

Regression Model Selection under Explicit Inference Costs

Nihal Singh Chauhan

nihalsinghchauhan11@gmail.com

University of Petroleum and Energy Studies

Research Article

Keywords:

Posted Date: October 29th, 2026

DOI: <https://doi.org/10.21203/rs.3.rs-11522485/v1>

License:   This work is licensed under a Creative Commons Attribution 4.0 International License.

[Read Full License](#)

Additional Declarations: The authors declare no competing interests.

Regression Model Selection under Explicit Inference Costs

Anonymous Authors¹

Abstract

Does selecting one regression expert per input justify its inference cost? On 35 previously unused benchmark tasks in 33 source families, our validation-selected one-query gate has a geometric MSE ratio of 1.055 to a calibrated global ensemble, at 0.410 times its measured warm latency. A retrained global boosting model has both lower MSE (1.038) and lower latency (0.264); AutoGluon improves MSE to 0.851 at substantially greater inference cost. The evaluated gate therefore adds no point to the aggregate accuracy–cost frontier, although it does on six individual source families. We also prove that bounded, pointwise dominated actions can consume every split of a depth-limited shared-contrast greedy tree while preserving a positive action margin; a scalar stump avoids this failure. The construction extends to specified forests, a leaf constraint, target normalization, and outcome noise. Real-data library growth from 2 to 16 experts, however, does not show consistent degradation from weak additions. Shrinkage, clipping and projection controls further limit a scalar-selection explanation. The combined evidence supports checking strong global alternatives before attributing practical value to regression routing, and distinguishes a provable learner failure from its observed empirical prevalence.

1. Introduction

A library of trained predictors creates two distinct decisions.¹ If all predictions are available, choosing one restricts a continuous ensemble. If inference must stop after

¹Anonymous Institution, Anonymous City, Anonymous Region, Anonymous Country. Correspondence to: Anonymous Author <anon.email@domain.com>.

Preliminary work. Under review by the International Conference on Machine Learning (ICML). Do not distribute.

¹Generative AI assisted experimental implementation, mathematical derivations, statistical analysis, and manuscript drafting. The human authors are responsible for independently validating and approving all content.

one expert query, a selector must choose before observing that expert’s output. This distinction makes the information and computation used by a regression router part of its evaluation.

Our development study provided little support for an accuracy advantage: nearest-output residual projection gave no clear improvement over a strong global ensemble in Library B, while retrained global models were stronger alternatives. Its seven actions also included means and stacks of the same four experts. We therefore evaluate genuinely distinct fitted models, complete serving paths, and strong alternatives on a replacement panel of previously unused tasks. The result is unfavorable to the studied router on the aggregate accuracy–cost comparison: global boosting is both faster and more accurate, and AutoGluon offers stronger accuracy at higher cost. Family-level results show that the aggregate conclusion has exceptions.

Under squared loss, the conditional mean determines every action-risk ordering. Outcome-first Bregman losses retain conditional-mean sufficiency (Banerjee et al., 2005). This established identity motivates scalar alternatives to many contrast estimates, but evaluating nearest outputs still requires acquiring those outputs. Our theoretical contribution instead concerns a specified shared-target greedy learner. Uniformly bounded, pointwise dominated actions can consume every split of a depth- d tree while its decision margin stays fixed. A redundant-target construction survives normalization; extensions cover forests of the specified trees, a minimum leaf probability, and outcome noise. On the same noiseless distribution, a scalar stump has an exponential finite-sample upper bound while the shared learner’s risk remains bounded away from zero. Standard margin techniques connect mean estimation error to regret for Bregman losses and known costs.

The empirical study grows 2/4/8/16-model libraries, adds independently permuted-target experts, measures split allocation, and separates target changes from decision-feature changes. A 35-task, 33-family primary panel uses all available policy labels, geometric MSE ratios and paired family comparisons; a generated control is separate. Shrinkage, clipping and projection ablations test competing explanations. The weak-library controls do not reproduce consistent degradation despite shifts toward prediction features.

A dated local protocol and explicit amendments document the study; no external preregistration exists. The resulting contribution is a scoped theoretical separation and a complete-system evaluation that identifies the limits of its practical implication.

2. Prediction, selection, and serving cost

Let $f_0, \dots, f_{K-1}$ be predictors fixed before policy fitting. A selector π outputs $f_{\pi(X)}(X)$. For nonnegative costs $c_a(X)$ known before querying the expert, define

$$\mathcal{R}_\lambda(\pi) = \mathbb{E}[\ell(Y, f_{\pi(X)}(X)) + \lambda c_{\pi(X)}(X)], \quad \lambda \geq 0. \quad (1)$$

For squared loss, write $m(x) = \mathbb{E}[Y \mid X = x]$ and define the reference contrasts

$$D_a = (f_a - Y)^2 - (f_0 - Y)^2 = f_a^2 - f_0^2 - 2Y(f_a - f_0). \quad (2)$$

Conditional action scores are $(f_a - m)^2 + \lambda c_a$, up to a common noise term. A shared multioutput forest estimates $\mathbb{E}[D \mid Z]$, where the decision features Z may be raw inputs X or augmented inputs $(X, F(X))$. Its split criterion averages squared error over contrast targets. The decision is $\arg \min_a \{\widehat{D}_a(Z) + \lambda c_a\}$, with $\widehat{D}_0 = 0$ imposed exactly. Action ties favor the lowest index throughout. Target scale therefore affects the fitted representation even when an added action is never optimal.

A scalar estimator instead estimates m directly, or estimates a residual $r = m - f_0$. We evaluate the anchored family

$$\widehat{m}_\gamma = f_0 + \gamma(\widehat{m} - f_0) \quad \text{or} \quad f_0 + \gamma\widehat{r}. \quad (3)$$

The three separate decisions are continuous prediction $\widehat{m}_\gamma$, clipping to the interval of available expert outputs, and nearest-output projection $\arg \min_a |f_a - \widehat{m}_\gamma|$. The last is a known dynamic-regressor-selection construction (Moura et al., 2019). Each decision has its own training-only choice of leaf size and γ .

In the fixed-bank experiments, all three scalar paths first constrain the estimate to the expert-role target range. The additional clipped path then restricts it to the input-specific interval between expert outputs. ‘‘Continuous’’ therefore denotes prediction without a discrete bank restriction, with the common training-range constraint retained.

Query contract. A raw-input gate runs before expert inference and executes only the selected expert. An augmented gate requires all inputs used in its feature vector; nearest-output projection ordinarily requires all K outputs. A residual continuous predictor uses the reference expert and correction model. A direct continuous predictor with $\gamma = 1$ uses only its mean estimator. When $\gamma = 0$, all

scalar variants reduce symbolically to the reference expert and bypass the correction and other experts. Counts follow these dependencies, including this simplification. Offline full prediction banks support counterfactual evaluation; they are not treated as free at serving time.

Our measured action prices are warm CPU expert-forward times on calibration inputs, normalized by the library’s median price. They define the fixed λ grid. The complete-path timing additionally includes encoding, gate execution, selected expert execution, and combining outputs. This overhead is reported rather than assumed to vanish. It is not included in the action-dependent training penalty. Timings concern this implementation, hardware, batch size, and cache state, rather than a hardware-independent complexity claim.

3. Scalar sufficiency and a greedy-learner separation

For a differentiable strictly convex ϕ , use the outcome-first divergence $D_\phi(y, f) = \phi(y) - \phi(f) - \langle \nabla \phi(f), y - f \rangle$. Assume finite displayed expectations and predictions and conditional means in the required differentiability domain. Conditional expectation gives the established identity (Banerjee et al., 2005)

$$\mathbb{E}[D_\phi(Y, f_a) \mid X] = \mathbb{E}[D_\phi(Y, m) \mid X] + D_\phi(m, f_a). \quad (4)$$

Known action costs can be added without changing the common term. Scalar Poisson deviance, with positive predictions, therefore has a scalar conditional-mean reduction. For categorical KL loss the sufficient statistic is a probability *vector*. Reversing the divergence arguments changes the optimal statistic. Equation (4) is not itself an algorithmic novelty.

Theorem 1 (Margin-dependent plug-in regret). *Let $q_a(x) = D_\phi(m(x), f_a(x)) + \lambda c_a(x)$ have a unique minimizer almost surely. Let Δ be the gap to the second-smallest score and assume $\Pr(\Delta \leq t) \leq Ct^\alpha$, $\alpha \geq 0$. Suppose $\sup_x \max_{a,b} \|\nabla \phi(f_a(x)) - \nabla \phi(f_b(x))\|_* \leq L$. Write $\mathcal{E}_\lambda = \mathcal{R}_\lambda(\widehat{\pi}) - \mathcal{R}_\lambda(\pi^*)$ for plug-in excess risk. Condition on the fitted action bank. For a fixed mean estimator, with expectation over a fresh input:*

$$\|\widehat{m} - m\|_\infty \leq \epsilon \implies \mathcal{E}_\lambda \leq C(L\epsilon)^{\alpha+1}, \quad (5)$$

$$\mathbb{E}\|\widehat{m}(X) - m(X)\|^2 \leq \eta \implies \mathcal{E}_\lambda \lesssim_{C,L,\alpha} \eta^{(\alpha+1)/(\alpha+2)}. \quad (6)$$

If, uniformly in x, t , training randomness satisfies $\Pr(\|\widehat{m}(x) - m(x)\| > t) \leq A \exp(-at^2/s_n^2)$, then expected excess risk is $O(s_n^{\alpha+1})$. With an almost-sure fixed positive gap, the same tail condition gives an exponential bound in $1/s_n^2$ (Appendix A).

The proof uses standard margin plug-in techniques (Audibert & Tsybakov, 2007): score-difference error is at most $L\|\hat{m} - m\|$, and any wrong action has Δ no larger than this error. These statements are conditional implications for an estimator, not a proof that a fitted ExtraTrees model attains the required estimation rates. Constants can depend on the action range and margin distribution; scalar sufficiency does not make every bound independent of library size. For squared loss with $|f_a| \leq M$, $L \leq 4M$. For Poisson deviance with $f_a \in [\ell, u]$, $\ell > 0$, $L \leq 2\log(u/\ell)$.

The next result concerns a specified learner: greedy axis-aligned squared-error trees with a shared sum/average target-impurity criterion and all coordinates eligible at each split. The depth-limited statement permits every nonempty binary split, with no additional leaf-mass restriction; Corollary 4 treats a leaf constraint separately. It is an algorithm-specific lower bound. It would be false as an impossibility result for all contrast learners, since a learner may exploit Equation (4).

Theorem 2 (Bounded dominated actions consume depth). *For squared loss and $\lambda = 0$, fix an integer $d \geq 1$ and $0 < p < 1$. Let $S \sim \text{Bernoulli}(p)$ and $Z_1, \dots, Z_d \sim \text{Bernoulli}(1/2)$ be independent, $X = (S, Z)$, and $Y = S$. Use useful actions $(p, 0, 1)$ and added actions*

$$f_{j+2}(X) = S + \sqrt{(p - S)^2 + 1 + 4Z_j}. \quad (7)$$

All predictions lie in $[0, 1 + \sqrt{6}]$, independently of d . Every added action is pointwise worse than $f_0 = p$, with contrast exactly $1 + 4Z_j$. A population shared-contrast greedy tree of maximum depth d uses only nuisance coordinates and routes every input to f_0 , with risk $p(1 - p)$. Any finite forest averaging such trees has the same risk. Without the added actions, a contrast stump attains zero risk; a scalar stump also attains zero risk with the added actions present. Depth $d + 1$ removes this construction's restriction.

For the empirical scalar stump followed by projection, using n i.i.d. training observations,

$$\mathbb{E}\mathcal{R}(\hat{\pi}_{\text{scalar}}) \leq p(1 - p)^n + (1 - p)p^n + 2d2^{-n}. \quad (8)$$

For the empirical depth- d shared greedy learner, put $m_0 = \min\{p, 1 - p\}$. Then

$$\mathbb{E}\mathcal{R}(\hat{\pi}_{\text{shared}}) \geq p(1 - p) \left[1 - 2^{d+2} \exp \left\{ -\frac{nm_0}{75d2^d} \right\} \right]. \quad (9)$$

This finite-sample bound also holds for any finite forest of these trees fitted to the same full sample without bootstrapping. A fixed finite bootstrap forest has the same persistent population limit.

The mechanism is the shared split objective. At every nuisance-only node, an unused nuisance split gains 4 in

summed target variance, while a signal split gains at most 1. Thus the learner spends its depth on targets belonging to unusable actions. The positive decision margin $\min\{p^2, (1 - p)^2\}$ and the action bound both remain fixed as d grows. The added experts are fixed functions of the input coordinate S ; they need no outcome query. The next construction isolates redundant-target weighting after normalization.

Corollary 3 (Normalization does not remove redundant-target weighting). *Replace each added action in Theorem 2 by three distinct actions*

$$f_{j,r} = S + \sqrt{(p - S)^2 + 1 + 4Z_j + r/4}, \quad r = 0, 1, 2.$$

All actions remain in $[0, 4]$, and every addition is pointwise dominated. Even after centering and standardizing each nonconstant contrast to unit population variance, a depth- d shared greedy tree, with inverse transformation before action selection, has risk $p(1 - p)$. Any fixed finite forest has the same risk. The same empirical limit holds with consistent sample target normalization.

Each nuisance now contributes three unit-variance targets, while S controls only two. Their respective split gains are 3 and 2. Normalization remedies scale imbalance but does not guarantee immunity to redundant targets. These statements concern the specified objective; feature subsampling and alternative target-sharing schemes require their own analysis.

Corollary 4 (Unrestricted depth with a leaf-mass constraint). *Fix $0 < \rho \leq p < 1/2$, put $k = \lfloor \log_2(1/\rho) \rfloor$, and use at least k nuisance coordinates in Theorem 2. A population greedy tree with minimum leaf probability ρ , regardless of its allowed depth, exhausts feasible nuisance splits before splitting on S . It routes to the reference with risk $p(1 - p)$, whereas a scalar root split on S is feasible and achieves zero risk. At non-boundary settings the corresponding fixed minimum empirical leaf fraction has this limit.*

Full proofs, conservative constants, and finite-support checks appear in Appendix A. This establishes a realizable failure mechanism; the empirical question is whether it is frequent or consequential in actual fitted libraries. Conditionally centered, finite-variance outcome noise preserves the raw shared-tree risk gap (Proposition 5). With constant conditional variance, the redundant-target result also survives normalization (Corollary 6).

4. Evaluation design

Panels and roles. The original 23-dataset study is development evidence: its 14-family primary subset includes six generated families, and its gas-turbine source overlaps a development domain. Its natural-only reanalysis is

post hoc. The replacement audit yields 35 previously unused natural-data tasks in 33 source families from pinned BeyondArena/TabArena metadata (Erickson et al., 2025), OpenML-CTR23 suite 353 (Fischer et al., 2023), and AutoML regression suite 269. MetaQSAR’s two assays share a family, as do two drug-design tasks. Known prior sources/aliases, known simulators, unresolved dependence, and direct target-component leakage are excluded. A simulated insurance table remains a secondary control. Its exclusion and the removal of NYC taxi’s total-amount leakage field are dated amendments before affected fits; the original local record is preserved. Source-quality and chemistry exclusions provide sensitivity panels (Appendix B).

We preserve official repeat-0/fold-0 outer tests and supplied group/time structure, except that MIP runtimes are explicitly regrouped by optimization problem. Outcome-independent caps retain at most 8,000 training and 4,000 test rows. Three seeds allocate training groups to expert fitting E (50%), calibration C (12.5%), and policy fitting P (37.5%); temporal roles are ordered. IDs are removed. Bank preprocessing and target standardization use E only. Curator-supplied target transformations are retained. The capped, modified protocol is distinct from a full published leaderboard reproduction.

Libraries and selection. Sixteen E -trained experts use two fixed recipes each of ridge, XGBoost, random forest, histogram boosting, ExtraTrees, MLP, Nyström RBF ridge, and nearest neighbors. The first recipes form nested 2/4/8-model libraries; the alternatives extend them to 16. A matched weak 16-model library adds eight independently permuted-target fits to the initial eight. Means/stacks are comparators, rather than additional experts. Prediction correlations and realized losses characterize the fitted diversity and weakness.

Shared forests have 80 trees, all features eligible, and minimum leaf $\max\{10, \lceil \ell n \rceil\}$ for $\ell \in \{.01, .05\}$. Three-fold grouped P -only validation selects a forest or a constant individual expert separately at each $\lambda \in \{0, .001, .01, .1, 1\}$. The full P budget and $\lambda = 0$ define the primary gate. All operating points are retained, including the corresponding constant-expert candidates, so conditional choice can be compared with simply choosing a cheaper fixed expert. Prefix-only validation is repeated for 100/300/1,000/3,000 labels when feasible and $1/8, 1/4, 1/2, 1$ fractions of P . Absolute curves report eligible counts; a fixed eligible panel supports the long-budget comparison.

Mechanism controls. Raw/output-augmented fits, inverse-transformed target normalization, and independent per-contrast forests probe the shared objective; the independent fits have larger gate/training compute. Weak-library controls either forbid weak decisions after fitting or keep

16-output features fixed while refitting only eight contrast targets. Split and impurity shares are compared with prediction features’ fraction of all features.

For direct/residual scalar fits, $\gamma \in \{0, .1, .25, .5, .75, 1, 1.25\}$ and leaf fraction have separate validation for continuous, clipped and projected prediction. Unshrunk versions isolate the decision transformation. For continuous estimate u and projected output v ,

$$\hat{R}(v) - \hat{R}(u) = \overline{(v - u)^2} + 2\overline{(u - Y)(v - u)}. \quad (10)$$

Projection improves risk only if the error-correction cross term offsets its displacement penalty. For separately CV-selected variants, we also separate postprocessing at matched hyperparameters from changes in shrinkage and leaf size relative to continuous CV (Appendix B.4).

Complete-system alternatives. The $C + P$ global comparator selects an individual, mean, ridge stack or convex stack by grouped validation. Global histogram boosting and a reference-residual GBM use all $E + C + P$ labels. Actual AutoGluon uses native categories, three-fold bagging and one stacking level, requesting one CPU and a nominal allowance based on the initial measured 16-expert fit times plus full-budget policy-forest fitting and prediction, including validation. Actual overruns are retained. This generous fit-time allowance differs from exact total compute: preprocessing, bank-wide prediction, calibration and global-stack selection are additional costs. An added serialized AutoGluon run repeats every allocation with the same issued allowance after other research fits finish, addressing logged memory skips; both runs are retained. This is a secondary execution control, recorded before aggregate accuracy review. A jointly trained four-expert neural mixture has soft and hardened serving paths; hard routing is a sensitivity of soft training. Public cached CPU TabPFN V2 has context $\min\{1000, \lfloor 100000/p_{\text{raw}} \rfloor\}$ and one estimator. The original nominal $\lceil p_{\text{raw}}/500 \rceil$ sensitivity can still omit inputs because preprocessing expansion reduces each estimator’s input budget. An amended input-complete variant chooses the smallest count from 1 to 16 that exposes all preprocessor inputs, using feature-plan coverage rather than prediction accuracy. These older, context-limited comparisons exclude pretraining cost. Recipes, resources, failures and amendment timing appear in Appendix B.

Estimation and inference. The principal effect is the geometric MSE ratio: average seed log ratios within datasets, then dataset means within source families, then families. Derived ratios use 10^{-12} stabilization in E -standardized MSE units, with raw risks retained. The two original primary paired comparisons are the full-policy one-query CV gate versus retrained global boosting and the original

shared-execution AutoGluon run; the serialized run is secondary. Holm adjustment covers the two primary family-level sign-flip tests: exact for at most 20 families, otherwise 200,000 fixed-seed flips with a plus-one correction. Sign-flip inference assumes symmetric/exchangeable paired family effects under its null. Exact two-sided sign tests are separate directional checks. Hierarchical family/dataset/seed resampling describes variation; overlapping fits and shared tests preclude treating seeds as independent replications. Family effects, comparison-specific coverage and every failure are retained. The selected panel is not a probability sample of future applications.

Costs and protocol status. Action prices use calibration features only. Complete warm batch-one timings include preprocessing and dispatch on eight fixed, outcome-independent test inputs, with one warmup and three repeats. Their median repeat mean and individual repeats are retained. Measurements use an Apple M1 MacBook Air CPU with 8 GiB memory. Shared host use creates timing variability. Selected-only serving is replayed on every test input against the offline bank; a separate forbidden-expert test checks accidental unselected calls. The local pre-fit lock and later amendments are explicit. No external OSF registration exists.

5. Empirical results

Coverage and primary accuracy. All 108 task/seed allocations, 540 constant cost points and 972 complete-system baseline outcomes finished; all baseline outcomes succeeded. The primary estimate uses the 105 natural-data allocations, not the three generated-control allocations. Prediction replay verifies 25,740 policy risks and the stored serving decisions. Table 6 gives the complete-system comparison; Figures 1 and 2 address cost and library growth.

At full policy budget, the one-query CV gate has geometric MSE ratio 1.055 to global CP (descriptive interval $[1.024, 1.095]$), versus 1.087 for a constant individual. Direct continuous estimation, continuous residual correction and residual projection have ratios 1.229, 1.355 and 1.339. The two fixed primary paired comparisons give gate/GBM ratio 1.016 (Holm-adjusted sign-flip $p = .796$) and gate/original-AutoGluon ratio 1.240 ($p < 10^{-4}$). The first provides no clear difference, rather than evidence of equivalence. AutoGluon’s original and serialized ratios to CP are 0.851 and 0.850; both improve on CP in 31/33 families. The input-complete public V2 comparator gives 1.060, under its declared CPU/context limits. Source-quality and chemistry exclusions preserve these broad orderings (Table 7).

A cheaper gate is not a better frontier. The unpenalized gate uses 0.410 times CP latency, but retrained GBM has both lower MSE (1.038 versus 1.055) and lower latency (0.264). GBM also dominates every tested gate cost penalty at the aggregate medians. Soft MoE trades MSE ratio 1.070 for latency 0.189; hardening raises MSE to 2.926. Serialized AutoGluon attains ratio 0.850 at latency 14.62 times CP. Thus the evaluated gate contributes no point to the aggregate nondominated set. On individual families it is non-dominated in 6/33, compared with 23/33 for GBM; this is descriptive, not statistical separation of frontiers. At $\lambda = 0$, the gate improves on the fixed constant candidate by ratio 0.971, but validation chooses a constant fallback in 42.9% of source-balanced allocations. Conditional choice helps within the frozen bank without establishing an advantage over rebuilding a global predictor.

Prediction-feature shifts do not imply degradation.

Relative to the same method at $K = 8$, raw shared-gate MSE ratios at $K = 2, 4, 16$ are 1.071, 1.027 and 0.983 for the diverse extension; the weak extension gives 1.000. Augmented counterparts are 1.096, 1.050, 0.994 and 0.999. Normalized shared targets give weak- $K16/K8$ ratio 1.017. Meanwhile, prediction-feature split share rises from 17.1% at $K = 2$ to 51.9% for diverse- $K16$ and 57.3% for weak- $K16$; corresponding feature-count fractions are 5.6% and 23.1%. The split shift is real, but the raw/augmented weak-target controls, decision masks and fixed-feature target refits show no consistent accuracy penalty. The theorem establishes possibility, not prevalence. Distinct recipes remain correlated: the mean pairwise prediction correlation is 0.776 for diverse- $K16$, versus 0.181 for weak- $K16$. Weak additions are worse than the reference on 64.8% of test action–input pairs, so they are not pointwise dominated as in the theorem.

Regularization and policy-label budgets. For raw residual fits, separately selected clipping improves on continuous CV by ratio 0.970, versus 0.988 for projection. Projection’s matched-hyperparameter factor is 0.976, offset by a hyperparameter-choice factor 1.012. For raw direct fits, the projection ratio is 0.995; after augmentation it is 1.005. These results distinguish postprocessing from separately chosen shrinkage and leaf size; neither scalar sufficiency nor quantization ensures a regression gain. On a fixed 12-task panel eligible for all absolute budgets, gate/CP ratios at 100, 300, 1,000 and 3,000 policy labels are 1.091, 1.084, 1.044 and 1.048; residual projection gives 1.711, 1.565, 1.400 and 1.391. Appendix C includes full curves, mechanism factors and family effects. The original natural-development Library B projection/CP ratio, 0.986 ($p = .508$), remains contextual evidence rather than the primary estimate.

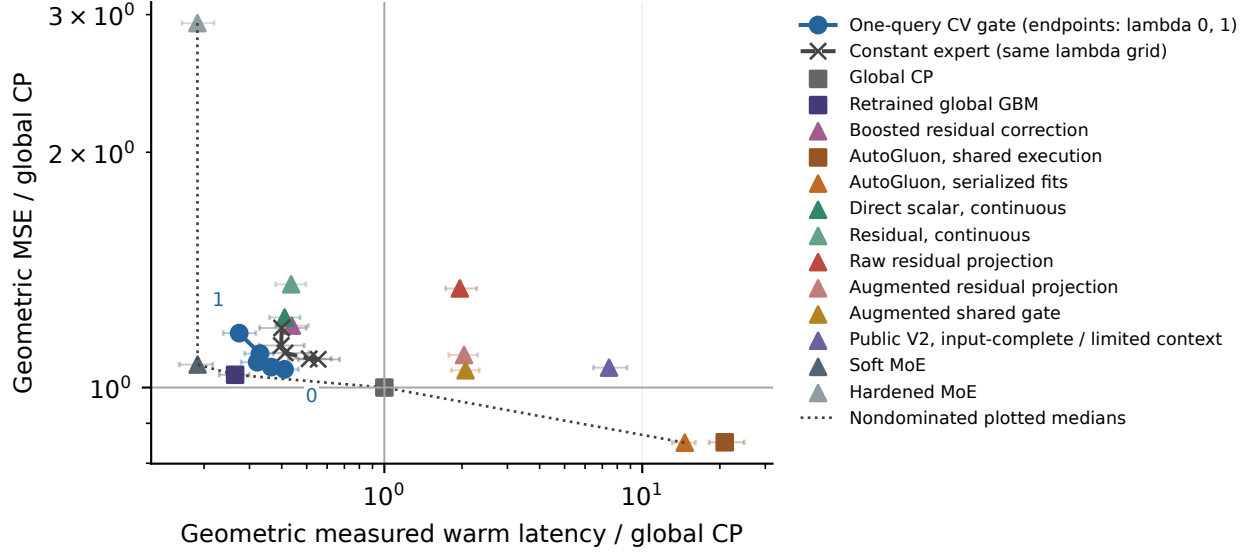


Figure 1. Accuracy versus complete-path warm batch-one latency on common coverage: 35 tasks, 33 families and 105 allocations. Both axes are family-balanced geometric ratios to global CP; lower is better. Gate and constant-expert curves use the five fixed cost penalties. Bars describe three observed timing repeats, not confidence intervals. The dotted set is nondominated only among plotted median-cost operating points on this shared M1 host.

6. Relationship to model-selection and deferral methods

Dynamic ensemble selection. Algorithm selection (Rice, 1976), ensemble selection (Caruana et al., 2004), stacking (Breiman, 1996; van der Laan et al., 2007; Sill et al., 2009), and dynamic regressor selection (Moura et al., 2019) already distinguish local competence from global prediction. META-DES (Cruz et al., 2015) combines several competence signals in a meta-classifier, including information derived from classifier behavior. Such signals can be useful while requiring multiple model outputs. Our raw/output-augmented comparison makes that serving dependency explicit. The nearest-output scalar projection is an existing competence rule; the question here is its relationship to shrinkage, the shared greedy objective, and inference cost.

Shared versus separate tree targets. The choice between joint and separate multi-output tree fits is established: Schmid et al. (2023) compare multivariate ensembles with separate univariate models; Nissenbaum & Painsky (2024) use cross-validation to choose shared or target-specific splits and explicitly normalize target scales. Breskvar et al. (2018) randomize output subsets within predictive-clustering ensembles. Our construction specializes this representation conflict to policy targets belonging to actions that are never optimal. Target normalization, independent fits, and limiting target sharing are existing remedies, rather than new methods introduced here. Their controls test whether the specified routing failure survives changes to

the shared objective.

Policy and cost-sensitive reductions. Doubly robust policy evaluation and learning (Dudík et al., 2011) addresses partially observed rewards under a logging policy; propensities and overlap matter there. In our offline fitting data, Y and all fixed expert predictions reveal every squared-loss action cost, so direct full-information cost estimation is available without importance weighting. The shared-contrast fit is a cost-estimation reduction, related in purpose to cost-sensitive classification reductions (Zadrozny et al., 2003). Its regression surrogate allocates capacity to target magnitudes, whereas policy regret depends on action ordering. Loss weighting and gradient interference are established multi-task concerns (Kendall et al., 2018; Yu et al., 2020); the contribution here is a routing-specific dominated-action construction and its learner-dependent separation. Our construction exhibits that mismatch for a particular greedy objective. It complements policy-tree optimization (Sverdrup et al., 2020; Zhou et al., 2023; Bertsimas & Peroni, 2024), rather than establishing a lower bound for all policy learners.

Multiple-expert deferral and model routing. Learning to defer to multiple experts (Verma et al., 2023) studies consistent classification surrogates, confidence calibration, and expert subsets. Its estimated correctness probabilities are different from the common conditional-mean structure of squared regression loss, although both require comparing expert-dependent risks. RouteLLM (Ong et al., 2025)

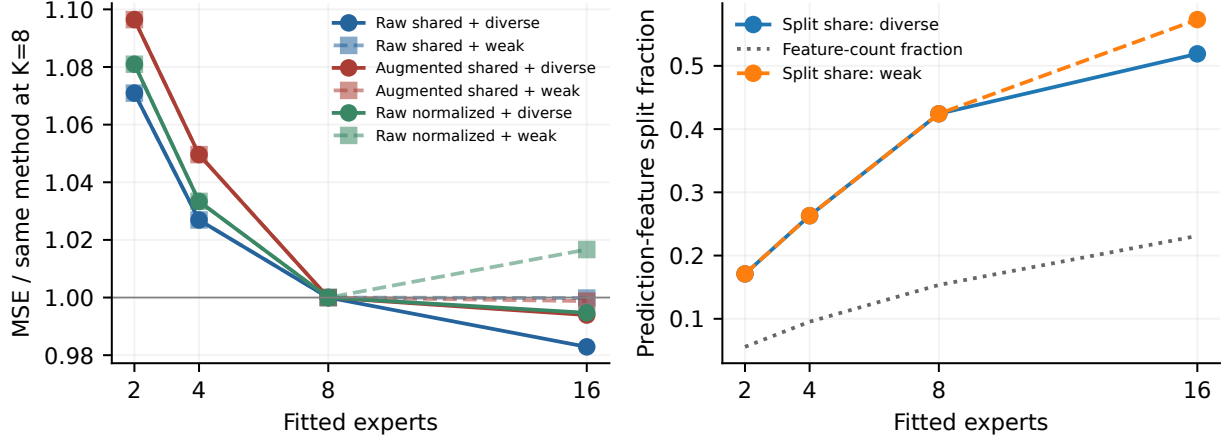


Figure 2. Full-policy library growth. Left: each method is normalized to its own $K = 8$ MSE, keeping its comparator fixed. Solid extensions add diverse recipes; dashed extensions add permuted-target models. Right: augmented-tree split shares and prediction-feature count fractions. Prediction-feature use grows without consistent raw/augmented degradation under weak additions. All summaries weight the same 33 natural-data families equally.

learns routing from preferences, and FrugalGPT (Chen et al., 2024) uses cost-aware language-model cascades. They motivate charging the information actually acquired before a decision. Their open-ended responses and sequential query options differ from our scalar full-information training problem, so their reported task scores are not regression baselines. Recent routing diagnostics and domain-mixture analyses (Shihab et al., 2026; Dann et al., 2025; Hao et al., 2026) likewise separate available oracle opportunity from what a trained selector can realize. Our principal empirical alternatives are complete tabular prediction systems, including AutoGluon (Erickson et al., 2020) and an explicitly resource-limited public TabPFN checkpoint (Hollmann et al., 2025).

Adaptive acquisition of predictions. Costly expert-advice models (Reyzin et al., 2026) include query acquisition in an online feedback protocol. Here training reveals all fixed-action losses, while deployment queries are restricted. LazyStack (Colaco et al., 2026) acquires predictions sequentially and learns when to stop and aggregate them. Adaptive query depth introduces a different serving decision from our gate, which commits to one expert before executing it. Our measured frontier compares the fixed systems evaluated here; extending the comparison to sequential acquisition would require a different serving policy and training protocol.

7. Implications and limitations

The studied one-query gate improves over a fixed individual expert but fails to improve the aggregate accuracy–cost frontier once a retrained global GBM is included. Auto-

Gluon provides a higher-cost accuracy alternative. This result concerns the declared frozen banks, limited gate grids, training-label allocations and CPU serving regime; it does not rule out other routing architectures, better expert specialization or different deployment costs. Family-level exceptions also matter. A practical evaluation should compare conditional selection with both a calibrated global ensemble and a predictor rebuilt using all allowed training labels.

The bounded construction identifies a failure of a shared greedy objective, including a redundant-target variant that survives normalization. The real-data controls do not establish that this failure is prevalent: weak additions have almost no aggregate effect on raw/augmented gates, even when prediction-feature split share increases. Their imperfect domination, correlated useful libraries and differing feature information separate these data from the theorem. The standard Bregman and margin tools describe statistical structure, while regularization and output acquisition determine what an implemented scalar alternative can gain.

The capped benchmark panel is a fixed selection rather than a random sample of future applications. Three seeds share outer tests. Source provenance, group/time constraints, native baseline validation, model clipping, and unequal component training labels limit generalization. The public V2 checkpoint is older and context-limited; AutoGluon’s issued wall-time allowance is not an exact compute match. Shared-host warm timing repeats are descriptive, and serializing fits did not isolate the machine. The local protocol and later amendments are transparent but supply no external preregistration. These limits favor the narrow supported conclusion: on this panel and at the evaluated operating

points, selecting an expert per input does not improve on strong complete-system alternatives.

Impact Statement

This work concerns regression benchmark methods and does not validate automated decisions about people. Application-specific costs, group effects, target meaning and prospective data collection need separate assessment before deployment. A calibration fallback is an empirical choice rather than a per-person or per-task risk guarantee.

References

- Audibert, J.-Y. and Tsybakov, A. B. Fast learning rates for plug-in classifiers. *The Annals of Statistics*, 35(2):608–633, 2007. doi: 10.1214/009053606000001217. URL <https://arxiv.org/abs/0708.2321>.
- Banerjee, A., Guo, X., and Wang, H. On the optimality of conditional expectation as a Bregman predictor. *IEEE Transactions on Information Theory*, 51(7):2664–2669, 2005. doi: 10.1109/TIT.2005.850145.
- Bertsimas, D. and Peroni, M. Policy trees for prediction: Interpretable and adaptive model selection for machine learning, 2024. URL <https://arxiv.org/abs/2405.20486>.
- Breiman, L. Stacked regressions. *Machine Learning*, 24(1):49–64, 1996. doi: 10.1007/BF00117832.
- Breskvar, M., Kocev, D., and Džeroski, S. Ensembles for multi-target regression with random output selections. *Machine Learning*, 107:1673–1709, 2018. doi: 10.1007/s10994-018-5744-y. URL <https://doi.org/10.1007/s10994-018-5744-y>.
- Caruana, R., Niculescu-Mizil, A., Crew, G., and Ksikes, A. Ensemble selection from libraries of models. In *Proceedings of the Twenty-First International Conference on Machine Learning*, 2004. doi: 10.1145/1015330.1015432.
- Chen, L., Zaharia, M., and Zou, J. FrugalGPT: How to use large language models while reducing cost and improving performance. *Transactions on Machine Learning Research*, 2024. URL <https://openreview.net/forum?id=cSimKw5p6R>.
- Colaco, A. G., Mehrotra, S., De Lucia, M. J., Hamlen, K., Kantarcioglu, M., Khan, L., Swami, A., Thuraisingham, B., and Jain, U. Unifying stacking and cascading for efficient ensemble inference. In *Proceedings of the 43rd International Conference on Machine Learning*, volume 306 of *Proceedings of Machine Learning Research*, pp. 21176–21211, 2026. URL <https://proceedings.mlr.press/v306/colaco26a.html>.
- Cruz, R. M. O., Sabourin, R., Cavalcanti, G. D. C., and Ren, T. I. META-DES: A dynamic ensemble selection framework using meta-learning. *Pattern Recognition*, 48(5):1925–1935, 2015. doi: 10.1016/j.patcog.2014.12.003.
- Dann, C., Mansour, Y., Marinov, T. V., and Mohri, M. Principled model routing for unknown mixtures of source domains. In *Advances in Neural Information Processing Systems*, 2025. URL https://papers.nips.cc/paper_files/paper/2025/hash/39c60dda48ebf0a2e5dda52ce08eb5c8-Abstract-Conference.html.
- Dudík, M., Langford, J., and Li, L. Doubly robust policy evaluation and learning. In *Proceedings of the 28th International Conference on Machine Learning*, pp. 1097–1104, 2011. URL https://icml.cc/2011/papers/554_icmlpaper.pdf.
- Erickson, N., Mueller, J., Shirkov, A., Zhang, H., Larroy, P., Li, M., and Smola, A. AutoGluon-Tabular: Robust and accurate AutoML for structured data, 2020. URL <https://arxiv.org/abs/2003.06505>.
- Erickson, N., Purucker, L., Tschalzev, A., Holzmüller, D., Desai, P. M., Salinas, D., and Hutter, F. TabArena: A living benchmark for machine learning on tabular data. In *Advances in Neural Information Processing Systems*, volume 38, 2025. URL <https://arxiv.org/abs/2506.16791>.
- Fischer, S., Harutyunyan, L., Feurer, M., and Bischl, B. OpenML-CTR23: A curated tabular regression benchmarking suite. In *Workshop at the International Conference on Automated Machine Learning*, 2023. URL <https://openreview.net/forum?id=HebAOoMm94>.
- Hao, S., Zeng, H., Wei, H., and Jing, B. RACER: Risk-aware calibrated efficient routing for large language models. In *Proceedings of the 43rd International Conference on Machine Learning*, volume 306, pp. 40499–40524, 2026. URL <https://proceedings.mlr.press/v306/hao26c.html>.
- Hollmann, N., Müller, S., Purucker, L., Krishnakumar, A., Körfer, M., Hoo, S. B., Schirrmeister, R. T., and Hutter, F. Accurate predictions on small data with a tabular foundation model. *Nature*, 637:319–326, 2025. doi: 10.1038/s41586-024-08328-6.
- Kendall, A., Gal, Y., and Cipolla, R. Multi-task learning using uncertainty to weigh losses for scene geometry and semantics. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pp. 7482–7491, 2018. URL <https://openaccess.thecvf>.

- com/content_cvpr_2018/html/Kendall_Multi-Task_Learning_Using_CVPR_2018_paper.html.
- Moura, T. J. M., Cavalcanti, G. D. C., and Oliveira, L. S. Evaluating competence measures for dynamic regressor selection. In *2019 International Joint Conference on Neural Networks (IJCNN)*, pp. 1–8, 2019. doi: 10.1109/IJCNN.2019.8851835.
- Nissenbaum, Y. and Painsky, A. Cross-validated tree-based models for multi-target learning. *Frontiers in Artificial Intelligence*, 7:1302860, 2024. doi: 10.3389/frai.2024.1302860. URL <https://doi.org/10.3389/frai.2024.1302860>.
- Ong, I., Almahairi, A., Wu, V., Chiang, W.-L., Wu, T., Gonzalez, J. E., Kadous, M. W., and Stoica, I. RouteLLM: Learning to route LLMs from preference data. In *International Conference on Learning Representations*, 2025. URL https://proceedings.iclr.cc/paper_files/paper/2025/file/5503a7c69d48a2f86fc0b3dc09de686-Paper-Conference.pdf.
- Reyzin, L., Saha, A., and Wu, S. Expert advice with costly observations. In *Proceedings of the 42nd Conference on Uncertainty in Artificial Intelligence*, volume 337 of *Proceedings of Machine Learning Research*, pp. 5700–5715, 2026. URL <https://proceedings.mlr.press/v337/reyzin26a.html>.
- Rice, J. R. The algorithm selection problem. *Advances in Computers*, 15:65–118, 1976.
- Schmid, L., Gerharz, A., Groll, A., and Pauly, M. Tree-based ensembles for multi-output regression: Comparing multivariate approaches with separate univariate ones. *Computational Statistics & Data Analysis*, 179: 107628, 2023. doi: 10.1016/j.csda.2022.107628. URL <https://doi.org/10.1016/j.csda.2022.107628>.
- Shihab, I. F., Ahsan, A. S.-A. M. M.-I. A., and Swaqeeb, M. N. Opportunity is not realizability: Selection-valid diagnostics for multi-LLM routing, 2026. URL <https://arxiv.org/abs/2608.08265>.
- Sill, J., Takács, G., Mackey, L., and Lin, D. Feature-weighted linear stacking, 2009. URL <https://arxiv.org/abs/0911.0460>.
- Sverdrup, E., Kanodia, A., Zhou, Z., Athey, S., and Wager, S. policytree: Policy learning via doubly robust empirical welfare maximization over trees. *Journal of Open Source Software*, 5(50):2232, 2020. doi: 10.21105/joss.02232.
- van der Laan, M. J., Polley, E. C., and Hubbard, A. E. Super learner. *Statistical Applications in Genetics and Molecular Biology*, 6(1):Article 25, 2007. doi: 10.2202/1544-6115.1309.
- Verma, R., Barrejon, D., and Nalisnick, E. Learning to defer to multiple experts: Consistent surrogate losses, confidence calibration, and conformal ensembles. In *Proceedings of the 26th International Conference on Artificial Intelligence and Statistics*, volume 206, pp. 11415–11434, 2023. URL <https://proceedings.mlr.press/v206/verma23a.html>.
- Yu, T., Kumar, S., Gupta, A., Levine, S., Hausman, K., and Finn, C. Gradient surgery for multi-task learning. In *Advances in Neural Information Processing Systems*, volume 33, 2020. URL https://proceedings.neurips.cc/paper_files/paper/2020/hash/3fe78a8acf5fda99de95303940a2420c-Abstract.html.
- Zadrozny, B., Langford, J., and Abe, N. Cost-sensitive learning by cost-proportionate example weighting. In *Third IEEE International Conference on Data Mining*, pp. 435–442, 2003. doi: 10.1109/ICDM.2003.1250950.
- Zhou, Z., Athey, S., and Wager, S. Offline multi-action policy learning: Generalization and optimization. *Operations Research*, 71(1):148–183, 2023. doi: 10.1287/opre.2022.2271.

A. Proofs and scope of the separation

A.1. Bregman decomposition

Put $m = \mathbb{E}[Y \mid X]$. Expanding the divergence yields

$$D_\phi(Y, f) - D_\phi(Y, m) - D_\phi(m, f) = \langle \nabla \phi(m) - \nabla \phi(f), Y - m \rangle.$$

Its conditional expectation is zero. This proves Equation (4), including for vector-valued outcomes. Squared loss has $\phi(u) = u^2$. Twice the divergence generated by $\phi(u) = u \log u - u$ gives Poisson deviance, with $0 \log 0 = 0$ and positive predictions. Negative entropy on a probability simplex gives categorical KL. The mean and estimated mean must remain in the relevant domain; projection onto that domain is part of the estimator if needed. These are applications of the established conditional-expectation characterization (Banerjee et al., 2005).

A.2. Proof of Theorem 1

Condition on the fitted action bank and mean estimator. Write $a^* = \arg \min_a q_a$ and $\hat{a}$ for the plug-in minimizer. The common $\phi(m)$ terms cancel in score differences. Because the estimated ordering favors $\hat{a}$,

$$\begin{aligned} 0 \leq g(x) &:= q_{\hat{a}}(x) - q_{a^*}(x) \\ &\leq |\langle \hat{m} - m, \nabla \phi(f_{\hat{a}}) - \nabla \phi(f_{a^*}) \rangle| \leq Le(x), \end{aligned}$$

where $e(x) = \|\hat{m}(x) - m(x)\|$. If $g > 0$, the chosen action is suboptimal, so $\Delta \leq g \leq Le$. Known action costs cancel from the estimation error because the same costs appear in both comparisons. Hence $g \leq Le \mathbf{1}\{\Delta \leq Le\}$.

If $e \leq \epsilon$ uniformly, integration gives $\mathbb{E}g \leq L\epsilon \Pr(\Delta \leq L\epsilon) \leq C(L\epsilon)^{\alpha+1}$. If instead $\mathbb{E}e^2 \leq \eta$, for every $t > 0$,

$$\begin{aligned} \mathbb{E}g &\leq Lt \Pr(\Delta \leq Lt) + L\mathbb{E}[e \mathbf{1}\{e > t\}] \\ &\leq CL^{\alpha+1}t^{\alpha+1} + L\eta/t. \end{aligned}$$

For $C, L, \eta > 0$, substituting $t = (\eta/((\alpha+1)CL^\alpha))^{1/(\alpha+2)}$ gives Equation (6). Zero-error cases follow directly.

For the conditional tail assumption, the region $e \leq s_n$ contributes at most $CL^{\alpha+1}s_n^{\alpha+1}$. On the j th dyadic event $2^j s_n < e \leq 2^{j+1} s_n$, bound g by $L2^{j+1}s_n$, restrict X to $\Delta \leq L2^{j+1}s_n$, and use the tail bound uniformly for every such x . The contribution is at most

$$ACL^{\alpha+1}s_n^{\alpha+1}2^{(j+1)(\alpha+1)} \exp(-a4^j).$$

Summing proves the result with finite constant $CL^{\alpha+1}[1 + A \sum_{j \geq 0} 2^{(j+1)(\alpha+1)} e^{-a4^j}]$. If $L = 0$, the score ordering is independent of the estimated mean and the regret is zero. Otherwise, if a fixed gap $\Delta \geq \delta > 0$ holds almost surely, the same conditional tail assumption gives an exponential bound. Put $t = \delta/L$. Then

$$\begin{aligned} \mathbb{E}g &\leq L\mathbb{E}[e \mathbf{1}\{e \geq t\}] \\ &\leq A \left[\delta + \frac{L^2 s_n^2}{2a\delta} \right] \exp\left(-\frac{a\delta^2}{L^2 s_n^2}\right). \end{aligned}$$

Here $\mathbb{E}[e \mathbf{1}\{e \geq t\}] = t \Pr(e \geq t) + \int_t^\infty \Pr(e > u) du$, and the Gaussian-tail integral is at most $s_n^2 e^{-at^2/s_n^2}/(2at)$. This is a conditional implication, with fixed constants and a uniform tail over inputs. It is not an exponential-rate claim for an arbitrary fitted forest.

The argument follows the plug-in/margin approach of Audibert & Tsybakov (2007); assumptions on the mean estimator are substantive. For squared loss, $\nabla \phi(f) = 2f$ gives $L \leq 4M$. For twice Poisson divergence, the gradient difference is $2 \log(f_a/f_b)$, giving the stated bounded-domain constant.

A.3. Proof of Theorem 2

Put $c_j = 1 + 4Z_j$. The added action is a fixed function of $X = (S, Z)$ and satisfies

$$(f_{j+2}(X) - Y)^2 - (p - Y)^2 = c_j \geq 1.$$

Thus domination holds at every support point. Since $(p - S)^2 \leq 1$ and $c_j \leq 5$, every added prediction is at most $1 + \sqrt{6}$, uniformly in the number of actions. The useful contrasts are

$$D_1 = 2pS - p^2, \quad D_2 = 1 - p^2 - 2(1 - p)S.$$

At a node that fixes some nuisance bits and leaves S unconditioned, all remaining bits remain independent. A split on an unused Z_j changes its own added contrast mean by 4 and changes no other contrast mean. Its sum-of-variance reduction is therefore $4^2/4 = 4$. Splitting on S leaves all added contrasts unchanged and gains

$$G_S = 4p(1 - p)\{p^2 + (1 - p)^2\} \leq 1.$$

A common average over targets preserves this ordering. Induction over depth shows that every path of a depth- d greedy tree spends all d splits on nuisances. In each leaf the useful mean contrasts are p^2 and $(1 - p)^2$, all added mean contrasts are at least 1, and the zero reference contrast is the unique minimum. The resulting risk is $\mathbb{E}(p - S)^2 = p(1 - p)$. Averaging any finite collection of these population score trees preserves positivity and the selected reference.

Without added targets, only a split on S has positive gain and the resulting useful action is correct on both leaves. A scalar fit also splits on S : $Y = S$ has zero conditional variance on its children, whereas every independent nuisance leaves its variance unchanged. One additional shared-tree level permits the signal split and recovers zero risk. For binary inputs, all-coordinate ExtraTrees considers the same candidate partitions, regardless of threshold draws within $(0, 1)$.

For the empirical scalar bound, if both values of S are observed, splitting on S has the maximum possible scalar impurity reduction. A nuisance can tie it only if its sampled column equals S or $1 - S$ on every training example. Conditional on the sampled S column, each such event has probability 2^{-n} , so their union has probability at most $2d2^{-n}$. Apart from that event the scalar split is uniquely S and generalizes perfectly. If only zeros are observed, the prediction is zero and test risk is p ; this occurs with probability $(1 - p)^n$. If only ones are observed, risk is $1 - p$ and the event probability is p^n . All scalar leaf means lie in $[0, 1]$; every added action exceeds 1, so nearest-output projection selects one of $(p, 0, 1)$ and squared loss is at most 1. Summing the three bounds proves Equation (8), for arbitrary tie rules.

A.4. An explicit finite-sample shared-learner bound

Write $q = 1 - p$, $m_0 = \min\{p, q\}$ and $\delta = 1/(5\sqrt{d})$. Let $\mathcal{E}$ be the event that every input atom a satisfies

$$|\hat{P}(a) - P(a)| \leq \delta P(a).$$

There are 2^{d+1} atoms, each with probability at least $m_0 2^{-d}$. Multiplicative Chernoff bounds and a union bound give

$$\Pr(\mathcal{E}^c) \leq 2^{d+2} \exp\left\{-\frac{nm_0}{75d2^d}\right\}. \quad (11)$$

On $\mathcal{E}$, every atom is observed. For any support cells $V \subseteq U$ of positive population mass, summing the atom bounds gives

$$\frac{1 - \delta}{1 + \delta} P(V | U) \leq \hat{P}(V | U) \leq \frac{1 + \delta}{1 - \delta} P(V | U).$$

In a nuisance-only node N , an unused Z_j has population probability $1/2$, both conditional on N and conditional on $(N, S = s)$. Each corresponding empirical probability therefore differs from $1/2$ by at most

$$e = \frac{\delta}{1 - \delta} \leq \frac{1}{4\sqrt{d}}.$$

All subsequent gains are conditional on N and use the summed target criterion. Splitting on Z_j reduces its own contrast's variance by

$$16\widehat{\text{Var}}(Z_j | N) \geq 4 - 16e^2 \geq 4 - 1/d \geq 3.$$

Other coordinates' contributions are nonnegative. In contrast, the useful coordinates contribute at most $4(p^2 + q^2)\widehat{\text{Var}}(S | N) \leq 1$ to a signal split. Each unused weak contrast contributes

$$16\widehat{\text{Var}}(S | N) [\widehat{\mathbb{E}}(Z_j | N, S = 1) - \widehat{\mathbb{E}}(Z_j | N, S = 0)]^2.$$

The difference in brackets is at most $2e$ in absolute value. Summing over at most d unused coordinates bounds their total contribution by $16de^2 \leq 1$. Already fixed nuisance contrasts contribute zero. Hence every nuisance-only node with an unused nuisance has a nuisance split of gain at least 3, while its signal split has gain at most 2. Induction forces all depth- d splits to use nuisances.

In each resulting leaf, the empirical signal probability is at least $2p/3$, and its complement is at least $2q/3$, since $(1-\delta)/(1+\delta) \geq 2/3$. The useful mean contrasts are therefore at least $p^2/3$ and $q^2/3$; every added contrast remains at least one. The reference is the unique selected action. Its fresh-input risk is pq throughout $\mathcal{E}$. Nonnegativity of risk off this event proves Equation (9). Any finite forest fitted to the same full sample obeys the same gain and positivity bounds on this common event, irrespective of nuisance tie choices. The displayed finite-sample bound does not cover bootstrap resampling.

For fixed p and d , finite-support cell moments and all gain comparisons converge; a fixed finite forest using size- n bootstrap samples has the same limiting selection and risk. The bound remains conservative and depends exponentially on d through the support size; it is not a dimension-uniform guarantee. Feature subsampling may expose S without its competitors and requires a different analysis.

A.5. Proof of Corollary 3

The added contrasts are $1 + 4Z_j + r/4$, so their means and standard deviations are $3 + r/4$ and 2. After population centering/scaling, all three targets associated with Z_j are $2Z_j - 1$. Each has variance one. The two useful normalized contrasts are $(S - p)/\sqrt{p(1-p)}$ and its negative, also of variance one. At every nuisance-only node, an unused nuisance split removes total variance three and a signal split removes total variance two. The greedy learner therefore exhausts its depth on nuisances. Inverse transformation recovers the original positive non-reference leaf means, so it chooses f_0 . The added predictions are bounded by $1 + \sqrt{13/2} < 4$ and their original contrasts are at least one. Thus the margin and action bounds remain independent of d . Consistent empirical means and nonzero scales, finite support, and the strict gain gap yield the same empirical limit. The fixed finite forest statement follows by averaging before inverse transformation, which commutes with a common affine transformation of each target.

A.6. Proof of Corollary 4

Let $k = \lfloor \log_2(1/\rho) \rfloor$. A nuisance-only cell after j splits has mass 2^{-j} . For $j < k$, another nuisance split is feasible and dominates the signal split by the preceding gain calculation. After k splits the leaf mass $q = 2^{-k}$ lies in $[\rho, 2\rho)$. Another nuisance would produce children of mass $q/2 < \rho$. Splitting on S would produce a rare child of mass $pq < \rho$, since $p < 1/2$. No split is feasible. All leaves still have signal probability p and select the reference. At the root, however, a scalar split on S is feasible because $\rho \leq p < 1/2$. It achieves zero risk. Away from equality boundaries, empirical feasibility and gain comparisons converge to these decisions. The checked examples use $p = .1$ and $(\rho, d) = (.01, 6), (.05, 4)$.

A.7. Finite-variance outcome noise

Proposition 5 (Noise preserves the raw shared-objective separation). *Retain X , the actions, and the raw depth-limited learners of Theorem 2, but let $Y = S + \xi$, where $\mathbb{E}[\xi | X] = 0$ and $v = \mathbb{E}\xi^2 < \infty$. The added actions are conditionally dominated by f_0 . The conditional action gap remains p^2 for $S = 0$ and $(1-p)^2$ for $S = 1$. The population shared learner and any fixed finite forest of these trees have risk $v + p(1-p)$; the population scalar stump followed by projection has risk v . For empirical fits from i.i.d. observations, the corresponding population risks converge to these values in probability and in expectation. Forests use a fixed number of trees and size- n bootstrap samples, if bootstrapping is used. These are asymptotic statements for the raw targets; the noiseless finite-sample inequalities and normalized-target construction are separate results.*

Proof. Let D_a^0 denote a contrast for the noiseless outcome S . Expanding the squared-loss contrast gives

$$D_a = D_a^0 - 2\xi(f_a - f_0), \quad \mathbb{E}[D_a | X] = D_a^0.$$

Thus every input cell and candidate split has the same conditional target means as in the noiseless construction. A squared-error split's gain is the sum of the between-child variances of those means; changes in the within-cell noise variances cancel by total variance. Every candidate gain and its strict nuisance-versus-signal ordering are therefore unchanged. The shared tree spends its depth on nuisances and selects f_0 , while the scalar stump splits on S , estimates its conditional mean, and

selects the action equal to S . For any input-measurable selected action,

$$\mathbb{E}[(f_{\pi(X)} - Y)^2] = v + \mathbb{E}[(f_{\pi(X)} - S)^2].$$

This proves the population risks, conditional domination and unchanged margin, including for heteroskedastic noise satisfying the displayed assumptions.

For empirical fits, the finite input support has positive atom probabilities. Cell frequencies and response/contrast means converge, and all raw contrasts have finite second moments because the actions are bounded. The finite set of split gains therefore converges. Strict comparisons exclude signal splits from shared trees of depth d , and the positive useful leaf scores eventually select the reference with probability tending to one. The scalar root's strict signal gain and converging leaf means give the correct projected action with probability tending to one. A fixed finite forest preserves these conclusions; bootstrap cell moments also converge in probability. Finally, the selected actions are uniformly bounded. Their excess risks above v are uniformly bounded, so convergence in probability implies convergence of their expected risks to the same limits. $\square$

Corollary 6 (Normalization with homoskedastic outcome noise). *Use the three-actions-per-nuisance library of Corollary 3 and $Y = S + \xi$, with $\mathbb{E}[\xi | X] = 0$ and $\mathbb{E}[\xi^2 | X] = v < \infty$. Center and standardize each nonconstant raw contrast using its population mean and variance. A depth- d shared greedy tree, followed by inverse transformation before selection, has risk $v + p(1 - p)$. A fixed finite forest has the same population risk, and, under the i.i.d. fitting conditions of Proposition 5, consistent empirical target normalization gives the same limit. This result assumes constant conditional variance; Proposition 5's arbitrary heteroskedastic extension concerns the raw targets.*

Proof. Put $q = 1 - p$, $t = S - p$, $c = 1 + 4Z_j + r/4$, and $A_r = \mathbb{E}(f_{j,r} - p)^2$. By Cauchy–Schwarz and $pq \leq 1/4$, $\mathbb{E}c \leq 7/2$,

$$\begin{aligned} A_r &= 2pq + \mathbb{E}c + 2\mathbb{E}[t\sqrt{t^2 + c}] \\ &\leq 2pq + \mathbb{E}c + 2\sqrt{pq(pq + \mathbb{E}c)} \\ &\leq 4 + \sqrt{15}/2 < 6. \end{aligned}$$

Total variance gives $\text{Var}(D_{j,r}) = 4 + 4vA_r$. At a nuisance-only cell, splitting on the unused Z_j therefore gains $\sum_{r=0}^2 (1 + vA_r)^{-1}$ across its three normalized targets. The useful contrasts have variances $4p^2(pq + v)$ and $4q^2(pq + v)$; their normalized signal-split gain is $2pq/(pq + v)$. Consequently,

$$G_{Z_j} \geq \frac{3}{1 + 6v} > \frac{2}{1 + 4v} \geq \frac{2pq}{pq + v} = G_S.$$

These strict inequalities persist at every nuisance-only node. The shared learner exhausts its depth on nuisances. Inverse transformation restores the original positive non-reference conditional means, so it selects the reference. The risk identity in Proposition 5 gives the conclusion. Finite second moments, positive atom probabilities and nonzero target variances ensure consistent empirical scales and the same limiting split/selection decisions. The fixed finite forest conclusion follows as before. $\square$

A.8. Exact-support verification

Weighted CART and 16-tree ExtraTrees fits verify the depth result for $d = 1, \dots, 8$ and $p \in \{.1, .25, .5, .75, .9\}$; allowed depth $d + 1$ recovers zero risk. The redundant-target normalized construction is checked for all eight depths and $p \in \{.1, .5, .9\}$. Separate weighted fits with no allowed-depth cap check the two leaf-fraction cases. For the noisy extension, conditionally symmetric two-point noise is checked at depths 1, 2, 4, 6, $p \in \{.1, .5, .9\}$, and base magnitudes .05, .5, 2, using both constant and input-dependent scales (72 cases). The 36 constant-variance cases also check the normalized redundant-target result. These are population calculations on the full weighted support. Separately, 80 integer-count weighted fixtures check the finite-sample relative-atom event implication at depths 1, $\dots, 8$ and $p \in \{.1, .25, .5, .75, .9\}$, with one multinomial sample and one bounded pairwise mass-transfer sample per setting. Weighted CART and non-bootstrap ExtraTrees fits select the reference; explicit node gains satisfy the bounds above. Integer weights encode sample multiplicities without expanding repeated rows. These fixtures do not estimate the probability of the relative-atom event. All these checks can detect algebra/implementation discrepancies but do not substitute for the proofs or establish that a real dataset has this mechanism.

B. Source panel, models, and counting units

B.1. Audit trail and source limitations

The pinned TabArena repository revision is 95820a2cec37174fe3e25da4df80e11ae76b8761; the public BeyondArena data snapshot is 2ecfe882ccfb814fc27c4de10a64ceefd5d7655c. OpenML suite IDs are 353 and 269; individual dataset/task versions, official split indices, processed-data hashes, source rows and role masks are retained. The prepared-data table includes the generated insurance control separately from the 35-task natural-data primary panel. Dataset-level source metadata contains the source’s license, citation and curation comments. A public benchmark’s availability does not establish scientific representativeness or resolve all dependence.

Two provenance/schema amendments affect the original local lock. First, the 1,338-row healthcare-insurance table has exactly the original textbook’s covariate-row multiset and log-charge multiset; the textbook describes its data as simulated. It was excluded from the natural primary before any model fit on that source, after three Fiat allocations had completed. Second, the NYC taxi description says to ignore total amount because it contains the tip target, but its dataset-version API metadata omits that ignore field. We remove total amount before any taxi fit. Original hashes, reconstructed identical pre-amendment source bytes, and the corrected schema are retained. Neither amendment depends on fitted scores for the affected source. They remain amendments rather than being concealed by replacing the original lock.

Food-delivery metadata gives a Kaggle origin and retains delivery-person information under an IID split. Its task concerns new deliveries with the supplied information, not generalization to unseen drivers. Some anonymized AMLB sources have limited documentation, and the CTR23 paper itself questions the source descriptions of Yolanda and quake and the dimension/provenance match for the two 41-compound tasks. These limitations motivate source-quality sensitivity as well as chemistry-family aggregation. The main 35-task panel is a fixed collection of benchmark tasks with known simulators removed; it is not a claim of independently established natural provenance for every row. Unknown compound-level overlap between MetaQSAR assays is not resolved by family aggregation. No scaffold-generalization result is asserted.

For curated targets, logarithms or reference normalization supplied by the source are part of the task. For example, the FPS task predicts log relative performance under the curator’s hardware reference. MSE ratios refer to that supplied task scale, rather than silently inverting a curated logarithm into a different application loss. Saved risks use the declared E -standardized units; multiplying by the stored target standard deviation squared recovers MSE on the supplied scale. The new scripts do no additional outcome transformation beyond E -fitted centering and scaling. Learned imputation and one-hot encoding use only E for the fixed expert bank and its gates; complete-system global alternatives may refit their preprocessing on their entire allowed $E + C + P$ training set.

B.2. Fixed model recipes and validation

The eight initial experts are ridge ($\alpha = 1$), XGBoost (100 trees, depth 3), random forest (100 trees, minimum leaf size 2), histogram boosting (100 rounds, 15 leaves), ExtraTrees (100 trees, minimum leaf size 2), an MLP (64/32 hidden units, $\alpha = .01$, 250 iterations), Nyström RBF ridge (up to 128 components, $\gamma = .1/p_{\text{enc}}$, ridge penalty 10), and standardized 10-nearest neighbors. Here p_{enc} is the encoded feature count. The alternatives use ridge penalty 100, XGBoost depth 6, forest leaf size 10, histogram boosting with 31 leaves, ExtraTrees leaf size 10, MLP $\alpha = 1$, RBF $\gamma = 1/p_{\text{enc}}$, and 30 neighbors. The learning rate is .05 for the boosted recipes; tree/linear fits use one numerical thread. KNN neighbor counts and Nyström components are capped by the available expert rows. Prediction clipping uses the expert-role target range in each estimator’s native output precision; XGBoost retains float32 values and correspondingly rounded thresholds. Warning messages, including incomplete MLP convergence, are retained.

The deliberately weak additions fit independent permutations of expert-role targets: four ExtraTrees models with 32 trees and minimum leaf sizes 2, 10, 30 and 50; two ridge models with penalties 1 and 100; a depth-2 tree; and histogram boosting with 64 rounds and seven leaves. They are real fitted models. Their correlation, realized test risk and hindsight-minimum contribution are reported; permuting labels does not mathematically guarantee pointwise domination. The theorem instead imposes exact domination.

The generic global comparator has candidate individual experts, their mean, four ridge penalties (.1, 1, 10, 100), and a simplex-constrained least-squares stack. Three-fold grouped validation on $C + P$ selects a candidate, which is refitted on the full allowed calibration/policy sample. A numerical weight threshold of 10^{-10} defines dependencies for serving; simplex weights are renormalized after that threshold. A failed convex solve is a retained failure, not an unreported estimator substitution.

Task	Role	E	C	P	Test
Fiat500 prices	I	512	128	385	513
Food delivery	I	4000	1000	3000	4000
Insurance (generated control)	I	446	111	335	446
Miami housing	I	4000	1000	3000	4000
MetaQSAR TID11	I	1913	478	1436	1914
Allstate claims	I	4000	1000	3000	4000
Santander transactions	I	1482	370	1112	1483
Mercedes manufacturing	T	662	165	497	320
Sberbank housing	T	2463	948	4589	2809
Rossmann sales	T	4063	1023	2914	4000
Garment productivity	T	599	131	443	24
Early learning	G	3932	927	2926	4000
5G energy	G	3981	948	2958	4000
SAT11 runtime	G	980	250	740	990
Video transcoding	G	3717	895	2225	4000
FPS hardware	G	3987	1006	2998	4000
Music origin	I	476	119	358	106
CPS wages	I	4000	1000	3000	2816
Social mobility	I	520	130	390	116
King County housing	I	4000	1000	3000	2162
Wives' work hours	I	4000	1000	3000	2228
FIFA	I	4000	1000	3000	1918
Moneyball	I	554	138	416	124
Cars	I	361	90	272	81
Space GA	I	1398	349	1049	311
Twitter buzz	I	4000	1000	3000	4000
Yolanda	I	4000	1000	3000	4000
Quake	I	980	245	735	218
Sensory	I	259	64	195	58
Black Friday	I	4000	1000	3000	4000
Drug design topo	I	3998	999	2999	889
Drug design yprop	I	3998	999	2999	889
Colleges	I	3178	794	2384	707
NYC taxi tips	I	4000	1000	3000	4000
MetaQSAR TID10980	I	2594	649	1946	577
MIP solver runtime	G	435	110	330	215

Table 1. Prepared source panel. I/G/T indicate IID rows, grouped entities, or chronological roles. Counts show the minimum across three seeds for E,C,P and the fixed outer test size; full role sizes and masks are retained. Insurance is generated and excluded from the natural primary.

Policy forests have 80 trees and minimum leaf $\max\{10, \lceil \ell n \rceil\}$ for $\ell \in \{.01, .05\}$. All encoded raw features are available at each split. For each cost coefficient, policy-only validation compares these fits with a constant expert selected using calibration plus the training-policy fold. It never selects a constant ensemble as the one-query fallback. Reference contrast zero is enforced at prediction time. For target normalization, training-policy target means/scales are inverted before choosing the action; scales below 10^{-8} are set to one. Direct/residual scalar fits have identical raw fitting problems across library sizes. A checked cache reuses those fitted models and OOF estimates without changing library-specific shrinkage or projection choices.

The five constant-expert operating points select $\arg \min_a \{\widehat{R}_{C+P}(a) + \lambda c_a\}$ on the same fixed coefficient grid as the gates. They fit no new expert or gate. Identical action choices reuse identical serving measurements. This descriptive curve tests whether conditional selection improves the tradeoff over a fixed cheap action.

The replacement global and residual GBM controls use 200 histogram-boosting rounds, 31 leaves, learning rate .05 and L_2 penalty 1, fitted with all training labels. AutoGluon 1.6.1 fits GBM, extra-randomized GBM, CatBoost, random forest, ExtraTrees and a Torch neural network under its nominal training-time allowance, with three bagging folds and one stacking level. The setting `num_cpus=1` is a resource request, not an operating-system CPU quota; native thread activity is not proven confined to one core. Actual training time, selected dependencies and failures are retained.

The four independently parameterized neural experts each have 64/32 ReLU hidden units. The gate has a 32-unit ReLU hidden layer and four softmax outputs. Joint training minimizes mixture MSE plus .01 times the squared deviation of mean gate weights from uniform. Adam uses learning rate .001 and weight decay .001, with minibatches of at most 256 rows. A 10% grouped validation split selects the epoch count, with at most 100 epochs, patience 15, and an end-of-epoch stopping check at 45% of the issued allowance. A fresh all-training fit uses that epoch count, with an end-of-epoch stopping check against the total allowance; actual epochs are retained. Hard inference calls only the gate’s chosen neural expert. It is a sensitivity of the soft-trained model, rather than a separately tuned hard-MoE competitor. The global GBM, residual correction and mixtures use the fixed expert-role target-range constraint; native AutoGluon and TabPFN predictions receive no additional clipping.

The fixed bank uses E and the learned policy forests use P , with C also available to the constant/global calibration candidates. Complete-system alternatives may refit on $E + C + P$. This compares selecting from a frozen bank with rebuilding a complete predictor; it does not equalize the labels used to fit every component. The declared fixed recipes and small validation grids cover only a limited set of gates and experts. AutoGluon’s internal bagging uses its default row folds rather than the policy/global comparator’s grouped folds; the outer grouped or temporal test remains held out for every system. These training and tuning differences constrain any conclusion about routing in general.

AutoGluon’s native memory warnings during shared execution motivated an additional serialized fit of every allocation. This control starts after all other recorded research fits finish and admits each AutoGluon-only fit when available memory exceeds 2.5 GB. It uses the same labels, recipes, three-fold bagging, stacking level, single-CPU request and original issued nominal allowance. Original shared-execution risks and timings remain separate; no run is chosen by test accuracy. Native completed dependencies and component failures are retained. This improves resource comparability without claiming exclusive control of unrelated host applications, equal executed compute, or identical latency conditions. The control was added during fitting before replacement-panel aggregate accuracy review; the original two primary comparisons are preserved. Only generated native model caches are removed after saved predictions independently reproduce their saved risks; recipes, outcomes, component failures, timing repeats and verification hashes remain.

Public TabPFN V2 initially uses one CPU estimator and a reusable context cache. The context cap is $\min\{1000, \lfloor 100000/p_{\text{raw}} \rfloor\}$, with at least two rows when possible; context rows are chosen without outcomes, using seed $s + 2600$. Native categorical columns and default model random state 0 are retained; test batching is capped at 64. Pretraining-feature-limit overrides, unsupported configurations and memory failures are recorded. Context and pretraining resources differ from methods fitted on all $E + C + P$ labels.

A library warning motivated a nominal $\lceil p_{\text{raw}}/500 \rceil$ -estimator sensitivity, with identical-fit reuse when raw dimension was at most 500. A later input-only audit found that this rule does not guarantee input coverage: the 500-feature budget includes preprocessing expansion. For example, reconstructed QSAR TID11 seed 10501 plans expose 257 of 1,024 original inputs with one estimator and 721 with three. The original predictions, risks and timings are retained; earlier metadata interpreting raw dimension at most 500 as complete coverage is superseded by this correction.

The input-complete and serialized AutoGluon cases run in fresh supervised child processes. Abnormal exits and partial

output files retain diagnostics rather than stopping the remaining panel. Cancellation remains an interrupted attempt, and exit codes alone are not interpreted as proof of an out-of-memory failure. This execution change preceded both secondary phases and leaves their recipes and statistical comparisons fixed. One completed replay interpreter remained during shutdown; its final queue marker, task-idle state and successful correction audit are recorded as an exception to requiring parent-process exit. It performs no fitting during the serialized phase, but its shutdown activity can still consume CPU. The author ended that completed process during the serialized phase; the first observed absence was after 73 completed cases, without measuring its exact exit time. These measurements therefore do not establish an otherwise idle host.

The additional input-complete variant retains the original context and chooses the smallest estimator count from 1 to 16 whose actual preprocessing plan exposes every input column. A low-memory executor reconstructs each candidate plan without test prediction; the chosen cached executor must reproduce that plan. Selection depends on input exposure, rather than validation or test risk. A verified complete one-estimator case reuses the identical original predictions and timing, while charging the additional plan work to fitting. Other cases are fitted and timed anew, with unsupported plans/fits retained as failures. This secondary correction was declared after the metadata audit, with no fresh accuracy aggregate reviewed, and leaves the two primary hypotheses unchanged. It remains an older, context-limited V2 comparison, not a latest-checkpoint or full-context claim.

B.3. What the counts mean

Table 2 reconciles allocations, calibrated libraries, configurations, and method-risk records. Repeated fits and additional role tracks do not create new independent datasets.

Stage	Count	Unit / derivation
Original grouped allocations	115	23 datasets $\times$ 5 seeds
Original chronological allocations	20	Four of the same datasets $\times$ 5 seeds
Original combined allocations	135	115+20 role-track/seed allocations
Original calibrated libraries	270	135 allocations $\times$ 2 libraries
Original main configurations	644	Allocation $\times$ library $\times$ feasible policy budget
Original method-risk records	21896	644 configurations $\times$ 34 methods
Controlled action configurations	648	$3p \times 2\sigma \times 3n \times 12$ seeds $\times$ 3 pools
Controlled scalar method records	2592	216 data blocks $\times$ 4 estimators $\times$ 3 projections
Post-review anchoring configurations	290	40 allocations $\times$ 2 libraries $\times$ feasible budgets
Post-review growth configurations	725	40 allocations $\times$ feasible budgets $\times$ 5 pools
Post-review strong baseline risks	160	40 allocations $\times$ 4 reported systems
Replacement natural allocations	105	35 tasks $\times$ 3 seeds; 33 source families
Replacement generated allocations	3	Insurance control, kept outside natural primary

Table 2. Different counting units, reconciled. A configuration contains several method-risk evaluations; it is not an independent dataset. Replacement-stage completed configuration and risk totals are reported with the final coverage audit.

B.4. Additional statistical and serving details

Policy budget	Tasks (all seeds)	Families	Eligible allocations
100	35	33	105
300	33	31	99
1,000	24	22	72
3,000	12	12	41
Full available P	35	33	105

Table 3. Design eligibility, determined from role masks before outcome analysis. A task enters an all-seed panel only if all three policy roles support the budget. Eligible allocations can include seeds from additional tasks. Fractional curves retain all 35 tasks; the absolute balanced curve uses the fixed 3,000-label-eligible panel at every point. These are planned evaluation units, not completed fits.

Separating postprocessing and hyperparameter choice. Let u_θ be the continuous scalar prediction with shrinkage/leaf setting θ , and let v_θ be its clipped or projected output. Continuous CV selects $\hat{\theta}_u$, while the decision variant selects $\hat{\theta}_v$,

both using policy-only validation. Put $A = \widehat{R}(v_{\widehat{\theta}_v})$, $B = \widehat{R}(u_{\widehat{\theta}_v})$, and $C = \widehat{R}(u_{\widehat{\theta}_u})$. Then

$$A - C = \underbrace{(A - B)}_{\text{matched postprocessing}} + \underbrace{(B - C)}_{\text{hyperparameter-choice difference}}, \quad \frac{A + \epsilon}{C + \epsilon} = \frac{A + \epsilon}{B + \epsilon} \frac{B + \epsilon}{C + \epsilon}.$$

The first difference obeys Equation (10); the second includes changes in both shrinkage and leaf size. Using $\epsilon = 10^{-12}$ and identical family weights preserves the product identity for geometric ratios. Stored continuous predictions before each decision supply the matched intermediate, preserving the original validation choices and requiring no refit. These diagnostics distinguish the contribution of the transformation from that of its separately selected regularization.

A batch-consistency correction. A complete-path replay on the health-insurance task, seed 10501, exposed a native KNN cutoff-tie discrepancy between batched and one-row queries. The serving contract is one row, so the correction retains the fitted KNN estimators and executes their native prediction one row at a time when constructing counterfactual banks as well. It introduces no new neighbor-selection rule or model fit. Among 61 existing banks, 23 had materially changed KNN outputs under this contract. Dependency checks restored 102 configurations whose actual library columns and calibration/policy-prefix/test rows were unchanged; 146 configurations required policy replay. Four changed banks affected only unused expert-fitting rows and required no policy replay. The incomplete allocation at detection was also resumed. These scopes were determined from prediction bytes and row/column dependencies, without comparing errors or rankings. The original arrays, fitted states, configuration records and failure trace are retained. Raw inputs, labels, target scaling, reference Ridge, other fitted experts and issued baseline allowances remain fixed; the correction audit checks that raw scalar estimates and continuous shrinkage decisions remain identical. Final results use the corrected banks and completed serving replays. Earlier development-stage fixed-bank diagnostics retain their original batched semantics and are not evidence of this replacement-stage serving contract.

For family g containing datasets $\mathcal{D}_g$ and paired seeds $\mathcal{S}_d$, the reported ratio is

$$\exp \left\{ \frac{1}{G} \sum_{g=1}^G \frac{1}{|\mathcal{D}_g|} \sum_{d \in \mathcal{D}_g} \frac{1}{|\mathcal{S}_d|} \sum_{s \in \mathcal{S}_d} \log \frac{\widehat{R}_{ds}(\pi) + 10^{-12}}{\widehat{R}_{ds}(b) + 10^{-12}} \right\}.$$

Raw MSE and every ratio remain available, including harmful and near-zero-risk cases. Hierarchical resampling draws families, then datasets within a drawn family, then seeds within a drawn dataset. Repeated copies receive independent inner draws. The resulting interval is descriptive under the selected task panel and overlapping seed fits. Sign tests target the probability of a directional family win, whereas sign-flip tests concern an average paired effect under stronger symmetry/exchangeability assumptions. They are not interchangeable confirmations of one null.

The saved policy field named `total_fit_seconds_including_cv` sums newly executed forest fit-and-prediction calls within a whole configuration. It includes diagnostic forests and grouped validation but excludes global-stack selection, preprocessing, expert-bank construction, and scalar fits reused from another configuration. It is shared across that configuration’s reported methods, rather than a per-method total. The AutoGluon allowance adds this initial full-budget field to the initial 16-expert estimator-fit sum, rounds up to ten seconds, and has a minimum of 60 seconds. Corrective serving replays preserve already issued allowances. Baseline timers have their own scopes: GBM estimator fits exclude preprocessing; the AutoGluon native fit call includes its internal validation and stacking; soft/hard MoE share one search/refit timer; TabPFN fitting includes context-cache construction, and the input-complete variant also charges feature-plan selection. These are component records, not an exact total-training-compute comparison.

Replacement-panel fits and timings use the CPU of an Apple M1 MacBook Air with 8 GiB memory, running macOS 26.0. Complete timing starts from a raw feature frame and includes the fitted encoder. Timed outputs are in the common E -standardized target units; the common inverse affine output map is outside these timings. A raw gate then predicts scores and queries only the chosen library expert. Output-augmented and projection paths compute required expert dependencies once. A global sparse stack executes only nonzero-weight dependencies. Continuous direct prediction with unit shrinkage bypasses the bank, and zero shrinkage bypasses correction/quantization. Timings are warm, batch-one, on a shared CPU; cached-context construction is charged to TabPFN fitting. Whole-test prediction time and timing repeats are retained. Horizontal timing bars use the observed minimum method repeat divided by the maximum CP repeat, and the converse maximum/minimum ratio, followed by the same family-balanced geometric aggregation. CP compared with itself is exactly one. These bars describe the three observed repeat means, rather than confidence intervals or bounds on future latency.

The reported nondominated set uses measured median-repeat costs and is not a claim of statistically established Pareto optimality. Cold starts, GPU serving, parallel experts, energy measurement, and latency confidence intervals under repeated isolated runs are outside this deployment measurement.

For routing and scalar diagnostics, expert-query counts concern members of the fixed bank: a direct continuous estimator can make zero bank queries while still doing substantial prediction work. For jointly trained mixtures, the counts instead describe neural expert branches; for TabPFN, they describe ensemble estimators. They are not interchangeable units of computation across architectures. Measured complete-path latency supplies the cross-system cost comparison.

C. Completed replacement-panel evidence

All nine reported baseline systems have 108 successful outcomes and no retained failures; 105 allocations enter the natural primary estimate. Internal component failures or warnings remain in their execution records. The generated-control allocations contribute to execution counts but not to the natural estimate. Full policy-label curves use fixed paired panels, and all method choices precede held-out outcome comparison.

Baseline system	Completed	Retained failures	Pending
Retrained global GBM	108	0	0
Boosted residual correction	108	0	0
AutoGluon, shared execution	108	0	0
AutoGluon, serialized fits	108	0	0
Soft MoE	108	0	0
Hardened MoE	108	0	0
Public V2, one estimator	108	0	0
Public V2, nominal coverage	108	0	0
Public V2, verified inputs	108	0	0

Table 4. All 108 planned allocations, including the generated control. A retained failure is not silently replaced by another estimator. Missing systems reduce comparison-specific coverage.

Replacement-stage unit	Completed count
Role/seed allocations	108
Library/budget configurations	1154
Policy-method risks	25740
Complete-system baseline risks	972
Constant-candidate cost points	540

Table 5. Replacement-stage execution counts. Natural primary and generated secondary-control allocations are both included in these execution totals; the primary estimand uses only the natural panel. Baseline systems can share fitted states, and repeated constant choices share timing records.

Mechanism controls. Raw weak addition, augmented weak addition, raw weak targets with weak decisions forbidden, and weak targets/features with weak decisions forbidden yield ratios 1.000, 0.999, 0.996 and 1.000 to their specified $K8$ comparators. Refitting eight targets versus masking sixteen, with sixteen fixed features, yields 1.002. Normalizing weak-library targets gives ratio 1.006 to the unnormalized weak-library learner. Independent contrast fits give ratio 1.078 to shared fits, despite additional training/gate computation; these controls do not support a general benefit from separating targets.

Matched decision decomposition. For raw direct clipping and projection, the matched-postprocessing / hyperparameter-choice / total ratios to continuous CV are 0.9888/0.9972/0.9860 and 0.9964/0.9990/0.9954. Raw residual counterparts are 0.9670/1.0035/0.9704 and 0.9759/1.0123/0.9879. Augmented direct projection gives 1.0025/1.0020/1.0045, and augmented residual projection gives 0.9958/0.9977/0.9935; products hold before rounding. These factors use the saved matched continuous vector defined in Appendix B.4, and change no validation choice. For unshrunk raw residual projection, mean displacement penalty 0.02548 plus error cross term -0.04583 gives standardized risk difference -0.02035 .

Regression Model Selection under Explicit Inference Costs

System	Tasks/families/allocations	MSE ratio [descriptive interval]	Sign p
Global CP ensemble	35/33/105	1.000 [1.000, 1.000]	1.000
Constant individual	35/33/105	1.087 [1.042, 1.144]	1.1e-04
One-query CV gate	35/33/105	1.055 [1.024, 1.095]	0.005
Direct scalar, continuous	35/33/105	1.229 [1.120, 1.369]	1.3e-07
Residual, continuous	35/33/105	1.355 [1.206, 1.567]	7.9e-09
Residual projection	35/33/105	1.339 [1.196, 1.539]	7.9e-09
Retrained global GBM	35/33/105	1.038 [0.926, 1.169]	0.728
Boosted residual correction	35/33/105	1.200 [1.106, 1.308]	0.005
AutoGluon, shared execution	35/33/105	0.851 [0.743, 0.925]	1.3e-07
AutoGluon, serialized fits	35/33/105	0.850 [0.744, 0.924]	1.3e-07
Soft MoE	35/33/105	1.070 [0.986, 1.149]	0.014
Hardened MoE	35/33/105	2.926 [2.040, 4.389]	7.9e-09
Public V2, verified inputs	35/33/105	1.060 [0.865, 1.274]	0.163

Table 6. Full-policy natural-panel accuracy relative to global CP. Ratios below one favor the reported system. Intervals describe hierarchical family/dataset/seed resampling; they are not guarantees for future tasks. Sign tests describe directional family effects against CP and are secondary. The two fixed primary tests instead compare the one-query gate with retrained GBM and the original shared-execution AutoGluon run, with Holm-adjusted sign-flip results in the text. Coverage is method-specific; the cost figure uses the common intersection.

Its sign can differ from a geometric effect because one averages MSE differences and the other log ratios. Full per-configuration and family-level values are included with the evidence.

Panel	System	Tasks/families	MSE ratio
Exclude chemistry families	AutoGluon, shared execution	31/31	0.848
Exclude chemistry families	AutoGluon, serialized fits	31/31	0.847
Exclude chemistry families	Retrained global GBM	31/31	1.042
Exclude chemistry families	One-query CV gate	31/31	1.055
Exclude chemistry families	Residual projection	31/31	1.333
Exclude chemistry families	Public V2, verified inputs	31/31	1.029
Exclude food delivery	AutoGluon, shared execution	34/32	0.848
Exclude food delivery	AutoGluon, serialized fits	34/32	0.847
Exclude food delivery	Retrained global GBM	34/32	1.040
Exclude food delivery	One-query CV gate	34/32	1.057
Exclude food delivery	Residual projection	34/32	1.328
Exclude food delivery	Public V2, verified inputs	34/32	1.062
Stricter source quality	AutoGluon, shared execution	30/29	0.835
Stricter source quality	AutoGluon, serialized fits	30/29	0.834
Stricter source quality	Retrained global GBM	30/29	1.041
Stricter source quality	One-query CV gate	30/29	1.061
Stricter source quality	Residual projection	30/29	1.361
Stricter source quality	Public V2, verified inputs	30/29	1.057

Table 7. Fixed source-quality and chemistry sensitivity panels. All ratios use the corresponding global CP comparator, with family balancing; these are secondary summaries.

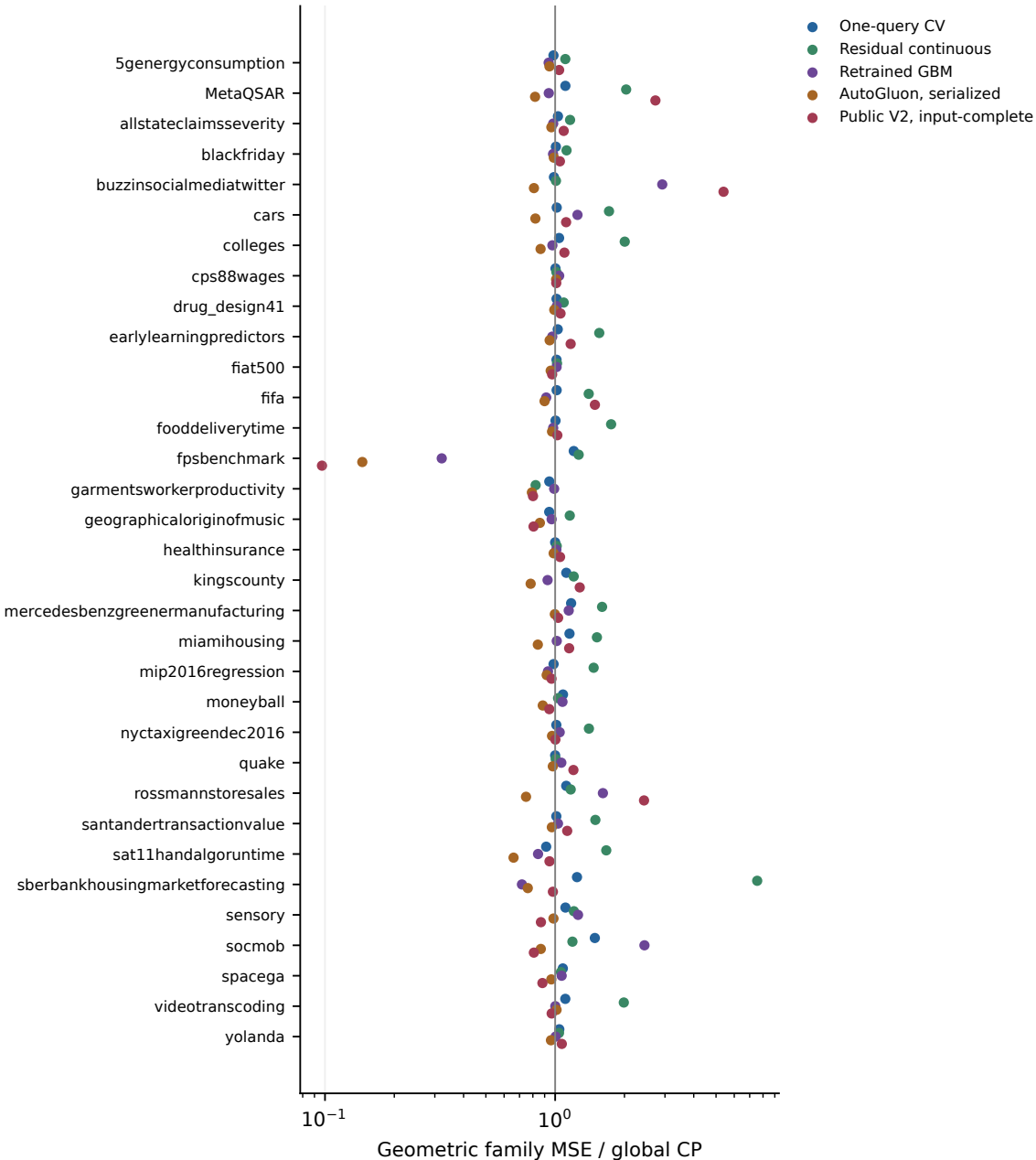


Figure 3. All 33 natural-data source families, in fixed alphabetical order. Values below one favor the reported system over CP. Three seeds are averaged within tasks and tasks within families. The comparison preserves unfavorable families as well as favorable ones.

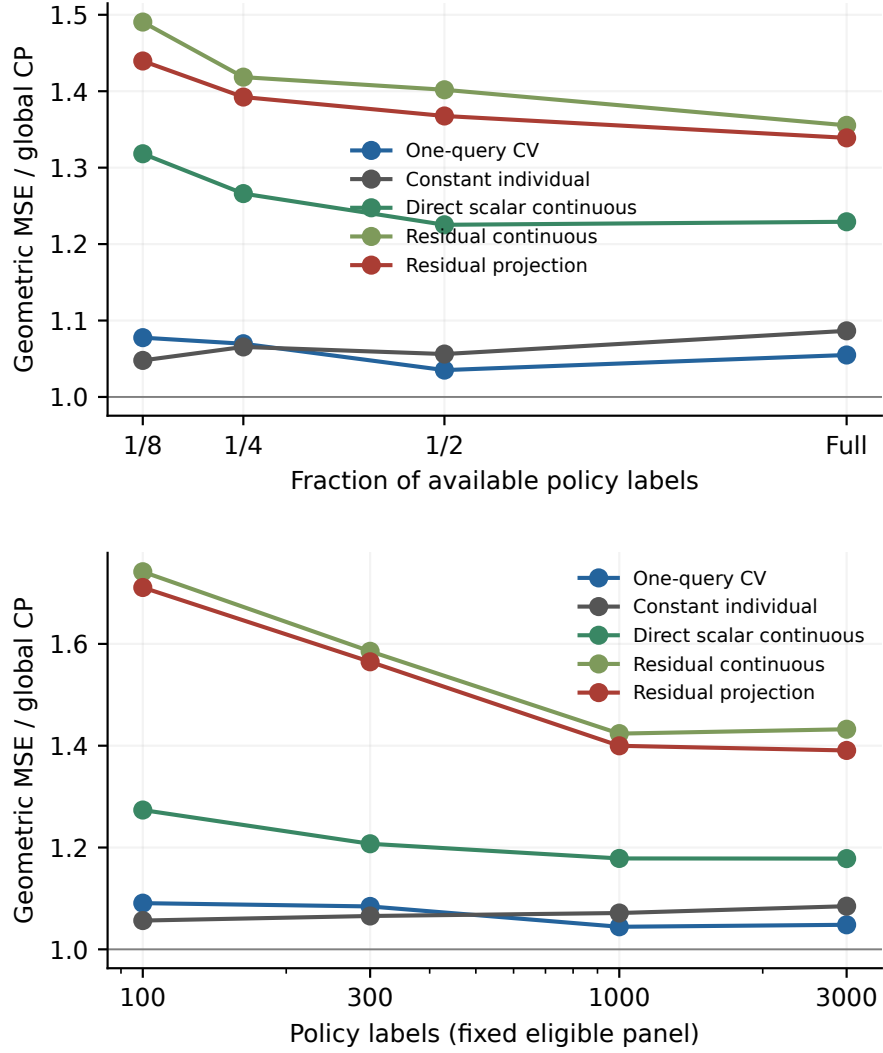


Figure 4. Complete policy-label curves. Top: matched policy fractions on all 35 primary tasks. Bottom: 100/300/1,000/3,000 labels on the same 12 eligible tasks in 12 families, each with all three seeds. The changing-panel eligibility counts are separately reported in Table 3. Lower geometric MSE ratios favor the reported method over CP at that budget.

D. Development evidence and the original estimands

These analyses predate the replacement panel or were added during review of the earlier study. The original 14-family primary set contains six generated/simulator families. The natural-only reanalysis and new shrinkage/library-growth experiments are post hoc development evidence; their positive cells are not promoted into a new confirmatory primary result.

The earlier grouped protocol used five seeds, a 12,000-row outcome-independent cap, and $E/C/P/T$ group fractions of 40%/10%/30%/20%; four datasets also had a chronological track. Expert settings were selected by three-fold grouped validation within E , using two settings per family, before refitting on E . Library A contained XGBoost, random forest, standardized ridge and an MLP. Library B contained histogram boosting, ExtraTrees, the same ridge and RBF-feature ridge. Each seven-action pool added the four experts' mean, a ridge stack with intercept and a convex stack, calibrated on C ; the reference was also chosen using C . The stronger global CP comparator refitted combinations and selected globally using C plus the same policy prefix. These correlated composite actions differ from the replacement panel's library of distinct fitted experts. The eight natural development tasks are Beijing PM2.5, gas turbine, Parkinsons, protein, SGEMM, SML2010, superconductivity and temperature bias.

D.1. The strong-library null and complete-system alternatives

At 1,000 policy labels, natural-family geometric MSE ratios to the recalibrated global CP ensemble are 0.986 in B and 0.953 in A for fixed 1% residual projection. B has no clear improvement: its family sign-flip $p = 0.508$, exact sign-test $p = 0.289$, and descriptive hierarchical interval $[0.942, 1.041]$. Retraining the calibration-chosen individual expert gives ratios 0.897 in B and 0.887 in A. Actual AutoGluon, fitted with the same labels and a nominal expert-construction-time allowance, gives 0.874 in B and 0.834 in A. It improves the comparator on seven of eight B tasks and all eight A tasks. The B AutoGluon family sign-flip $p = 0.109$ illustrates the instability of small-panel inference despite a favorable average; no universal improvement is inferred.

The public V2 one-estimator, 1,000-context-row CPU TabPFN control gives ratios 1.725 (B) and 1.646 (A), with large dataset heterogeneity. This is not evidence against the latest full-context foundation model. The joint soft mixture gives 1.211/1.156 (B/A), and hardening its soft-trained gate gives 4.600/4.389. Hardening is particularly poor and is not a competitive hard-routing training algorithm. These controls are retained rather than used to imply that outperforming a weak neural baseline establishes routing value. All 160 stored baseline risks were independently replayed. AutoGluon training budgets, overshoots, individual task effects, and timing repeats are retained in the evidence.

D.2. Full learning curves, including unfavorable budgets

For direct comparison with the earlier report, Table 8 retains its arithmetic percent-gain estimand on the same 14 families at each budget. In B, fixed residual projection loses to CP at both 100 and 300 labels (-1.96% and -1.26%), then gains 1.57% at 1,000 labels. These are not an across-budget improvement guarantee. The main text instead uses geometric paired ratios and the full available policy role on the replacement panel.

Library	Labels	Shared	Indep.	Screen/refit	Residual	CV res.	FW stack
A	100	-4.30	-4.81	-4.13	-0.55	+0.13	-10.55
A	300	-0.85	-0.63	-1.20	+2.05	+1.91	-3.14
A	1000	+1.96	+2.59	+0.40	+5.57	+4.77	-2.34
B	100	-2.72	-3.24	-2.13	-1.96	-0.76	-7.77
B	300	-1.74	-2.21	-1.24	-1.26	-0.54	-4.11
B	1000	-0.85	+0.12	-0.96	+1.57	+0.34	-3.62

Table 8. Secondary budget comparison on the same 14 primary families at every budget. Entries are percent MSE reduction versus the corresponding CP control, which also receives each additional label budget. The common panel avoids confusing changes in task eligibility with a learning curve.

D.3. Anchoring, shrinkage, and quantization

At 1,000 labels on the eight natural families, an unshrunk direct forest is substantially worse than the CP ensemble: continuous/projection ratios are 1.142/1.039 in A and 1.197/1.199 in B. Prefix-only validation of its shrinkage and leaves changes direct projection to 0.941 in A and 1.000 in B. CV residual projection gives 0.957/0.989 (A/B). Thus the failure

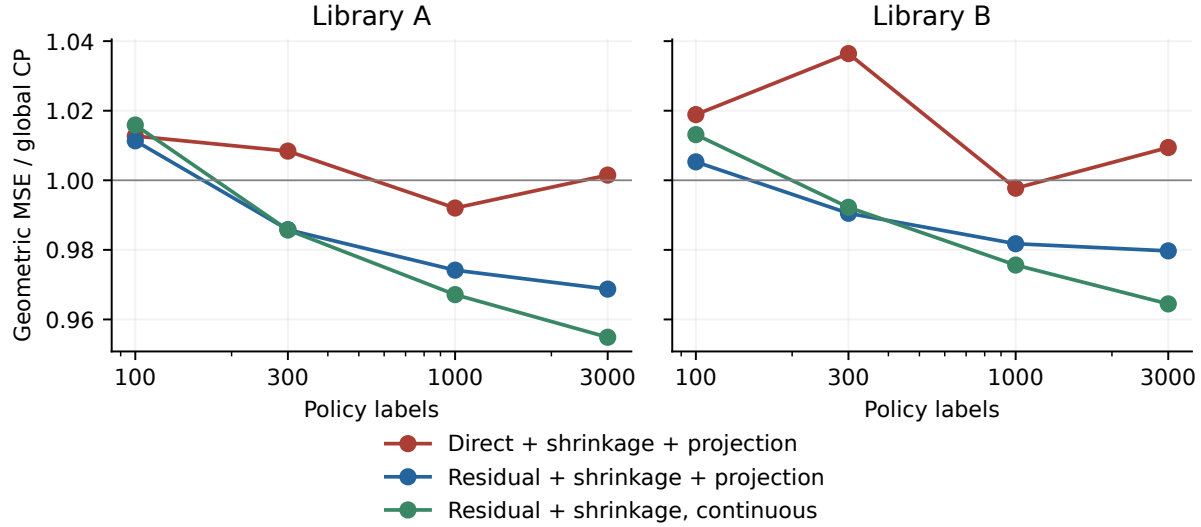


Figure 5. Post-hoc, prefix-validated shrinkage curves on the same five natural families eligible for 3,000 labels at every point: superconductivity, protein, SGEMM, gas turbine, and Beijing PM2.5. These use geometric MSE ratios and CV shrinkage, so they estimate a different quantity from the original fixed-router, 14-family arithmetic summary above. The five-family composition remains fixed through 3,000.

of an unregularized direct-mean estimate is not evidence against scalar sufficiency, and the improvement from anchored fitting cannot be assigned to scalar structure alone.

For fixed 1% residual fits at $\gamma = 1$, continuous, clipped-to- E -bounds, and projected ratios are respectively 1.0003, 0.9995, and 0.9535 in A, and 1.0396, 1.0389, and 0.9860 in B. This older clipping control uses the expert-training target range; the replacement stage additionally tests clipping to the per-input expert-output interval. The older bound clipping explains little of projection’s observed advantage. Equation (10) measures the remainder: projection’s nonnegative displacement term can be outweighed by its signed covariance with estimation errors. This is an empirical regularization effect, not a general theorem that discretization improves MSE. A boosted-residual control with separately validated shrinkage gives continuous/projected ratios 0.9869/0.9573 in A and 1.0160/0.9966 in B. All gamma-grid risks and independently selected decisions are retained.

D.4. Real-data library growth does not give uniform degradation

The 40 development allocations fit 24 actual models: eight expert-family recipes, eight alternatives, and eight independently permuted-target fits. At 1,000 labels, diverse 16-versus-8 model growth changes raw shared-router MSE by a geometric factor 1.0108, and output-augmented shared MSE by 0.9999. Weak 16-versus-8 growth changes these factors to 0.9898 and 1.0168, respectively. Weak additions therefore improve one raw learner while slightly hurting the augmented learner. Normalization changes the comparison again. These mixed results do not support a general monotone-degradation claim.

Prediction features receive a larger fraction of augmented shared-forest splits as the library grows, exceeding their feature-count fraction in this panel. This is evidence of a representation change, not by itself proof that the change causes worse risk. The replacement-panel decision-mask and fixed-feature/refit-target controls address that distinction. Development-stage timing estimates sum calibration expert-forward prices and a batched per-row gate cost; they are explicitly component proxies, not measured end-to-end single-query latency. The replacement stage executes the complete serving path.

D.5. Earlier synthetic intervention and global screening

The original controlled intervention crosses three rare-region probabilities, two noise levels, three policy budgets, twelve seeds and three action-pool sizes: 648 configurations. Added predictors are conditionally dominated by construction. The scalar extension fits four estimators on 216 data blocks and projects each onto three pools, giving 2,592 method records. These are different units from the public-data configurations.

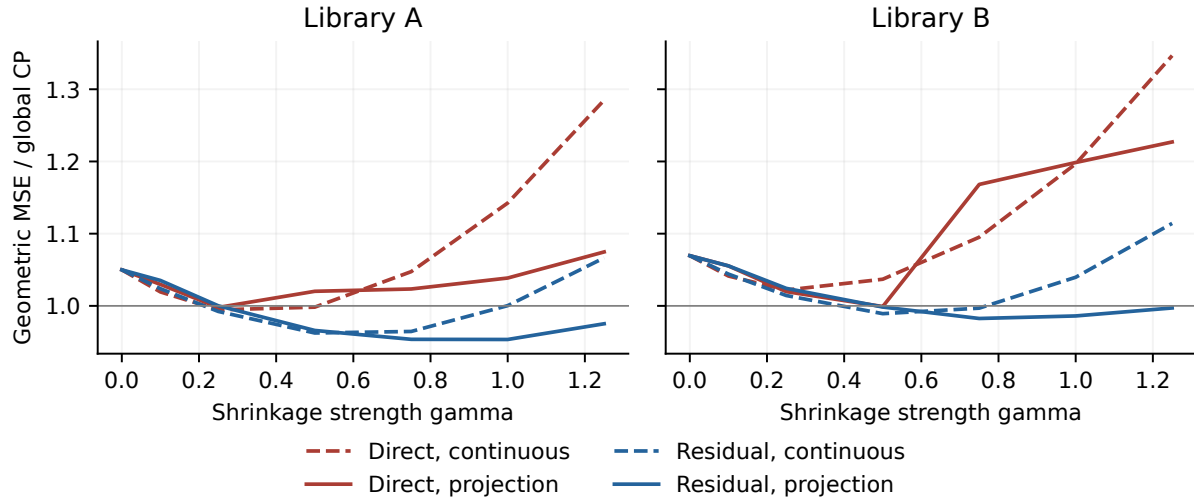


Figure 6. All fixed shrinkage strengths for the 1% direct and residual forests on eight natural families at 1,000 policy labels. Curves describe the full grid on held-out outcomes; a visually favorable grid point is not a new test-selected method. Separate policy-only validation results are stated in the text.

In the 2%-rare, low-noise, 3,000-label condition with sixteen added nuisance actions, shared fitting recovers -0.48% of conditional opportunity, independent contrasts 57.95% , and residual projection 60.64% . Removing known nuisance targets before refitting restores the scalar result, whereas masking them only after fitting does not. The nuisance-removal control uses unavailable DGP knowledge. Global screening instead removes the rare specialist and retains little opportunity. Under a high-noise null, screening can protect against erroneous switching while scalar projection still loses. This establishes possible mechanisms under known specialization, not their prevalence among natural tasks.

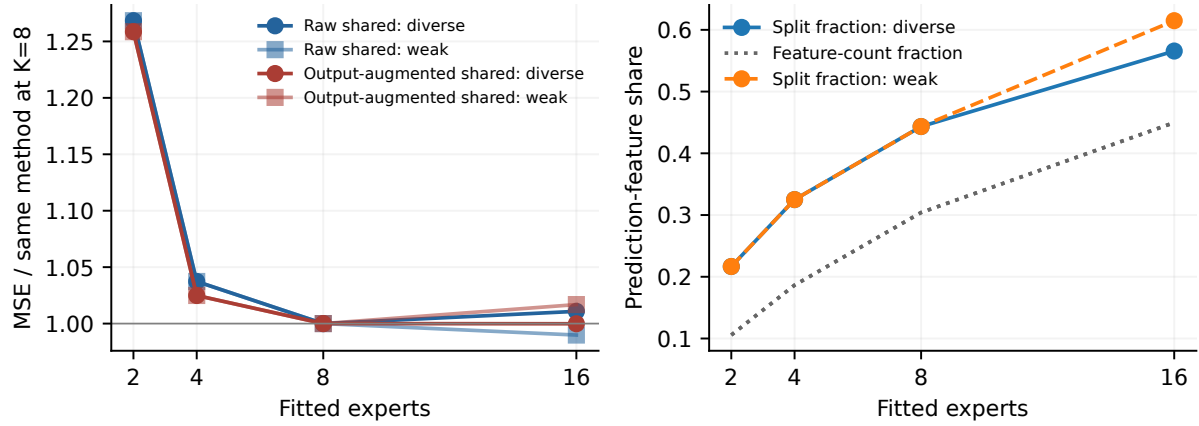


Figure 7. Left: real-library growth on the same eight natural development families, normalized to each method’s own eight-model risk. The weak and diverse branches share the 2/4/8 models. Right: augmented prediction-feature split fraction compared with prediction features’ fraction of available features. Counts averaged over seeds and tasks do not establish causal mediation.

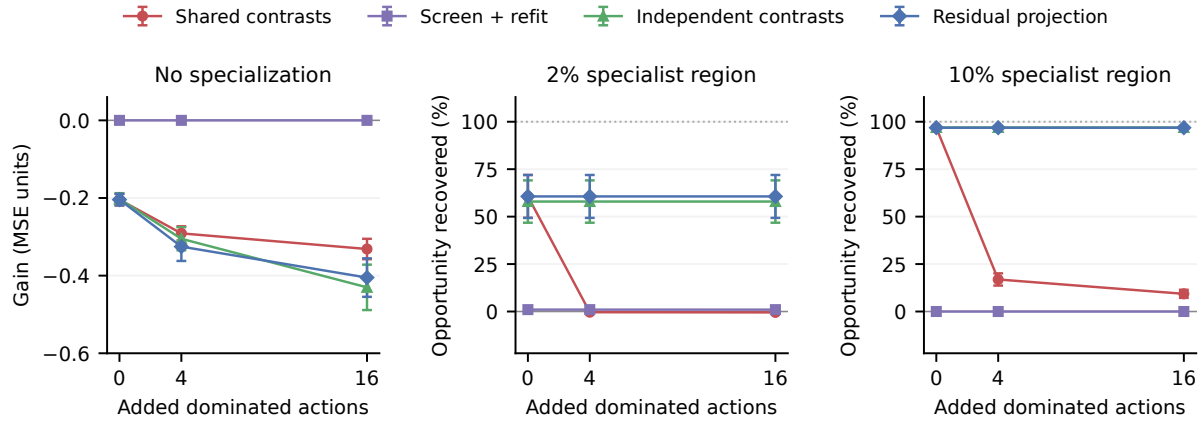


Figure 8. Original controlled action intervention. Specialist panels report opportunity recovery, while the null panel reports absolute MSE gain; their vertical scales are different. Error bars are one Monte Carlo SE across twelve paired replications.