When Does Diversity Help? Realizable Complementarity in Neural Operator Ensembles

Anonymous authors
Paper under double-blind review

Abstract

Neural operators are commonly developed and evaluated as individual predictors, with improvements largely pursued through architectural advances. Yet our experiments reveal a surprising result: a simple average of independently trained models can deliver substantial accuracy gains. Across six PDE tasks, simple equal-weight averaging reduces error by up to 18.46%, and the effect persists across different model families and controlled settings. However, the benefit varies sharply across expert pools, revealing that predictive diversity alone does not imply useful aggregation. We show that what matters is whether differences among models produce complementary errors that actually improve the combined prediction. This motivates the notion of *useful complementarity*. We further uncover a second gap: useful structure may exist within an expert pool without being reliably exploitable from finite development data. Increasing the flexibility of the combination rule does not necessarily translate into larger held-out gains, indicating a distinction between complementarity that is available and complementarity that can actually be realized. We call the latter *realizable complementarity*. Finally, our results show that information useful for constructing a strong aggregate need not remain equally useful for refining it afterwards. Together, these findings provide a framework for understanding when diversity among neural operators translates into reliable accuracy gains.

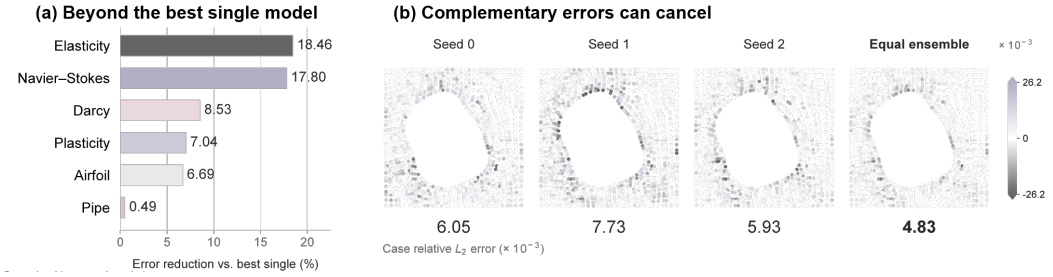

Figure 1: **Simple averaging recovers accuracy beyond the best individual neural operator.** **(a)** Error reduction from equal-weight averaging of three independently trained LinearNO models across six PDE tasks, using the best individual model in each pool as the reference. **(b)** Prediction errors of the three models and their ensemble on an illustrative Elasticity case, showing partial error cancellation after averaging. Colours represent signed normalized prediction error with a shared symmetric scale. Numbers are case-relative L_2 errors in units of 10^{-3}.

1 Introduction

Neural operator learning aims to predict the solution fields of partial differential equations (PDEs) from physical inputs such as coefficient fields, problem geometries, or histories of physical states (Kovachki et al., 2023). Trained on paired input–solution data, these models provide reusable surrogates for predicting spatial fields or their temporal evolution on new instances of a given task. Improving prediction accuracy has motivated architectural advances, including LinearNO (Hu et al., 2026) and low-rank spatial attention (LRSA) (Yang et al., 2026a).

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.

Yet our experiments reveal a simple but strong source of additional accuracy: *averaging independently trained neural operators can substantially outperform even the best individual model*. With three independently trained LinearNO models and fixed equal weights, error decreases by 18.46% on Elasticity and 17.80% on Navier–Stokes relative to the best model in each pool (See Fig. 1a). No joint training, learned aggregation weights, or input-dependent routing is required. Fig. 1b illustrates how prediction errors from different models can partially cancel after averaging. These results reveal substantial *recoverable accuracy* beyond what any individual predictor achieves.

The benefit, however, varies markedly across tasks. Under the same three-model LinearNO setting, the gain ranges from 18.46% on Elasticity to only 0.49% on Pipe. Substantial averaging gains also appear with LRSA (Yang et al., 2026a) on Elasticity, suggesting that the phenomenon is not specific to a single neural-operator architecture. The large variation across expert pools nevertheless shows that simply having different predictions is not enough. This motivates the central question of this work: *when does predictive diversity become useful complementarity?*

Predictive ensembling is well established (Lakshminarayanan et al., 2017), and prior work has explored neural-operator ensembles, explicit diversity construction, and uncertainty-aware improvement (Mouli et al., 2024). Here, we identify independent-model ensembling itself as a strong accuracy resource for neural operators and study why its effectiveness varies so widely. A controlled Darcy comparison in Sec. 3.2 further shows that the gain is not reproduced by the tested single-model scaling strategies at comparable training compute.

Our analyses show that disagreement alone is not sufficient. Averaging is useful when cancellable errors outweigh differences in member quality and when the resulting gains occur on cases that matter to the overall risk. A small number of difficult cases can therefore offset improvements across many others. Cross-family and explicit-diversity controls lead to the same conclusion: greater predictive variation does not necessarily yield greater accuracy. We refer to differences that actually improve the combined prediction as *useful complementarity*.

Useful complementarity, however, need not be reliably exploitable from finite data. Better combinations can exist within a fixed expert pool without being identified by a learned procedure, and greater combination flexibility does not necessarily improve held-out performance. We use *realizable complementarity* to describe the held-out gain that a specified procedure can recover from finite development data. Moreover, information that is useful for constructing an aggregate need not remain equally useful for improving it afterwards.

Our contributions are threefold:

- **Recoverable accuracy through simple ensembling.** We establish independent-model averaging as a strong accuracy resource for neural operators across multiple PDE tasks and backbones, beyond what can be explained by increased compute or model size.
- **From diversity to useful complementarity.** We characterize when predictive differences translate into aggregation gains, showing that useful complementarity depends on error cancellation, member quality, and where the remaining risk is concentrated.
- **From useful to realizable complementarity.** We distinguish complementarity available in an expert pool from what can be reliably recovered with finite development data, and show that the value of expert-specific information can change after a strong aggregate has been formed.

2 Related Work

2.1 Neural Operators

Neural operators learn mappings between function spaces and provide a general framework for learning solution operators across families of PDE problems (Kovachki et al., 2023). Representative approaches include DeepONet (Lu et al., 2021), graph- and spectral-based neural operators (Li et al., 2020b;a; Gupta et al., 2021), and architectures designed for multiscale representations and general geometries (Li et al., 2023; Rahman et al., 2022). Transformer and attention mechanisms have further expanded this family, including Galerkin-style attention (Cao, 2021), GNOT (Hao et al., 2023), and Physics-Attention in Transolver (Wu et al., 2024). More recent architectures continue to improve

the scalability and representation of physical interactions: Transolver++ extends Physics-Attention to large-scale physical systems (Luo et al., 2025), LinearNO revisits it through the lens of linear attention (Hu et al., 2026), LRSA introduces low-rank spatial attention (Yang et al., 2026a), and MoNo develops multiscale latent representations through optimal-transport-based assignments (Yang et al., 2026b). Our study complements this line of architectural development by asking a different question: how much additional accuracy can be recovered from independently trained, frozen neural operators without modifying their architectures or training objectives?

2.2 Ensembles and Predictive Diversity

Deep ensembles are a standard approach to improving predictive accuracy and uncertainty estimation (Lakshminarayanan et al., 2017), and independent initializations can lead to functionally distinct solutions (Fort et al., 2019). For PDEs, Mouli et al. (2024) study ensemble uncertainty, shared-backbone multi-head DiverseNO, and uncertainty-informed Operator-ProbConserv under distribution shift. Probabilistic neural operators instead learn distributions over output functions using proper scoring rules (Bülte et al., 2025). Our focus is different: we study when differences among deterministic operator predictions translate into accuracy gains, and how much of that complementarity can be realized from finite development data. Our LinearNO-adapted DiverseNO control isolates one explicit diversity-construction mechanism rather than reproducing the full DiverseNO/Operator-ProbConserv pipeline.

2.3 Refinement after Prediction

Iterative refinement has also been used to improve neural PDE predictions, including diffusion-inspired multi-step refinement for long-horizon forecasting (Lippe et al., 2023). IRNO learns iterative updates to pretrained neural operators and interprets them as learned fixed-point solvers (Liu et al., 2026). Whereas these methods ask whether an existing prediction can be refined, we additionally ask whether such refinement benefits specifically from access to the individual expert predictions. Our matched centre-only and expert-aware controls separate correctable residual error from the incremental value of expert information.

3 Recoverable Accuracy Beyond Individual Operators

We first establish the empirical effect underlying our study. For a fixed pool of M independently trained neural operators $\mathcal{F} = \{f_1, \ldots, f_M\}$, which we call an *expert pool*, we use the equal-weight prediction $\bar{f}(x) = M^{-1} \sum_{m=1}^{M} f_m(x)$, without learned combination weights or routing rules.

3.1 Simple Averaging Reveals Recoverable Accuracy

Tab. 1 evaluates this rule on six fixed three-model LinearNO pools. For each task, the equal average is compared with the best individual model in the same pool on the same evaluation cases. Equal averaging yields a lower error point estimate on all six tasks, with reductions ranging from 0.49% on Pipe to 18.46% on Elasticity. The gains are substantial on several tasks, while Pipe provides a contrasting regime in which the improvement is marginal. Thus, even fixed equal weights can recover accuracy that is not achieved by any individual member of the pool.

Table 1: **Performance of simple equal-weight averaging across six PDE tasks.** The best individual is the member with the lowest overall evaluation error. Errors are reported in units of 10^{-3}.

	Elasticity	Navier–Stokes	Darcy	Plasticity	Airfoil	Pipe
Best individual error ↓	5.244	70.828	4.862	1.036	4.955	2.634
3-model equal-average error ↓	**4.277**	**58.221**	**4.447**	**0.963**	**4.623**	**2.621**
Error reduction vs. best individual ↑ (%)	**18.46**	**17.80**	**8.53**	**7.04**	**6.69**	**0.49**

Table 2: **Compute-controlled Darcy comparison.** All configurations use LinearNO and approximately matched counted training FLOPs. Ensemble parameters and FLOPs are summed across members.

Predictor	Params (M)	Train PFLOPs	Rel. L_2 ($\times 10^{-3}$)
Longer training	1.77	131.08	5.585
Wider model	5.37	130.89	5.774
Larger latent size	1.88	131.04	5.178
Three-model equal mean	5.30	131.08	**4.428**

3.2 Beyond One Backbone and Compute Controls

The averaging effect is not confined to the original three-model LinearNO pools. On Elasticity, averaging seven independently trained LRSA models reduces error by 23.80% relative to the best LRSA member, while a separate eight-member LinearNO pool yields a 26.84% reduction. These results show that substantial averaging gains persist across backbones and larger pool sizes, although the two pools are not intended as a compute-matched architecture comparison.

We also test whether the gain can be explained simply by increased compute or model size. In a separate Darcy study using LinearNO throughout, a three-model ensemble is compared with single LinearNO models trained longer, made wider, or given a larger latent dimension under approximately matched training FLOPs. As shown in Tab. 2, the ensemble achieves the lowest error among the tested configurations, while the width-scaled model provides an additional control with a similar parameter count. Within these controls, the averaging advantage therefore cannot be explained simply by increased training compute or model size.

Taken together, these results establish independent-model averaging as a strong source of recoverable accuracy across the settings studied. Its magnitude, however, varies sharply across expert pools. We next ask what distinguishes predictive differences that merely create diversity from those that become *useful complementarity*.

4 When Does Predictive Diversity Become Useful Complementarity?

Sec. 3 shows that simple averaging can recover substantial accuracy, but that the magnitude of the gain varies sharply across expert pools. We now ask what makes differences among model predictions useful for aggregation.

4.1 Cancellation Must Overcome the Quality Gap

The first requirement for useful diversity is that differences among model predictions actually lead to error cancellation. Two expert pools may contain similarly different predictions, yet averaging helps only when their errors offset one another rather than simply reflecting weaker members.

To quantify this effect, for case i we use the relative error $\ell_i(f) = \|f(a_i) - u_i\|_2 / \|u_i\|_2$ and define the normalized squared risk as $R(f) = |\mathcal{D}|^{-1} \sum_i \ell_i(f)^2$. Let R_{avg} denote the average risk of the individual members, R_{best} the risk of the best member, and R_{eq} the risk of their equal average. The gain over the best individual can then be written as

$$R_{\mathrm{best}} - R_{\mathrm{eq}} = \underbrace{R_{\mathrm{avg}} - R_{\mathrm{eq}}}_{V\text{: cancellable error}} - \underbrace{R_{\mathrm{avg}} - R_{\mathrm{best}}}_{Q\text{: quality gap}}. \tag{1}$$

Here, V measures how much average member risk is removed by averaging, whereas Q measures how much better the strongest member is than an average member. Averaging improves on the strongest individual under this squared-risk diagnostic when $V > Q$. We use this standard identity as a diagnostic rather than as a new theoretical result.

This identity concerns normalized squared risk rather than mean relative L_2. The member minimizing squared risk need not be the member minimizing mean relative L_2. Moreover, the normalized

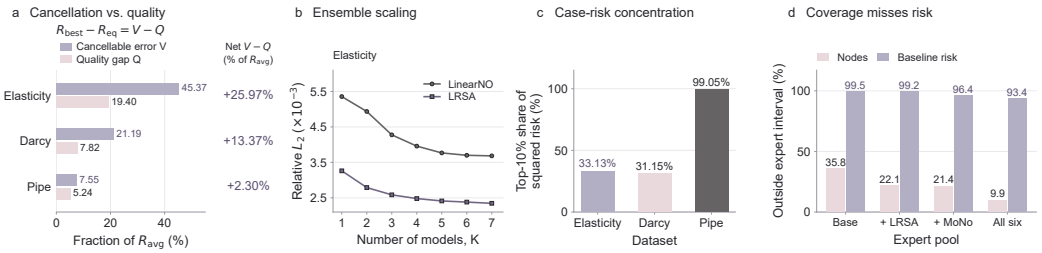

Figure 2: **Useful complementarity depends on more than predictive diversity. (a)** Cancellable error V and quality gap Q, each normalized by average-member risk R_{avg}. The annotated net margin is $(V - Q)/R_{\text{avg}}$, not the relative-L_2 gain reported in Table 1. **(b)** Fixed-order Elasticity ensemble curves for independently trained LinearNO and LRSA models, with $K = 1, \ldots, 7$. **(c)** Fraction of the equal ensemble's normalized squared risk contributed by its highest-error 10% of cases. **(d)** Fraction of Pipe nodes whose targets lie outside the expert prediction interval and the share of the original LinearNO mean's squared risk located at those nodes. After expansion to six experts, uncovered nodes still account for 93.38% of that original risk. This measures the location of baseline risk, not the remaining error of an optimized combination.

margin $(V - Q)/R_{\text{avg}}$ uses average-member risk as its denominator. It therefore differs from the percentage error reduction relative to the best individual reported in Table 1. The diagnostic uses observed targets and is not a label-free rule for predicting ensemble performance on a new task.

Fig. 2a illustrates three representative regimes. The normalized margin $(V - Q)/R_{\text{avg}}$ is 25.97% on Elasticity, 13.37% on Darcy, and 2.30% on Pipe. In these pools, the larger averaging gains coincide with a larger cancellable component relative to the member-quality gap. The comparison illustrates why prediction differences alone do not guarantee an improvement over the best member.

4.2 More Diversity Does Not Necessarily Help

Increasing the number of models and explicitly encouraging disagreement are two natural ways to increase predictive diversity. Fig. 2b shows fixed-order Elasticity ensemble curves for independently trained LinearNO and LRSA models. Both backbones benefit substantially from the first few members, while later additions yield diminishing gains. Because these curves follow one fixed ordering rather than averaging over all subsets, they show diminishing returns within the tested pools rather than a fixed benefit from every additional model.

We further test explicit diversity using a LinearNO-adapted DiverseNO (Mouli et al., 2024). On Pipe, the validation-selected $\lambda = 10^{-3}$ increases prediction disagreement but yields error 2.648×10^{-3}, compared with 2.390×10^{-3} for the matched $\lambda = 0$ control. On Elasticity, validation selects $\lambda = 0$. Thus, increased disagreement does not by itself guarantee improved aggregation in the tested adaptation. This control isolates the diversity mechanism of DiverseNO rather than reproducing the full Operator-ProbConserv pipeline.

4.3 Shared High-Risk Errors Limit Useful Complementarity

Useful complementarity also depends on *where* cancellation occurs. An ensemble may improve many cases while gaining little overall if the remaining errors are concentrated in a small difficult subset. Fig. 2c shows that the highest-error 10% of cases contribute 33.13% and 31.15% of the equal-ensemble squared risk on Elasticity and Darcy, respectively, but 99.05% on Pipe. Indeed, the three-model Pipe average improves on its fixed best member for 177 of 200 cases, yet the overall mean relative-L_2 gain is only 0.49%. Thus, improvements across most cases can be largely offset by errors on a small high-risk tail.

Pipe further reveals that these difficult errors are spatially shared. The remaining risk is strongly concentrated near the downstream end of the domain, where different model families exhibit similar deviations (Fig. 3). Adding experts substantially improves pointwise coverage: the fraction of locations whose targets lie outside the expert prediction interval falls from 35.76% with three LinearNO

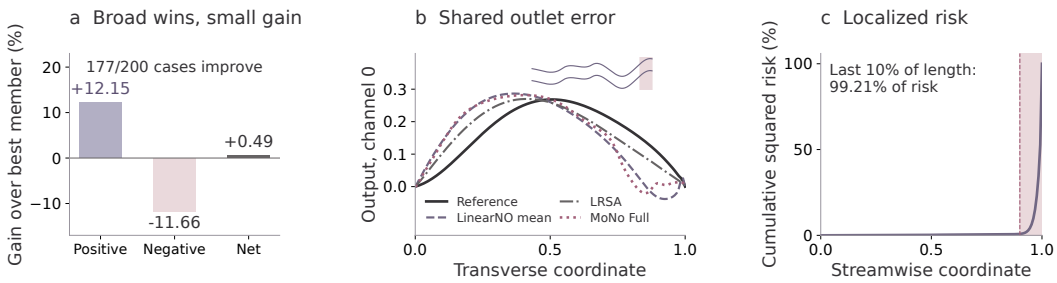

Figure 3: **Shared errors on difficult Pipe cases limit the benefit of aggregation.** **(a)** Positive and negative case-wise contributions nearly cancel, leaving the 0.49% net gain of the original three-LinearNO mean. **(b)** Cross-family prediction profiles on a fixed difficult case show similar downstream deviations. **(c)** The equal ensemble's squared risk is strongly concentrated near the downstream end of the domain. More details can be found in Appendices B and D.

members to 9.86% with all six experts. However, those remaining uncovered locations still account for 93.38% of the original LinearNO mean's squared risk (Figure 2d). Thus, broader coverage does not necessarily resolve the errors that dominate overall performance.

Together, these results show that useful complementarity depends not only on how much models differ, but on whether their differences reduce the errors that dominate the evaluation risk. This motivates the next question: how much of the available complementarity can be reliably realized from finite data?

5 How Much Complementarity Can Be Realized from Finite Data?

The previous section identifies predictive differences that can make aggregation useful. We now ask how much of this complementary structure can actually be recovered from a limited number of labelled development cases. We use *realizable complementarity* to describe the held-out gain that a specified combination or correction procedure can recover from a fixed expert pool using finite development data, relative to a stated reference. This distinguishes the existence of better predictions within the pool from our ability to identify and exploit them reliably.

5.1 Potential and Realization Are Different

We first ask whether a more flexible combination rule can recover more of the complementary structure available in a fixed expert pool. We compare four levels of spatial flexibility. The Global rule predicts one input-dependent weight vector for each case and shares it across all spatial points. The regional rules predict one weight vector for each cell of a 2×2 or 4×4 partition. The Point rule allows weights to vary at every point. These rules differ from the fixed convex reference used below. That reference fits a single weight vector on the fitting cases and applies the same weights to every case. Thus, Global is spatially uniform within a case, whereas the fixed convex reference is also shared across cases. As shown in Fig. 4a, greater flexibility does not translate monotonically into better held-out performance. On Elasticity, the learned gain increases gradually with finer combinations, while on Pipe it peaks at an intermediate granularity and decreases again for the pointwise rule.

The modest learned gains do not imply that the expert pools contain little complementary information. Fig. 4b compares the best learned combinations with a target-aware pointwise oracle. The oracle achieves gains of 64.6% on Elasticity and 30.4% on Pipe, compared with only 2.8% and 3.4% for the best learned combinations. Because the oracle uses the evaluation targets, it is not deployable; rather, it reveals the combination potential already present in the fixed expert predictions. The large gap therefore shows that available complementarity can far exceed what the finite-data learner actually recovers.

A strong reference further changes the apparent benefit of learning a more elaborate combination. Relative to a fixed convex combination, the validation-selected strategy adds only 0.54% on Elastic-

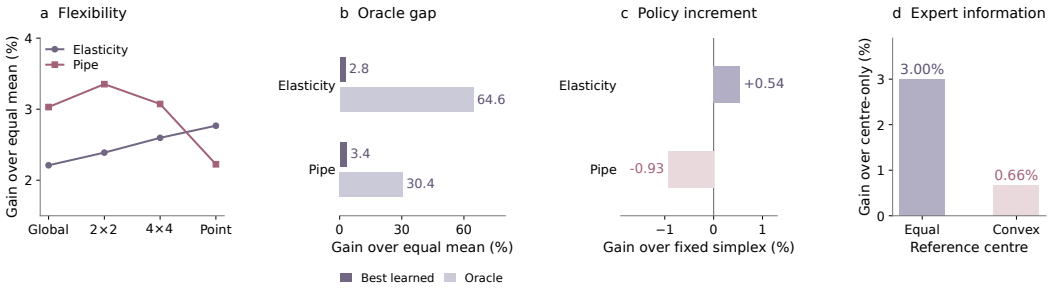

Figure 4: **Available complementarity can substantially exceed what is realized from finite data.** **(a)** Held-out gain over equal averaging as the spatial flexibility of the combination rule increases. **(b)** Best learned gain compared with a nondeployable target-aware pointwise oracle. **(c)** Increment of a validation-selected combination strategy over a strong fixed convex combination. **(d)** Incremental gain of expert-aware over centre-only correction after an equal-average centre and a stronger fixed convex centre.

Table 3: **Cross-case transfer of learned reweighting.** Values report the additional held-out gain over equal averaging (%); higher is better.

Task	Random holdout	Input-statistic holdout
Elasticity	**+2.83**	**+2.82**
Darcy	-0.11	+0.11
Pipe	-2.82	-0.79

ity and is 0.93% worse on Pipe (Fig. 4c). Thus, gains measured relative to equal averaging should not automatically be attributed to additional combination flexibility: a strong fixed rule can already capture much of the finite-data benefit. Because the candidate strategies also differ in how they are learned, we interpret this comparison as the increment of strategy selection over a strong fixed reference, rather than as an isolated causal effect of input adaptivity.

A learned combination must also remain useful beyond the cases used to estimate it. We therefore perform a separate five-fold cross-case analysis using frozen three-model pools on Elasticity, Darcy, and Pipe. Combination weights are estimated on one subset of cases, calibrated toward equal averaging on a second subset, and evaluated on held-out cases. We consider both random case splits and splits along an input-statistic axis.

The three tasks reveal different realizability regimes. Elasticity retains an additional gain of approximately 2.8% under both splits, indicating that the learned combination direction transfers across held-out cases. Darcy remains close to equal averaging, suggesting that little additional global benefit is recovered by reweighting. On Pipe, the learned reweighting does not retain a positive gain under either split. These results are mechanism-oriented cross-validation analyses using frozen historical predictions rather than new official benchmark scores.

Taken together, the experiments separate *available potential* from *realized gain*. A fixed expert pool may contain much better combinations than finite data can identify, and even a combination that appears beneficial during fitting is useful only if that benefit persists relative to a strong reference and transfers to unseen cases.

5.2 Predicting Error Is Not the Same as Choosing a Useful Action

Realizing additional accuracy requires more than identifying when the current prediction is likely to be inaccurate. A useful selector must also determine which available alternative will improve it. Predicting error magnitude and predicting the benefit of an action are therefore different problems.

A heterogeneous Pipe study provides a positive example. A validation-selected risk-switching rule retains the LRSA/MoNo Full equal average on most cases and switches to MoNo Full on predicted high-risk cases. Relative to the validation-selected fixed predictor, risk switching reduces error

Table 4: **Action selection in a heterogeneous Pipe pool.** Errors are mean case-relative L_2 in units of 10^{-3}. Gains are relative to the validation-selected single model.

Method	Error ↓	Gain (%) ↑
Validation-selected single model	2.189	0.00
Fixed convex combination	2.067	5.58
Risk-based switching	1.976	9.73
Case-level advantage selection	**1.933**	**11.67**

Table 5: **Expert information adds less after a stronger aggregate is formed.** Mean relative L_2 errors ($\times 10^{-3}$) across three corrector seeds. Gain compares EA with CO for the same centre.

Centre	CO ↓	EA ↓	Shuffled ↓	EA vs. CO (%)
Single-model anchor	5.545	3.774	5.546	31.95
Equal combination	3.888	3.772	3.891	3.00
Fixed convex	3.799	3.774	3.801	0.66

by 9.73%, while case-level advantage selection achieves an 11.67% reduction (Tab. 4). Both also outperform the fitted fixed convex combination.

Although advantage selection gives the lowest point estimate, its paired improvement over risk switching is not statistically conclusive. The results therefore support useful case-dependent selection without establishing that the more elaborate selector is better than the simpler risk rule. Additional stability analyses are reported in Appendix F.

Accurate risk prediction, however, does not guarantee a useful action. On Elasticity, high-risk cases can also be predicted accurately, yet the tested fallback worsens accuracy relative to the strong fixed combination. The relevant question is therefore not only whether a case is difficult, but whether an available alternative improves upon the current prediction.

These results sharpen the notion of realizability: useful complementary structure must support a beneficial decision, not merely an accurate diagnosis of where the current predictor fails.

5.3 Does Expert Information Remain Useful after Aggregation?

A strong aggregate can still contain correctable residual error. As a general post-hoc refinement control, we apply IRNO (Liu et al., 2026) to already combined predictions. After validation calibration, IRNO improves the Darcy centre by about 1.00% and a heterogeneous Pipe centre by 1.39%. Thus, aggregation does not exhaust all correctable accuracy. We next ask a more specific question: does correcting this residual still benefit from access to the individual expert predictions?

To isolate the incremental value of expert information, we compare matched centre-only (CO) and expert-aware (EA) correctors. Both variants receive the same physical features and starting prediction c, while EA additionally receives expert-dependent information D:

$$\widehat{f}_{\mathrm{CO},c} = c + \eta_{\mathrm{CO},c} g_{\theta_{\mathrm{CO},c}}(x, c, 0), \qquad \widehat{f}_{\mathrm{EA},c} = c + \eta_{\mathrm{EA},c} g_{\theta_{\mathrm{EA},c}}(x, c, D). \tag{2}$$

Here, D contains differences among the individual expert predictions and summaries of their spread. The CO variant sets the standardized expert-information channels to zero, so the two variants share the same architecture. A shuffled-information control tests whether any benefit depends on the expert information being aligned with the current case.

The incremental value of expert information depends strongly on the starting prediction. With an equal-average centre, EA improves over CO by 3.00%, whereas the gain falls to 0.66% after replacing the centre with a stronger fixed convex combination (Fig. 4d and Tab. 5). The shuffled control does not reproduce these gains. Importantly, the smaller EA advantage does not imply that the stronger centre is no longer correctable. The IRNO control above shows that an aggregated prediction can still be improved. Rather, expert differences become less incrementally useful once part of their structure has already been incorporated into a strong aggregate.

Table 6: **Full expert-conditioned correctors do not consistently outperform centre-only correction.** Errors are reported in units of 10^{-3} and averaged over three corrector seeds.

Task	Centre ↓	CO ↓	Full ↓	Full vs. centre (%)	Full vs. CO (%)
Elasticity	3.3305	3.2892	3.2916	1.169	-0.073
Plasticity	0.3676	0.3632	0.3636	1.101	-0.119
Darcy	4.4472	4.3477	4.3510	2.164	-0.076

We further test this distinction using full same-centre correctors on Elasticity, Plasticity, and Darcy. Unlike the strict input-mask comparison above, these controls compare complete centre-only and expert-conditioned correctors and therefore also differ in other expert-dependent components.

Both CO and Full improve on their respective uncorrected centres, but CO achieves slightly lower mean error in all three comparisons. The differences are small, and their signs are not uniform across individual seeds on Elasticity and Plasticity. Because these full-model controls modify additional expert-dependent components, they complement rather than duplicate the strict input-mask comparison above. Together, the results separate *residual correctability* from the *incremental value of expert information*: a strong aggregate can remain correctable even when individual expert predictions provide little additional benefit beyond the aggregate itself.

6 Discussion and Conclusion

Our results distinguish three questions that are often conflated: whether neural operators make different predictions, whether those differences improve aggregation, and whether the resulting complementarity can be reliably exploited from finite data. Simple averaging can recover substantial accuracy, but disagreement alone is not sufficient. Useful complementarity requires cancellable errors to outweigh member quality differences and to reduce the errors that dominate the overall risk. This explains why more models, broader prediction coverage, or explicit diversity need not produce better aggregates.

The same distinction continues beyond equal averaging. More flexible combination rules can have much greater oracle potential without producing correspondingly larger held-out gains, and their advantage can shrink relative to a strong fixed combination. Likewise, predicting that a case is difficult does not by itself identify a beneficial action. After aggregation, residual error can still be corrected, but access to individual expert predictions need not provide the same incremental value. Together, these results motivate evaluating learned combination and refinement methods against strong fixed references and matched centre-only controls.

Our conclusions are conditioned on the expert pools and protocols studied here. Some mechanism analyses reuse historical predictions or released checkpoints, uncertainty estimates do not include backbone-retraining variation, and several follow-up studies use different development budgets. Within these limits, our experiments show that the value of model diversity lies not in disagreement itself, but in whether complementary errors can be converted into reliable held-out accuracy. We refer to this finite-data benefit as *realizable complementarity*.

AI use statement

In this work, we used generative AI tools for language polishing, related-work drafting assistance, code debugging, and assistance in analyzing experimental results.All AI-assisted text and code were manually reviewed and verified by the authors. The authors take full responsibility for the final content of this work.

Ethics statement

This study involves neither human participants nor animal experiments. The datasets used in our experiments are publicly available benchmarks for PDE surrogate modeling; they were obtained and used in accordance with their respective usage guidelines and licenses, and their use raises no

privacy concerns. Throughout the research process, we have made every effort to prevent biases or discriminatory outcomes. Our experiments involve no personally identifiable information and give rise to no privacy or security concerns. We remain committed to transparency and integrity in every stage of this research.

Reproducibility statement

To ensure the full reproducibility of our empirical results, the complete source code of the proposed framework, together with instructions for downloading the datasets used in this work, will be made publicly available in a public GitHub repository upon acceptance.

References

Christopher Bülte, Philipp Scholl, and Gitta Kutyniok. Probabilistic neural operators for functional uncertainty quantification. *arXiv preprint arXiv:2502.12902*, 2025.

Shuhao Cao. Choose a transformer: Fourier or galerkin. *Advances in neural information processing systems*, 34:24924–24940, 2021.

Stanislav Fort, Huiyi Hu, and Balaji Lakshminarayanan. Deep ensembles: A loss landscape perspective. *arXiv preprint arXiv:1912.02757*, 2019.

Gaurav Gupta, Xiongye Xiao, and Paul Bogdan. Multiwavelet-based operator learning for differential equations. *Advances in neural information processing systems*, 34:24048–24062, 2021.

Zhongkai Hao, Zhengyi Wang, Hang Su, Chengyang Ying, Yinpeng Dong, Songming Liu, Ze Cheng, Jian Song, and Jun Zhu. Gnot: A general neural operator transformer for operator learning. In *International conference on machine learning*, pp. 12556–12569. PMLR, 2023.

Wenjie Hu, Sidun Liu, Peng Qiao, Zhenglun Sun, and Yong Dou. Transolver is a linear transformer: Revisiting physics-attention through the lens of linear attention. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 40, pp. 408–416, 2026. doi: 10.1609/aaai.v40i1.37003. URL https://ojs.aaai.org/index.php/AAAI/article/view/37003.

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.

Balaji Lakshminarayanan, Alexander Pritzel, and Charles Blundell. Simple and scalable predictive uncertainty estimation using deep ensembles. *Advances in neural information processing systems*, 30, 2017.

Zongyi Li, Nikola Kovachki, Kamyar Azizzadenesheli, Burigede Liu, Kaushik Bhattacharya, Andrew Stuart, and Anima Anandkumar. Fourier neural operator for parametric partial differential equations. *arXiv preprint arXiv:2010.08895*, 2020a.

Zongyi Li, Nikola Kovachki, Kamyar Azizzadenesheli, Burigede Liu, Andrew Stuart, Kaushik Bhattacharya, and Anima Anandkumar. Multipole graph neural operator for parametric partial differential equations. *Advances in neural information processing systems*, 33:6755–6766, 2020b.

Zongyi Li, Daniel Zhengyu Huang, Burigede Liu, and Anima Anandkumar. Fourier neural operator with learned deformations for pdes on general geometries. *Journal of Machine Learning Research*, 24(388):1–26, 2023.

Phillip Lippe, Bas Veeling, Paris Perdikaris, Richard Turner, and Johannes Brandstetter. Pde-refiner: Achieving accurate long rollouts with neural pde solvers. *Advances in Neural Information Processing Systems*, 36:67398–67433, 2023.

Xiaotian Liu, Shuyuan Shang, Xiaopeng Wang, Pu Ren, and Yaoqing Yang. Iterative refinement neural operators are learned fixed-point solvers: A principled approach to spectral bias mitigation. In *Forty-third International Conference on Machine Learning*, 2026. URL https://openreview.net/forum?id=4lAS8tI23u.

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(3):218–229, 2021.

Huakun Luo, Haixu Wu, Hang Zhou, Lanxiang Xing, Yichen Di, Jianmin Wang, and Mingsheng Long. Transolver++: An accurate neural solver for pdes on million-scale geometries. *arXiv preprint arXiv:2502.02414*, 2025.

S Chandra Mouli, Danielle C. Maddix, Shima Alizadeh, Gaurav Gupta, Andrew Stuart, Michael W. Mahoney, and Bernie Wang. Using uncertainty quantification to characterize and improve out-of-domain learning for PDEs. In Ruslan Salakhutdinov, Zico Kolter, Katherine Heller, Adrian Weller, Nuria Oliver, Jonathan Scarlett, and Felix Berkenkamp (eds.), *Proceedings of the 41st International Conference on Machine Learning*, volume 235 of *Proceedings of Machine Learning Research*, pp. 36372–36418. PMLR, 21–27 Jul 2024. URL https://proceedings.mlr.press/v235/mouli24a.html.

Md Ashiqur Rahman, Zachary E Ross, and Kamyar Azizzadenesheli. U-no: U-shaped neural operators. *arXiv preprint arXiv:2204.11127*, 2022.

Haixu Wu, Huakun Luo, Haowen Wang, Jianmin Wang, and Mingsheng Long. Transolver: A fast transformer solver for pdes on general geometries. *arXiv preprint arXiv:2402.02366*, 2024.

Zherui Yang, Haiyang Xin, Tao Du, and Ligang Liu. Simple yet effective: Low-rank spatial attention for neural operators. *ICML*, 2026a.

Zijiang Yang, Xiaomeng Wu, and Dongmei Fu. Mono: Multiscale optimal transport neural operator for solving pdes on general geometries. *arXiv preprint arXiv:2608.09764*, 2026b.

Appendix Overview

This appendix provides supplementary experimental details and analyses supporting the main results. Appendix A summarizes the experimental cohorts and evaluation conventions used throughout the paper. Appendix B provides additional field visualizations and spatial diagnostics for the Elasticity, Darcy, and Pipe experiments. Appendix C gives further details of the compute-controlled Darcy comparison. Appendix D examines cross-family error structure on Pipe using LinearNO, LRSA, and MoNo predictions. Appendix E reports additional correction results beyond strong aggregate predictions. Finally, Appendix F reports decision stability and development-label analyses, including the cross-case transfer study in Appendix F.1.

A Experimental Protocols

Our experiments use different expert pools to answer distinct questions about neural-operator aggregation. The six-task study establishes the averaging effect, while subsequent experiments examine its generality, mechanisms, finite-data realization, action selection, and post-aggregation refinement. These experiments are separate comparisons rather than successive stages of a single system. Table 7 summarizes their roles.

Evaluation conventions. The six-task benchmark follows the evaluation metric used for each task. Elasticity, Darcy, Airfoil, and Pipe use mean case-relative L_2, Plasticity averages relative errors over 20 steps, and Navier–Stokes uses relative L_2 over the full 10-step rollout. Comparisons within each task use the same metric and evaluation cases. When we refer to the best individual model, we use one fixed member selected by its overall evaluation performance, rather than choosing the best model separately for each case. Most mechanism analyses use mean case-relative L_2. The squared-risk quantity used for the cancellation analysis is defined separately in Sec. 4.

Table 7: **Overview of the experimental cohorts.** Different analyses use separate expert pools to answer different questions and should be interpreted within their own evaluation setting.

Analysis	Experimental setup	Question addressed
Six-task averaging	3 LinearNO models per task	How much accuracy can averaging recover?
Ensemble scaling	Larger LinearNO and LRSA pools	Does the effect persist with more models and another backbone?
Darcy compute control	Fresh LinearNO runs at matched training FLOPs	Is the gain explained by compute or model size?
Spatial combination	Separate Elasticity and Pipe pools	Does more flexible combination realize more complementarity?
Cross-case transfer	Frozen 3-model pools on three tasks	Does learned reweighting transfer to held-out cases?
Pipe action selection	Heterogeneous LRSA/MoNo pool	Can predicted risk support beneficial selection?
Expert-information control	Five-expert Elasticity pool	Does expert information remain useful after aggregation?

B Data Fields and Spatial Diagnostics

Elasticity illustration in Fig. 1. The displayed Elasticity case was selected from the frozen 200-case pool using member-error and residual-correlation statistics, without using ensemble gain as a selection criterion. It is therefore intended as a descriptive example rather than evidence for the population-level benefit of averaging. Signed errors are visualized as $(\widehat{u} - u)/\operatorname{RMS}(u)$ using a common symmetric colour scale for all models and the equal mean. The display limit is set by the 99.5th percentile of the absolute values and affects visualization only, not the reported errors.

Pipe spatial-risk diagnostics. Fig. 3(b) shows case 2141, the highest-error case under the original three-LinearNO mean, fixed before the cross-family comparison. Fig. 3(c) instead aggregates all 200 Pipe cases. Pointwise squared errors are normalized by the target mean-square amplitude, so summing over nodes is proportional to the normalized squared risk used in the main text. The transverse and streamwise coordinates are normalized for visualization.

For Fig. 2(d), node coverage is defined by whether the target lies inside the interval spanned by the expert predictions. Risk is always measured relative to the original LinearNO mean. The $+\mathrm{LRSA}$ and $+\mathrm{MoNo}$ settings are separate four-member augmentations of the same three-LinearNO base, while the six-expert pool contains all three LinearNO models, LRSA, MoNo Full, and MoNo Light.

Additional field views. Fig. 5 compares median-error and highest-error cases from the audited Elasticity, Darcy, and Pipe pools. Each row shows the reference field, three member errors, the equal-mean error, and expert spread. Examples are selected by the equal mean's error rank rather than by ensemble improvement. Error scales are shared within each row but not across tasks.

Additional Pipe checks. The hardest 10% of Pipe cases account for 99.05% of normalized squared risk and 99.06% of raw squared error, indicating that the observed tail concentration is not caused by unusually small target norms. The pointwise convex oracle still retains approximately 58.08% of the equal centre's squared risk, showing that limited expert coverage does not make all residual error irreducible.

C Compute-Controlled Darcy Details

The compute-controlled Darcy study uses $1{,}000$ training cases and 200 evaluation cases from the smooth1/smooth2 split, with stride 5 and 85×85 spatial points. All configurations use the same LinearNO backbone and training setup.

The reference model uses hidden width $C = 128$ and latent size $M_{\mathrm{lat}} = 64$ and is trained for 500 epochs. We compare the three-model ensemble with three single-model scaling strategies: longer training, increased width, and increased latent size. The longer-training control uses the reference architecture for $1{,}500$ epochs. The width-scaled model increases C from 128 to 224, while the larger-latent model increases M_{lat} from 64 to 512. The ensemble consists of three independently initialized reference models trained for 500 epochs each.

All three single-model controls use approximately the same total training compute as the ensemble. The wider model also closely matches the ensemble in parameter count and forward FLOPs, while having substantially lower serial inference latency. Thus, the comparison isolates training-compute allocation without claiming that the ensemble is preferable under every computational cost measure.

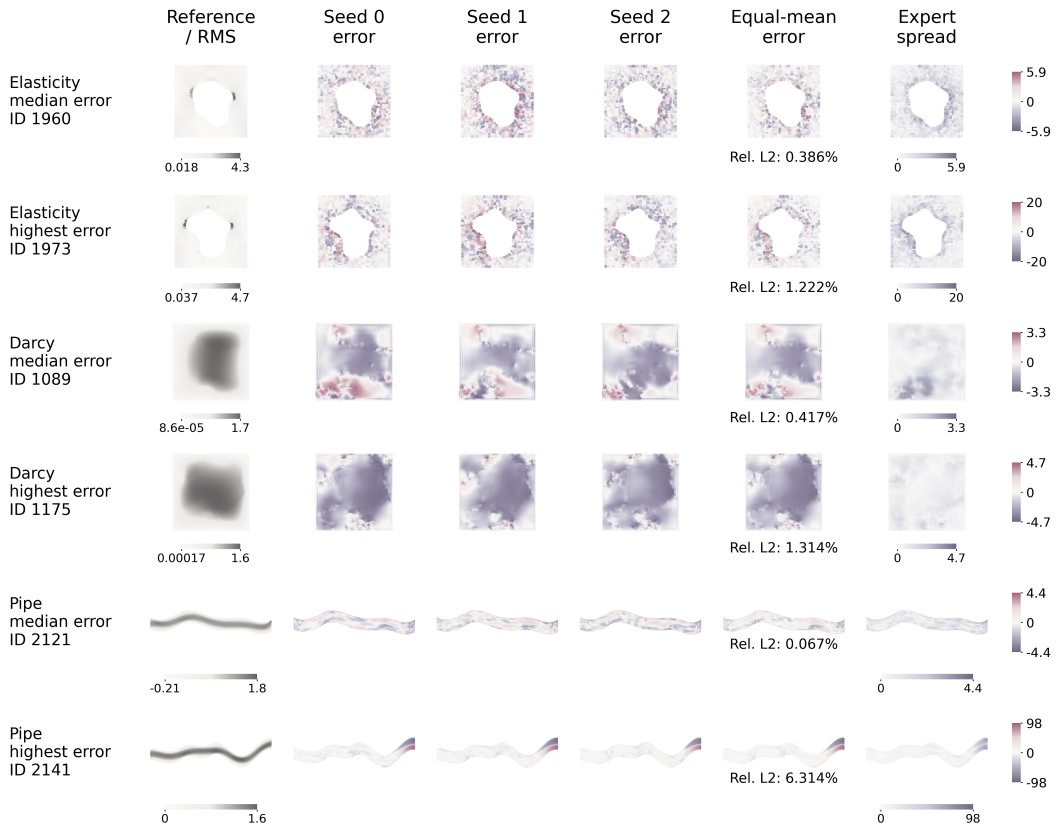

Figure 5: **Median and difficult cases reveal different error-sharing patterns.** Fixed examples from Elasticity, Darcy, and Pipe showing the reference, three member errors, equal-mean error, and expert spread. Signed errors and spread are normalized by the target RMS. Scales are shared within each row but are not quantitatively comparable across tasks.

Table 8: **Compute-controlled Darcy configurations.** Parameters and forward FLOPs for the ensemble are summed across members. Ensemble latency assumes serial execution.

Configuration	Epochs	Params	Forward GFLOPs	Train PFLOPs	Latency (ms)
Reference $(128, 64)$	500	1.77M	29.13	43.69	–
Longer $(128, 64)$	1,500	1.77M	29.13	131.08	21.84
Wider $(224, 64)$	518	5.37M	84.23	130.89	24.55
Larger latent $(128, 512)$	785	1.88M	55.65	131.04	119.13
Three-model mean	3×500	5.30M	87.40	131.08	65.53

This controlled study uses a separate Darcy expert pool reported in Tab. 1. Its absolute error values should therefore be compared only within the compute-controlled experiment.

D Cross-Family Pipe Evidence

We test whether the downstream error pattern observed in the three-LinearNO pool persists across model families. The same 200 Pipe cases are evaluated using the original LinearNO members together with LRSA, MoNo Full, and MoNo Light. The models are not compute-matched and are used only as a diagnostic of cross-family error structure. The difficult subset is fixed as the original LinearNO mean's highest-error 20 cases before adding the new models.

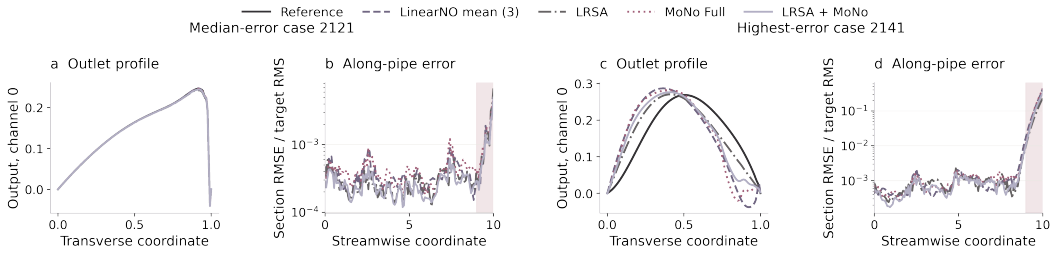

Figure 6: **Cross-family profiles on fixed Pipe examples.** The median-error and highest-error cases are fixed before the cross-family comparison. Different model families agree closely on the median case, whereas the difficult case shows a shared increase in error near the outlet.

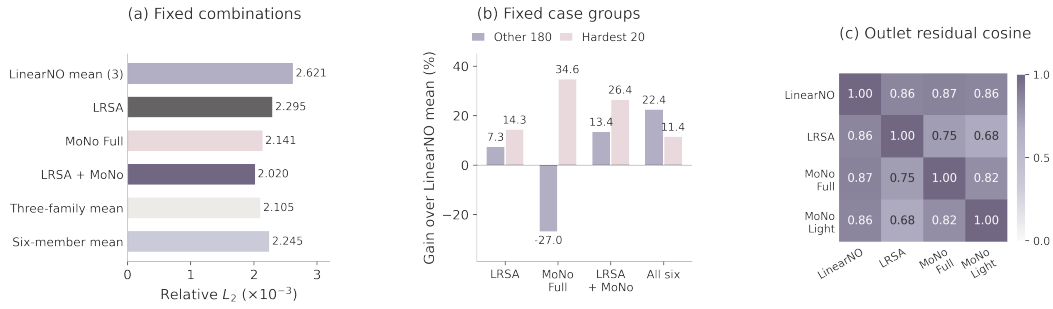

Figure 7: **Cross-family accuracy and downstream error structure.** **(a)** Fixed combinations on the same 200 cases. **(b)** Gains on the fixed 180 ordinary and 20 hardest cases. **(c)** Pairwise normalized-residual cosine similarity in the final 10% of the streamwise domain.

Relative to MoNo Full, the LRSA/MoNo Full mean improves the fixed 180 ordinary cases by 31.80% but worsens the fixed 20 hardest cases by 12.68%. Its overall point-estimate gain is 5.63%, with a paired-case interval crossing zero. Thus, cross-family diversity can improve many cases without resolving the high-risk downstream errors.

Table 9: **Additional gains beyond strong combinations.** Gain is measured relative to the centre in the same row. Historical best results and repeated-run means are reported separately.

Task and reported result	Centre ↓	Corrected ↓	Gain (%) ↑
Elasticity: signed + grouped, historical best	0.003330538	0.003258146	2.174
Darcy: geometry-aware correction, historical result	0.004337782	0.004223617	2.632
Plasticity: multiscale correction, best run	0.000367632	0.000361401	1.695
Plasticity: same strategy, three-run mean	0.000367632	0.000362701	1.341
Pipe: signed correction, best run	0.002110196	0.002081599	1.355
Pipe: same strategy, three-run mean	0.002110196	0.002093638	0.785

E Additional Gains beyond Strong Combinations

We report several previously completed task-specific combination-and-refinement results to illustrate the remaining correctable error beyond strong centres. These experiments use different expert pools and development protocols, so comparisons should be made within each row rather than across tasks.

The Elasticity entry is a historical full-pipeline result whose centre already contains refined predictions, so its 2.174% gain should not be attributed solely to the final grouped module. For Plasticity, the repeated runs also vary the internal fitting split and therefore do not isolate initialization variability.

Overall, these task-specific results show that strong combinations can retain correctable residual error. They do not establish that expert-specific information is necessary for the improvement or that one correction strategy is consistently superior across tasks.

F Decision Stability and Label Budget

The heterogeneous Pipe decision study uses frozen LRSA, MoNo Full, and MoNo Light predictions, with 600 fitting, 200 validation, and 310 evaluation cases. The main-text results use this fixed evaluation cohort.

To test stability, we repeat the fitting procedure over ten predefined overlapping splits. Risk switching gives a positive gain in all ten splits, with median gain 8.79%, while case-level advantage selection is also positive in all ten, with median gain 10.40%. Because the splits reuse evaluation cases, these results should be interpreted as repeated-fit stability rather than independent replications.

We also vary the number of fitting cases while keeping an additional 200 validation cases fixed. Risk-switching gains largely saturate after 200 fitting cases, whereas advantage selection continues to improve more gradually (Fig. 8).

F.1 Cross-Case Combination Stability

The cross-case analysis in Table 3 uses frozen three-model pools and five predefined folds. Each fold uses 120 cases to fit input-independent convex weights, 40 cases to calibrate the step from equal averaging toward the fitted combination, and 40 held-out cases for evaluation. The experts themselves are not retrained across folds.

Elasticity shows a stable positive increment under both random and input-statistic holdouts, whereas Darcy remains near zero and Pipe does not retain a positive gain. In four of five random Pipe folds, the fitted corrective direction becomes negatively aligned on held-out cases. This supports the interpretation that the additional reweighting learned on Pipe transfers poorly across cases.

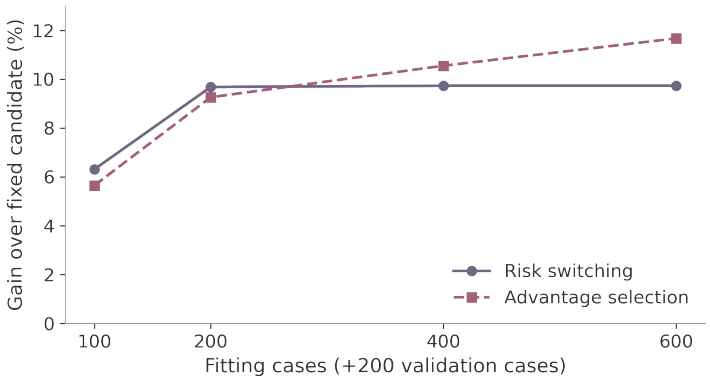

Figure 8: **Effect of development-label budget on Pipe action selection.** Held-out gains of risk switching and case-level advantage selection as the number of fitting cases increases. Each setting uses an additional fixed set of 200 validation cases.