CONSENSUSCOVER: ADAPTIVE INPUT SELECTION FOR NEURAL OPERATORS UNDER DISTRIBUTION SHIFT

Anonymous authors
Paper under double-blind review

ABSTRACT

Training neural operators requires deciding which candidate partial differential equation (PDE) inputs merit expensive solution labels. Under distribution shift, the best selection geometry depends on both the equation and the operator architecture: target-distribution coverage is stronger in five of six DeepONet/FNO settings, whereas field information is stronger for Darcy/FNO. We introduce *ConsensusCover*, an adaptive acquisition rule that resolves this conflict without target solutions or equation-specific switching. At each round, ConsensusCover ranks candidate inputs using two independent views: Fourier field modes and the current operator's readout Jacobian. Rank-consistent overlap beyond chance enters the next batch, while the remaining selections preserve target-distribution coverage through local probability transfer. We evaluate Darcy flow, Allen–Cahn and viscous Burgers with DeepONet and FNO. In a separate confirmation with 20 paired trials per setting and 32–256 labelled simulations, ConsensusCover cuts the worst-case gap in out-of-distribution learning-curve error from 55.44% for target coverage and 30.92% for field information to 6.62%. The paired-bootstrap 95% upper bound for this gap is 8.55%, and the geometric-mean gap across all six settings is 2.07%. The agreement signal activates most strongly on Darcy/FNO, showing how information consensus adapts PDE input selection across neural operators.

1 INTRODUCTION

Each neural-operator training label requires a numerical PDE solution. When the deployment inputs differ from the available source catalogue, selecting which source inputs to solve is consequential even if unlabelled deployment fields are available. The appropriate selection geometry is not universal. In the six architecture–equation settings studied here, projected target coverage has lower out-of-distribution (OOD) learning-curve error than linear field information in five, while field information is much stronger for Darcy with an FNO. Committing to either fixed rule therefore creates a substantial worst-setting loss.

We introduce *ConsensusCover*, a single adaptive acquisition rule. At each round it independently ranks candidate fields by their Fourier-channel information and by the response of the currently trained operator. Only ranked overlap exceeding its random expectation contributes information-led selections. The remaining positions use target-probability coverage on source representatives. This decision uses unlabelled source and target inputs and the model trained on already acquired source labels. It does not use target solutions, unselected source solutions, or a hand-coded PDE or architecture switch.

Our main evidence is an independent confirmation on Darcy, Allen–Cahn and viscous Burgers with both DeepONet and FNO. Twenty paired trials per setting compare ConsensusCover with the two fixed constituent rules. Relative to the better fixed rule separately in each setting, the largest excess OOD learning-curve error is 6.62%, versus 55.44% for fixed target coverage and 30.92% for fixed field information. The paired-bootstrap 95% upper bound for the adaptive worst-setting excess is 8.55%. Agreement activates most strongly on Darcy/FNO, where field information is needed. We report the selection-time cost of this adaptation alongside its predictive benefit.

Related work. DeepONet and Fourier Neural Operators (FNOs) learn solution maps from sampled fields [3, 5, 7]. Active operator learning can use predictive uncertainty [10], input functions 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.

numerical resolution [4], or solver-in-the-loop benchmarks [8]. Geometric acquisition includes CoreSet [9], LCMD [2], and target-query designs derived from distribution-shift risk bounds [1]. Nearest-neighbour weighting transfers target mass to source Voronoi cells [6]. ConsensusCover uses target mass as a coverage fallback and asks whether two distinct information views agree on additional source labels. It does not estimate prediction error from unavailable target outputs.

2 SETTING AND TARGET-COVERAGE COMPONENT

Let a be a coefficient or initial-condition field and $\mathcal{G}(a)$ its PDE solution or terminal state. A source catalogue $\mathcal{P} = \{a_j\}_{j=1}^{N_P}$ is eligible for simulation; a sample $\mathcal{Q} = \{\widetilde{a}_i\}_{i=1}^{N_Q}$ describes the unlabelled deployment inputs. Only a selected source set S receives solutions. A neural operator $\widehat{\mathcal{G}}_S$ is trained on those pairs and evaluated on independent target inputs. For a field loss ℓ, the population objective is

$$R_Q(S) = \mathbb{E}_{a\sim Q_A}[\ell(\widehat{\mathcal{G}}_S(a), \mathcal{G}(a))]. \tag{1}$$

Acquisition cannot evaluate this risk because target solutions are withheld. All methods begin from the same 32 labelled source inputs within a paired trial.

TargetCover supplies the fallback geometry. Fit a 16-dimensional proper orthogonal decomposition (POD) to the source input fields and denote source and target coordinates by z_j and $\widetilde{z}_i$. Give each target input weight $w_i = 1/N_Q$ and transfer its mass to its nearest source coordinate, sharing exact ties. The resulting coupling and source demand are

$$\Gamma_{ij} = \frac{w_i \mathbf{1}\{j \in J_i\}}{|J_i|}, \quad J_i = \arg\min_j \|\widetilde{z}_i - z_j\|_A, \qquad \mu_j = \sum_i \Gamma_{ij}. \tag{2}$$

Thus $\sum_j \Gamma_{ij} = w_i$. Each row also minimizes its source approximation cost, giving

$$A_\Gamma = \sum_{i,j} \Gamma_{ij} \|\widetilde{z}_i - z_j\|_A^2, \qquad \min_{\Gamma\mathbf{1}=w,\, \Gamma\geq 0} A_\Gamma = \sum_i w_i \min_j \|\widetilde{z}_i - z_j\|_A^2. \tag{3}$$

From the common initial set, projected target coverage greedily decreases

$$C_\mu(S) = \sum_j \mu_j \min_{s\in S} \|z_j - z_s\|_A^2. \tag{4}$$

Writing $d_j^2(S) = \min_{t\in S} \|z_j - z_t\|_A^2$, the gain of adding an unsolved candidate s is

$$\Delta_\mu(s \mid S) = \sum_j \mu_j \left[d_j^2(S) - \min\{d_j^2(S), \|z_j - z_s\|_A^2\}\right]. \tag{5}$$

Greedy coverage chooses the largest current gain. This is both the fixed *target coverage* comparator and ConsensusCover's fallback; it uses no trained model. The earlier kernel-density control instead maximizes

$$F_{Q\to P}(S) = \sum_j \bar{q}_j \max_{s\in S} K(a_j, a_s), \qquad \bar{q}_j \propto \sum_i w_i K(\widetilde{a}_i, a_j). \tag{6}$$

The appendix proves the coverage-component risk decomposition and greedy geometric statement (Appendix A). Those statements describe the fallback objective, not the information-selected portion of a ConsensusCover batch.

What the coverage objective controls. The transfer in (2) preserves each target input's probability before compressing it onto the finite source catalogue. This is different from globally normalizing a kernel density: a target input with weak affinity to every source field still contributes its full mass. The nearest-source transfer minimizes the coupling's projected squared distance for the prescribed target row weights. In particular, changing positive reference weights on the fixed source catalogue does not change its nearest source, demand vector, or nested greedy selection. These properties matter under shift because low-density target regions should not disappear from the acquisition objective simply because they have poor source affinity.

There is a precise, limited prediction interpretation. Let $e_S = \widehat{\mathcal{G}}_S - \mathcal{G}$ be L_e-Lipschitz on the relevant inputs. Appendix A proves that empirical target RMS error is at most

$$\sqrt{E_\Gamma(S)} + L_e\big(\sqrt{A_\Gamma} + \sqrt{C_\mu(S)} + \sqrt{T_\Gamma(S)}\big), \tag{7}$$

where E_Γ measures residual error at selected representatives and T_Γ measures the discarded POD directions. Of these quantities, greedy acquisition directly reduces only C_μ; training quality, projection error, and source-pool mismatch remain separate. The reduction $C_\mu(S_0) - C_\mu(S_0 \cup B)$ is monotone submodular for a fixed nonempty initial set S_0, so greedy coverage has the usual cardinality guarantee for that geometric term. Neither statement supplies a bound for the information-selected portion of a ConsensusCover batch or for the relative error used in our experiments. We use the result to specify the fallback's role, and test the combined rule empirically.

Independent evidence for the fallback. An earlier fixed-selector study, retained in Appendix C, compares TargetCover with random selection, CoreSet, source LCMD, density weighting, uniform-demand coverage, source-distance coverage, and direct target coverage. It uses 20 paired DeepONet trials on each of Darcy, Poisson, Allen–Cahn, and Burgers, with a five-count label schedule. The fixed TargetCover rule has lower target-law AUC than source LCMD on all four operators, whereas source LCMD has lower source-law AUC on all four. This opposed ordering is evidence that deployment weights affect which inputs are useful under shift; it is not an overall superiority claim for TargetCover. Direct target-distance coverage is often close, showing that the source-demand representation approximates the more expensive source-restricted target objective on these finite pools.

The earlier study also separates target weighting from distance geometry. With squared projected distance fixed, transferred target demand improves on density weights on Darcy, Poisson, and Allen–Cahn in crossed-control comparisons. Changing only the source reference weights on a fixed support does not change nearest-source TargetCover selections, whereas a globally normalized density transfer changes the effective target measure. These controls test distinct properties of the fallback: whether target probabilities survive transport, whether source representatives preserve the target-distance design, and whether a deployment-aware objective trades source-law accuracy for target-law accuracy. They do not evaluate information consensus. The independent six-setting experiment below is necessary because a reliable coverage component can still miss the architecture-dependent information need on Darcy/FNO.

3 CONSENSUSCOVER: AGREEMENT-LED ACQUISITION

At each subsequent label count, let U be the unsolved source candidates and b the number of new solutions. The two information designs independently produce ordered batches of size b, conditional on the currently labelled set S. Both use the unlabelled target sample to measure information where the operator will be deployed.

Fourier-field information. Standardize input fields with source-only statistics and retain Fourier modes through frequency eight in each spatial direction. For each field and mode, let x_{jk} be its complex linear Fourier channel and t_{ik} the analogous target channel. With prior precision one, the field design greedily reduces the target-integrated linear-model variance

$$V_{\mathrm{field}}(S) = \sum_k m_k \,\mathrm{tr}\left[Q_k \left(I + \sum_{j\in S} x_{jk}^* x_{jk}\right)^{-1}\right], \qquad Q_k = \frac{1}{N_Q}\sum_i t_{ik}^* t_{ik}, \tag{8}$$

where m_k accounts for conjugate Fourier frequencies. This is a ranking criterion for input fields; it does not use PDE outputs.

Current-model response information. After training the operator on S, evaluate its physical-output readout Jacobian on unlabelled source and target inputs. A fixed two-sided sketch compresses output locations and readout parameters to eight and sixteen dimensions, respectively. Source-only scaling yields response features $g_\theta(a)$ and the kernel $k_\theta(a, a') = 1 + \langle g_\theta(a), g_\theta(a')\rangle$. The second

ordered batch greedily reduces mean target posterior variance for this frozen kernel conditional on S, using regularization 0.1. For FNO the readout is its final scalar layer; for DeepONet it is the branch readout and scalar bias. This ranking can change between rounds because the fitted operator changes. It requires only the already labelled source solutions.

Agreement and fallback. Let F and J be the field and Jacobian ordered batches, with $o = |F \cap J|$. Under independent uniform batches from $|U|$ candidates, the expected overlap is $e = b^2/|U|$. ConsensusCover accepts

$$k = \min\{o, \max(0, \lceil o - e \rceil)\} \tag{9}$$

common candidates, prioritized by the sum of their ranks in F and J. It fills the other $b - k$ positions by greedy decrease of C_μ in (4), skipping already accepted indices. The target-to-source demand μ stays fixed for a trial, while the response ranking is recomputed after training at each label count. When agreement does not exceed chance, the batch follows target coverage. This is one rule for every equation and architecture; no task identity enters the decision. The implementation uses a common initial labelled set and nested selections.

Round-by-round implementation. The source-fitted POD map, target-to-source coupling, and source demand are constructed before any new PDE solution is requested. After training at a given budget, each information rule scores only the unsolved candidates and returns an ordered batch of the next budget increment. Rank-sum ordering breaks ties within their intersection; a fixed index order resolves any remaining ties. The k accepted points are removed from the candidate set before coverage fill, so a batch always contains exactly b distinct new source inputs. Their PDE solutions are computed once, the labelled set grows, and the operator is trained for the next stage. Thus the method uses the same number of expensive solution labels as either fixed comparator.

Chance correction matters most when the later batches are large relative to the remaining catalogue. With $|U| = N$ and two independent uniform size-b sets, intersection size has mean b^2/N; raw overlap therefore increases even without shared information. Equation (9) reserves only the overlap beyond this baseline. It is a deterministic gating rule, not a hypothesis test: rank dependence, shared inputs, and the current model can all alter the null distribution. The resulting accepted count measures agreement between two rankings rather than a calibrated probability that a candidate improves target accuracy. When the views disagree, the fallback continues to serve the full target distribution.

4 Independent confirmation protocol

We test Darcy permeability-to-pressure and the nonlinear initial-to-terminal Allen–Cahn and scalar two-dimensional viscous Burgers maps on 32×32 grids. Darcy has zero Dirichlet boundaries; the evolution problems are periodic. Source fields are spectrally filtered Gaussian inputs. The unlabelled acquisition target halves the source correlation length while holding marginal variance and PDE parameters fixed. Evaluation uses new held-out target fields at the same factor 0.5; other shift factors and the source law are secondary. Appendix B gives the numerical reference checks.

The six settings cross three equations with DeepONet and FNO. Each of 20 independent paired trials has 1,024 source candidates and 1,024 unlabelled target inputs. The common schedule is $32, 64, 128, 256$ source solutions, with 200 Adam epochs at each stage. Source and target fields, initial labels, initialization and training randomization are shared by the three rules in a setting: projected target coverage, linear field information, and ConsensusCover. The independent confirmation seeds are distinct from the four development seeds. All 360 learning curves passed checks on label counts, nested selections, training completion and paired inputs.

The primary outcome is the normalized trapezoidal area under the target relative-L^2 learning curve at shift factor 0.5:

$$\mathrm{AUC}(e) = \frac{1}{224} \sum_{r=1}^{3} \frac{e(m_r) + e(m_{r+1})}{2} (m_{r+1} - m_r). \tag{10}$$

Here $(m_1, m_2, m_3, m_4) = (32, 64, 128, 256)$. Lower is better. In each setting, the *expert envelope* is the smaller of the two fixed-rule mean AUCs. We divide each method's mean AUC by that envelope, then report the maximum ratio across six settings and their equal-setting geometric mean. Uncertainty resamples complete paired seeds 20,000 times within each setting. These cross-setting summaries and

their acceptance criteria were specified before the independent confirmation. Per-setting comparisons and computational time are secondary.

Numerical and training controls. The nonlinear input law holds marginal variance fixed while shortening only the correlation length; the elliptic law standardizes each sampled field before exponentiating permeability. The Allen–Cahn references combine an exact reaction flow with spectral diffusion through symmetric splitting. Burgers references use dealiased Fourier fluxes and a fourth-order Runge–Kutta integrator. Both evolution solvers use time steps no larger than 0.005; the Burgers step also obeys advection and diffusion stability limits. The reference checks include invariant constant states, temporal refinement, and spatial refinement. The independent convergence figure, numerical details, and random-input errors are in Appendix B.

The model comparison changes both the architecture and the readout used by the response-information view, while holding candidate fields, target law, label schedule and training budget fixed within a paired trial. FNO uses four Fourier layers of width 32 with twelve retained modes per spatial direction and GELU activations. Each architecture is trained for 200 Adam epochs per label count, with learning rate 10^{-3} and batch size 32. Consequently, the comparison assesses acquisition under a specified training recipe, not the best attainable model after architecture-specific retuning. The common initial labels anchor all three nested curves at the same 32-label error; subsequent differences arise from selection and training on the selected fields. Each held-out evaluation pool is sampled independently of the unlabelled acquisition target pool.

Pre-specified decision. The development study used four seeds, which are excluded from the twenty confirmation seeds. Before confirmation, we fixed five pass conditions: maximum adaptive ratio at most 1.10, equal-setting geometric-mean ratio at most 1.04, maximum regret at most half that of the better fixed rule, paired-bootstrap 95% upper bounds at most 1.15 and 1.06 for those two ratios, and complete validation of all 360 curves. The per-setting envelope is computed from the two fixed-rule *means*, rather than selecting the winning rule separately for each seed. It is a retrospective benchmark for cross-setting robustness, not information available to ConsensusCover at selection time. This protocol makes a near-envelope claim testable without requiring the adaptive rule to outperform each fixed expert in every cell. The paired bootstrap resamples the entire four-budget curve of a seed together, preserving dependence between successive fits and between methods run on the same inputs. The maximum is recalculated over all six settings in each resample, so its upper bound reflects uncertainty in which setting is worst. A pooled average over all runs could hide a large loss in one architecture–equation cell. Equal weighting across the six settings also prevents their different absolute error scales from dictating the robustness summary.

5 Results

Target coverage is the stronger fixed rule in five settings; field information is stronger on Darcy/FNO. On that setting, committing to target coverage raises mean AUC from 0.1769 to 0.2750. Conversely, on Burgers/FNO, field information raises it from 0.2930 to 0.3836. ConsensusCover records 0.1886 and 0.2995 in these two settings, respectively. Table 1 gives all six independent comparisons.

5.1 One rule approaches the stronger fixed geometry

Table 1: Independent held-out OOD relative-L^2 learning-curve AUC over 20 paired trials per setting; lower is better. Coverage is projected TargetCover, field info. is linear Fourier-field information, and Ratio is ConsensusCover AUC divided by the smaller of the two fixed-rule AUCs.

Setting	Coverage	Field info.	ConsensusCover	Ratio
DeepONet / Darcy	0.2624	0.2967	0.2700	1.0287
DeepONet / Allen–Cahn	0.6963	0.7030	0.6951	0.9983
DeepONet / Burgers	0.7661	0.7786	0.7666	1.0007
FNO / Darcy	0.2750	0.1769	0.1886	1.0662
FNO / Allen–Cahn	0.0487	0.0532	0.0491	1.0093
FNO / Burgers	0.2930	0.3836	0.2995	1.0222

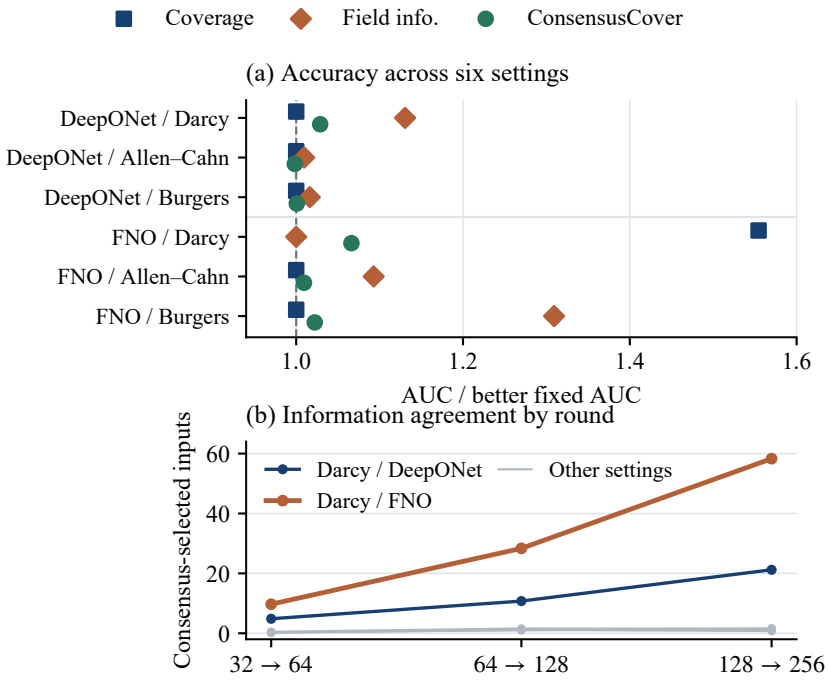

Figure 1: Independent six-setting confirmation. Left: mean OOD AUC divided by the better of the two fixed rules separately in each architecture–equation setting; the horizontal line marks parity. Right: mean number of candidates accepted through chance-corrected information agreement in the three active acquisition rounds. Each setting has 20 paired trials.

For method m and setting s, write $A_{s,m}$ for the mean AUC over the 20 paired trials and $B_s = \min\{A_{s,\mathrm{cov}}, A_{s,\mathrm{field}}\}$. The two reported cross-setting excesses are

$$R_{\max}(m) = \max_s \left(\frac{A_{s,m}}{B_s} - 1 \right), \tag{11}$$

$$R_{\mathrm{geo}}(m) = \exp\left(\frac{1}{6} \sum_{s=1}^{6} \log \frac{A_{s,m}}{B_s} \right) - 1. \tag{12}$$

The worst excess over the per-setting expert envelope is 55.44% for target coverage, 30.92% for field information, and 6.62% for ConsensusCover (Figure 1). The adaptive rule's equal-setting geometric-mean excess is 2.07%. Its paired-bootstrap 95% upper bounds are 8.55% for the maximum and 2.55% for the geometric mean. The largest adaptive gap occurs on Darcy/FNO; ConsensusCover is not claimed to beat the better fixed rule in every setting.

Coverage's maximum excess comes from Darcy/FNO; field information's comes from Burgers/FNO. ConsensusCover has AUC 0.1886 and 0.2995 on these cells, respectively, without switching on a stored PDE or architecture label. It gives up a little accuracy to each cell's stronger fixed rule while avoiding the much larger loss of its weaker rule.

The DeepONet results show a different scale of conflict. Its coverage AUC is lower than field information on all three equations, and ConsensusCover remains within 2.87% of coverage in each of them. On Allen–Cahn/DeepONet the adaptive mean is slightly lower than either fixed mean; that 0.17% advantage is not treated as systematic dominance. On FNO/Allen–Cahn, all three AUCs are small in absolute terms, so ratios show the acquisition ordering more clearly than raw differences. The geometric-mean ratio gives each cell one multiplicative contribution without letting a high-error PDE dominate through its output scale.

Figure 2 shows all four label counts. The separate four-PDE, five-count TargetCover study is reported as coverage-component evidence in Appendix C.

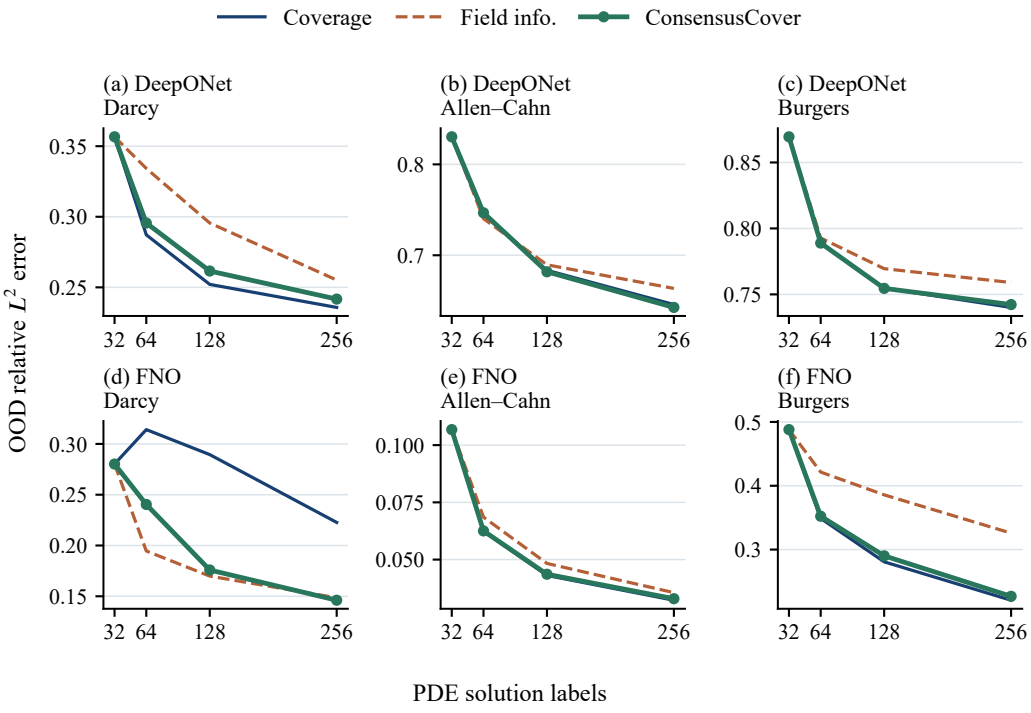

Figure 2: Held-out OOD relative-L^2 learning curves for the three fixed or adaptive selection rules across DeepONet and FNO. Each point averages 20 paired trials at $32, 64, 128, 256$ source solution labels. The target input correlation length is half the source value. The same inputs and training randomization are used within each setting.

5.2 Endpoint behavior complements integrated accuracy

Table 2 reports the 256-label mean for every confirmed setting. The first 32-label point is common within a paired trial, so the AUC gives substantial weight to early and middle improvements. The endpoint answers a related but distinct question: what is obtained after the full 224 additional source solves? On Darcy/FNO, ConsensusCover has the lowest endpoint mean, 0.1461 versus 0.1485 for field information, although its integrated AUC is higher. The adaptive sequence catches up late; the 6.62% AUC gap is therefore not a final-budget deficit. On Burgers/FNO, the adaptive endpoint 0.2267 remains close to coverage's 0.2206, while field information is 0.3258.

Table 2: Mean held-out OOD relative-L^2 error at 256 labels in the independent confirmation. These endpoints complement the prespecified four-budget AUC in Table 1.

Setting	Coverage	Field info.	ConsensusCover
DeepONet / Darcy	0.2358	0.2552	0.2417
DeepONet / Allen–Cahn	0.6458	0.6635	0.6426
DeepONet / Burgers	0.7401	0.7590	0.7423
FNO / Darcy	0.2227	0.1485	0.1461
FNO / Allen–Cahn	0.0322	0.0356	0.0329
FNO / Burgers	0.2206	0.3258	0.2267

The endpoint ranks do not uniformly match the AUC ranks: DeepONet/Allen–Cahn and FNO/Darcy favor the adaptive method at 256 labels, but only the former has an adaptive mean AUC below both fixed means. This distinction matters when PDE solutions are purchased sequentially. A selector can have a strong final model while spending more labels to reach that quality. The primary metric integrates the observed trajectory, without extrapolating beyond 256 labels or claiming monotone error reduction for every seed.

Paired contrasts. The cross-setting ratios above compare setting-level means. To show how much of each difference is stable across the matched runs, Table 3 reports mean AUC differences between ConsensusCover and each fixed selector with pointwise 95% paired-bootstrap intervals. A negative entry favors ConsensusCover. Each resample preserves the seed's source and target fields, initial labels, and training randomization across methods. The intervals are descriptive secondary comparisons; they are not simultaneous tests across the twelve contrasts.

Table 3: Paired mean OOD AUC differences (ConsensusCover minus comparator), in units of 10^{-3}, and pointwise 95% bootstrap intervals over 20 complete seeds per setting. The prespecified cross-setting decision uses Table 1.

Setting	vs. coverage	vs. field info.
DeepONet / Darcy	$+7.5\ [6.3, 8.8]$	$-26.7\ [-28.7, -24.6]$
DeepONet / Allen–Cahn	$-1.2\ [-4.1, 1.1]$	$-7.9\ [-11.2, -4.8]$
DeepONet / Burgers	$+0.5\ [-1.8, 3.1]$	$-12.0\ [-16.7, -7.4]$
FNO / Darcy	$-86.4\ [-96.3, -77.5]$	$+11.7\ [8.2, 15.8]$
FNO / Allen–Cahn	$+0.5\ [0.1, 0.9]$	$-4.1\ [-5.1, -2.9]$
FNO / Burgers	$+6.5\ [2.9, 10.5]$	$-84.1\ [-92.0, -76.0]$

The paired differences make the trade-off explicit. On Darcy/FNO, the adaptive rule gains 0.0864 AUC relative to coverage but gives up 0.0117 relative to field information. On Burgers/FNO, it gives up 0.0065 to coverage and gains 0.0841 over field information. These are asymmetric costs: the adaptive rule incurs a small loss relative to the stronger fixed selector in the two conflicting FNO cells while avoiding much larger losses relative to the weaker one. On Allen–Cahn and Burgers with DeepONet, the intervals against coverage include zero, so the small mean differences there should not be over-interpreted. The intervals also show that the cross-setting result is not driven solely by averaging away large paired errors within a cell.

5.3 Agreement responds to the architecture-specific conflict

The mean information-agreement counts on Darcy/FNO are 9.70, 28.35 and 58.30 in the three active acquisition rounds; the corresponding chance-corrected strengths are 0.312, 0.470 and 0.529. Darcy/DeepONet has intermediate counts 4.85, 10.75 and 21.20. In each of the other four settings, the mean accepted count is at most 1.65 in any round. The method therefore draws heavily on field information where the independent fixed-field rule wins and mostly retains target coverage elsewhere. Agreement is measured from inputs and the currently trained source-labelled model, without examining held-out target solutions.

At the first active round, 32 new labels are chosen from 992 unsolved candidates, so independent uniform batches would overlap by only $32^2/992 \approx 1.03$ inputs on average. At the next two rounds the corresponding chance overlaps are $64^2/960 \approx 4.27$ and $128^2/896 \approx 18.29$. The observed Darcy/FNO accepted counts exceed these baselines by large margins; even after correction, they grow from 9.70 to 58.30. The other four settings remain close to zero accepted inputs, including FNO/Allen–Cahn and FNO/Burgers, where coverage is the stronger fixed rule. Thus the mechanism diagnostic and the AUC comparison tell a consistent story, although neither proves that each accepted individual candidate has higher value than a coverage candidate.

Table 4 gives the complete setting-level mechanism record. The first two Darcy rows show that agreement is not determined by the input PDE alone: the FNO readout produces more accepted information candidates than the DeepONet readout on the same equation. The remaining four rows stay near the coverage fallback across all three rounds. The table reports accepted counts, not raw

intersections; the chance expectation has already been removed by (9). Counts are means across 20 paired trials and can therefore be fractional even though each trial selects whole fields.

Table 4: Mean number of chance-corrected information-agreement selections at each acquisition round, and mean cumulative selection time per curve. Round sizes are 32, 64 and 128 new source solutions.

Setting	$32 \to 64$	$64 \to 128$	$128 \to 256$	Selection (s)
DeepONet / Darcy	4.85	10.75	21.20	4.95
DeepONet / Allen–Cahn	0.20	1.10	0.75	4.88
DeepONet / Burgers	0.35	1.00	1.45	4.88
FNO / Darcy	9.70	28.35	58.30	5.43
FNO / Allen–Cahn	0.20	1.35	1.65	5.28
FNO / Burgers	0.50	1.60	1.40	5.32

Adaptation has a cost. Selection averages roughly 4.9–5.4 seconds for ConsensusCover across settings, compared with about 0.49 seconds for projected coverage and 0.75–0.78 seconds for field information. Training dominates the end-to-end curve time. This timing describes the reported implementation and hardware, rather than a hardware-independent complexity guarantee.

5.4 Secondary evaluation across the shift ladder

The primary acquisition pool has correlation-length factor 0.5 relative to the source law. The same trained operators were also evaluated on new inputs at factors 0.3, 0.7, and 1.0. No method selects labels using solutions from these evaluation pools. Table 5 recomputes the per-setting better-fixed mean AUC at each factor and reports the adaptive ratio's maximum and equal-setting geometric mean. This is a secondary assessment of transfer away from the acquisition target, not an additional confirmation criterion.

Table 5: ConsensusCover AUC relative to the better fixed-rule mean at each evaluation shift factor, summarized over six architecture–equation settings. Only factor 0.5 is the prespecified primary outcome.

Correlation-length factor	0.3	0.5	0.7	1.0
Maximum ratio	1.0540	1.0662	1.0556	1.0581
Geometric-mean ratio	1.0199	1.0207	1.0129	1.0171

The largest excess stays between 5.40% and 6.62% on the observed ladder, and the geometric-mean excess stays between 1.29% and 2.07%. The factor-1.0 inputs are newly sampled from the source law rather than the training catalogue; their risk is therefore an independent in-law evaluation. These results are compatible with the primary robustness finding, while leaving open other types of shift such as amplitude changes and changes in PDE coefficients. The older coverage-component experiment tests amplitude shift separately (Appendix D); it does not supply an amplitude-shift result for the adaptive rule.

6 Discussion and conclusion

ConsensusCover combines chance-corrected agreement between field and current-model response rankings with target-probability coverage for the remaining labels. Across three PDEs and two architectures, independent paired trials reduce the largest AUC excess over the better fixed rule to 6.62%, versus 55.44% for coverage and 30.92% for field information. Agreement activates most strongly on Darcy/FNO, where field information is the stronger fixed design.

Agreement does not guarantee individual label value, and the fallback's geometric bound excludes information-selected labels. Other shifts, architectures, catalogue sizes and weaker initial models remain untested.

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

[1] Antoine de Mathelin, Francois Deheeger, Mathilde Mougeot, and Nicolas Vayatis. Discrepancy-based active learning for domain adaptation. In *International Conference on Learning Representations*, 2022. URL https://openreview.net/forum?id=p98WJxUC3Ca.

[2] David Holzmüller, Viktor Zaverkin, Johannes Kästner, and Ingo Steinwart. A framework and benchmark for deep batch active learning for regression. *Journal of Machine Learning Research*, 24(164):1–81, 2023. URL https://jmlr.org/papers/v24/22-0937.html.

[3] Nikola Kovachki, Zongyi Li, Burigede Liu, Kamyar Azizzadenesheli, Kaushik Bhattacharya, Andrew Stuart, and Anima Anandkumar. Neural operator: Learning maps between function spaces with applications to PDEs. *Journal of Machine Learning Research*, 24(89):1–97, 2023. URL https://jmlr.org/papers/v24/21-1524.html.

[4] Shibo Li, Xin Yu, Wei Xing, Robert M. Kirby, Akil Narayan, and Shandian Zhe. Multi-resolution active learning of Fourier neural operators. In *Proceedings of the 27th International Conference on Artificial Intelligence and Statistics (AISTATS)*, volume 238 of *PMLR*, pp. 2440–2448, 2024. URL https://proceedings.mlr.press/v238/li24k.html. arXiv:2309.16971 (MRA-FNO).

[5] 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. URL https://arxiv.org/abs/2010.08895. arXiv:2010.08895.

[6] Marco Loog. Nearest neighbor-based importance weighting. In *IEEE International Workshop on Machine Learning for Signal Processing*, pp. 1–6, 2012. doi: 10.1109/MLSP.2012.6349714. URL https://arxiv.org/abs/2102.02291.

[7] Lu Lu, Pengzhan Jin, Guofei Pang, Zhongqiang Zhang, and George Em Karniadakis. Learning nonlinear operators via DeepONet based on the universal approximation theorem of operators. *Nature Machine Intelligence*, 3:218–229, 2021. doi: 10.1038/s42256-021-00302-5. URL https://arxiv.org/abs/1910.03193. arXiv:1910.03193.

[8] Daniel Musekamp, Marimuthu Kalimuthu, David Holzmüller, Makoto Takamoto, and Mathias Niepert. Active learning for neural PDE solvers. In *International Conference on Learning Representations*, 2025. URL https://openreview.net/forum?id=x4ZmQaumRg. AL4PDE; LCMD / Core-Set benchmark.

[9] Ozan Sener and Silvio Savarese. Active learning for convolutional neural networks: A core-set approach. In *International Conference on Learning Representations*, 2018. URL https://openreview.net/forum?id=H1aIuk-RW.

[10] Nickolas Winovich, Mitchell Daneker, Lu Lu, and Guang Lin. Active operator learning with predictive uncertainty quantification for partial differential equations. *Journal of Computational Physics*, 555:114791, 2026. doi: 10.1016/j.jcp.2026.114791. URL https://arxiv.org/abs/2503.03178. arXiv:2503.03178.

A MASS-PRESERVING TRANSPORT AND PREDICTION RISK

This section supplies the proofs for mass-preserving TargetCover. It separates the effect of transferring deployment mass from the approximation error of the trained operator. The density comparator concerns (6); the distance analysis concerns (4).

Preserving the deployment measure. For the target-to-source kernel matrix, define its row sums $d_i = \sum_j K_{ij} > 0$. The original demand weights can be written as

$$\bar{q}_j = \sum_i \widetilde{w}_i \frac{K_{ij}}{d_i}, \qquad \widetilde{w}_i = \frac{w_i d_i}{\sum_r w_r d_r}. \tag{13}$$

Thus global normalization changes the relative mass of target inputs when their aggregate source affinities differ. A row-normalized transfer instead uses a coupling $\Gamma_{ij} = w_i K_{ij}/d_i$ and source weights $\mu_j = \sum_i \Gamma_{ij}$. It satisfies $\sum_j \Gamma_{ij} = w_i$ exactly. The nearest-source version places each target input's mass on its nearest source feature, sharing exact ties equally. Both constructions use input fields. Density-ratio weighting provides a separate importance-weighting comparison.

For positive reference weights m_j, replacing K_{ij} by $m_j K_{ij}$ changes both the effective probabilities and the conditional transfer. Nearest-source transfer solves (3): each row places its prescribed mass on a minimum-cost source site, so summing the row minima gives the displayed optimum. Reference weights are absent from this optimization. With fixed support, representation and deterministic tie handling, the coupling and its source marginal are therefore unchanged by reference reweighting. Every greedy gain is then unchanged, proving invariance of the complete nested selected set by induction for the fixed source support.

Let $z_j = Pa_j$ and $\widetilde{z}_i = P\widetilde{a}_i$, where P is an orthogonal projection onto a fixed input subspace. The input and output Hilbert norms are denoted by $\|\cdot\|_A$ and $\|\cdot\|_U$, respectively; on a grid, we use root-mean-square norms and scale POD coordinates accordingly. For a nonempty labelled set S, let $s_S(j) \in S$ minimize $\|z_j - z_s\|_A$, with deterministic tie breaking. Define

$$C_\mu(S) = \sum_j \mu_j \|z_j - z_{s_S(j)}\|_A^2, \qquad A_\Gamma = \sum_{i,j} \Gamma_{ij} \|\widetilde{z}_i - z_j\|_A^2. \tag{14}$$

The acquisition rule greedily decreases C_μ, the squared geometric term in the following risk estimate. The coupling, projection, and source pool stay fixed across rounds.

Theorem 1 (Prediction-risk decomposition)**.** *Let $\Gamma \geq 0$ have row sums w_i and column sums μ_j. Suppose the residual map $e_S = \widehat{\mathcal{G}}_S - \mathcal{G}$ is L_e-Lipschitz in the full input norm on the target and selected source inputs. Define*

$$E_\Gamma(S) = \sum_{i,j} \Gamma_{ij} \|e_S(a_{s_S(j)})\|_U^2, \tag{15}$$

$$T_\Gamma(S) = \sum_{i,j} \Gamma_{ij} \|(I-P)(\widetilde{a}_i - a_{s_S(j)})\|_A^2, \tag{16}$$

$$\widehat{R}_Q^{\mathrm{sq}}(S) = \sum_i w_i \|e_S(\widetilde{a}_i)\|_U^2. \tag{17}$$

Then

$$\sqrt{\widehat{R}_Q^{\mathrm{sq}}(S)} \leq \sqrt{E_\Gamma(S)} + L_e \left(\sqrt{A_\Gamma} + \sqrt{C_\mu(S)} + \sqrt{T_\Gamma(S)} \right). \tag{18}$$

In particular, $L_e \leq L_{\mathcal{G}} + L_{\widehat{\mathcal{G}}_S}$ when the true and learned operators have these Lipschitz constants.

Proof. For each pair (i, j), the triangle inequality and residual stability give

$$\|e_S(\widetilde{a}_i)\|_U \leq \|e_S(a_{s_S(j)})\|_U + L_e \|\widetilde{a}_i - a_{s_S(j)}\|_A,$$
$$\|\widetilde{a}_i - a_{s_S(j)}\|_A \leq \|\widetilde{z}_i - z_j\|_A + \|z_j - z_{s_S(j)}\|_A + \|(I-P)(\widetilde{a}_i - a_{s_S(j)})\|_A.$$

Take the weighted $\ell^2(\Gamma)$ norm and apply Minkowski's inequality. The row sums recover the empirical target risk, and the column sums recover $C_\mu(S)$. This proves (18). □

The terms correspond to four sources of prediction error: approximation by the source pool, coverage by selected inputs, discarded representation directions, and fitting of the selected labels. Acquiring labels directly reduces C_μ; the selected representatives determine T_Γ, while operator training determines E_Γ and L_e. The theorem bounds squared absolute field error. A relative-error version additionally requires a positive lower bound on solution norms; the Gaussian evolution inputs allow arbitrarily small solution norms.

Theorem 2 (Greedy reduction of the risk's geometric term)**.** *Fix Γ, P, and a nonempty initial labelled set S_0. The function $H(B) = C_\mu(S_0) - C_\mu(S_0 \cup B)$ is normalized, monotone and submodular. If S_g adds b greedy candidates and S_* minimizes C_μ among sets extending S_0 by at most b candidates, then*

$$C_\mu(S_g) \le e^{-1} C_\mu(S_0) + (1 - e^{-1}) C_\mu(S_*). \tag{19}$$

Proof. For each source demand site j, its improvement is the maximum, over newly selected candidates s, of $[d(z_j, S_0)^2 - \|z_j - z_s\|_A^2]_+$. A nonnegative weighted sum of these maxima is submodular. Apply the cardinality-constrained greedy theorem and rearrange. □

Why the target-distance objective matters. When $P = I$ and the coupling uses nearest source sites, write $C_Q(S) = \sum_i w_i d(\widetilde{a}_i, S)^2$. The distance-to-set function is one-Lipschitz, hence Minkowski's inequality gives the two-sided estimate

$$\left| \sqrt{C_Q(S)} - \sqrt{C_\mu(S)} \right| \le \sqrt{A_\Gamma}. \tag{20}$$

The projection cost therefore quantifies how accurately compressed source demand represents direct target coverage, independently of the selected set.

For scalar L-Lipschitz response functions and noiseless values observed on a fixed set S, the minimax empirical squared prediction error is exactly $L^2 C_Q(S)$. Nearest-neighbour interpolation gives the upper bound. For the lower bound, consider the two indistinguishable functions $f_\pm(a) = \pm L d(a, S)$, both zero on S. For any predictions h_i, the average of their two risks is $\sum_i w_i h_i^2 + L^2 C_Q(S) \ge L^2 C_Q(S)$. The minimax statement concerns the scalar Lipschitz class on a fixed observation set. The neural-operator estimate is Theorem 1.

Theorem 3 (Comparison with the best source-restricted target design)**.** *Use the full input norm, so $P = I$, and let the target weights be nonnegative and sum to one. Fix a finite source pool, a nonempty initial set S_0, and $1 \le b \le N_P - |S_0|$ additional queries. Let S_Q^* minimize C_Q over source sets containing S_0 with at most b additions, and let S_g be the TargetCover greedy extension. With $\rho_b = (1 - 1/b)^b$,*

$$\sqrt{C_Q(S_g)} \le \sqrt{A_\Gamma} + \sqrt{\rho_b C_\mu(S_0) + (1 - \rho_b)\big(\sqrt{C_Q(S_Q^*)} + \sqrt{A_\Gamma}\big)^2}. \tag{21}$$

For noiseless scalar L-Lipschitz responses, multiplying the two sides by L compares the corresponding optimal worst-case RMS prediction errors. The same statement holds for responses in any nonzero Hilbert space.

Proof. Write $B_* = S_Q^* \setminus S_0$. At each greedy step, submodularity implies that the remaining improvement $H(B_*) - H(B_t)$ is at most b times the largest available marginal gain. Iterating gives $C_\mu(S_g) \le \rho_b C_\mu(S_0) + (1 - \rho_b) C_\mu(S_Q^*)$. Apply (20) first to S_Q^* and then to S_g to obtain (21). For the prediction interpretation, nearest-neighbour interpolation has pointwise error at most $L d(a, S)$ for every L-Lipschitz response. Conversely, choose a unit output vector v and the indistinguishable responses $f_\pm(a) = \pm L d(a, S) v$. Their labels on S are zero. For any reconstruction, the average of its two squared target risks is its own squared prediction norm plus $L^2 C_Q(S)$. Thus the optimal worst-case squared risk is exactly $L^2 C_Q(S)$. □

This comparison separates the greedy error from replacing target sites by source sites. For an arbitrary trained neural operator, the fitting and stability terms in Theorem 1 remain relevant.

Physical stability for a nonlinear evolution operator. Consider periodic Allen–Cahn evolution $\partial_t u = \nu \Delta u + \gamma(u - u^3)$ with $\nu, \gamma \ge 0$. For two solutions u, v, multiply their difference equation by

$u - v$ and integrate over the periodic domain. Since $(u^3 - v^3)(u - v) \geq 0$, integration by parts yields

$$\frac{1}{2}\frac{d}{dt}\|u - v\|_{L^2}^2 \leq \gamma\|u - v\|_{L^2}^2, \qquad \|\mathcal{G}_T(a) - \mathcal{G}_T(b)\|_{L^2} \leq e^{\gamma T}\|a - b\|_{L^2}. \tag{22}$$

Thus the physical part of L_e has an explicit constant for this nonlinear, time-dependent problem. The implemented reaction substep has derivative at most $e^{\gamma \Delta t}$ and the Fourier diffusion substep is a contraction; the same stability constant holds for their symmetric composition in the discrete root-mean-square norm.

For the discrete Poisson source operator with zero boundary conditions, the corresponding constant is the inverse smallest eigenvalue of the positive interior discrete Laplacian. Adding the network constant gives L_e. A population-risk extension requires a concentration argument and suitable tail or boundedness assumptions.

For the implemented ReLU DeepONet, a network constant is computable from its weights. Let B_θ be the branch network, let $W_1, \ldots, W_L$ be its weight matrices, and let T_θ contain the trunk features evaluated at the output grid. If the scalar input and output normalization scales are $s_a > 0$ and $s_u > 0$, then on equal-size input and output grids

$$L_{\widehat{\mathcal{G}}_S} \leq \frac{s_u}{s_a}\|T_\theta\|_2 \prod_{\ell=1}^{L} \|W_\ell\|_2. \tag{23}$$

This follows by composing affine maps and the one-Lipschitz ReLU, then applying the trunk matrix. Equal input/output grid sizes cancel the conversion between Euclidean and root-mean-square norms. The estimate supplies an explicit stability constant; Appendix B measures its tightness.

Computational implications. A coupling from N_Q target inputs to N_P source sites is computed once. After its column sums are retained, greedy acquisition uses N_P demand sites, independent of the target-pool size. A dense implementation requires $O(N_P N_Q d + N_P^2 d)$ distance-construction work and $O(bN_P^2)$ greedy work. The source-source distances and transferred weights can be reused across acquisition rounds when the input representation is fixed.

B Numerical verification and coverage-component diagnostics

Elliptic inputs and references. The elliptic grids include the boundary and use spacing $1/(n-1)$. Gaussian grid noise is Fourier-filtered with amplitude proportional to $\exp(-\pi^2\ell^2|k|^2)$, where k is measured in cycles per unit length. Each resulting field is centred and divided by its spatial standard deviation. Writing the standardized field as g, the Darcy input is $a = \exp(g)$ and the Poisson forcing is $f = 1 + g$. These per-realization standardized laws differ from the ensemble-normalized evolution law below. Darcy uses a conservative five-point flux discretization with harmonic face conductivities; Poisson uses the constant-coefficient five-point Laplacian. Both impose zero Dirichlet data and solve the resulting sparse linear systems.

The nonlinear benchmarks use the periodic square $[0, 1)^2$ with nodes $(j/n, k/n)$ for $0 \leq j, k < n$. White Gaussian grid noise is filtered by $\exp(-\ell^2|\xi|^2/4)$ in Fourier space, where ξ is the angular wave-number vector. A deterministic factor makes the ensemble marginal variance equal to the specified value, using one ensemble-level scaling for all realizations. Changing the correlation length leaves terminal time, diffusivity, reaction rate, grid, and boundary conditions unchanged.

Allen–Cahn. The reaction flow over a step h is evaluated exactly:

$$\mathcal{R}_h(v) = \frac{v}{\sqrt{v^2 + (1 - v^2)e^{-2\gamma h}}}. \tag{24}$$

The diffusion flow multiplies each Fourier coefficient by $e^{-\nu|\xi|^2 h}$. The reference is the symmetric composition $\mathcal{R}_{h/2}\mathcal{D}_h\mathcal{R}_{h/2}$, with $h \leq 0.005$ and an integer number of steps reaching the specified terminal time. Constant equilibria $-1, 0, 1$, the exact spatially constant reaction trajectory, temporal refinement, and the stability estimate in (22) are tested separately from acquisition accuracy.

Viscous Burgers. The scalar equation contains nonlinear advection in both spatial directions. A Fourier two-thirds cutoff is applied to the initial condition and nonlinear flux, and a fourth-order Runge–Kutta method advances the resolved state. The step is at most 0.005 and is further bounded by conservative advection and diffusion stability limits. Tests check constant solutions, conservation of the spatial mean, temporal refinement and spatial refinement. Training labels are solutions of the 32×32 dealiased evolution map.

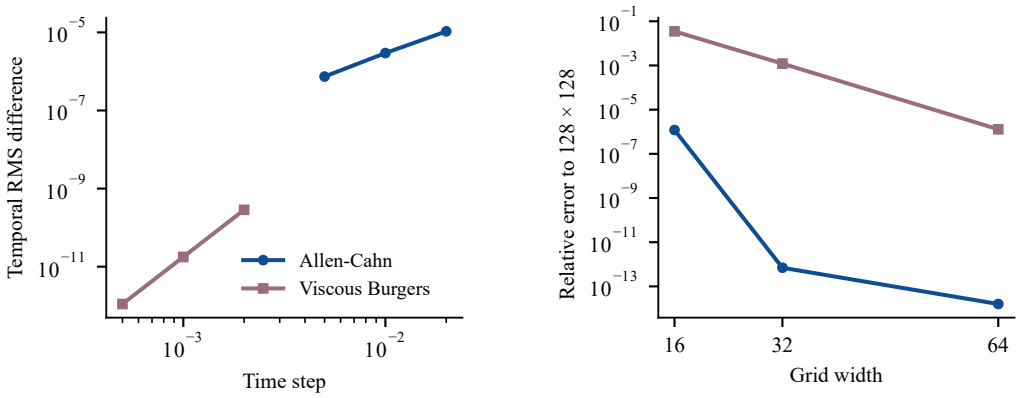

Figure 3: Reference-solver verification for the prescribed smooth initial field $0.1 + 0.4\sin(2\pi x) + 0.3\cos(4\pi y) + 0.2\sin(2\pi(x+y))$. Left: RMS differences of solutions at successive time refinements, plotted against the coarser step. Right: relative spatial differences against a 128×128 reference restricted to coincident grid nodes. These solver convergence experiments use $T = 0.25$, $\nu = 0.02$, and $\gamma = 2$ for Allen–Cahn.

The asymptotic temporal orders are approximately two for the symmetric Allen–Cahn splitting and four for the Burgers integrator. For the smooth field in Figure 3, the Burgers 32^2 solution differs from the restricted 128^2 reference by approximately 0.12%. A separate five-field random-input check gives maximum relative differences of 2.73% at source length 0.12 and 0.44% at target length 0.06. For Allen–Cahn these maxima are below 4.4×10^{-8} and 1.8×10^{-5}, respectively. These refinement errors quantify spatial discretization accuracy for the five random fields and the smooth reference field.

Neural stability bounds and risk measurements. Five additional trials on each of Poisson and Allen–Cahn use 1,024 source candidates and the five label counts from 32 to 256. For each trained DeepONet, we compute every branch matrix spectral norm, the spectral norm of the trunk matrix evaluated on the output grid, and the actual normalization scales. These give the analytic neural constant in (23); the physical constant is supplied by the PDE. All 50 resulting finite-grid inequalities are satisfied. Their upper bounds are approximately 278–995 times the observed target RMS error, showing that products of spectral norms substantially overestimate the error.

The earlier four-PDE DeepONet coverage study evaluates all four decomposition terms for every method, trial and training-set size. Held-out target inputs define a nearest-source coupling in the acquisition POD geometry. This measures coverage on new inputs separately from coverage on the unlabelled acquisition sample. The regression centres logged risk terms within task, trial and label count. Bootstrap resampling preserves complete trial clusters within each PDE, including all methods and training sizes. The resulting associations are exploratory analyses of prediction error.

The standardized association between logged held-out coverage and logged target MSE is 0.393, with a trial-cluster bootstrap 95% interval $[0.354, 0.434]$ after accounting for the other measured terms and task, trial and label-count means. This is an association for fixed TargetCover and its controls, not a predictive guarantee for ConsensusCover.

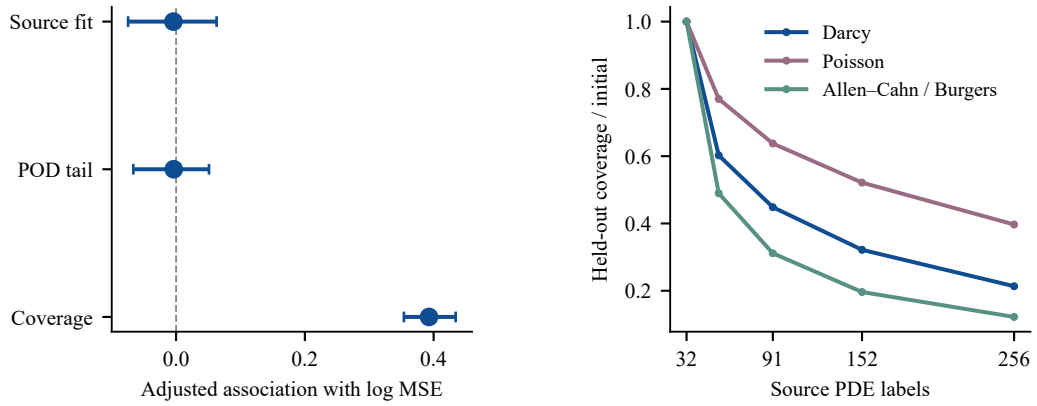

Figure 4: Coverage-component geometry and held-out prediction in the earlier four-PDE DeepONet study. Left: standardized associations with log target MSE, with intervals resampling complete trial clusters. Right: coverage relative to its 32-label value. Allen–Cahn and Burgers share input laws and selected inputs but have different prediction errors.

Acquisition settings and simulation cost.

Kernel controls. Every radial basis function (RBF) control uses $\exp(-\|z - z'\|^2/(2\sigma^2))$. Its automatic scale is the median of the strictly positive pairwise Euclidean distances in the concatenated target and source feature arrays, with a lower bound of 10^{-12} and fallback 1 for a degenerate array. The original density, uniform-RBF and direct-target rules recompute this scale each round using the remaining source candidates. The crossed density–distance and projected–RBF controls instead use the full source pool and target sample, fixing the weights and costs throughout selection. All scales are determined from input distances. Projected distance coverage and both LCMD rules require no kernel bandwidth.

The development experiments select projected transfer before the main evaluation. The original coverage comparisons use a Linux CPU environment; the supplementary comparisons use a macOS CPU environment with their own TargetCover reference. Each method difference is computed within one environment on shared inputs and training randomization.

Held-out reference solutions are shared within comparisons and never enter selection. Training-label costs count every selected source solution once. Acquisition-plus-source-solve time and training-plus-ID-evaluation time are recorded separately. Shared reference caches reduce repeated evaluation work; label costs count distinct source solutions.

A zero reference norm is replaced by one in the relative-error denominator.

Large unlabelled target samples. Five independent Darcy input samples use 1,024 source candidates, 32 initial labels and 224 additions. Target sample sizes are 1,024, 4,096, 16,384 and 32,768, with source/target correlation lengths $0.12/0.06$ and source-fitted POD dimension 16. TargetCover, direct target distance and fixed-scale direct RBF coverage use one common exact lazy-greedy implementation. Source-demand sites with zero weight are omitted. We permute method execution order and time each complete selection, including distance construction. We then solve the 256 selected source PDEs without reusing labels from other methods.

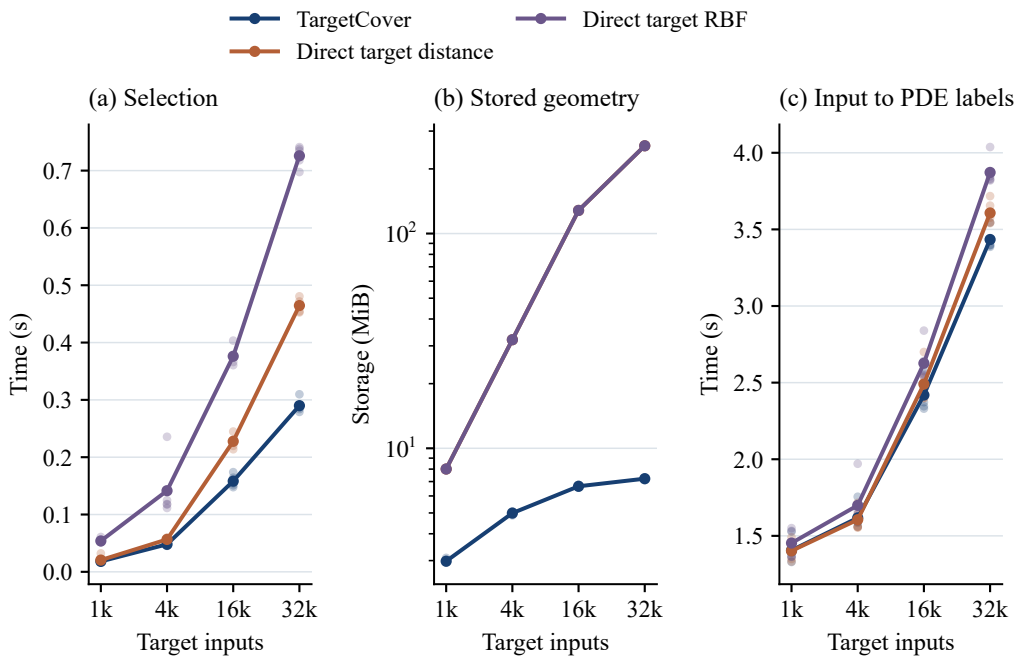

Figure 5: Computational effect of representing target demand on source sites. Lines give means over five independently sampled inputs; points show every trial. Left: selection, including distance construction. Middle: retained distance and weight arrays, not peak process memory. Right: source/target input generation, POD, acquisition and 256 source PDE solutions. Neural training is not part of this timing experiment.

At 32,768 target inputs, mean selection times are 0.290 s for TargetCover, 0.465 s for direct distance and 0.726 s for direct RBF. Retained geometry occupies 7.22 MiB for TargetCover and 256.25 MiB for either direct method. Including input generation, POD and source solves, TargetCover lowers training-data construction time by 4.8% against direct distance. Its direct target coverage cost is 0.37% larger on the same input samples. This geometric comparison measures compression error separately from the neural prediction experiments.

C Coverage-component prediction and mechanism tests

This earlier study evaluates fixed TargetCover as the coverage component of ConsensusCover. It compares eight methods on Darcy, Poisson, Allen–Cahn and Burgers with DeepONet, 20 paired trials per task, and label counts $32, 54, 91, 152, 256$. These results contain no adaptive ConsensusCover curve; the independent three-PDE, two-architecture confirmation uses $32, 64, 128, 256$ labels and is reported in Section 5. Table 6 gives the coverage study's mean AUCs. Figure 6 shows its learning curves. Table 7 consolidates every numerical curve and source/target AUC. Figure 7 shows the magnitude and uncertainty of the 28 method comparisons. Figure 8 evaluates the same trained operators across the complete roughness ladder, including the source distribution.

Table 6: OOD learning-curve error on four PDE operators. Entries are AUC (mean ± standard deviation over 20 trials), using 1,024 candidates and 32–256 labels. Bold marks the lowest mean; paired inference is shown in Figure 7.

Method	Darcy	Poisson	Allen–Cahn	Burgers
TargetCover	**0.2595** ± 0.0034	0.2113 ± 0.0053	**0.6897** ± 0.0060	0.7677 ± 0.0094
Random	0.2774 ± 0.0048	0.2268 ± 0.0072	0.6996 ± 0.0069	0.7793 ± 0.0094
CoreSet	0.2997 ± 0.0047	0.2131 ± 0.0050	0.7096 ± 0.0086	0.7948 ± 0.0105
LCMD	0.2824 ± 0.0060	0.2182 ± 0.0065	0.6990 ± 0.0067	0.7810 ± 0.0101
Density	0.2684 ± 0.0035	0.2259 ± 0.0045	0.6951 ± 0.0080	0.7707 ± 0.0084
Uniform demand	0.2786 ± 0.0038	0.2274 ± 0.0048	0.7050 ± 0.0083	0.7766 ± 0.0104
Source distance	0.2789 ± 0.0043	0.2274 ± 0.0052	0.7045 ± 0.0076	0.7764 ± 0.0118
Direct target	0.2607 ± 0.0040	**0.2097** ± 0.0050	0.6906 ± 0.0068	**0.7639** ± 0.0094

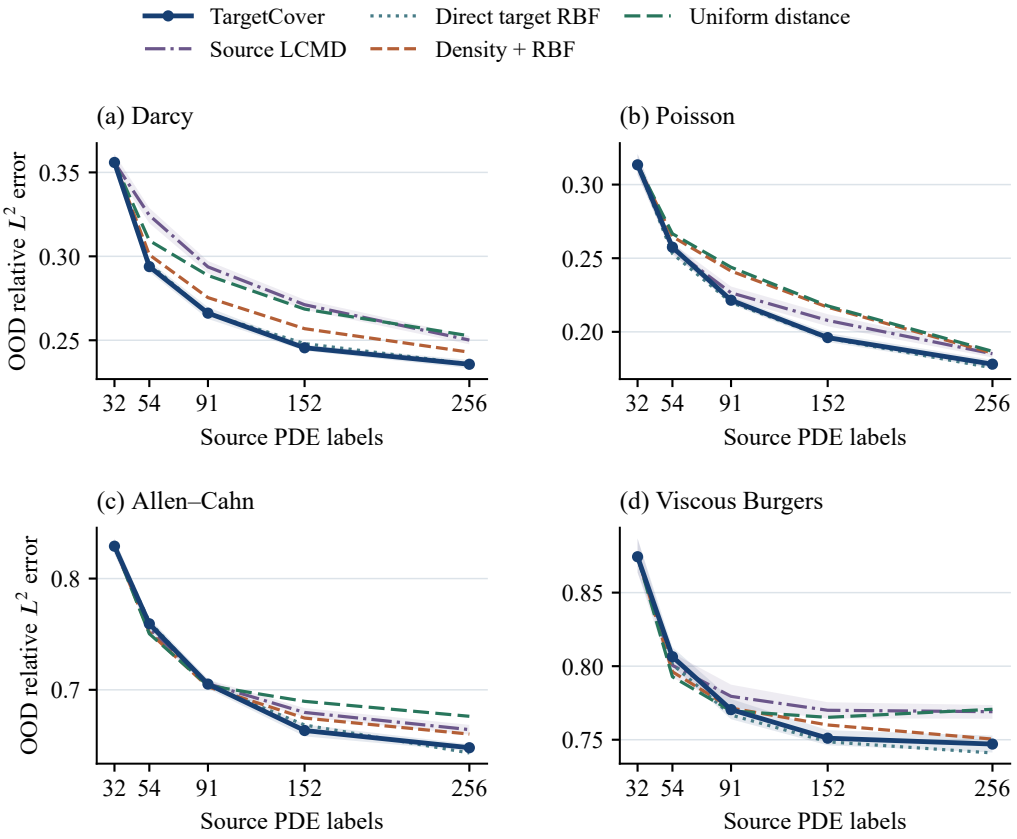

Figure 6: Fixed TargetCover in the earlier four-PDE DeepONet coverage study. Each curve averages 20 paired trials at $32, 54, 91, 152, 256$ source labels. Shading gives pointwise bootstrap 95% intervals for TargetCover and source LCMD. These curves are separate from the six-setting ConsensusCover confirmation in Figure 2.

Table 7: Learning curves at every training size, with source and target AUCs. Entries average 20 trials for each of the eight methods. Methods share the initial 32-label prediction within each trial. Table 6 gives OOD AUC standard deviations; Figure 7 gives paired comparison intervals.

Method	32	54	91	152	256	ID AUC	OOD AUC
Darcy							
TargetCover	0.3559	0.2939	0.2662	0.2455	0.2357	0.2493	0.2595
Random	0.3559	0.3184	0.2879	0.2659	0.2460	0.2465	0.2774
CoreSet	0.3559	0.3400	0.3201	0.2898	0.2614	0.2546	0.2997
Source LCMD	0.3559	0.3244	0.2937	0.2712	0.2501	0.2454	0.2824
Density + RBF	0.3559	0.3009	0.2754	0.2569	0.2429	0.2430	0.2684
Uniform + RBF	0.3559	0.3073	0.2859	0.2705	0.2520	0.2432	0.2786
Uniform + distance	0.3559	0.3094	0.2886	0.2686	0.2526	0.2426	0.2789
Direct target RBF	0.3559	0.2955	0.2659	0.2480	0.2359	0.2487	0.2607
Poisson							
TargetCover	0.3134	0.2575	0.2215	0.1961	0.1781	0.1405	0.2113
Random	0.3134	0.2692	0.2411	0.2146	0.1903	0.1431	0.2268
CoreSet	0.3134	0.2480	0.2230	0.2008	0.1823	0.1370	0.2131
Source LCMD	0.3134	0.2563	0.2265	0.2078	0.1851	0.1374	0.2182
Density + RBF	0.3134	0.2647	0.2413	0.2167	0.1856	0.1380	0.2259
Uniform + RBF	0.3134	0.2679	0.2444	0.2158	0.1885	0.1378	0.2274
Uniform + distance	0.3134	0.2666	0.2439	0.2177	0.1868	0.1384	0.2274
Direct target RBF	0.3134	0.2531	0.2200	0.1957	0.1755	0.1405	0.2097
Allen–Cahn							
TargetCover	0.8291	0.7594	0.7051	0.6634	0.6480	0.6015	0.6897
Random	0.8291	0.7596	0.7115	0.6765	0.6635	0.5818	0.6996
CoreSet	0.8291	0.7596	0.7220	0.6936	0.6695	0.5797	0.7096
Source LCMD	0.8291	0.7549	0.7055	0.6797	0.6642	0.5768	0.6990
Density + RBF	0.8291	0.7516	0.7027	0.6746	0.6602	0.5886	0.6951
Uniform + RBF	0.8291	0.7490	0.7024	0.6912	0.6781	0.5886	0.7050
Uniform + distance	0.8291	0.7504	0.7038	0.6896	0.6762	0.5879	0.7045
Direct target RBF	0.8291	0.7611	0.7045	0.6684	0.6433	0.6037	0.6906
Burgers							
TargetCover	0.8744	0.8064	0.7705	0.7511	0.7470	0.6563	0.7677
Random	0.8744	0.8040	0.7749	0.7732	0.7589	0.6454	0.7793
CoreSet	0.8744	0.8153	0.7910	0.7878	0.7809	0.6450	0.7948
Source LCMD	0.8744	0.8005	0.7795	0.7700	0.7691	0.6425	0.7810
Density + RBF	0.8744	0.7962	0.7715	0.7600	0.7504	0.6426	0.7707
Uniform + RBF	0.8744	0.7943	0.7711	0.7651	0.7695	0.6424	0.7766
Uniform + distance	0.8744	0.7928	0.7696	0.7652	0.7707	0.6415	0.7764
Direct target RBF	0.8744	0.8013	0.7668	0.7486	0.7410	0.6550	0.7639

The complete curves show that the target-shift advantage is present across the label schedule rather than only at its endpoint. At 256 labels, TargetCover has lower OOD error than source LCMD on all four PDEs: 0.2357 versus 0.2501 on Darcy, 0.1781 versus 0.1851 on Poisson, 0.6480 versus 0.6642 on Allen–Cahn, and 0.7470 versus 0.7691 on Burgers. Direct target RBF is close to TargetCover on Darcy and has the lower endpoint error on the other three tasks. The AUC column captures the full five-count trajectory, so an endpoint ordering need not imply the same integrated ordering.

The ID AUC column records the complementary source-law behavior. Source LCMD has lower ID AUC than TargetCover on each PDE, whereas TargetCover has lower OOD AUC than source LCMD on each PDE. Thus the same selected sets can trade source-law accuracy for target-law accuracy;

these coverage-component results alone do not establish an adaptive advantage. The independent ConsensusCover comparison in Table 1 addresses that separate claim.

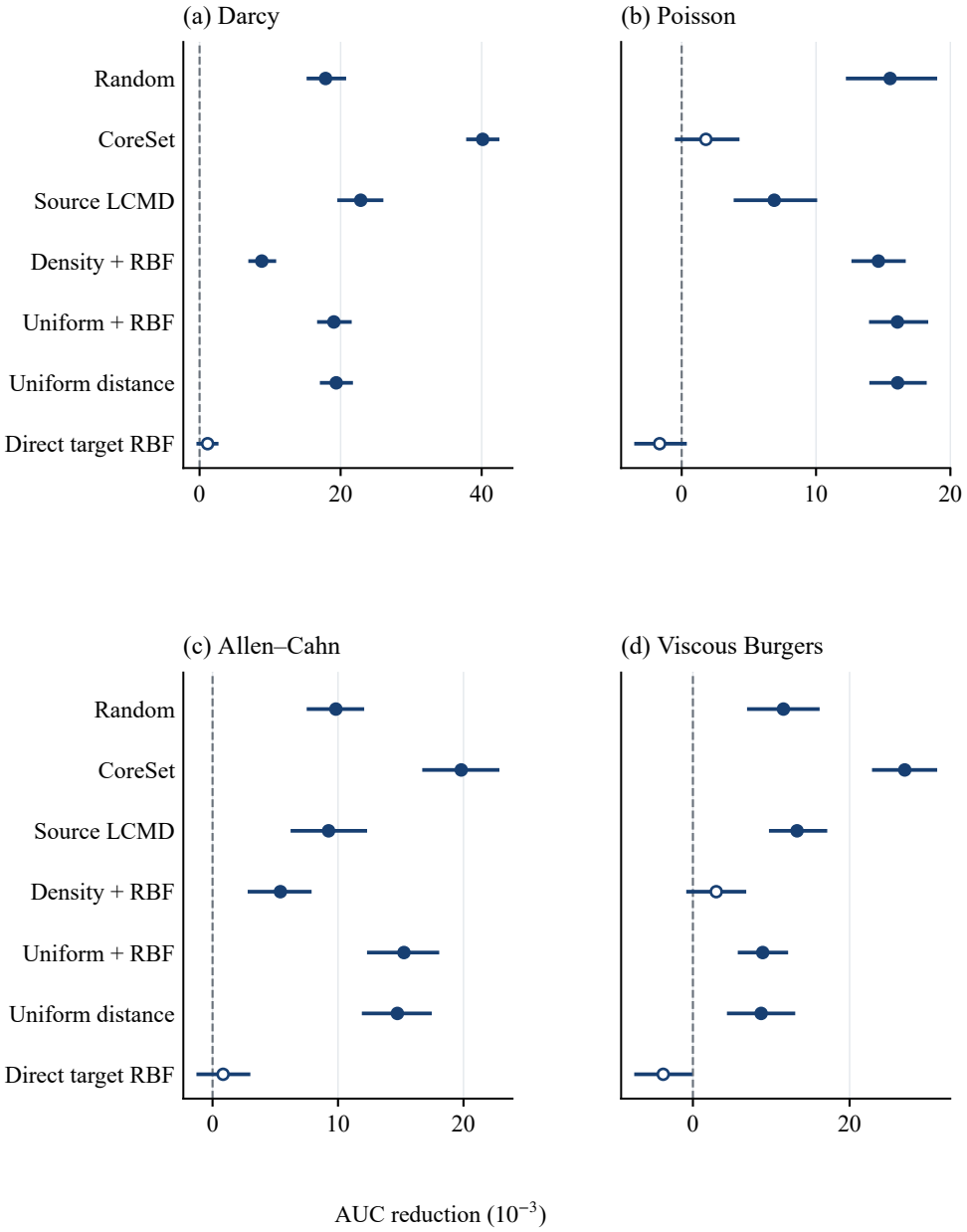

Figure 7: All 28 coverage-component method comparisons. Positive values mean lower AUC for TargetCover. Horizontal intervals are pointwise paired bootstrap 95% intervals; filled markers denote two-sided sign-randomization $p_{\mathrm{Holm}} < 0.05$ across the full 28-test family, and open markers denote the remaining comparisons. Interval exclusion of zero and corrected significance are different criteria. All use 20 paired trials.

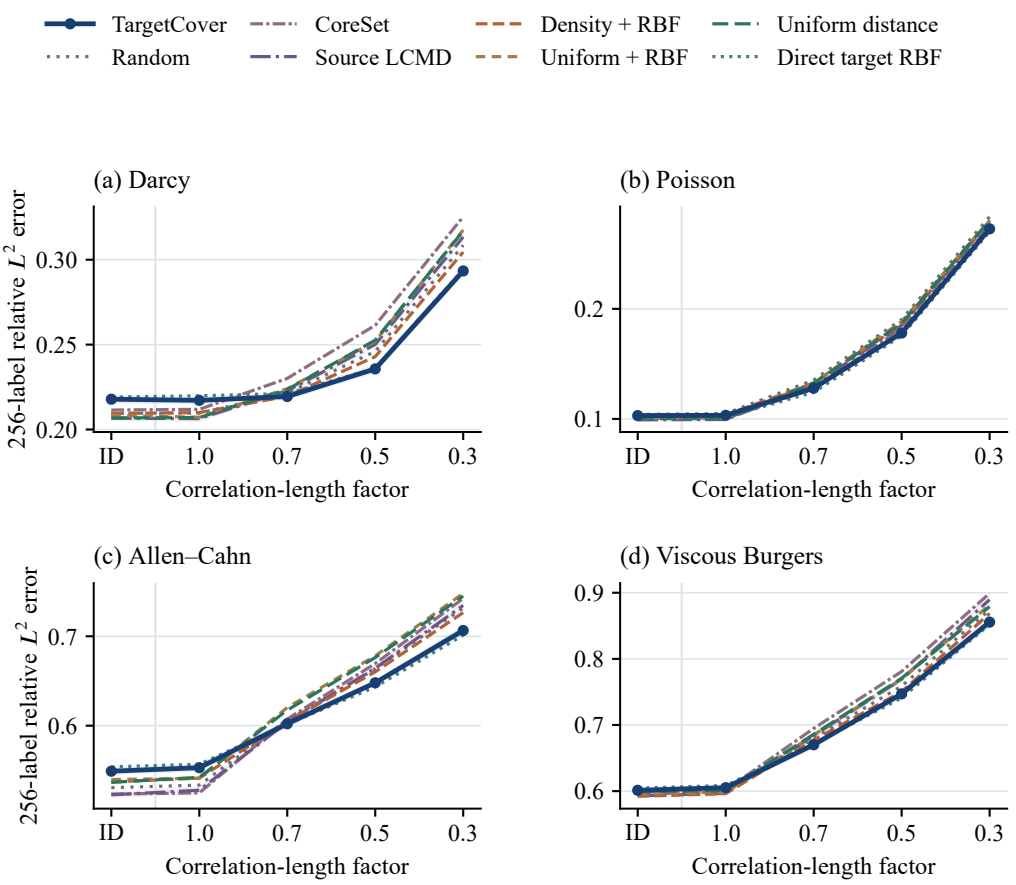

Figure 8: Prediction profiles across the full input-shift ladder at 256 labels. Each line averages 20 trials; all eight methods are shown. The target pool used in acquisition has correlation-length factor 0.5, while evaluation uses independently sampled fields at every plotted factor. The isolated ID point and factor-1.0 point are distinct samples from the same law. Shading gives pointwise bootstrap 95% intervals for TargetCover and source LCMD. These profiles expose the source/target trade-off as well as performance away from the acquisition target law.

Demand construction and independent mechanism tests. Table 8 brings together the supplementary projected reference, density–distance, projected–RBF and target-aware LCMD comparisons. Each task uses 20 paired trials; Figure 9 gives the comparison intervals. The two crossed controls form an eight-test family; target-aware LCMD, added after the original coverage analysis, forms its own four-test family. Each comparison uses its own TargetCover reference and common training conditions.

Table 8: Weight and distance ablations and target-aware cluster selection. Entries are OOD AUC (mean ± standard deviation over 20 trials). Bold marks the lowest mean. A dagger marks larger AUC than TargetCover after Holm correction within the eight-test ablation or four-test target-LCMD family. Figure 9 gives paired comparison intervals.

Demand / cost	Darcy	Poisson	Allen–Cahn	Burgers
TargetCover	0.2601 ± 0.0042	0.2112 ± 0.0054	0.6899 ± 0.0071	0.7667 ± 0.0093
Density + distance	$0.2714 \pm 0.0039^{\dagger}$	$0.2274 \pm 0.0050^{\dagger}$	$0.6965 \pm 0.0086^{\dagger}$	0.7703 ± 0.0091
Projected + RBF	$\mathbf{0.2596} \pm 0.0040$	$\mathbf{0.2105} \pm 0.0044$	$\mathbf{0.6894} \pm 0.0087$	$\mathbf{0.7648} \pm 0.0086$
Target LCMD	$0.2735 \pm 0.0040^{\dagger}$	$0.2191 \pm 0.0061^{\dagger}$	$0.6986 \pm 0.0078^{\dagger}$	$0.7765 \pm 0.0128^{\dagger}$

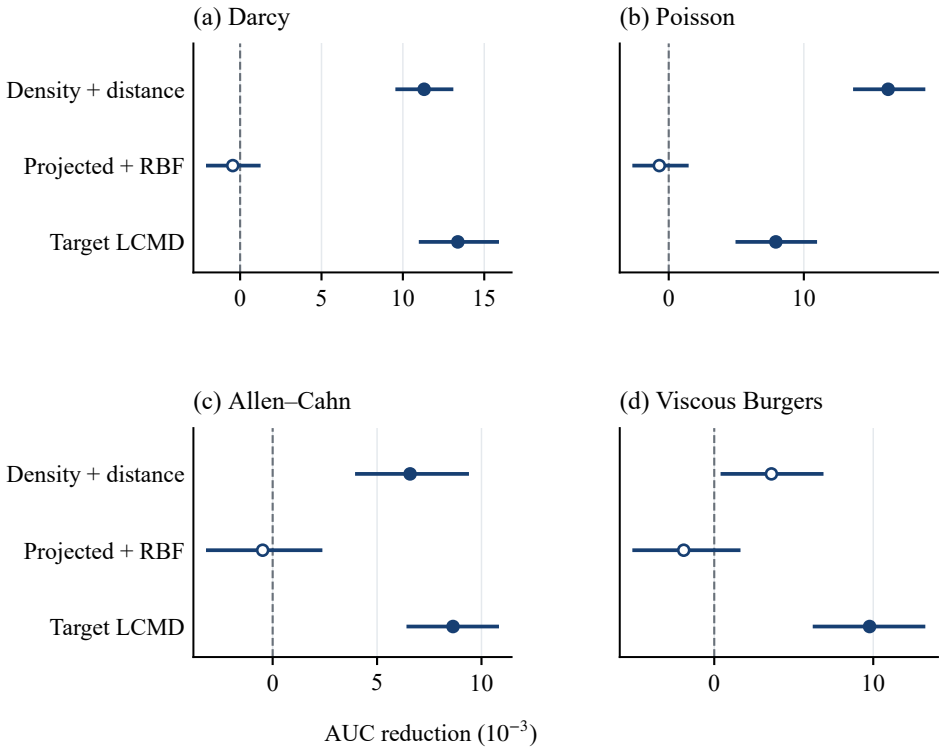

Figure 9: Weight and distance ablations. Positive reductions favour TargetCover; intervals are pointwise paired bootstrap 95% intervals over 20 trials. Filled markers indicate corrected significance in the separate eight-test crossed-control or four-test target-LCMD family. Transferred weights improve on density weights with squared distance fixed on Darcy, Poisson and Allen–Cahn. Projected–RBF and TargetCover differences are statistically unresolved.

Fixed-support reference-weight intervention. The reference-weight experiment uses 20 new trials on each of Darcy and Allen–Cahn. For every trial, the source and target pools each contain 1,024 fields, evaluation uses 256 held-out fields per law, and the complete sequence of label counts is $32, 54, 91, 152, 256$. Methods share input fields, initial labels, normalization and training randomization. Each DeepONet is trained for 200 epochs at each label count.

Let ξ_j be the standardized first source POD coordinate. We set $m_j = \exp(\eta_j)$, where $\eta_j = \mathrm{clip}(\alpha\xi_j, -30, 30)$ and $\alpha \in \{0, 2, 4\}$. These weights enter only the target-to-source transfer. The source support, POD basis, target sample and available labels are fixed. Globally normalized demand uses $\widetilde{w}_i(m) = w_i d_i(m) / \sum_r w_r d_r(m)$, with $d_i(m) = \sum_j m_j K_{ij}$. Row-normalized transfer retains w_i but sends it according to $m_j K_{ij}/d_i(m)$. Both use a fixed full-pool median bandwidth. We measure total variation $\frac{1}{2}\sum_i |\widetilde{w}_i(m) - w_i|$ and prediction AUC. Projected transfer is independent of the reference weights. Figure 10 therefore repeats one prediction curve per trial across all three tilts.

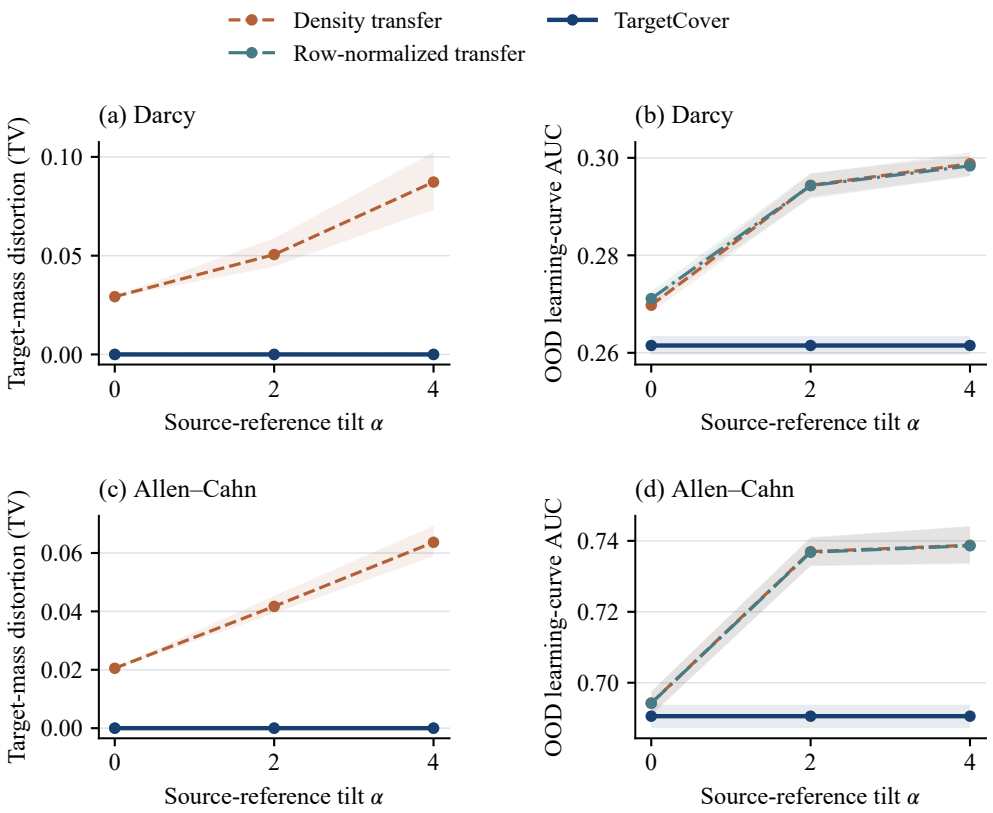

Figure 10: Reference-weight intervention for the fixed TargetCover component. Left: total variation of effective target probability. Right: OOD AUC as source-reference weights vary; bands are bootstrap 95% intervals over 20 trials. Source fields and deployment inputs remain fixed. TargetCover's curve repeats one result per trial because its selection does not depend on the reference weights.

Table 9: Reference-weight interventions and weight/distance ablations. Entries are OOD AUC (mean ± standard deviation over 20 trials). All rows use squared distance except those marked RBF. TargetCover is computed once per trial and is invariant to reference reweighting.

Transfer / coverage	Darcy	Allen–Cahn
TargetCover	0.2615 ± 0.0044	0.6906 ± 0.0078
Density, $\alpha = 0$	0.2698 ± 0.0035	0.6943 ± 0.0077
Row-normalized, $\alpha = 0$	0.2711 ± 0.0033	0.6942 ± 0.0079
Density + RBF, $\alpha = 0$	0.2690 ± 0.0045	0.6936 ± 0.0072
Projected + RBF	0.2610 ± 0.0042	0.6894 ± 0.0068
Density, $\alpha = 2$	0.2944 ± 0.0055	0.7370 ± 0.0096
Row-normalized, $\alpha = 2$	0.2943 ± 0.0062	0.7368 ± 0.0092
Density, $\alpha = 4$	0.2988 ± 0.0057	0.7388 ± 0.0119
Row-normalized, $\alpha = 4$	0.2984 ± 0.0051	0.7386 ± 0.0126

The experiment contains 320 control learning curves and 40 shared TargetCover reference curves. Six paired contrasts compare row-normalized with density transfer at the three tilts on two tasks. Four further contrasts compare TargetCover with density–distance at zero tilt and with projected–RBF. All ten were specified before these runs and share a Holm correction. Table 9 also completes the fixed-kernel 2×2 demand/cost comparison at zero tilt. The intervention varies reference weights on a fixed source support and deployment distribution.

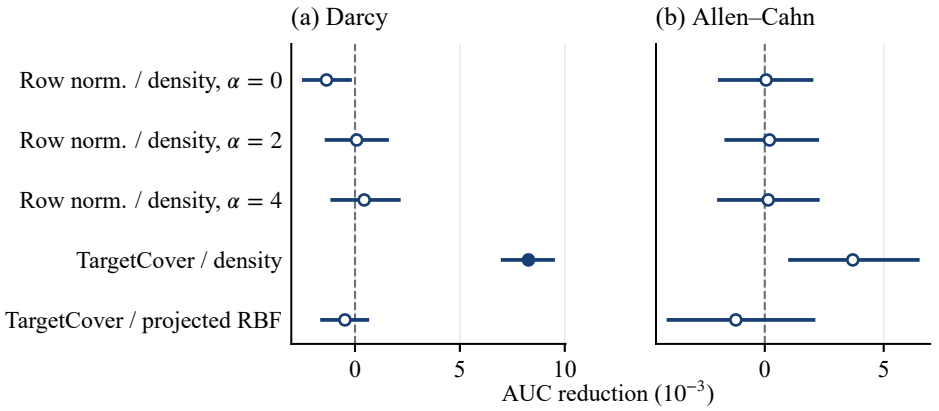

Figure 11: All ten pre-specified paired mechanism contrasts. The first three rows in each panel compare row-normalized transfer with globally normalized density transfer at the indicated tilt; the last two compare TargetCover with the stated control. Positive values favour the first-named method. Intervals are pointwise paired bootstrap 95% intervals; filled markers indicate Holm $p < 0.05$ across all ten tests. The six row-normalized versus density-transfer differences are statistically unresolved.

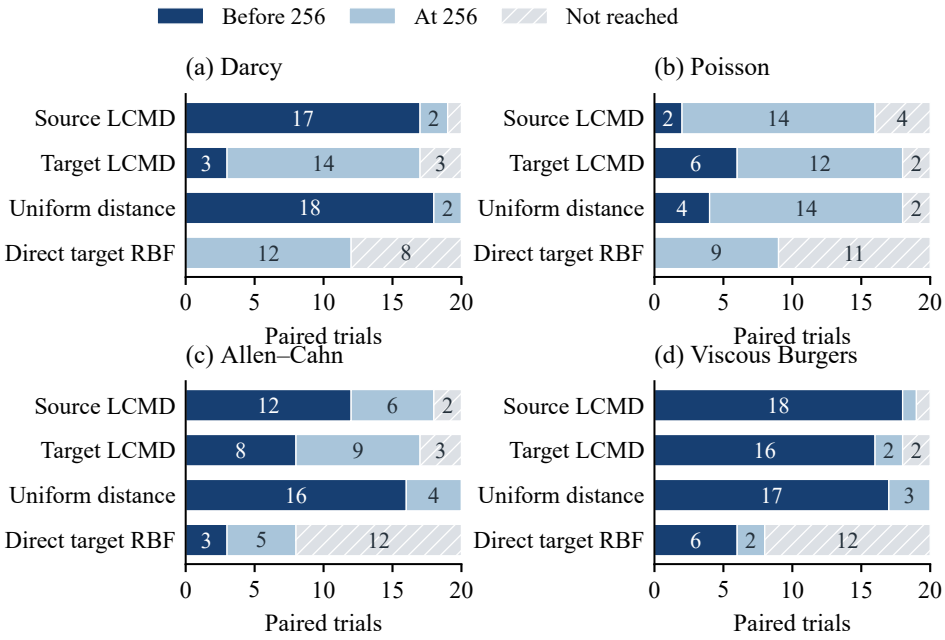

Figure 12: PDE solutions required to reach comparator accuracy. Each bar contains all 20 paired trials for one comparator. Blue counts trials where TargetCover reaches the comparator's 256-label error using strictly fewer labels; light blue counts first attainment at 256; hatched grey counts nonattainment by 256. Counts appear in segments wide enough to hold them. The direct-target comparator is included alongside both LCMD variants and uniform-distance coverage.

Label efficiency and nonlinear predictions. For each paired trial, the comparator's error at 256 labels defines the threshold. We record the first TargetCover count on $\{32, 54, 91, 152, 256\}$ reaching it. Figure 12 places every trial in one of three categories: earlier attainment, attainment at 256, or nonattainment. The statistic measures first crossing on the observed learning curve. Target LCMD uses the supplementary reference; the other comparisons use the coverage-component cohort. This

post-hoc endpoint is comparator-relative and uses no interpolation. Restricted label costs in the numerical records cap nonattainment at 256; they summarize label use within the observed training sizes. For an unreached threshold, the total number of solutions needed remains unknown.

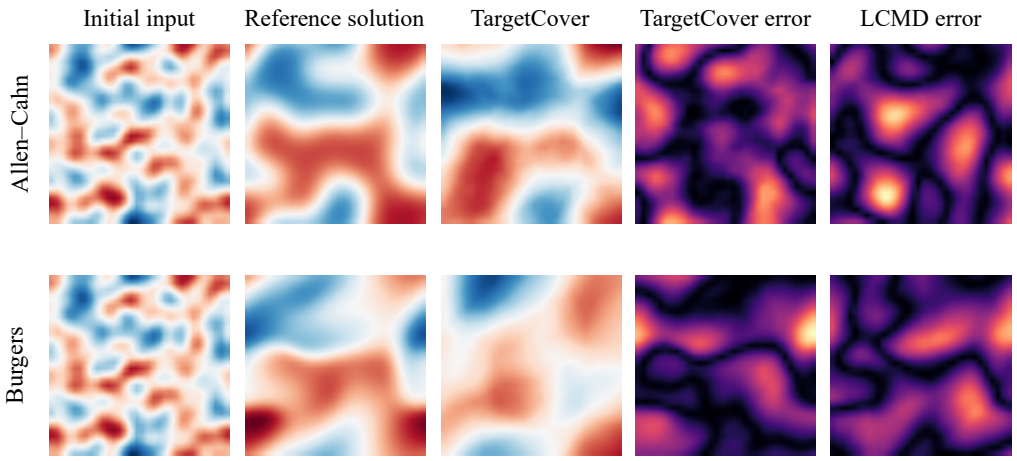

Figure 13: Nonlinear terminal-state prediction at 256 source labels. The first held-out input from the first coverage-component trial is shown for Allen–Cahn and scalar 2D Burgers. The initial input has its own colour scale; reference and prediction share a second scale within each row, while the two absolute-error maps share a nonnegative row scale. The sampling unit is a complete trial. Solver refinement and reference accuracy are assessed in Appendix B.

For the displayed fields, the pointwise error maps also permit a direct numerical check. With reference field u and prediction $\hat{u}$ on the 32×32 plotting grid, define

$$E_{\mathrm{rel}}(\hat{u}, u) = \frac{\left(\sum_{i,j} |\hat{u}_{ij} - u_{ij}|^2\right)^{1/2}}{\left(\sum_{i,j} |u_{ij}|^2\right)^{1/2}}. \tag{25}$$

The values computed from the two error maps in each row of Figure 13 are

Displayed field	TargetCover	Source LCMD
Allen–Cahn	0.757	0.732
Viscous Burgers	0.840	0.891

The ordering changes between the two examples. These numbers describe one held-out field per PDE and are not additional trial averages; method comparisons in the main text use the paired trials. Here the maps and these fieldwise errors serve to interpret the particular solutions shown below.

Figure 14 expands the two held-out terminal-state examples in Figure 13. The upper plate shows Allen–Cahn and the lower plate shows Burgers, using the first held-out input from the first coverage-component trial at 256 labels. These are visual examples rather than additional evaluation samples.

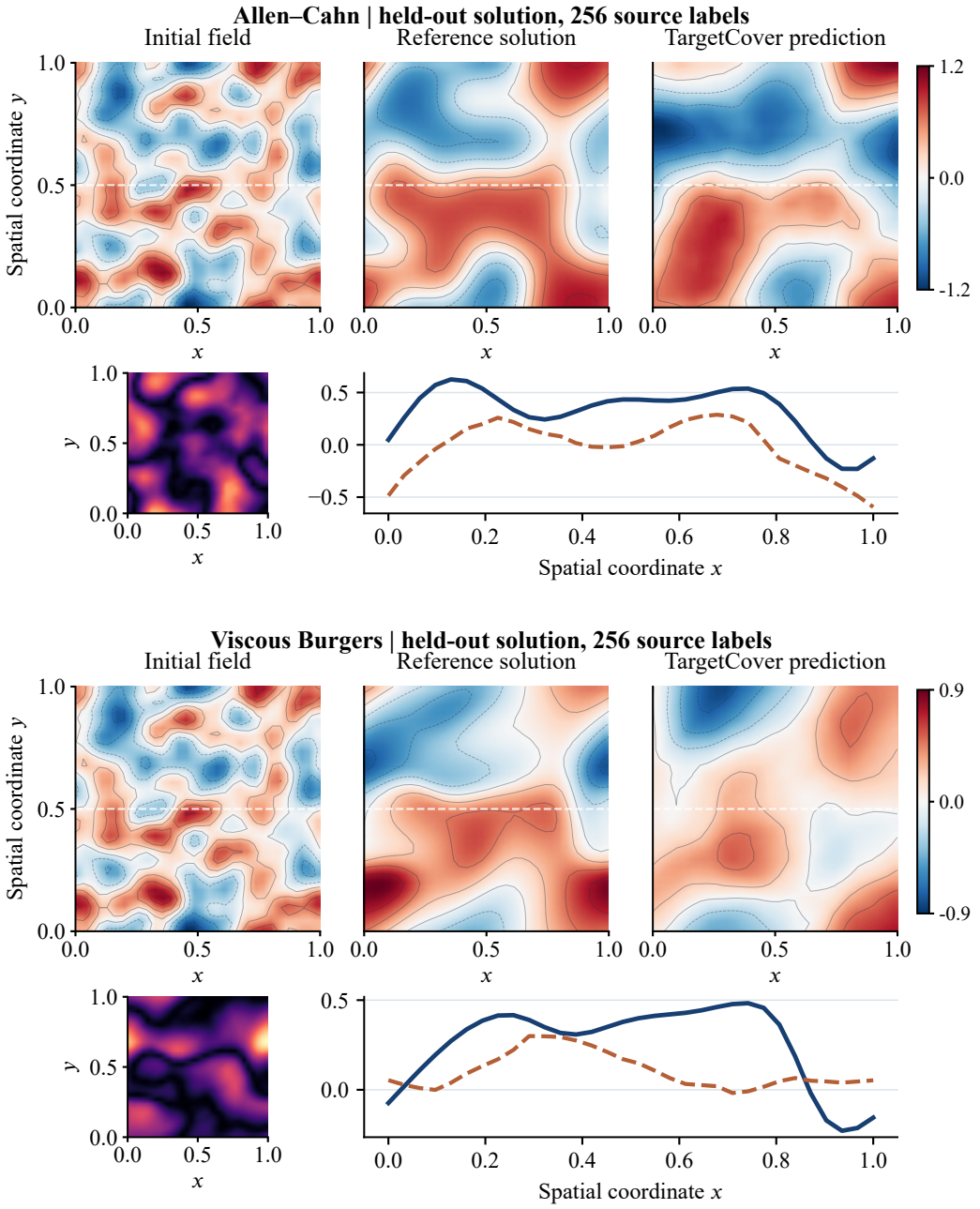

Figure 14: Held-out terminal-state solutions for Allen–Cahn (upper) and viscous Burgers (lower). Each plate shows the initial field, reference, TargetCover prediction, absolute error and horizontal centreline. Reference and prediction share a symmetric scale within each PDE; the input and nonnegative error use separate scales. The input range is ± 2.85; solution/error ranges are $\pm 1.16/[0, 1.04]$ and $\pm 0.87/[0, 0.96]$, respectively. Blue solid and orange dashed centreline curves denote reference and TargetCover; white dashed lines locate the centreline in the field maps.

D Coverage-component architecture and training checks

The additional experiment contains 480 learning curves: two PDEs, three architecture/shift settings, four acquisition methods and 20 paired input trials. Each curve uses $32, 54, 91, 152, 256$ source labels, with 1,024 source candidates, 1,024 unlabelled target inputs and 256 evaluation fields per law.

The roughness setting halves the source correlation length at fixed input variance. The amplitude setting multiplies input variance by four while keeping the correlation length and PDE parameters fixed.

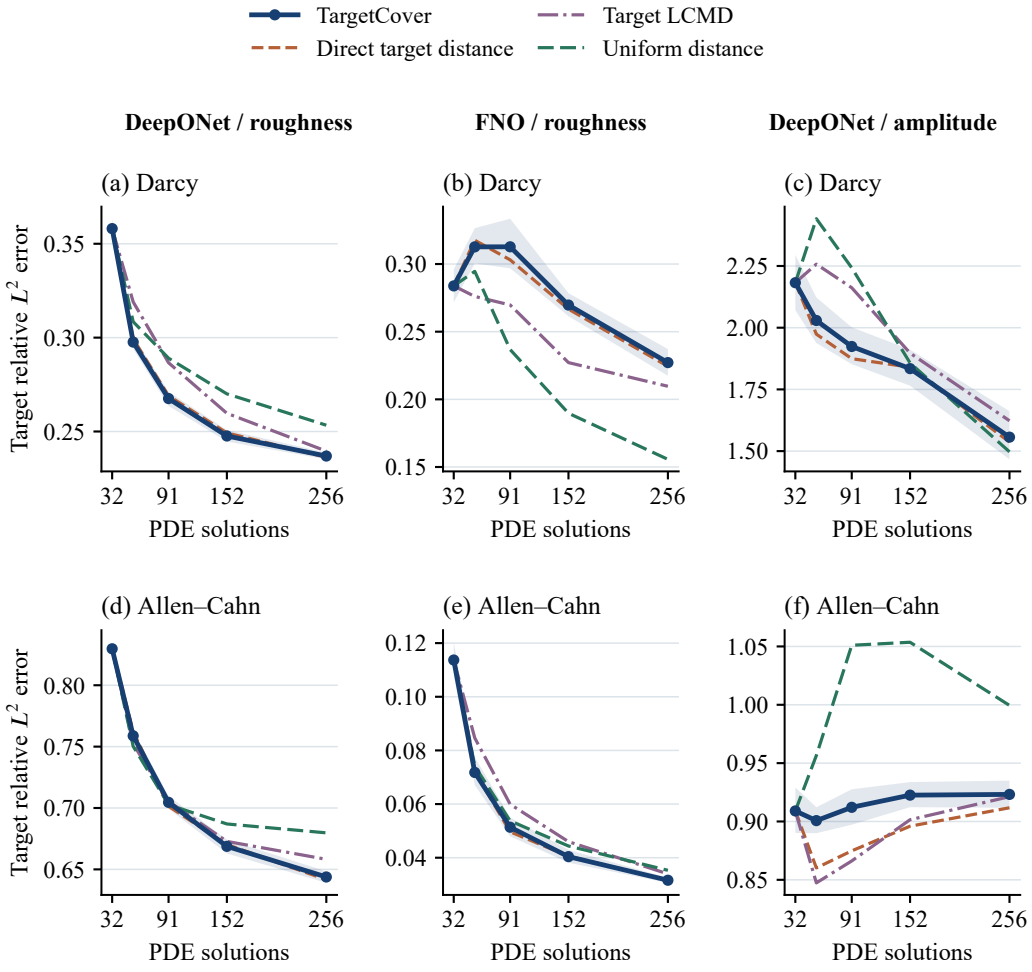

Figure 15: Full learning curves across architectures and shift types. All four methods appear in every setting; curves average 20 paired trials and shading gives pointwise bootstrap 95% intervals. TargetCover and direct distance track closely under roughness shift. FNO attains substantially smaller Allen–Cahn errors than DeepONet, while the Darcy acquisition ordering changes with architecture. Amplitude-shift panels retain errors above one. Axis ranges differ between panels.

All four methods use the same source-fitted 16-dimensional POD representation. Direct target distance greedily minimizes C_Q in that representation with facilities restricted to the source catalogue. Uniform distance and TargetCover minimize the source-site objective with different demand weights; target LCMD uses target-weighted source clusters. Methods share input fields, initial labels, normalization and training randomization within each trial. These comparisons isolate selection within each architecture.

DeepONet uses the original coverage-study architecture and training settings. FNO has four Fourier layers, width 32, twelve modes per spatial direction and Gaussian error linear unit (GELU) activations. Both architectures receive 200 epochs of Adam training at each label count, with learning rate 10^{-3} and batch size 32. The input-only selections are unchanged by the architecture. FNO is evaluated under roughness shift; DeepONet is evaluated under both roughness and amplitude shift.

The eighteen pre-specified comparisons pair TargetCover with each of the three alternatives in every task/setting. Two-sided sign-randomization tests share a Holm correction over all eighteen contrasts. The methods' individual standard deviations in Table 10 describe trial variability; significance is computed from paired AUC differences.

For DeepONet roughness shift, TargetCover improves on target LCMD and uniform distance on Darcy, and on uniform distance on Allen–Cahn ($p_{\mathrm{Holm}} = 0.00036$). Its target-LCMD difference on Allen–Cahn has $p_{\mathrm{Holm}} = 0.18828$. The two direct-distance comparisons have $p_{\mathrm{Holm}} = 1$.

Table 10: Architecture and input-shift comparisons over 20 paired trials. Each upper row gives target AUC (mean $\pm$ standard deviation); the lower row gives mean relative L^2 error at 256 labels. Bold marks the smallest mean in each row. All four methods use squared-distance geometry and source-only queries.

Setting	TargetCover	Direct distance	Target LCMD	Uniform distance
DeepONet / roughness				
Darcy AUC	$\mathbf{0.2615} \pm 0.0044$	0.2625 ± 0.0039	0.2735 ± 0.0041	0.2797 ± 0.0035
256 labels	0.2369	**0.2363**	0.2395	0.2533
Allen–Cahn AUC	0.6906 ± 0.0078	$\mathbf{0.6903} \pm 0.0065$	0.6947 ± 0.0085	0.7041 ± 0.0086
256 labels	0.6438	**0.6416**	0.6582	0.6798
FNO / roughness				
Darcy AUC	0.2756 ± 0.0170	0.2724 ± 0.0132	0.2416 ± 0.0163	$\mathbf{0.2105} \pm 0.0183$
256 labels	0.2272	0.2244	0.2097	**0.1556**
Allen–Cahn AUC	0.0485 ± 0.0036	$\mathbf{0.0482} \pm 0.0033$	0.0547 ± 0.0040	0.0516 ± 0.0038
256 labels	**0.0316**	0.0316	0.0340	0.0353
DeepONet / amplitude				
Darcy AUC	1.8322 ± 0.1815	$\mathbf{1.8116} \pm 0.1714$	1.9529 ± 0.2014	1.9509 ± 0.2101
256 labels	1.5566	1.5369	1.6230	**1.4972**
Allen–Cahn AUC	0.9169 ± 0.0214	$\mathbf{0.8909} \pm 0.0197$	0.8916 ± 0.0189	1.0206 ± 0.0240
256 labels	0.9232	**0.9117**	0.9212	0.9995

With FNO, TargetCover improves on target LCMD and uniform distance on Allen–Cahn; both orderings reverse on Darcy ($p_{\mathrm{Holm}} = 0.00036$ for all four contrasts). Both direct-distance comparisons have $p_{\mathrm{Holm}} = 1$.

Under amplitude shift, TargetCover improves on target LCMD and uniform distance on Darcy and on uniform distance on Allen–Cahn ($p_{\mathrm{Holm}} = 0.00036$). Direct distance and target LCMD improve on TargetCover on Allen–Cahn, with the same corrected value. The remaining direct-distance comparison on Darcy has $p_{\mathrm{Holm}} = 0.18828$. These are the complete eighteen contrasts.

The direct comparison shows how the target-to-source representation affects prediction at a fixed distance cost. The architecture comparison addresses a different question: how well the trained operator uses the selected fields. On Allen–Cahn, the 256-label TargetCover error is 0.6438 with DeepONet and 0.0316 with FNO. Thus the large DeepONet error does not reflect an equivalent accuracy floor for the selected training inputs.

Larger training sets and longer optimization. Twenty additional paired input trials on each of Darcy and Allen–Cahn use 2,048 source and 2,048 target fields, with $32, 64, 128, 256, 512, 1024$ source labels. TargetCover, direct target distance and uniform distance are each trained for 200 and 1,000 epochs, yielding 240 complete DeepONet learning curves. Architectures, learning rate, normalization and input laws are unchanged. The selected sets are identical between training durations.

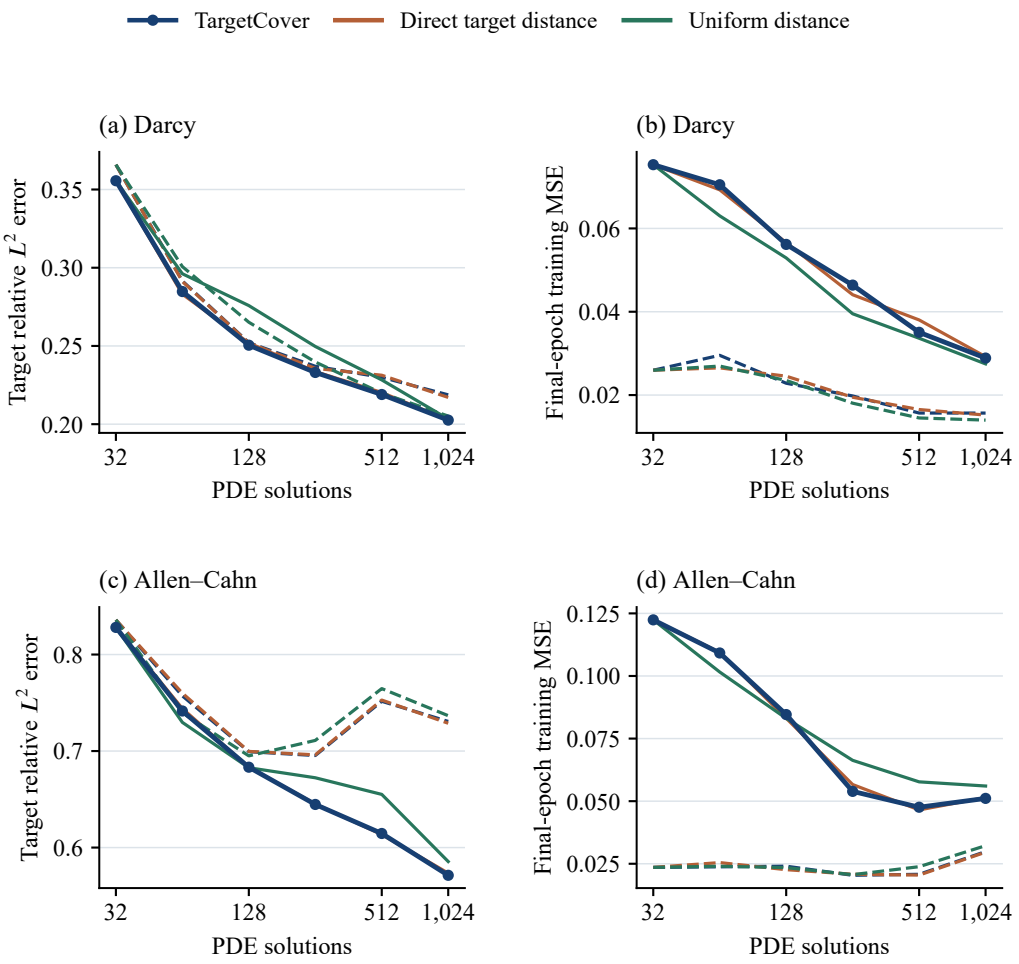

Figure 16: Training-set size and optimization duration have different effects. Solid lines use 200 epochs and dashed lines use 1,000. Every curve averages 20 paired trials; shading gives pointwise bootstrap 95% intervals for TargetCover. Right panels show normalized final-epoch minibatch MSE, not validation error. Increasing duration improves source fitting but increases target prediction error. All three selectors and both durations are shown.

At 1,024 labels and 200 epochs, TargetCover reaches mean target relative errors 0.2026 on Darcy and 0.5713 on Allen–Cahn. At 1,000 epochs the respective errors are 0.2187 and 0.7307, while training MSE falls from 0.02888 to 0.01567 and from 0.05111 to 0.02986. Direct distance and uniform distance show the same separation between fitting improvement and target error. These measurements distinguish optimization duration from predictive accuracy under shift.

Fixed target-error thresholds were specified before this experiment: $0.10, 0.20, 0.25$ for Darcy and $0.05, 0.10, 0.30, 0.50$ for Allen–Cahn. In each Darcy method/duration group, all 20 trials reach 0.25 and none reaches 0.10. At threshold 0.20, the counts for TargetCover, direct distance and uniform distance are $4, 3, 5$ at 200 epochs and $0, 0, 2$ at 1,000 epochs. None of the Allen–Cahn trials reaches the four fixed thresholds. Attainment means first crossing on the six observed label counts; unsuccessful trials remain in the totals.