 24, the smallest recorded development seed, is fixed across operators and selectors. The renderer evaluates saved programs without refitting or reranking them. Fields use 201×201 display grids. Quantitative errors come from the separate 4,096-point evaluation. Figure 9 covers all eight development seeds, including the stationary Allen–Cahn, NLS amplitude and Keller–Segel cases. Their absolute errors remain in Table 6; similar solution shapes need not imply comparable numerical accuracy.

Two-dimensional solution fields. The Poisson and Helmholtz manufactured targets share the same sinusoidal surface while specifying different operators. Figure 5 shows the analytic field, the returned

Method	Heat	Burg.	AC	NLS	KS	PM	gCLM	Pois.	Helm.	Wave
Baseline	2.4e-07	2.3e-07	3.1e-08	7e-08	2.4e-05	1.8e-06	0.98	8.2e-07	7.9e-07	1.4e-07
Held-out	2.4e-07	2.3e-07	3.1e-08	7.6e-08	1.8e-05	1.7e-06	0.99	8.2e-07	7.8e-07	1e-07
Ensemble	2.4e-07	2.4e-07	3e-08	8.1e-08	2.6e-05	1.7e-06	0.99	8.2e-07	7.8e-07	7.4e-08
Multi-slice	2.4e-07	2.5e-07	3e-08	8.1e-08	2.6e-05	1.9e-06	0.99	8.2e-07	7.8e-07	1.2e-08
Higher-order	2.3e-06	4.1e-05	3.1e-08	9.5e-08	0.0011	0.062	0.99	8.7e-07	7.8e-07	1.1e-05
Variation-weighted	2.4e-07	2.5e-07	2.9e-08	8.4e-08	3.3e-05	1.7e-06	0.75	8.2e-07	7.8e-07	8.2e-08
Rank combination	2.4e-07	2.1e-07	3.1e-08	8.1e-08	1.8e-05	1.7e-06	0.98	8.3e-07	7.9e-07	1.2e-07
Operator router	2.4e-07	2.3e-07	3.1e-08	7e-08	0.0011	1.8e-06	0.98	8.2e-07	7.9e-07	1.4e-07
SAAS	2.4e-07	2.4e-07	3e-08	8.4e-08	1e-06	1.7e-06	0.79	8.2e-07	7.8e-07	3.4e-07
SAAS-mean	2.4e-07	2.3e-07	2.9e-08	7.3e-08	9.4e-06	1.5e-06	0.99	8.3e-07	7.8e-07	1.1e-07
SAAS-no-metadata	2.4e-07	2.4e-07	3e-08	6.9e-08	1.1e-05	1.7e-06	0.99	8.3e-07	7.9e-07	3.6e-07
SAAS-no-holdout	2.4e-07	2.6e-07	3e-08	8.3e-08	9.4e-07	1.7e-06	0.98	8.3e-07	7.9e-07	3.7e-07

Table 6: Median returned-candidate relative error on the independent 4,096-point grid. All 12 selectors and ten operators from the diagnostic study (960 development searches). AC: Allen–Cahn; KS: Keller–Segel; PM: porous medium; Pois./Helm.: two-dimensional controls.

Method	Heat	Burg.	AC	NLS	KS	PM	gCLM	Pois.	Helm.	Wave
Baseline	8/8	8/8	8/8	8/8	8/8	8/8	0/8	8/8	8/8	8/8
Held-out	8/8	8/8	8/8	8/8	8/8	8/8	0/8	8/8	8/8	8/8
Ensemble	8/8	8/8	8/8	8/8	8/8	8/8	0/8	8/8	8/8	8/8
Multi-slice	8/8	8/8	8/8	8/8	8/8	8/8	0/8	8/8	8/8	8/8
Higher-order	8/8	8/8	8/8	8/8	8/8	1/8	0/8	8/8	8/8	8/8
Variation-weighted	8/8	8/8	8/8	8/8	8/8	8/8	0/8	8/8	8/8	8/8
Rank combination	8/8	8/8	8/8	8/8	8/8	8/8	0/8	8/8	8/8	8/8
Operator router	8/8	8/8	8/8	8/8	8/8	8/8	0/8	8/8	8/8	8/8
SAAS	8/8	8/8	8/8	8/8	8/8	8/8	0/8	8/8	8/8	8/8
SAAS-mean	8/8	8/8	8/8	8/8	8/8	8/8	0/8	8/8	8/8	8/8
SAAS-no-metadata	8/8	8/8	8/8	8/8	8/8	8/8	0/8	8/8	8/8	8/8
SAAS-no-holdout	8/8	8/8	8/8	8/8	8/8	8/8	0/8	8/8	8/8	8/8

Table 7: Solved counts out of eight development seeds; success means $E < 0.05$.

Method	Heat	Burg.	AC	NLS	KS	PM	gCLM	Pois.	Helm.	Wave
Baseline	42.0	128.5	58.2	56.7	380.2	232.1	193.1	118.5	130.4	115.0
Held-out	41.7	136.1	63.4	57.9	375.0	214.6	197.7	121.3	130.1	110.7
Ensemble	43.4	139.8	62.4	63.2	395.4	194.8	226.1	115.6	127.0	117.3
Multi-slice	41.0	133.0	62.4	63.2	395.4	221.0	226.1	115.6	127.0	105.7
Higher-order	74.1	180.7	50.8	50.4	455.0	374.3	226.1	143.4	169.0	152.0
Variation-weighted	40.9	141.8	62.2	48.0	379.7	239.8	142.3	124.7	118.6	111.1
Rank combination	41.7	144.7	63.4	59.3	389.2	211.9	245.7	116.6	131.9	107.1
Operator router	42.0	128.5	58.2	56.7	455.0	232.1	193.1	118.5	130.4	115.0
SAAS	44.9	126.9	59.8	52.2	355.3	227.4	260.6	115.6	134.6	113.7
SAAS-mean	45.0	142.6	61.0	59.4	390.9	240.7	173.4	121.3	127.4	119.3
SAAS-no-metadata	46.3	126.8	58.3	54.2	388.2	199.5	227.6	110.1	130.9	103.1
SAAS-no-holdout	47.8	113.1	47.8	52.9	360.3	212.8	190.4	115.8	124.7	111.3

Table 8: Median actual candidate-prediction points, in millions.

SAAS field and their absolute difference. This directly displays what a returned program computes over the two spatial coordinates.

The surfaces show the recovered spatial mode, while the error maps resolve amplitude differences at a finer scale. Although the target fields coincide, Poisson and Helmholtz apply different differential operators to them. Their paired display provides a control for evaluating the same spatial structure under distinct public equations and boundary specifications.

Method	Median (s)	Q1 (s)	Q3 (s)	Total (h)
Baseline	51.7	35.3	95.6	1.66
Held-out	51.8	36.1	97.3	1.73
Ensemble	52.0	38.5	101.2	1.74
Multi-slice	50.5	38.0	101.3	1.75
Higher-order	79.5	42.2	107.0	2.35
Variation-weighted	55.2	36.6	100.0	1.76
Rank combination	51.5	36.0	91.1	1.80
Operator router	51.2	35.4	101.5	1.73
SAAS	54.2	35.4	91.1	1.75
SAAS-mean	56.1	37.7	100.9	1.79
SAAS-no-metadata	52.2	33.9	92.0	1.62
SAAS-no-holdout	50.2	34.1	90.3	1.61

Table 9: Search-time distributions across 80 diagnostic searches per selector. Quartiles use linear interpolation. Total hours sum recorded search times; they are not elapsed parallel wall time. These implementation-dependent measurements complement the prediction-point matrix in Table 8.

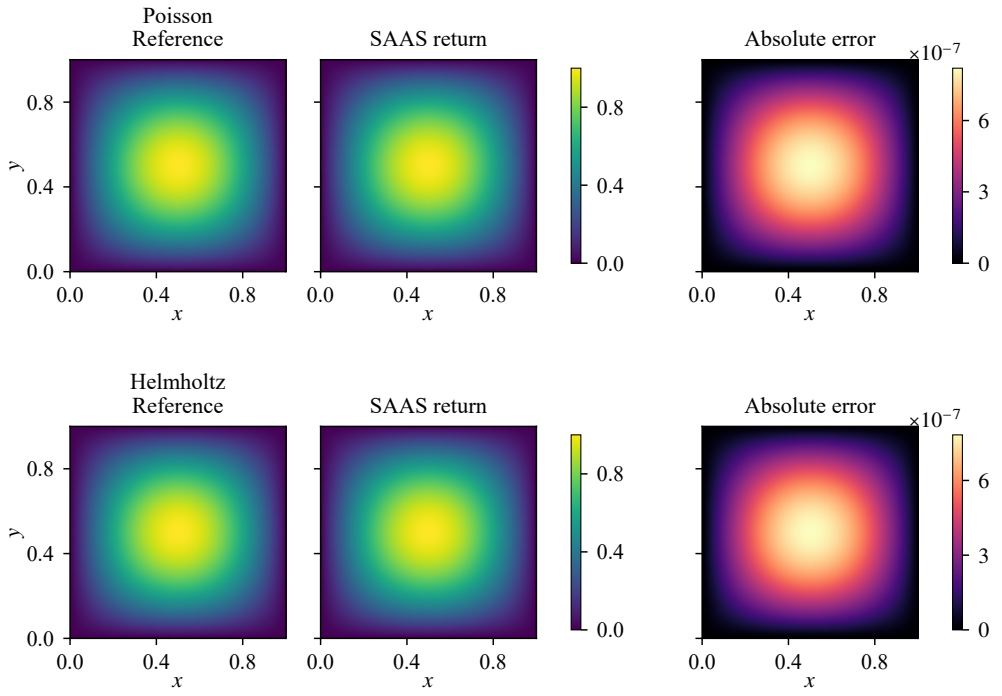

Figure 5: Reference, SAAS return, and absolute error for Poisson and Helmholtz. Reference and prediction share a color range within each row; error maps have separate labeled scales.

Moving-support geometry. Figure 6 shows the porous-medium returns over the full space–time domain, including the support interface. The surfaces show the spreading support and decreasing peak, while the four error maps resolve differences that are not visible at the scale of the solution.

The annotated E_{dense} values are the recorded 4,096-point relative errors, not errors recomputed on the display grid. For this seed, three returns are visually close to the reference, whereas higher-order selection has a much larger error. These are views of existing search outcomes, not additional fitted solutions or a new aggregate comparison.

Nonlocal space–time geometry. Figure 7 shows the same seed-24 gCLM programs over the complete display interval. The reference develops increasingly concentrated positive and negative

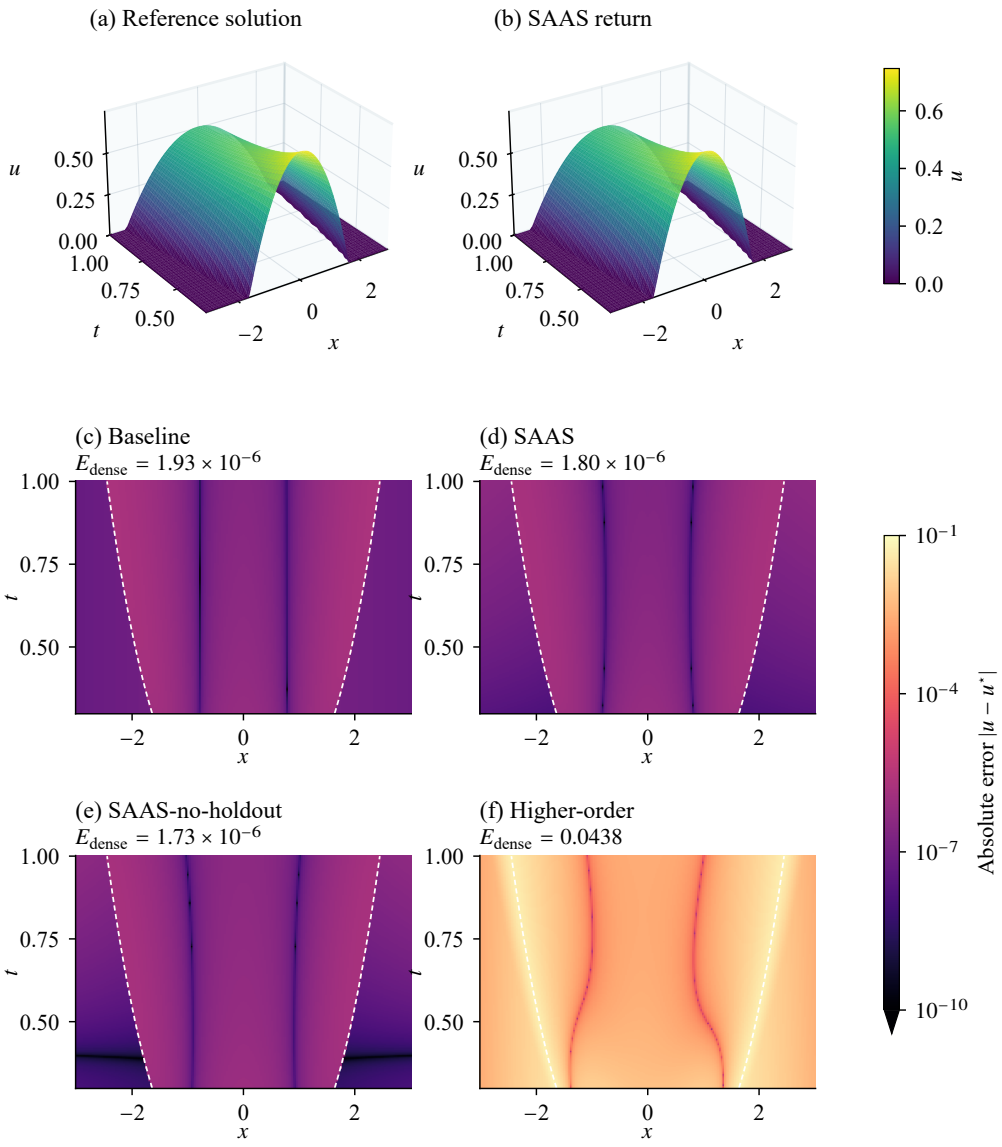

Figure 6: Porous-medium solution geometry and selector errors. Top: reference and SAAS surfaces share amplitude limits and viewing angle. Bottom: all four returns use one logarithmic absolute-error scale; values below 10^{-10} occupy the lowest color, without modifying the data. White dashed curves mark the exact support boundary $|x| = \sqrt{6}\, t^{1/3}$. Surface meshes are thinned for rendering; error maps use the full 201×201 display grid.

structures that the SAAS return does not reproduce; the error maps show where this discrepancy accumulates.

The large late-time error is shared across selectors, rather than confined to a narrow interface. SAAS has a larger recorded error than baseline and SAAS-no-holdout on this fixed example. The geometric view therefore preserves the unfavorable outcome and complements, rather than replaces, the paired multi-seed comparisons.

Space–time fields and localized error. Figure 8 shows the full display domain for heat, Burgers, and wave solutions. The heat profile decays while preserving its spatial mode, the Burgers front translates, and the wave changes phase.

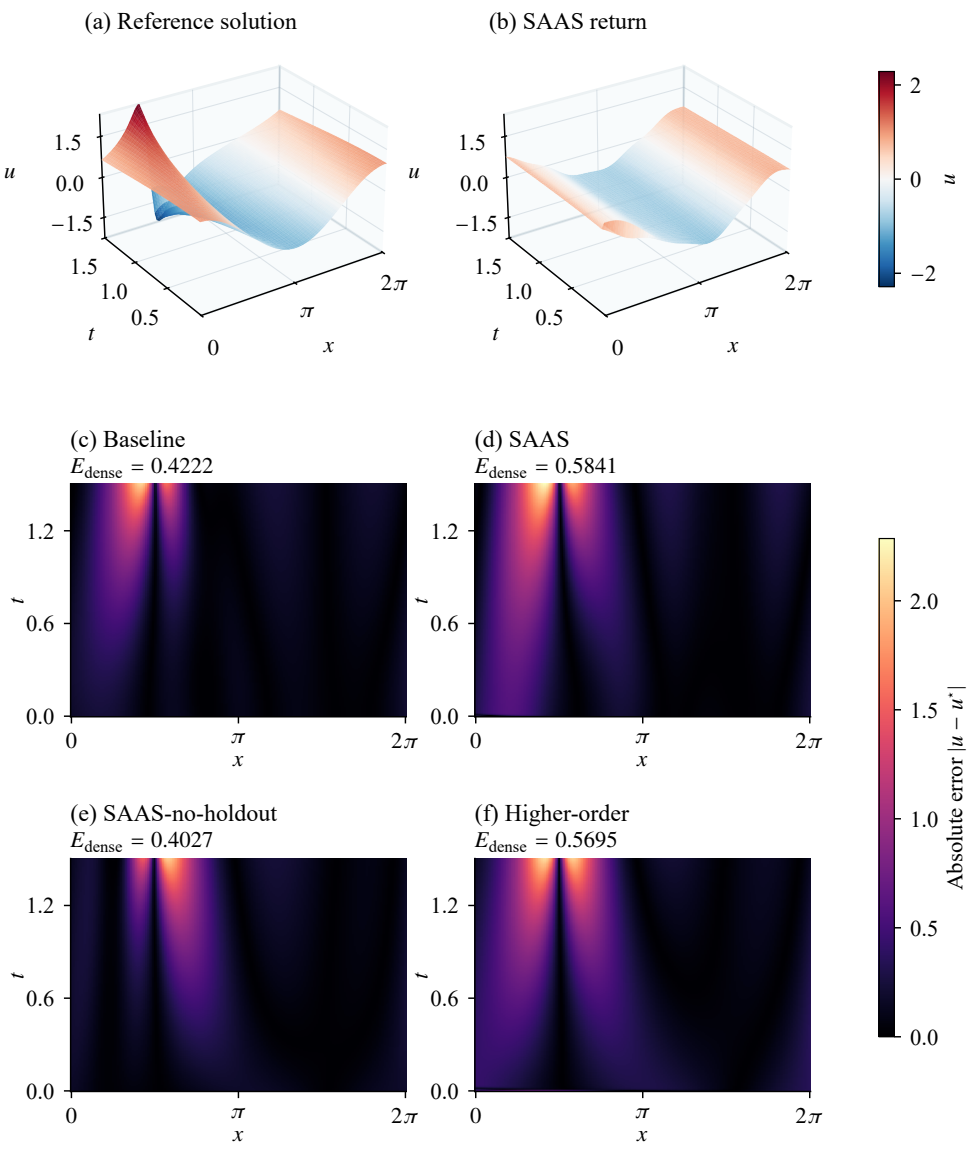

Figure 7: gCLM solution geometry and selector errors. Reference and SAAS surfaces share symmetric amplitude limits and viewing angle. The four absolute-error maps use a common linear color scale on the full 201×201 grid. The annotations retain the recorded 4,096-point relative errors. Only the surface meshes are thinned for rendering.

The maps locate errors relative to a moving front or oscillation. A narrow band of error near a front can be difficult to distinguish from a position shift using a scalar norm alone; a broader pattern can indicate amplitude or phase error. The source-log and figure checksums in the artifact connect these spatial diagnostics to the returned programs used in the quantitative tables.

All development seeds: gains and failures together. Figure 9 displays the paired contrasts for every operator and all eight development seeds. It complements the median tables by showing whether a method improves consistently, changes little, or has occasional large failures.

The log-ratios span approximately -1.85 to 4.83. Saturating the color scale keeps moderate differences visible alongside the largest failures. The radial improvement and higher-order failures appear in the same view; nearly white cells identify operators with little change in returned error. Because a

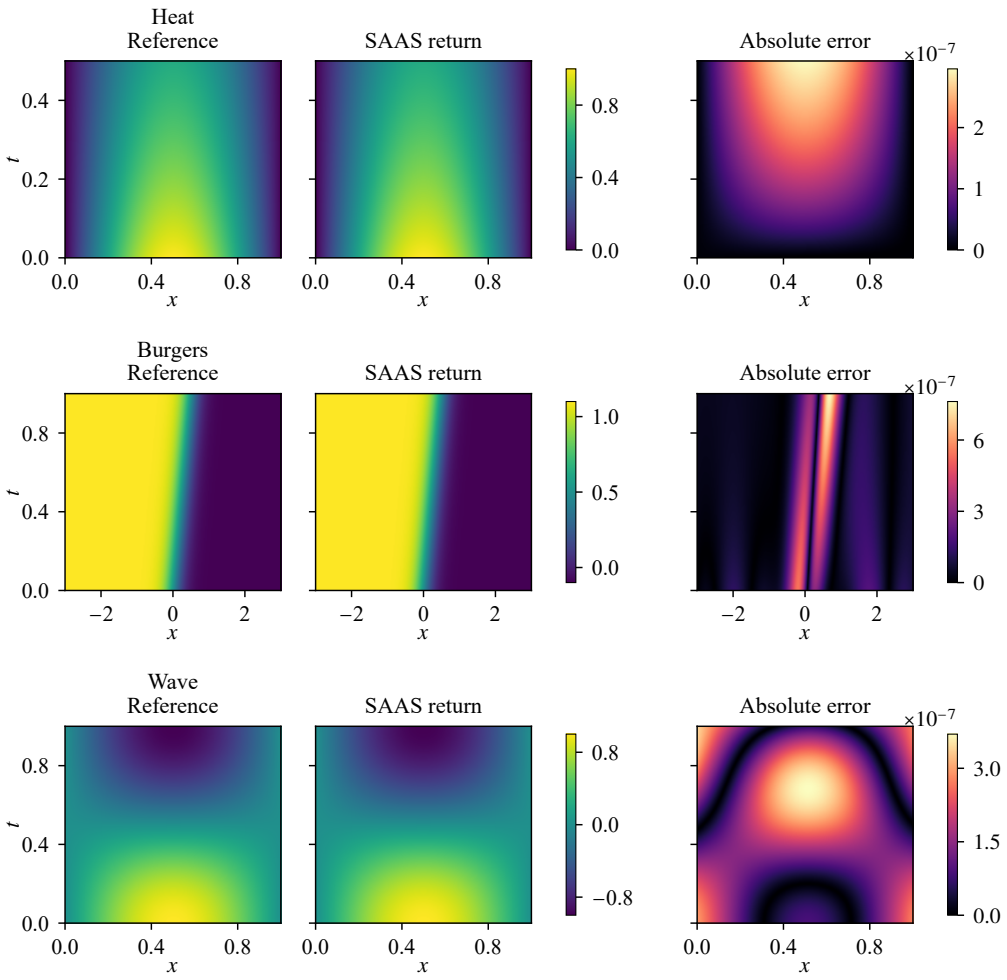

Figure 8: Reference, SAAS return, and absolute error over space (horizontal) and time (vertical). Each row shares a reference/prediction color scale and has a separately labeled error scale.

log-ratio describes relative change, it should be read with the absolute errors in Table 6: a large ratio between two small errors need not change the solved status.

D Prospective Experimental Protocol

D.1 Replication and sensitivity

This predeclared study is pending. It uses the ten manufactured cases and the fixed selector definitions from Section 3. Table 10 separates replication, sensitivity, and fitting-budget comparisons. The last arm is specified in Section D.2; none of these planned runs contributes to the completed results.

Replication uses seeds 10000–10015 with the standard 12 rounds and nominal grid 96. Sensitivity uses seeds 11000–11003 and compares baseline, SAAS, SAAS-mean, SAAS-no-metadata and SAAS-no-holdout under five matched settings: standard; six rounds; grid 48; grid 192; and random-restart proposals. Each comparison changes only the named setting. A standard arm on the same sensitivity seeds provides the paired baseline.

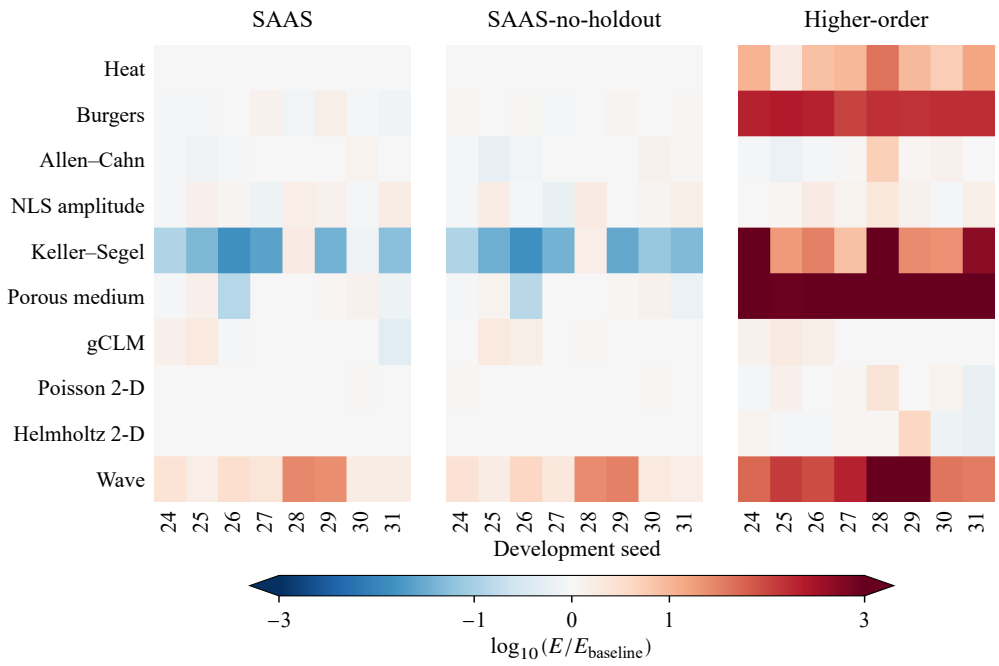

Figure 9: Paired log-ratios of the recorded dense error to baseline error, for SAAS, SAAS-no-holdout and higher-order selection. Blue means lower error, white indicates parity and red means higher error. The 240 cells use the same underlying searches as Tables 6–8. Colors saturate beyond $[-3, 3]$, as marked by the colorbar extensions; numerical values are retained without clipping.

Study	Operators	Selectors	Seeds	Searches
Fresh-seed replication	10	12	16	1,920
Matched sensitivity (five settings)	10	5	4	1,000
Fitting budgets (three settings)	10	4	4	480

Table 10: Planned search allocation. The total excludes engineering checks and ten full-configuration pilot runs.

Primary endpoints are solved and catastrophic rates. SAAS is compared with each of the other 11 selectors, defining 22 two-sided paired binary tests with Holm correction. Conditional exact tests assume independent case–seed pairs and equiprobable discordance directions under the null. Intervals resample paired seeds within each fixed operator; they do not generalize to a population of unseen PDEs. Continuous errors, costs, per-operator comparisons and sensitivity effects remain descriptive. Nonfinite returned errors count as unsolved and catastrophic. Execution exceptions retain partial logs and stop the stage for diagnosis.

Final errors use 4,096 independent evaluation points. Logs retain all candidate events, returned and best-pool errors, prediction counts, timing, configurations, and source hashes. The deployment plan in the artifact specifies resource limits and execution checks separately from this analysis protocol.

D.2 FITTING-BUDGET DIAGNOSTICS

The fitting-budget arm of Section D.1 varies the maximum number of L-BFGS iterations while fixing the selector and search settings. The paired design is given in Table 11. It measures whether additional fitting improves the available candidates, their selection, or both; results are pending.

Fixed factor or intervention	Setting
Operators	All 10 diagnostic cases
Selectors	Baseline, SAAS, SAAS-mean, SAAS-no-holdout
Fresh seeds	12000–12003
Inner iteration limits	25, 100, 200
Outer search rounds / nominal grid	12 / 96
Searches in this arm	$10 \times 4 \times 4 \times 3 = 480$

Table 11: Matched optimization-budget design. The 100-iteration arm is rerun on the same seeds as both alternatives. These descriptive comparisons are separate from the primary testing family.

Separate returned error from pool quality. For the final pool $\mathcal{U}$, let $E_{\mathrm{best}} = \min_{u \in \mathcal{U}} E(u)$ on the common offline grid and let E_{ret} be the error of its publicly selected return. Whenever these errors are finite,

$$E_{\mathrm{ret}} = E_{\mathrm{best}} + \Delta_{\mathrm{select}}, \qquad \Delta_{\mathrm{select}} = E_{\mathrm{ret}} - E_{\mathrm{best}} \geq 0. \tag{17}$$

In equation 17, E_{best} measures the attained pool and Δ_{select} measures the error added by its terminal selection. Changing fitting effort can also change later proposals, so paired budget runs measure its full effect on search. The reference-based quantities are computed offline.

Predeclared reporting. For each paired seed, report returned and best-pool errors, selection gaps, solved and catastrophic counts, prediction points, and wall time. A terminal selection failure occurs when the return is unsolved but the final pool contains a candidate with $E < 0.05$. Nonfinite returns count as failures and have undefined gaps. These descriptive comparisons retain every operator and use the fixed selector definitions.

Interpreting the paired outcomes. If best-pool error falls while returned error stays fixed, the selection gap has increased: the selector has not captured the pool improvement. If both errors fall, additional fitting helps the returned solution. If neither falls, the tested budget increase is insufficient for that search. Prediction counts and wall time quantify the work associated with each outcome.