WHEN FEATURE BAGGING HELPS ADAPTIVE SEARCH: SNR THEORY AND PROSPECTIVE MTRY SELECTION

Anonymous authors
Paper under double-blind review

ABSTRACT

Feature bagging averages predictors fitted after restricting an adaptive search to random feature subsets. We study when this improves prediction and use the same search structure to tune `mtry`, the number of features searched at each forest split. Under an isotropic Gaussian signal model, an exact restriction–diversity identity gives a posterior signal-to-noise ratio (SNR) boundary for grouped nonlinear search with arbitrary candidate correlations. A complementary fixed-signal bound permits general feature covariance; under orthogonal weak signals, smaller feature fractions have lower asymptotic risk. A common-projection result extends the SNR identity to completed recursive fits. For tuning, we couple full-depth trees across budgets to share split searches and selectively refine promising candidates. On a development benchmark of 12 wide-feature regression tasks and five outer folds, with 39–85 candidate budgets, the selector attains 1.336% mean regret relative to the best 256-tree test-grid candidate, compared with 1.385% for successive halving. Its measured tuning and total wall-time ratios to halving are 0.3710 and 0.5123, respectively; total time includes the selected 256-tree forest.

1 INTRODUCTION

Random forests combine instance bagging, which averages trees fitted to bootstrap samples, with feature randomization, which restricts each split search to a random subset of variables. We use *feature bagging* for averaging these feature-randomized predictors. Its main control is `mtry`, the number of features searched at a split, often fixed by an implementation rule or tuned by fitting a separate full-depth out-of-bag (OOB) forest at every candidate value (Breiman, 2001; Probst et al., 2019). We ask when feature bagging improves prediction and how to select a useful budget at less cost.

The statistical difficulty is adaptivity. At a classification and regression tree (CART) node, the learner searches cutpoints within each feature and then compares the featurewise winners. Restriction can exclude signal or prevent noise-driven splits; an early split also changes all later searches. Existing studies connect feature randomization to regularization at low signal-to-noise ratio (SNR), complementary bias reduction at high SNR, and SNR-dependent choices of `mtry` (Mentch & Zhou, 2020; Liu & Mazumder, 2025; Chen et al., 2026; Ma & Sun, 2026). An exact account for correlated adaptive candidates must distinguish the cost of restriction from the diversity gained by averaging.

The computational difficulty has the same origin. Exhaustive OOB tuning repeats similar split searches for every budget; shorter forests reduce each fit's cost but still repeat the search. Our starting point is a grouped pilot: a single fitted candidate library supplies prediction moments for every budget through the rank probabilities of grouped search. We carry this reuse principle into recursive forests, where budgets whose trees reach the same node can share featurewise split searches. This local sharing lets us compare a broad grid before concentrating computation on the surviving candidates.

Under squared loss, feature bagging changes risk by the restriction cost minus the fraction $1 - B^{-1}$ of prediction diversity removed by averaging B independent fits. Appendix A.3 gives this standard ambiguity identity (Krogh & Vedelsby, 1994). Our theory turns it into a one-search SNR boundary and a conditional recursive account; the search structure also supplies the computational reuse used for budget selection.

Reviewers: please read the Reviewer Guidelines (iclr.cc/Conferences/2027/ReviewerGuidelines) and the AI Policy for Reviewers (iclr.cc/Conferences/2027/AIPolicyForReviewers). If you used AI to expand, edit, or polish your review, please provide the input text to the LLM. Better still, consider skipping the LLM and submitting your original text: we, and the authors, are much more interested in your unedited thoughts than in what an LLM has to say. AI-assisted or not, you are putting your name and reputation behind your review: LLM-generated falsehoods, hallucinations or misrepresentations are subject to disciplinary action, which may include desk-rejecting all papers you have authored.

Contributions. First, we derive an exact posterior SNR identity for grouped adaptive search with arbitrary within- and between-group correlations. A complementary fixed-signal alignment bound permits general Gaussian feature covariance, and an orthogonal weak-signal limit explicitly orders feature budgets.

Second, we extend the account to recursive adaptive search. Completed hybrid fits and a martingale decomposition attribute restriction and terminal diversity to individual decisions; a common least-squares projection geometry gives the corresponding SNR identity.

Third, starting from all-budget evaluation in a grouped pilot, we develop shared-search selection for full-depth forests, using a target-size OOB surrogate and selective refinement of surviving budgets. On 12 wide-feature regression tasks, its descriptive accuracy comparison with successive halving is close while its directly measured total wall time is about half, including the selected 256-tree forest.

1.1 Related work

Bagging and random subspaces stabilize fitted rules through instance and feature randomization (Breiman, 1996; Bühlmann & Yu, 2002; Ho, 1998; Breiman, 2001). Mentch & Zhou (2020) and Liu & Mazumder (2025) study low-SNR regularization and high-SNR bias reduction, respectively. The closest rank-based foundation is Chen et al. (2026, Sections 3–4): our subset rank probability is their one-step orthogonal-search boundary case. We establish posterior risk results for correlated nonlinear candidate families and a separate fixed-signal alignment bound for one-feature regression fits.

Stability analyses of instance bagging (Soloff et al., 2024) and feature-axis randomization (Ma & Sun, 2026, Section 4.2), and population CART theory with independent covariates (Mei et al., 2026), provide complementary accounts. Nonadaptive feature-subset OLS and ridge apply a mask before a fixed convex fit (LeJeune et al., 2020; Ruben & Pehlevan, 2023); our candidates are selected by adaptive search.

For tuning, random search, Bayesian optimization, and successive halving allocate evaluations across candidate configurations (Bergstra & Bengio, 2012; Snoek et al., 2012; Karnin et al., 2013; Jamieson & Talwalkar, 2016; Li et al., 2018). Extrapolated cross-validation (ECV) estimates larger-ensemble risk from fewer base learners (Du et al., 2024). Our selector combines a finite-ensemble OOB score with paired bootstrap paths and recursive search sharing. The main experiments evaluate shared search, OOB references, and conventional fixed budgets on the same wide-feature tasks, reserving outer test data for evaluation (Janitza & Hornung, 2018).

Notation and roadmap. For $q \geq 1$, $[q] = \{1, \ldots, q\}$. There are p feature groups; $m \in \mathcal{M} \subseteq [p]$ is the number exposed to a search, $G = |\mathcal{M}|$ is the grid size, and B is the target ensemble size. A forest draws a fresh size-m subset at every node. For an estimator g, $\mathcal{R}(g \mid \mathcal{C})$ denotes conditional expected squared error given information $\mathcal{C}$; each model specifies its target and norm. Sections 2 and 3 develop the one-search and recursive theory; Sections 4 and 5 present selection and its accuracy–cost evaluation. Proofs and mechanism studies are in the appendix.

2 SNR theory for grouped adaptive search

This section analyzes one adaptive search, the statistical unit repeated at forest nodes. We first average exactly over feature subsets for a fixed fitted candidate library. An isotropic Gaussian model then turns its fitted energies into a posterior risk boundary. A separate fixed-signal result shows which part of that conclusion survives general feature covariance. None of these one-search statements assumes that a recursively grown forest makes independent node decisions.

2.1 Grouped search and exact mask integration

A grouped search first finds the best candidate within each exposed group and then compares those group winners. For a CART node, a group is one feature and its candidates are the admissible cutpoints; in forward regression, a group can contain one residualized direction. These are examples of the same two-level search, detailed in Appendix A.4.

Let $\mathcal{H}$ be a real Hilbert space and let p be the number of groups. Group $j \in [p]$ has L_j unit-norm candidate atoms $\{\phi_{j\ell} : \ell \in [L_j]\} \subset \mathcal{H}$. We allow arbitrary correlations within and across groups. The atoms are fixed throughout the subset calculation; for design-dependent atoms, all statements condition on that design. Given a response-derived target $Z \in \mathcal{H}$, choose the best atom in each group and its one-dimensional least-squares fit:

$$\ell_j \in \arg\max_{\ell \le L_j} |\langle Z, \phi_{j\ell}\rangle|, \qquad h_j = \langle Z, \phi_{j\ell_j}\rangle \phi_{j\ell_j}. \tag{1}$$

The squared correlation is the training-loss improvement. The selected atom may depend on Z, yet the fitted vector obeys the following projection identity for every realized search:

$$\langle h_j, Z\rangle = \|h_j\|^2. \tag{2}$$

Fix deterministic tie rules. Full search chooses the winner h_F with the largest fitted energy $\|h_j\|^2$. Restricted search first draws ξ uniformly among the size-m subsets of $[p]$ and returns the strongest group winner h_ξ within ξ. Rank all p fitted winners by decreasing energy, writing $j_{(r)}$ for the group at rank r and $h_{(r)}$ for its fit. Let J_ξ be the selected group. Conditional on the fitted library, its rank probability is

$$\Pr_\xi\{J_\xi = j_{(r)}\} = \pi_r(m) := \binom{p-r}{m-1} \Big/ \binom{p}{m}, \qquad r \in [p], \tag{3}$$

Here $\binom{a}{b} = 0$ when $b > a$: rank r wins precisely when that group is exposed and the other $m-1$ exposed groups have lower ranks. Thus a single fitted library can be averaged over every feature subset without enumerating subsets. Ties in fitted predictions do not alter this group probability because the tie rule fixes a rank. Conditional on Z, define

$$D = \|h_F\|^2 - \mathbb{E}_\xi \|h_\xi\|^2 \ge 0, \qquad V = \mathbb{E}_\xi \|h_\xi - \mathbb{E}_\xi h_\xi\|^2. \tag{4}$$

The nonnegative quantity D is the loss of fitted energy from restriction; V is the variation of the resulting fitted predictions. They need not move together: correlated groups can have different ranks yet nearly identical predictions. For example, with $p = 2$, $m = 1$, and group 1 ranked first, $D = \{\|h_1\|^2 - \|h_2\|^2\}/2$ whereas $V = \|h_1 - h_2\|^2/4$. Even if the fitted energies differ, nearly identical predictions leave little diversity for averaging to exploit.

2.2 An exact weak–strong SNR boundary

Fitted energy alone is a training quantity. To translate it into a prediction-risk statement, use a Gaussian sequence model on the same space:

$$Z = F + \omega W, \qquad F \sim N_{\mathcal{H}}(0, a^2 I), \quad W \sim N_{\mathcal{H}}(0, I), \quad F \perp W, \qquad \kappa = \frac{a^2}{a^2 + \omega^2}. \tag{5}$$

For this model, $\mathcal{H}$ is a fixed finite-dimensional space containing the candidate atoms; for one search it may be their span. Assume $a \ge 0$, $\omega > 0$, and let I be the identity on $\mathcal{H}$. The latent signal is F; a^2/ω^2 is its variance-to-noise ratio and $\mathbb{E}(F \mid Z) = \kappa Z$. If the atoms depend on a design, the stated Gaussian law must hold conditional on that design, which supplies no additional information about F or W. Draw the subsets independently of (F, W), and let $\bar{h}_B = B^{-1}\sum_{b=1}^B h_{\xi_b}$. For a fitted predictor g, define its posterior risk by $\mathcal{R}(g \mid Z) = \mathbb{E}[\|g - F\|^2 \mid Z]$. This expectation averages over $F \mid Z$ and any feature-subset randomness in g, holding the design and observed Z fixed. It gives a posterior comparison, rather than a guarantee for every fixed signal.

Proposition 2.1 (Exact grouped-search SNR identity) For every atom Gram matrix, $1 \le m \le p$, and $B \ge 1$, conditionally on Z,

$$\mathcal{R}(\bar{h}_B \mid Z) - \mathcal{R}(h_F \mid Z) = (2\kappa - 1)D - (1 - B^{-1})V. \tag{6}$$

Consequently, feature bagging has no larger posterior risk whenever $\kappa \le 1/2$, equivalently $a^2/\omega^2 \le 1$. If $D > 0$, the exact boundary is

$$\kappa < \frac{1}{2} + (1 - B^{-1})\frac{V}{2D}. \tag{7}$$

The identity separates two effects. With one fit ($B = 1$), only restriction remains: when $a^2/\omega^2 \le 1$, searching fewer candidates cannot raise posterior risk. Averaging $B > 1$ fits removes the additional amount $(1 - B^{-1})V$ and can extend the benefit above that SNR threshold. Correlation changes D and V, but requires no correction to the identity. This explains the low-SNR behavior seen in empirical and stability studies (Mentch & Zhou, 2020; Ma & Sun, 2026) without asserting that every high-SNR restricted search improves risk.

2.3 Fixed signals and general covariance

The preceding boundary uses isotropic signal and noise. For a fixed signal β, prediction instead weights fitted coefficients by the test-feature covariance. The following result keeps that covariance general and identifies the signal-alignment term that prevents a universal SNR threshold. Let $(X_i, Y_i)_{i=1}^n$ be the training sample, with independent $X_i \sim N(0, \Sigma)$ and $Y_i = X_i^\top \beta + \varepsilon_i$. Assume $\mathbb{E}(\varepsilon_i \mid X_i) = 0$ and $\mathrm{diag}(\Sigma) = 1$. Define

$$\widehat{G} = X^\top X/n, \quad z = X^\top y/n, \quad s_j = z_j^2/\widehat{G}_{jj}, \quad b_j = z_j/\widehat{G}_{jj}.$$

Here X is the $n \times p$ design matrix, y is the response vector, and e_j is coordinate vector j. The coefficient $b_j e_j$ is the one-feature least-squares fit and s_j its training-loss improvement. Each $\widehat{G}_{jj}$ is positive almost surely under the stated nondegenerate Gaussian marginal law. Use the same fixed tie rule to select the full-search winner j_* and the size-m subset winner J_ξ. Set $\widehat{\beta}_F = b_{j_*} e_{j_*}$, $\widehat{\beta}_\xi = b_{J_\xi} e_{J_\xi}$, and let $\bar{\beta}_B$ average B independent subset fits. Write $\langle u, v\rangle_\Sigma = u^\top \Sigma v$ and $\|u\|_\Sigma^2 = u^\top \Sigma u$, $M = s_{j_*}$, $\lambda_- = \lambda_{\min}(\Sigma)$, $\gamma_p = m/p$, and

$$A_\beta = \langle b_{j_*} e_{j_*}, \beta\rangle_\Sigma - \mathbb{E}_\xi \left\langle b_{J_\xi} e_{J_\xi}, \beta\right\rangle_\Sigma. \tag{8}$$

The signed quantity A_β compares how well the unrestricted and restricted fits align with the true signal on fresh covariates. It can be negative. Conditional on the realized training sample $\mathcal{D} = (X, y)$, the risks below average only the fresh feature subsets; the Σ norm integrates over a fresh test covariate. An o_P term refers to variation in the training sample as n and p grow.

Theorem 2.2 (General-covariance alignment bound) Assume $\log p = o(n)$. For $\mathcal{R}(\widehat{\beta}_F \mid \mathcal{D}) = \left\|\widehat{\beta}_F - \beta\right\|_\Sigma^2$ and $\mathcal{R}(\bar{\beta}_B \mid \mathcal{D}) = \mathbb{E}_{\xi_{1:B}}\{\left\|\bar{\beta}_B - \beta\right\|_\Sigma^2 \mid \mathcal{D}\}$,

$$\mathcal{R}(\bar{\beta}_B \mid \mathcal{D}) - \mathcal{R}(\widehat{\beta}_F \mid \mathcal{D}) \le 2A_\beta - (1 - B^{-1})\lambda_- \gamma_p(1 - \gamma_p)M + o_P(M). \tag{9}$$

At $\beta = 0$, $A_\beta = 0$. Hence, if $\inf_p \lambda_{\min}(\Sigma_p) > 0$, $m/p \to \gamma \in (0, 1)$, and $B > 1$, and $\Pr(M > 0) \to 1$, the fixed-null risk gap is negative with probability tending to one.

The alignment term is the possible cost of missing signal; the second term bounds ensemble diversity from below. Gaussian concentration makes the feature-normalization remainder negligible relative to M, giving a fixed-null conclusion without orthogonal features or a signal prior. For a column-normalized fixed design, Appendix B.2 gives the deterministic bound without a remainder; Appendix B.4 bounds the generally unobserved A_β. Under orthogonal weak signals, the next result explicitly orders budgets.

Corollary 2.3 (Orthogonal weak-signal budget ordering) Use the preceding fits and risks with a fixed design and geometry $\Sigma_p = X^\top X/n = I_p$. Let $\varepsilon \sim N(0, \sigma^2 I_n)$, $\sigma > 0$, and $B > 1$ be fixed. Set $q_p = \mathbb{E}\left\|X^\top \varepsilon/n\right\|_\infty$ and assume $\|\beta\|_\infty / q_p \to 0$ as $p \to \infty$. Let $\bar{\beta}_B(m)$ denote the ensemble using budget m. If $m_k/p \to \gamma_k$ with $0 < \gamma_1 < \gamma_2 \le 1$, then

$$\frac{\mathcal{R}\big(\bar{\beta}_B(m_k) \mid \mathcal{D}\big) - \|\beta\|^2}{q_p^2} \xrightarrow{P} c_{\gamma_k, B} := \frac{1}{B} + \left(1 - \frac{1}{B}\right)\frac{\gamma_k}{2 - \gamma_k}, \qquad k = 1, 2. \tag{10}$$

For $B > 1$, $c_{\gamma, B}$ is strictly increasing in γ; hence the smaller feature fraction has lower asymptotic risk.

Here risk uses the fitted-value geometry $\|u\|_{\Sigma_p}^2 = n^{-1}\|Xu\|_2^2$; orthogonality requires $p \le n$. The largest marginal signal is negligible relative to the largest noise correlation scale q_p. Smaller

fractions then reuse the leading noise-driven coordinates less often. At $B = 1$ the leading constant is one for every fraction, so this limit does not order their risks. For example, at large B, fractions $\gamma = 1/4, 1/2, 1$ have leading constants close to $1/7, 1/3, 1$, respectively. These constants multiply q_p^2 after subtracting the common signal term $\|\beta\|^2$. The conclusion therefore asymptotically orders feature budgets for the same fixed weak signal, complementing the posterior comparison that averages over signals conditional on Z. Appendix B.3 gives the general normalized fixed-design sufficient condition and proves the ordering.

3 From one search to recursive feature bagging

A tree makes many adaptive choices, and an early restriction can change every later node. Comparing *completed* recursive fits separates the cumulative effect of restriction from the diversity retained in final predictions. A general squared-risk identity gives this decomposition; a common-projection model then yields its SNR form.

For the general identity, fix $f^* \in \mathcal{H}$ and let $\mathcal{A}$ contain the training sample and shared non-feature randomness. For a fitted predictor g, set $\mathcal{R}(g \mid \mathcal{A}) = \mathbb{E}[\|g - f^*\|_{\mathcal{H}}^2 \mid \mathcal{A}]$. The Hilbert space may be $L_2(P_X)$ for population risk or the fitted-value space on one sample. Expectations below are over feature subsets given $\mathcal{A}$. Fix a budget m, a node-expansion order, and a deterministic bound T on the number of searches; draw independent uniform size-m subsets $\xi_1, \dots, \xi_T$ of $[p]$.

For $t = 0, \dots, T$, the hybrid $H^{(t)}$ uses the common subsets for its first t searches and all p features afterward, always refitting to completion. Thus $H_F := H^{(0)}$ is the full-search fit and $H := H^{(T)}$ is the fully randomized fit. Assume all hybrids are measurable and conditionally square integrable. Let $\bar{H}_B$ average B conditionally independent copies of H. Define the signed risk change and terminal diversity attributable to search t by

$$\begin{aligned} \delta_t &= \mathcal{R}(H^{(t)} \mid \mathcal{A}) - \mathcal{R}(H^{(t-1)} \mid \mathcal{A}), \\ v_t &= \mathbb{E}\,\|\mathbb{E}(H \mid \xi_1, \dots, \xi_t) - \mathbb{E}(H \mid \xi_1, \dots, \xi_{t-1})\|^2 . \end{aligned} \tag{11}$$

The inner mean before any subset is revealed is $\mathbb{E}H$. For two searches, δ_1 compares finished fits whose first decisions differ, including downstream changes; v_1 measures how much the first subset shifts the expected final prediction after averaging over the second. Thus δ_t is signed, whereas $v_t \geq 0$.

For the posterior comparison, fix the design and the observed score Z in Equation (5). Expectations defining the energy increments and diversity now average over feature subsets given Z, including those in v_t. Set

$$d_t = \mathbb{E}\left\|H^{(t-1)}\right\|^2 - \mathbb{E}\left\|H^{(t)}\right\|^2, \quad D_T = \sum_{t=1}^{T} d_t, \quad V_T = \sum_{t=1}^{T} v_t. \tag{12}$$

For this specialization, $\mathcal{R}(g \mid Z) = \mathbb{E}[\|g - F\|^2 \mid Z]$ averages both posterior uncertainty and feature randomization, as in the one-search model.

Theorem 3.1 (Recursive restriction, diversity, and SNR) For every $B \geq 1$:
(i) In the general setting, conditional on $\mathcal{A}$,

$$\mathcal{R}(\bar{H}_B \mid \mathcal{A}) - \mathcal{R}(H_F \mid \mathcal{A}) = \sum_{t=1}^{T} \{\delta_t - (1 - B^{-1}) v_t\}. \tag{13}$$

(ii) Suppose the Gaussian sequence model Equation (5) holds on a fixed finite-dimensional $\mathcal{H}$, and the feature subsets are independent of (F, W). Assume every completed hybrid is $H^{(t)} = P_t Z$, with $P_t^* = P_t = P_t^2$ under the same inner product for every realization; P_t may depend on Z and the subsets. Conditional only on the design and Z,

$$\mathcal{R}(\bar{H}_B \mid Z) - \mathcal{R}(H_F \mid Z) = (2\kappa - 1) D_T - (1 - B^{-1}) V_T. \tag{14}$$

Under part (ii) of Theorem 3.1, if $D_T \geq 0$, feature bagging has no larger posterior risk when $a^2/\omega^2 \leq 1$. If $D_T > 0$, it strictly improves risk precisely when

$$\kappa < \frac{1}{2} + (1 - B^{-1})\frac{V_T}{2D_T}. \tag{15}$$

The cumulative fitted-energy loss D_T can be negative because a restricted early split may enable a better later fit, while V_T counts diversity that persists in completed predictions. Thus the weak-SNR conclusion requires $D_T \geq 0$; the one-search threshold cannot be applied node by node.

An adaptive least-squares tree on one fixed fitting sample meets the projection condition: its leaf means project Z onto the selected partition. Different row-bootstrap samples generally use different geometries, so the posterior SNR formula does not automatically give a population-risk guarantee for the deployed forest; the general identity retains its stated scope. Appendix D gives the proof, projection argument, and coupled estimators, while Figure 5 illustrates the stagewise attribution.

4 ALL-BUDGET REUSE FOR FOREST SELECTION

A grouped pilot shows how one fitted candidate library can evaluate all budgets. This motivates reuse at common forest nodes. The pilot is an illustrative prediction model; the implemented forest selector starts with the full budget grid and selects directly through completed-tree OOB predictions.

4.1 FROM THE GROUPED PILOT TO RECURSIVE REUSE

Fit one predictor g_j per feature group $j \in [p]$, such as a depth-two univariate tree, and rank candidates by training-loss reduction with a fixed tie rule. Write $g_{(r)}$ for rank r. The *grouped pilot* draws a uniform size-m subset ξ and returns the highest-ranked exposed candidate, with prediction $T_m(x;\xi)$. Conditional on the fitted library, Equation (3) gives

$$\begin{aligned} \mu_m(x) &:= \mathbb{E}_\xi T_m(x;\xi) = \sum_{r=1}^{p} \pi_r(m) g_{(r)}(x), \\ q_m(x) &:= \mathbb{E}_\xi T_m(x;\xi)^2 = \sum_{r=1}^{p} \pi_r(m) g_{(r)}(x)^2. \end{aligned} \tag{16}$$

Changing m changes only the weights. Draw B independent subsets from this fixed library and let $\bar{T}_{B,m}(x) = B^{-1}\sum_{b=1}^{B} T_m(x;\xi_b)$. The conditional expected loss at an assessment pair (x, y) is

$$\mathbb{E}_{\xi_{1:B}}\{y - \bar{T}_{B,m}(x)\}^2 = \{y - \mu_m(x)\}^2 + \frac{q_m(x) - \mu_m(x)^2}{B}. \tag{17}$$

Averaging over held-out observations supplies the entire pilot loss curve after one library fit. These moments follow from the subset law and fixed ranking, with no Gaussian or projection assumption.

At a forest node, each feature supplies its best admissible cutpoint, just as each pilot group supplies a candidate. An early split changes the candidate libraries at later nodes, as Section 3 describes. Reuse therefore follows the evolving partitions: budgets share calculations while their node states coincide and continue together after the same split; differing splits create separate computations (Figure 1). The following OOB score evaluates the resulting completed forest predictions.

4.2 A TARGET-SIZE OOB SCORE FROM FEWER TREES

Fix the training sample $\mathcal{D} = \{(X_i, Y_i)\}_{i=1}^{n}$ and budget m. A *prefix* consists of the first b full-depth trees; row-bootstrap and feature randomizations are independent across tree indices. Let r_i count trees omitting row i, and let $\bar{T}_i(m)$ and $S_i^2(m)$ be the mean and unbiased sample variance of their predictions. Write $\mu_i(m), v_i(m)$ for the mean and variance of one tree prediction conditional on $\mathcal{D}$ and on omitting row i. The expected loss of B independent trees from this row-omitting distribution is

$$L_{B,i}(m) = \{Y_i - \mu_i(m)\}^2 + \frac{v_i(m)}{B}. \tag{18}$$

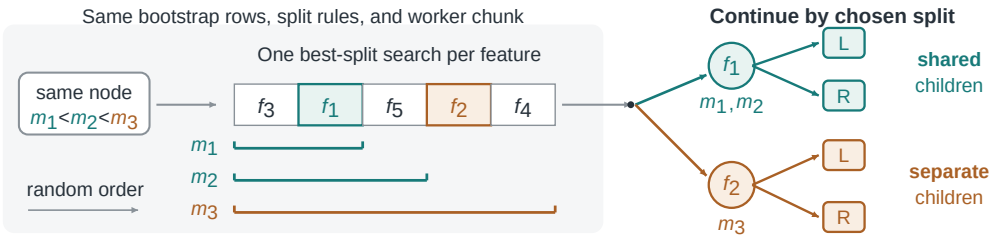


Figure 1: Shared node searches for budgets $m_1 < m_2 < m_3$. Nested feature prefixes reuse one candidate library within a worker chunk. An identical chosen split permits shared child searches; different splits produce separate child states.

Raw prefix loss instead contains $v_i(m)/r_i$ in conditional expectation. Replacing this variance contribution gives

$$\widehat{L}_{B,i}(m) := \{Y_i - \bar{T}_i(m)\}^2 + \left(\frac{1}{B} - \frac{1}{r_i}\right) S_i^2(m), \qquad r_i \geq 2. \tag{19}$$

For fixed budgets and prefix sizes, Proposition C.1 establishes conditional unbiasedness for $L_{B,i}(m)$. This training-side target averages B trees all omitting row i. Its relation to full-sample test risk and adaptive selection is discussed in Appendix C.1.

For $B = 256$, average Equation (19) over eligible rows $\mathcal{E}_b = \{i : r_i \geq 2\}$ to obtain $\widehat{L}_B(m)$. Shared bootstrap samples give every budget the same eligible rows, making comparisons paired; at least two eligible rows are required. This uses the finite-ensemble principle of ECV (Du et al., 2024).

4.3 Shared-search refinement

At each common node, budgets expose prefixes of one fresh uniform feature permutation. Within a worker chunk, featurewise split fits are reused by budgets with identical ordered bootstrap rows and split rules (Figure 1). This coupling preserves each budget's uniform subset law and ordinary tree predictions.

Screen all G budgets with four trees using $\widehat{L}_B$, retain half, extend to eight trees, retain half again, and extend to 16. For example, $G = 40$ gives $40 \to 20 \to 10$ active budgets. The 16-tree default minimizes the average of raw and target-size OOB scores. This half correction moderates the estimated variance adjustment (Equation (38)).

Disagreement between raw and half-corrected leaders, or a ranking reversal across two eight-tree blocks, nominates two budgets for refinement. Add an available interior grid value closest to their geometric mean on a log scale, then compare the nominees at 50 trees using raw OOB loss. Otherwise, return the 16-tree default. Retain the selected trees and complete the forest to 256 trees. These training-side stability rules concentrate computation on unsettled rankings. Appendices C.2 and E.1 give rounding, the exact block criterion, tie rules, and charged midpoint reconstruction.

5 Experiments

We evaluate the shared-search selector with geometric refinement from Section 4. Complementary studies in Appendices E.4 and E.5 illustrate fixed-null gains and signal-strength effects in one search, and compare recursive attribution on calibration and untouched test data.

5.1 Protocol and comparators

The main panel contains all 12 regression tasks in our collection with at least 100 retained predictors, evaluated on five outer folds. Training-only preprocessing leaves 100–4,251 predictors and a common grid of 39–85 budgets per task–fold case. Methods share the splits, preprocessing, bootstrap and feature-permutation streams, and squared-error tree code. Each chooses one global budget $\widehat{m}$ from

Table 1: Accuracy and wall time of tuning methods on 12 tasks with five folds each. Time ratios use SH = 1 and task-level aggregation. Ours and SH are measured with eight workers; italic entries are provisional estimates. Total time includes selection and the delivered 256-tree forest, starting from prepared inputs. Appendix E.2 gives CPU costs, intervals, and estimation details.

Method	Regret (%)	Tuning wall time	Total wall time
Exhaustive OOB selector			
Full-grid OOB256	0.98	*14.95*	*11.68*
Tuning methods			
Geometric refinement (ours)	1.34	0.37	0.51
Full-grid OOB50	1.22	*2.92*	*2.56*
Successive halving	1.39	1.00	1.00
Corrected halving	1.56	*1.04*	*1.06*
ECV25	1.72	*1.45*	*1.42*

training data and delivers a 256-tree forest. Write $\widehat{f}_m$ for the forest using budget m and $\widehat{\mathcal{R}}_{\text{test}}(\widehat{f}_m)$ for its held-out mean-squared error against observed responses. Population response MSE equals risk against the regression function plus irreducible noise; $\mathcal{R}$ uses the signal target specified by each theoretical model. Relative target regret is

$$100\left\{\frac{\widehat{\mathcal{R}}_{\text{test}}(\widehat{f}_{\widehat{m}})}{\min_{m\in\mathcal{M}}\widehat{\mathcal{R}}_{\text{test}}(\widehat{f}_m)}-1\right\}.$$

The denominator is an empirical test-grid oracle used only for evaluation. We average regret over folds and then equally over tasks; paired 95% intervals resample tasks, retaining their folds together. The selector was developed on this task library, so these intervals are descriptive and do not account for method exploration. Appendix E.1 specifies the grid, tree parameters, and preprocessing.

Table 1 compares geometric refinement with full-grid OOB50, 8–16–32 successive halving (SH), corrected halving, and 25-tree ECV. Full-grid OOB256 is the exhaustive training-side tuning reference. The tuning-free rules choose the grid value nearest $p/3$ or $\sqrt{p}$, all features, or a uniform grid value; each delivers a 256-tree forest, and the uniform rule is averaged exactly over its possible choices.

5.2 Prediction accuracy and computation

Geometric refinement has 1.336% mean regret, compared with 1.385% for SH and 1.216% for OOB50. Its paired differences are -0.049 percentage points (95% interval $[-0.243, 0.154]$) and 0.120 points ($[-0.407, 0.692]$), respectively. Both include zero; the library does not establish an accuracy ordering among these tuners. Exhaustive OOB256 has 0.976% regret. The $p/3$, $\sqrt{p}$, all-feature, and uniform-grid rules have mean regrets of 1.72%, 8.55%, 4.66%, and 13.44%, respectively.

In 60 paired runs using the same split-search kernel and eight workers, tuning and total wall-time ratios to SH are 0.3710 and 0.5123, with task-bootstrap 95% intervals [0.3280, 0.4231] and [0.4743, 0.5585]. For each time scope we average foldwise cost ratios within tasks and take the geometric mean across tasks. Total time includes model retention, completion to 256 trees, and prediction, starting from prepared inputs; its ratio is below one on every task. Figure 2(b) shows the corresponding tuning and total speedups: their geometric means are $2.70\times$ and $1.95\times$. Completing the selected forest reduces the relative saving because both methods must deliver 256 trees. The saving that remains measures the practical benefit of the full selection and delivery procedure.

Other tuners' time entries combine historical timings with work counts calibrated to SH; they are provisional estimates, not matched complete executions. Thus the measured speed comparison concerns SH, while the broader frontier in Figure 2(a) provides exploratory context. Appendix E.2 details this distinction and the complete cost accounting.

5.3 Selective refinement

Selection ends at 16 trees in 36 of 60 cases; the other 24 receive a 50-tree comparison. Raw and half-corrected leaders disagree in seven cases, and block rankings reverse in 17. Thus the extra

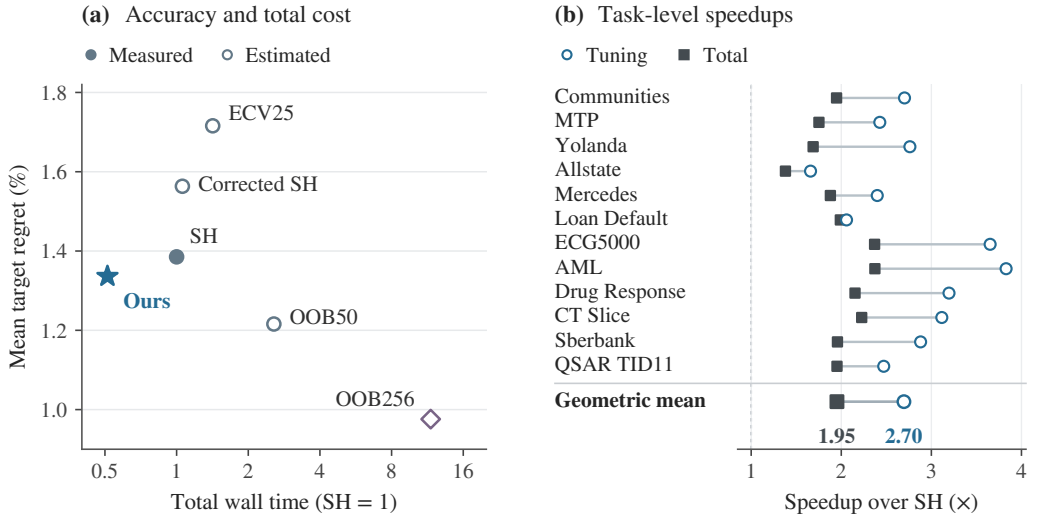


Figure 2: Accuracy and computation on 12 tasks with five folds each. (a) Test-grid regret versus total wall time relative to SH (log scale). Filled markers are measured; open markers use provisional time estimates. (b) Task-level tuning and total speedups over SH; gray segments pair the scopes, and the dashed line marks equal time. Each speedup is the reciprocal of a task's mean foldwise cost ratio. The bottom row gives geometric means; task order follows Table 3, and AML denotes acute myeloid leukemia.

computation is concentrated on cases flagged by the training-side stability checks. An interior midpoint is available in 18 of the 24 comparisons and wins in two. Holding screening and the two-endpoint comparison fixed, adding an available geometric midpoint changes mean regret from 1.416% to 1.336%, a paired change of -0.080 points (95% interval $[-0.235, 0.000]$). This localized ablation isolates the midpoint's accuracy contribution; the measured cost saving reflects allocation, search sharing, and retained-tree continuation together. Appendix E.3 gives trigger counts and task-level results.

6 Discussion

Feature bagging balances the effect of restricting adaptive search against the diversity removed by averaging. Our posterior and fixed-signal results quantify this balance for one search, and a common-projection model extends the SNR identity to completed recursive fits. Shared split searches then make broad-budget selection cheaper: the development benchmark shows accuracy close to practical OOB tuners and substantially lower measured total cost than SH. Together these results support comparing a broad range of feature budgets, including small fractions in weak-signal settings, while using shared computation and selective refinement to control the cost of that comparison.

The recursive SNR specialization requires a common projection geometry, and fixed-budget OOB centering does not guarantee an adaptive selector's risk. Empirical conclusions concern one repeatedly examined task library and forest implementation; other tuners' costs remain provisional. New task collections, depth-dependent budgets, and adaptive risk guarantees are natural next steps.

AI use statement

OpenAI Codex was used to assist with developing experimental code, checking mathematical proofs, searching the literature, and revising the manuscript. The author takes full responsibility for the correctness, originality, and integrity of the final work, including all AI-assisted text, code, mathematical claims, and reported results.

REFERENCES

James Bergstra and Yoshua Bengio. Random search for hyper-parameter optimization. *Journal of Machine Learning Research*, 13(10):281–305, 2012.

Leo Breiman. Bagging predictors. *Machine Learning*, 24(2):123–140, 1996.

Leo Breiman. Random forests. *Machine Learning*, 45(1):5–32, 2001.

Peter Bühlmann and Bin Yu. Analyzing bagging. *The Annals of Statistics*, 30(4):927–961, 2002.

Xin Chen, Jason M. Klusowski, Yan Shuo Tan, and Chang Yu. Revisiting randomization in greedy model search. *arXiv preprint arXiv:2506.15643*, 2026.

Jin-Hong Du, Pratik Patil, Kathryn Roeder, and Arun Kumar Kuchibhotla. Extrapolated cross-validation for randomized ensembles. *Journal of Computational and Graphical Statistics*, 33(3):1061–1072, 2024.

Sebastian Felix Fischer, Liana Harutyunyan, Matthias Feurer, and Bernd Bischl. OpenML-CTR23: A curated tabular regression benchmarking suite. In *AutoML Conference 2023 Workshop*, 2023.

Léo Grinsztajn, Edouard Oyallon, and Gaël Varoquaux. Why do tree-based models still outperform deep learning on typical tabular data? In *Advances in Neural Information Processing Systems*, volume 35, pp. 507–520, 2022.

Tin Kam Ho. The random subspace method for constructing decision forests. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 20(8):832–844, 1998.

Kevin Jamieson and Ameet Talwalkar. Non-stochastic best arm identification and hyperparameter optimization. In *Proceedings of the 19th International Conference on Artificial Intelligence and Statistics*, volume 51 of *Proceedings of Machine Learning Research*, pp. 240–248. PMLR, 2016.

Silke Janitza and Roman Hornung. On the overestimation of random forest's out-of-bag error. *PLOS ONE*, 13 (8):e0201904, 2018.

Zohar Karnin, Tomer Koren, and Oren Somekh. Almost optimal exploration in multi-armed bandits. In *Proceedings of the 30th International Conference on Machine Learning*, volume 28 of *Proceedings of Machine Learning Research*, pp. 1238–1246. PMLR, 2013.

Anders Krogh and Jesper Vedelsby. Neural network ensembles, cross validation, and active learning. In *Advances in Neural Information Processing Systems*, volume 7, 1994.

Daniel LeJeune, Hamid Javadi, and Richard Baraniuk. The implicit regularization of ordinary least squares ensembles. In *Proceedings of the Twenty Third International Conference on Artificial Intelligence and Statistics*, volume 108 of *Proceedings of Machine Learning Research*, pp. 3525–3535. PMLR, 2020.

Lisha Li, Kevin Jamieson, Giulia DeSalvo, Afshin Rostamizadeh, and Ameet Talwalkar. Hyperband: A novel bandit-based approach to hyperparameter optimization. *Journal of Machine Learning Research*, 18(185): 1–52, 2018.

Brian Liu and Rahul Mazumder. Randomization can reduce both bias and variance: A case study in random forests. *Journal of Machine Learning Research*, 26(150):1–49, 2025.

Yuheng Ma and Qiang Sun. Feature bagging provides stability. *arXiv preprint arXiv:2607.26964*, 2026.

Tianxing Mei, Yingying Fan, Mingming Leng, and Jinchi Lv. CART random forests as sequential allocation over random opportunity sets: A stochastic-control theory of ensemble risk. *arXiv preprint arXiv:2605.26675*, 2026.

Lucas Mentch and Siyu Zhou. Randomization as regularization: A degrees of freedom explanation for random forest success. *Journal of Machine Learning Research*, 21(171):1–36, 2020.

Philipp Probst, Marvin N. Wright, and Anne-Laure Boulesteix. Hyperparameters and tuning strategies for random forest. *Wiley Interdisciplinary Reviews: Data Mining and Knowledge Discovery*, 9(3):e1301, 2019.

Ivan Rubachev, Nikolay Kartashev, Yury Gorishniy, and Artem Babenko. TabReD: Analyzing pitfalls and filling the gaps in tabular deep learning benchmarks. In *The Thirteenth International Conference on Learning Representations*, 2025.

Ben Ruben and Cengiz Pehlevan. Learning curves for noisy heterogeneous feature-subsampled ridge ensembles. In *Advances in Neural Information Processing Systems*, volume 36, pp. 50041–50093, 2023.

Jasper Snoek, Hugo Larochelle, and Ryan P. Adams. Practical bayesian optimization of machine learning algorithms. In *Advances in Neural Information Processing Systems*, volume 25, 2012.

Jake A. Soloff, Rina Foygel Barber, and Rebecca Willett. Bagging provides assumption-free stability. *Journal of Machine Learning Research*, 25(131):1–35, 2024.

Joaquin Vanschoren, Jan N. van Rijn, Bernd Bischl, and Luis Torgo. OpenML: Networked science in machine learning. *ACM SIGKDD Explorations Newsletter*, 15(2):49–60, 2014.

APPENDIX CONTENTS

The supplement follows the three claims in the main text.

1. Appendix A maps the assumptions and derives the grouped search representation for forward regression and CART.
2. Appendix B proves the one-search posterior, general-covariance, and fixed-design weak-signal results.
3. Appendix C proves the target-size OOB correction and specifies shared-search screening and selective continuation.
4. Appendix D proves the recursive SNR theorem and the coupled stagewise estimators.
5. Appendix E reports the 12-task geometric-refinement benchmark, its resource measurements and midpoint ablation, and the one-search and recursive mechanism studies.

Guide to the proofs of the main results. The following list maps the main results to their proofs, in main-text order.

1. Proposition 2.1 is proved in Appendix B.1 by combining the ensemble second moment, the projection relation, and the Gaussian posterior mean.
2. Theorem 2.2 is proved in Appendix B.2. The argument bounds lost signal alignment, retains the full-search winner's contribution to diversity, and controls feature normalization by Gaussian concentration.
3. Corollary 2.3 is proved in Appendix B.3 by comparing the leading signal and noise correlations and summing squared rank probabilities. That subsection also gives the general fixed-design sufficient condition in Proposition B.1.
4. Theorem 3.1 is proved in Appendix D.1 by telescoping completed-hybrid risks and decomposing terminal diversity into martingale increments; the common projection relation gives the SNR specialization.
5. The target-size OOB identity used in Section 4.2 is stated as Proposition C.1 and proved in Appendix C.1. Conditional on the OOB count, the correction replaces the variance of an average of r_i trees by that of B trees.

A MODELS, ASSUMPTIONS, AND EXAMPLES

A.1 ASSUMPTION MAP

Table 2 separates the statistical regimes used by the main results. The exact one-search posterior result permits arbitrary candidate correlations but uses an isotropic signal prior in prediction space. The fixed-signal result permits a general feature covariance but treats one-step forward regression. The target-size correction concerns a fixed OOB row and budget; it requires conditionally iid OOB trees but permits dependence across budgets. The recursive SNR theorem is general in its first identity and adds a common-score projection condition only for the SNR specialization.

Table 2: Assumptions and conclusions of the main theoretical layers.

Result	Dependence	Signal model	Scope and conclusion
Equation (20)	Arbitrary fitted predictors	Fixed regression target	Exact restriction–diversity accounting
Proposition 2.1	Arbitrary atom Gram matrix	Isotropic Gaussian prior, or fixed null	Exact conditional SNR boundary for one grouped adaptive search
Theorem 2.2	General positive-definite feature covariance	Fixed β	One-step forward regression; alignment bound and fixed-null dominance
Corollary 2.3	Orthogonal fixed design	Fixed β, Gaussian errors	Explicit limiting budget ordering under weak signals
Proposition B.1	General normalized fixed Gram	Fixed β, Gaussian errors	Weak-marginal-signal sufficient condition
Proposition C.1	Independent tree randomizations conditional on training data; shared bootstrap samples across budgets	Positive OOB probability and finite conditional second moments	Exact rowwise correction to the B-tree OOB risk for fixed budgets
Theorem 3.1	Arbitrary adaptive recursion; common-score projection for SNR statement	Fixed target for general identity; sequence model for SNR statement	Exact recursive accounting and weak–strong SNR boundary

A.2 What is fixed and what is averaged

The same expectation symbol refers to different probability laws in the three parts of the paper. The following conventions make these laws explicit.

- *One-search posterior risk.* The design and atom dictionary are fixed before drawing F and W in Equation (5). After observing Z, the risk averages over $F \mid Z$ and fresh feature subsets, independently conditional on Z. The atoms selected by the response are not conditioned on separately: they are functions of Z and the fixed dictionary.
- *Fixed-signal risk.* In Theorem 2.2, β and Σ are fixed. Conditional on $\mathcal{D} = (X, y)$, only the feature masks are averaged. The Σ norm integrates over a fresh test covariate. The training-sample probability controls the remainder and the event of improvement. Corollary 2.3 fixes X as well, so its probability statements concern only the Gaussian training errors.
- *OOB correction.* Fix $\mathcal{D}$, one row i, budget m, and number b of generated trees. Conditioning on $r_i(b) = r$ fixes a count, not the bootstrap samples, feature subsets, or fitted predictions. It does not condition on the budget having survived a data-dependent elimination.
- *Recursive risk.* The general identity fixes a background $\mathcal{A}$ containing training data and non-feature randomization, then averages over feature subsets. For the posterior specialization, the fixed background is instead the design and common geometry, followed by conditioning on Z; every hybrid uses only Z and independent subsets. Adding other information about F to this conditioning could change its posterior mean.

Conditional identities can be integrated over their conditioning variables by the tower property when the relevant risks are integrable. Likewise, a conditional probability bound that holds almost surely with the same failure probability also holds unconditionally. These operations preserve the estimand: for example, a bound around a training-dependent risk curve remains a bound around that random curve, not automatically around its training-sample average. An identity at each fixed budget does not become unbiased after minimizing over noisy scores.

A.3 Ambiguity identity in the paper's notation

For a fixed target f^* in a Hilbert space, let $\mathcal{R}(g \mid \mathcal{D}) = \mathbb{E}[\|g - f^*\|^2 \mid \mathcal{D}]$, where the training sample $\mathcal{D}$ is held fixed and the expectation averages feature randomization in the estimator g. Let h_F be the full-search fit and h_ξ its random-subset counterpart. Define

$$\Delta_{\mathrm{res}} = \mathcal{R}(h_\xi \mid \mathcal{D}) - \mathcal{R}(h_F \mid \mathcal{D}), \qquad V = \mathbb{E}_\xi \left\| h_\xi - \mathbb{E}_\xi h_\xi \right\|^2 .$$

For $\bar{h}_B = B^{-1} \sum_{b=1}^{B} h_{\xi_b}$ with independent subsets, the squared-loss ambiguity identity (Krogh & Vedelsby, 1994) gives

$$\mathcal{R}(\bar{h}_B \mid \mathcal{D}) - \mathcal{R}(h_F \mid \mathcal{D}) = \Delta_{\mathrm{res}} - (1 - B^{-1})V. \tag{20}$$

Proof of Equation (20). Condition on the training data, so only the independent feature subsets $\xi_1, \ldots, \xi_B$ remain random. Write $H_b = h_{\xi_b}$, $\bar{H} = B^{-1}\sum_b H_b$, and $\mu = \mathbb{E}_\xi h_\xi$. For every realization, expanding $H_b - f^* = (H_b - \bar{H}) + (\bar{H} - f^*)$ and summing over b gives

$$\frac{1}{B}\sum_{b=1}^{B}\|H_b - f^*\|^2 = \left\|\bar{H} - f^*\right\|^2 + \frac{1}{B}\sum_{b=1}^{B}\left\|H_b - \bar{H}\right\|^2. \tag{21}$$

Independence of the H_b implies

$$\mathbb{E}\left\{\frac{1}{B}\sum_b \left\|H_b - \bar{H}\right\|^2\right\} = \mathbb{E}_\xi\|h_\xi\|^2 - \mathbb{E}\left\|\bar{H}\right\|^2$$
$$= (1 - B^{-1})\{\mathbb{E}_\xi\|h_\xi\|^2 - \|\mu\|^2\} = (1 - B^{-1})V.$$

Taking expectations in Equation (21) and subtracting $\mathcal{R}(h_F \mid \mathcal{D})$ proves Equation (20). □

A.4 Forward regression and CART as grouped searches

In both examples, Z is the response-derived object currently being fitted, $\phi_{j\ell}$ is a unit-norm candidate direction, and $|\langle Z, \phi_{j\ell}\rangle|^2$ is its least-squares improvement.

Forward regression. Let $X = (x_1, \ldots, x_p) \in \mathbb{R}^{n\times p}$ and $y \in \mathbb{R}^n$ be centered. If the variables in $S \subset [p]$ have already been selected, let P_S project onto $\mathrm{span}(X_S)$ and define

$$Z = (I - P_S)y, \qquad u_{j|S} = (I - P_S)x_j, \qquad \phi_{j1} = u_{j|S}/\left\|u_{j|S}\right\|. \tag{22}$$

Features with $u_{j|S} = 0$ are discarded. Raw feature j is one group and has one candidate.

Proof of the forward-regression representation. The Frisch–Waugh–Lovell identity gives the fitted increment $h_j = \langle Z, \phi_{j1}\rangle\,\phi_{j1}$ and

$$\|Z\|^2 - \min_{a\in\mathbb{R}}\|Z - a\phi_{j1}\|^2 = |\langle Z, \phi_{j1}\rangle|^2 = \|h_j\|^2.$$

Forward regression therefore chooses the group maximizing Equation (1). At its first step, $S = \varnothing$, $Z = y$, and $\phi_{j1} = x_j/\|x_j\|$.

For the first-step notation in Section 2, put $z_j = x_j^\top y/n$ and $\widehat{G}_{jj} = x_j^\top x_j/n$. The average training loss of coefficient a on feature j is

$$\ell_j(a) = n^{-1}\|y - ax_j\|^2 = n^{-1}\|y\|^2 - 2az_j + a^2\widehat{G}_{jj}.$$

Its minimizer and improvement over $a = 0$ are

$$b_j = z_j/\widehat{G}_{jj}, \qquad \ell_j(0) - \ell_j(b_j) = z_j^2/\widehat{G}_{jj} = s_j.$$

The corresponding p-dimensional coefficient vector is $b_j e_j$. □

CART split search. Consider a node containing observations $i \in A$, $n_A = |A|$, and use the empirical inner product $\langle u, v\rangle_A = n_A^{-1}\sum_{i\in A} u_i v_i$. Let $Z_i = Y_i - \bar{Y}_A$. For feature j, let $t_{j1}, \ldots, t_{jL_j}$ be its admissible cutpoints and set

$$q_{j\ell} = n_A^{-1}\sum_{i\in A}\mathbf{1}\{X_{ij} \le t_{j\ell}\}, \qquad \phi_{j\ell,i} = \frac{\mathbf{1}\{X_{ij} \le t_{j\ell}\} - q_{j\ell}}{\{q_{j\ell}(1 - q_{j\ell})\}^{1/2}}. \tag{23}$$

Admissibility gives $0 < q_{j\ell} < 1$, and the atom is centered with unit norm.

Proof of Equations (24) *and* (25). If $\bar{Z}_L$ and $\bar{Z}_R$ are the two child means, then

$$\langle Z, \phi_{j\ell}\rangle_A = \{q_{j\ell}(1 - q_{j\ell})\}^{1/2}(\bar{Z}_L - \bar{Z}_R), \tag{24}$$

and

$$\|Z\|_A^2 - \min_a\|Z - a\phi_{j\ell}\|_A^2 = q_{j\ell}(1 - q_{j\ell})(\bar{Z}_L - \bar{Z}_R)^2. \tag{25}$$

Multiplying by n_A gives the usual residual-sum-of-squares improvement. □

Thus group j is a raw feature, its members are the admissible cutpoints, the within-group search selects the best cutpoint, and the outer search compares the featurewise winners. Feature randomization samples feature groups while retaining the complete cutpoint family of every exposed feature.

B PROOFS FOR THE ONE-SEARCH SNR THEORY

B.1 GROUPED-SEARCH SNR IDENTITY

Proof of Proposition 2.1. Condition on the fixed dictionary and Z throughout. Gaussian conditioning gives $\mathbb{E}(F \mid Z) = a^2(a^2+\omega^2)^{-1}Z$: equivalently, $F - \kappa Z$ is jointly Gaussian with Z and has zero cross-covariance, so it is independent of Z and has mean zero. Since the feature subsets are independent of (F, W), conditioning additionally on their values does not change this posterior mean. Let $H = h_\xi$, $\mu = \mathbb{E}_\xi H$, and let $H_1, \dots, H_B$ be independent copies. Direct expansion gives

$$\mathbb{E}\left\|B^{-1}\textstyle\sum_b H_b\right\|^2 = \mathbb{E}_\xi \|H\|^2 - (1 - B^{-1})V. \tag{26}$$

By Equation (2),

$$\langle \mu, Z\rangle = \mathbb{E}_\xi \|H\|^2, \qquad \langle h_F, Z\rangle = \|h_F\|^2. \tag{27}$$

The posterior covariance of F given Z contributes the same constant to both risks. Since $\mathbb{E}(F \mid Z) = \kappa Z$,

$$\begin{aligned}\mathcal{R}(\bar{h}_B \mid Z) - \mathcal{R}(h_F \mid Z) &= \mathbb{E}\left\|\bar{h}_B\right\|^2 - \|h_F\|^2 - 2\kappa\langle \mu - h_F, Z\rangle \\ &= (1-2\kappa)\{\mathbb{E}_\xi\|H\|^2 - \|h_F\|^2\} - (1-B^{-1})V \\ &= (2\kappa - 1)D - (1 - B^{-1})V.\end{aligned}$$

The full search maximizes energy, so $D \geq 0$, and $V \geq 0$ by definition. The dominance and boundary statements follow. At $F = 0$, the same calculation compares prediction energies directly and does not use a prior. □

B.2 GENERAL-COVARIANCE ALIGNMENT BOUND

Proof of Theorem 2.2. Fix $\mathcal{D} = (X, y)$ first; all feature-subset probabilities and expectations in the finite-sample argument are conditional on this sample. Let J_ξ be the winner in one uniformly random size-m feature subset and let $\eta = \max_j |\widehat{G}_{jj} - 1|$. The coefficient energy of candidate j is $b_j^2 = s_j/\widehat{G}_{jj}$. On $\eta < 1$,

$$b_{J_\xi}^2 - b_{j_*}^2 \leq \frac{M}{1-\eta} - \frac{M}{1+\eta} = \frac{2\eta M}{1-\eta^2}.$$

Let $\Delta_{\mathrm{res}}(\beta) = \mathcal{R}(\widehat{\beta}_\xi \mid \mathcal{D}) - \mathcal{R}(\widehat{\beta}_F \mid \mathcal{D})$. Expanding prediction risk in the Σ norm yields

$$\Delta_{\mathrm{res}}(\beta) \leq \frac{2\eta M}{1-\eta^2} + 2A_\beta. \tag{28}$$

Let $C = b_{J_\xi} e_{J_\xi}$ and $\pi_j = \Pr(J_\xi = j \mid X, y)$. Since $\Sigma \succeq \lambda_- I$,

$$\begin{aligned}V &= \mathbb{E}_\xi(C - \mathbb{E}_\xi C)^\top \Sigma (C - \mathbb{E}_\xi C) \\ &\geq \lambda_- \sum_{j=1}^p \pi_j(1-\pi_j) b_j^2.\end{aligned} \tag{29}$$

The full-search winner wins the restricted search exactly when it is exposed, so $\pi_{j_*} = m/p = \gamma_p$. Therefore

$$V \geq \frac{\lambda_- \gamma_p(1-\gamma_p)M}{1+\eta}. \tag{30}$$

Combining Equation (20), Equation (28), and Equation (30) gives the finite-sample bound

$$\mathcal{R}(\bar{\beta}_B \mid \mathcal{D}) - \mathcal{R}(\widehat{\beta}_F \mid \mathcal{D}) \leq 2A_\beta + \frac{2\eta M}{1-\eta^2} - (1 - B^{-1})\frac{\lambda_-\gamma_p(1-\gamma_p)M}{1+\eta}. \tag{31}$$

For a column-normalized fixed design, $\eta = 0$ and this statement is deterministic. This does not require the empirical and population off-diagonal Gram entries to agree: the candidate energies use $\mathrm{diag}(\Sigma) = 1$, and the diversity bound uses Σ.

For the Gaussian random design, $\widehat{G}_{jj} \stackrel{d}{=} \chi^2_n/n$. A standard chi-square tail bound and a union bound give, for $0 < t < 1$,

$$\Pr(\eta > t) \le 2p \exp(-nt^2/8).$$

Thus $\log p = o(n)$ implies $\eta = o_P(1)$, and the two scale corrections in Equation (31) are $o_P(M)$. At the fixed null, $A_\beta = 0$; if the diversity coefficient stays bounded away from zero, the gap is strictly negative on $\{M > 0\}$ with probability tending to one. Therefore $\Pr(M > 0) \to 1$ gives the stated conclusion; this holds, for example, under conditionally continuous nondegenerate errors. If $M = 0$, all one-variable fits vanish and the procedures coincide. Finally, Cauchy–Schwarz gives $s_j \le n^{-1} \|y\|^2$, hence $M \le n^{-1} \|y\|^2$. If $\beta^\top \Sigma \beta$ and the errors' second moments are uniformly bounded, Markov's inequality yields $M = O_P(1)$ without needing independence between feature columns. □

B.3 Fixed-design weak signals and budget ordering

The following sufficient condition extends the fixed-null conclusion to weak signals under a general normalized fixed design.

Proposition B.1 (Weak marginal signal under a fixed design) Use the coefficient fits and risks of Theorem 2.2, now with fixed X and test-risk geometry $\Sigma = \Sigma_p = X^\top X/n$. Assume $\mathrm{diag}(\Sigma_p) = 1$, $\inf_p \lambda_{\min}(\Sigma_p) \ge \lambda_0 > 0$, and $m/p \to \gamma \in (0, 1)$. As $p \to \infty$, let $B > 1$ be fixed, $\sigma > 0$, and $\varepsilon \sim N(0, \sigma^2 I_n)$. Define

$$\tau_p = \|\Sigma_p \beta\|_\infty, \qquad q_p = \mathbb{E} \left\|X^\top \varepsilon/n\right\|_\infty, \qquad \rho_p = \tau_p/q_p.$$

If

$$\limsup_{p \to \infty} \rho_p < \frac{(1 - B^{-1})\lambda_0 \gamma}{4 + (1 - B^{-1})\lambda_0 \gamma}, \tag{32}$$

then $\mathcal{R}(\bar{\beta}_B \mid \mathcal{D}) < \mathcal{R}(\widehat{\beta}_F \mid \mathcal{D})$ with probability tending to one.

The quantity τ_p is the largest marginal signal, while q_p is the expected largest noise correlation. The condition therefore compares the signal available to any feature with the noise that an adaptive search can exploit. The positive eigenvalue condition implies $p \le n$.

Proof of Proposition B.1. Under $\widehat{G} = \Sigma_p$ with unit diagonal, write $g = \Sigma_p \beta$ and $r = X^\top \varepsilon/n$, so $b = g + r$ and $M = \|g + r\|^2_\infty$. Let $\pi_j = \Pr_\xi(J_\xi = j \mid X, y)$ and $u = \pi \odot b$. Exact expansion gives

$$D = M - \sum_j \pi_j b_j^2, \qquad V = \sum_j \pi_j b_j^2 - u^\top \Sigma_p u,$$

and

$$A_\beta = (1 - \gamma_p) b_{j_*} g_{j_*} - \sum_{j \ne j_*} \pi_j b_j g_j.$$

Because $|b_j| \le \sqrt{M}$, $|g_j| \le \tau_p$, and $\sum_{j \ne j_*} \pi_j = 1 - \gamma_p$, $|A_\beta| \le 2(1 - \gamma_p)\tau_p \sqrt{M}$. Also,

$$V \ge \lambda_0 \sum_j \pi_j (1 - \pi_j) b_j^2 \ge \lambda_0 \gamma_p (1 - \gamma_p) M.$$

Dropping $D \ge 0$ from the exact ambiguity expansion gives

$$\mathcal{R}(\bar{\beta}_B \mid \mathcal{D}) - \mathcal{R}(\widehat{\beta}_F \mid \mathcal{D}) \le (1 - \gamma_p)\sqrt{M}\{4\tau_p - a_p \sqrt{M}\}, \qquad a_p = (1 - B^{-1})\lambda_0 \gamma_p. \tag{33}$$

Let $\zeta_p = \|r\|_\infty$. As a function of a standard Gaussian vector, ζ_p is $\sigma/\sqrt{n}$-Lipschitz. Gaussian concentration gives $\zeta_p - q_p = O_P(\sigma/\sqrt{n})$. To bound q_p from below, write in distribution $r = (\sigma/\sqrt{n})(\sqrt{\lambda_0} U + V_0)$, where $U \sim N(0, I_p)$ and $V_0 \sim N(0, \Sigma_p - \lambda_0 I_p)$ are independent. Conditional

Jensen's inequality over V_0 gives $q_p \geq (\sigma/\sqrt{n})\sqrt{\lambda_0}\mathbb{E}\,\|U\|_\infty \gtrsim \sigma\sqrt{\log(p)/n}$. Therefore $\zeta_p/q_p \to 1$ in probability. Since $\sqrt{M} \geq \zeta_p - \tau_p$, condition Equation (32), equivalently

$$\limsup_p \frac{\tau_p}{q_p} < \frac{(1-B^{-1})\lambda_0\gamma}{4+(1-B^{-1})\lambda_0\gamma},$$

implies $\sqrt{M} > 4\tau_p/a_p$ with probability tending to one. The right-hand side of Equation (33) is then negative. This is exactly the threshold in Equation (32) because $\rho_p = \tau_p/q_p$. □

Proof of Corollary 2.3. Under orthogonality, set $r = X^\top\varepsilon/n$, $\tau_p = \|\beta\|_\infty$, and $\zeta_p = \|r\|_\infty$. The argument above gives $\zeta_p/q_p \to 1$ in probability, and the weak-signal assumption is $\tau_p/q_p \to 0$. The coordinates of r are independent $N(0,\sigma^2/n)$ variables. Let $W_j = (\beta_j + r_j)^2$ and let $W_{(r)}$ and $R_{(r)}$ be the decreasing order statistics of W_j and r_j^2. Then

$$\max_j |W_j - r_j^2| \leq 2\tau_p\zeta_p + \tau_p^2 = o_P(q_p^2),$$

and $W_{(r)}/q_p^2 \to 1$ for every fixed r. By the rank representation,

$$\mathbb{E}_\xi\left\|\bar{\beta}_B\right\|^2 = \sum_{r=1}^{p}\{\pi_r^2 + B^{-1}\pi_r(1-\pi_r)\}W_{(r)}. \tag{34}$$

If $m/p \to \gamma$, then $\pi_r \to w_r = \gamma(1-\gamma)^{r-1}$. For any fixed cutoff k, the omitted rank mass is exactly

$$\sum_{r>k}\pi_r(m) = \frac{\binom{p-k}{m}}{\binom{p}{m}} \leq (1-m/p)^k.$$

The coefficient of $W_{(r)}$ in Equation (34) is at most π_r. Since $W_{(1)}/q_p^2 = O_P(1)$, the normalized tail is bounded by $O_P(1)(1-\gamma/2)^k$ for large p. Taking $p \to \infty$ at fixed k and then $k \to \infty$ justifies passage to the limit, yielding

$$\frac{\mathbb{E}_\xi\left\|\bar{\beta}_B\right\|^2}{q_p^2} \xrightarrow{P} \sum_{r\geq 1}\{w_r^2 + B^{-1}w_r(1-w_r)\} = \frac{1}{B} + \left(1-\frac{1}{B}\right)\frac{\gamma}{2-\gamma}.$$

The signal cross term is $o_P(q_p^2)$ because $|\sum_j \beta_j b_j \pi_j| \leq \tau_p(\tau_p + \zeta_p)$. Expanding prediction risk proves Equation (10). Finally, $\partial c_{\gamma,B}/\partial\gamma = 2(1-B^{-1})/(2-\gamma)^2 > 0$ for $B > 1$. □

B.4 Controls for the signal-alignment term

Let $w = b_{j_*}e_{j_*} - \mathbb{E}_\xi(b_{J_\xi}e_{J_\xi})$.

Proof of Equations (35) *and* (36). Cauchy–Schwarz gives

$$A_\beta = \langle w, \beta\rangle_\Sigma \leq \|w\|_\Sigma\,\|\beta\|_\Sigma\,. \tag{35}$$

Thus a valid signal-norm radius and the population covariance yield a deterministic upper bound. Replacing the latter by an estimate also requires controlling its estimation error. Alternatively, on an independent labeled sample C,

$$\widehat{A}_C = |C|^{-1}\sum_{i\in C}(X_i^\top w)Y_i, \qquad \mathbb{E}(\widehat{A}_C \mid X, y) = A_\beta. \tag{36}$$

□

The second route requires an additional tail assumption to turn Equation (36) into a one-sided confidence bound. Theorem 2.2 itself is structural: without either signal information or an independent estimate, A_β is not observable from the feature-bag replicates alone.

C TARGET-SIZE OOB SCORING AND SHARED SEARCH

C.1 PROOF AND DIAGNOSTICS FOR TARGET-SIZE OOB CORRECTION

Fix $\mathcal{D} = \{(X_i, Y_i)\}_{i=1}^n$ and m. For tree k, let O_{ik} indicate that its bootstrap sample omits row i and let $Z_{ik}(m)$ be its prediction at X_i. Write $r_i = \sum_{k=1}^b O_{ik}$ and list the retained predictions as $T_{i1}(m), \ldots, T_{ir_i}(m)$. Their sample mean and unbiased variance are

$$\bar{T}_i(m) = r_i^{-1} \sum_{\ell=1}^{r_i} T_{i\ell}(m), \qquad S_i^2(m) = \frac{1}{r_i - 1} \sum_{\ell=1}^{r_i} \{T_{i\ell}(m) - \bar{T}_i(m)\}^2.$$

Let $\mu_i(m), v_i(m)$ be the mean and variance under the conditional law given $(\mathcal{D}, O_{i1} = 1)$, and let $\bar{T}_{B,i}(m)$ average B fresh independent draws from that law. Then $L_{B,i}(m)$ in Equation (18) is $\mathbb{E}[\{Y_i - \bar{T}_{B,i}(m)\}^2 \mid \mathcal{D}]$.

Proposition C.1 (Target-size OOB correction) Fix $\mathcal{D}$, row i, budget m, and deterministic integers $b \geq 2$ and $B \geq 1$. Suppose the pairs $(O_{ik}, Z_{ik}(m))$, $k \in [b]$, are conditionally i.i.d. given $\mathcal{D}$, with $\Pr(O_{i1} = 1 \mid \mathcal{D}) > 0$ and $\mathbb{E}\{Z_{i1}(m)^2 \mid \mathcal{D}, O_{i1} = 1\} < \infty$. For every $r \in \{2, \ldots, b\}$ having $\Pr(r_i = r \mid \mathcal{D}) > 0$, the retained predictions are conditionally i.i.d. given $(\mathcal{D}, r_i = r)$, with mean $\mu_i(m)$ and variance $v_i(m)$. Consequently,

$$\mathbb{E}\{\widehat{L}_{B,i}(m) \mid \mathcal{D}, r_i = r\} = L_{B,i}(m). \tag{37}$$

For two fixed budgets using the same bootstrap samples, the same count r_i applies and the conditional expectation of the difference of their estimates equals $L_{B,i}(m) - L_{B,i}(m')$. Independence between budgets is unnecessary.

Proof of Proposition C.1. Work throughout under the conditional law given the fixed training sample $\mathcal{D}$. Fix i, m, b, B as in the proposition and a count $r \in \{2, \ldots, b\}$ with $\Pr(r_i = r \mid \mathcal{D}) > 0$. Let $Q_{i,m,\mathcal{D}}$ denote the conditional distribution of $Z_{i1}(m)$ given $(\mathcal{D}, O_{i1} = 1)$. The positive omission probability and finite conditional second moment ensure that this distribution and its moments $\mu_i(m), v_i(m)$ are well defined.

Conditional law of the retained predictions. Write $O_i = (O_{i1}, \ldots, O_{ib})$. For any $o \in \{0,1\}^b$ with $\sum_k o_k = r$ and $\Pr(O_i = o \mid \mathcal{D}) > 0$, let $k_1 < \cdots < k_r$ be its indices with $o_{k_\ell} = 1$. On $O_i = o$, $T_{i\ell}(m) = Z_{ik_\ell}(m)$ for $\ell \in [r]$. Conditional independence of the pairs $(O_{ik}, Z_{ik}(m))$ implies, for any Borel sets $A_1, \ldots, A_r \subseteq \mathbb{R}$,

$$\Pr\{T_{i1}(m) \in A_1, \ldots, T_{ir}(m) \in A_r \mid \mathcal{D}, O_i = o\} = \prod_{\ell=1}^r Q_{i,m,\mathcal{D}}(A_\ell).$$

Here identical distributions of the pairs make every retained factor the same conditional law $Q_{i,m,\mathcal{D}}$. The product does not depend on which indices are retained. The conditional law of total probability, summing over o with weights $\Pr(O_i = o \mid \mathcal{D}, r_i = r)$, therefore gives

$$\Pr\{T_{i1}(m) \in A_1, \ldots, T_{ir}(m) \in A_r \mid \mathcal{D}, r_i = r\} = \prod_{\ell=1}^r Q_{i,m,\mathcal{D}}(A_\ell).$$

Thus the retained predictions are conditionally i.i.d. given $(\mathcal{D}, r_i = r)$. This argument conditions on the omission indicators without assuming independence between a tree's indicator and its prediction.

Conditional moments and correction. Suppress i, m from $\bar{T}_i(m), S_i^2(m), \mu_i(m), v_i(m)$, writing $\bar{T}, S^2, \mu, v$. The conditional iid law gives

$$\mathbb{E}(\bar{T} \mid \mathcal{D}, r_i = r) = \mu, \qquad \mathrm{Var}(\bar{T} \mid \mathcal{D}, r_i = r) = \frac{v}{r}.$$

Using $\sum_{\ell=1}^r (T_{i\ell}(m) - \bar{T})^2 = \sum_{\ell=1}^r (T_{i\ell}(m) - \mu)^2 - r(\bar{T} - \mu)^2$ yields

$$\mathbb{E}(S^2 \mid \mathcal{D}, r_i = r) = \frac{rv - r(v/r)}{r - 1} = v.$$

All these conditional expectations are finite by the second-moment assumption. The observed response Y_i is fixed under conditioning on $\mathcal{D}$, so

$$\mathbb{E}\{(Y_i - \bar{T})^2 \mid \mathcal{D}, r_i = r\} = (Y_i - \mu)^2 + v/r.$$

Likewise, the average $\bar{T}_{B,i}(m)$ of B fresh independent draws from $Q_{i,m,\mathcal{D}}$ has conditional mean μ and variance v/B. This proves the target identity in Equation (18). Substituting these moments into the correction gives

$$\begin{aligned}\mathbb{E}\{\widehat{L}_{B,i}(m) \mid \mathcal{D}, r_i = r\} &= \mathbb{E}\{(Y_i - \bar{T})^2 \mid \mathcal{D}, r_i = r\} + \left(\frac{1}{B} - \frac{1}{r}\right)\mathbb{E}(S^2 \mid \mathcal{D}, r_i = r) \\ &= \{Y_i - \mu\}^2 + \frac{v}{r} + \left(\frac{1}{B} - \frac{1}{r}\right)v \\ &= \{Y_i - \mu\}^2 + \frac{v}{B} = L_{B,i}(m).\end{aligned}$$

For a second fixed budget m' satisfying the same assumptions, the preceding argument gives its marginal conditional identity as well. Shared bootstrap samples imply identical omission indicators and hence the same event $\{r_i = r\}$ for both budgets. Both per-observation risk estimates are integrable under this conditional law. Subtracting their conditional expectations proves the claimed paired identity, irrespective of dependence between budgets.

Since $L_{B,i}(m)$ does not depend on r, averaging the identity over $r \geq 2$ also gives $\mathbb{E}\{\widehat{L}_{B,i}(m) \mid \mathcal{D}, r_i \geq 2\} = L_{B,i}(m)$. This averages a single row's statistic over its eligible counts, not a random ratio obtained by aggregating different rows. □

Why the bootstrap coupling matters. All candidate budgets use the same bootstrap inclusion indicators. Thus r_i and the eligible-observation set are identical across budgets, while the tree predictions may remain strongly dependent. For two budgets m and m', define the difference between their risk estimates for observation i,

$$D_i(m, m') = \widehat{L}_{B,i}(m) - \widehat{L}_{B,i}(m').$$

The implemented score $\widehat{L}_B(m)$ averages the per-observation risk estimates $\widehat{L}_{B,i}(m)$ from Equation (19) over $\mathcal{E} = \{i : r_i \geq 2\}$, giving each eligible row equal weight. Its paired comparison is $|\mathcal{E}|^{-1}\sum_{i\in\mathcal{E}} D_i(m, m')$ when $|\mathcal{E}| > 0$. Proposition C.1 establishes rowwise centering for fixed budgets and a fixed prefix. The eligibility set and its denominator are random and share bootstrap draws with the predictions, so this identity alone does not establish exact unbiasedness of the aggregate or a conditional guarantee after nominee selection. In particular, conditioning on a budget having survived earlier scores, or stopping at a score-dependent tree count, is not the conditioning used in the proposition. The common rows make comparisons paired; the outer test folds in Section 5 evaluate the resulting selection procedure.

The target $L_{B,i}(m)$ concerns trees conditioned to omit row i. It is not an exact identity for the population test risk of a forest trained on all rows. Whenever the relevant expectations are finite, averaging over the training sample preserves this row-omitting target; relating it to the full-sample target requires a separate argument.

Relation to raw-prefix and infinite-forest risk. For an idealized comparison in which each observation receives exactly r row-omitting predictions that are conditionally i.i.d. given $\mathcal{D}$, the conditional risk curves have the schematic form

$$F_r(m) = F_\infty(m) + \frac{V(m)}{r}, \qquad F_B(m) = F_\infty(m) + \frac{V(m)}{B},$$

where $F_\infty(m)$ averages the squared deviations $\{Y_i - \mu_i(m)\}^2$ and $V(m)$ averages $v_i(m)$ over a fixed set of rows. Actual OOB counts vary by row, giving terms $v_i(m)/r_i$ rather than a single $V(m)/r$. If $V(m)$ differs across budgets, the roundwise ordering need not match that of F_B. Equation (19) estimates the second curve at every round, which is the reason a short full-depth comparison can be informative about a 256-tree target despite substantial Monte Carlo variation.

C.2 SHARED SCREENING AND HALF-CORRECTED SCORING

The geometric-refinement selector uses shared 4–8–16 screening and a half-corrected terminal score. This subsection defines these components; Appendix E.1 specifies the two refinement triggers, geometric nomination, and final comparison.

Allocation and scoring. For target size $B = 256$, start with all $G = |\mathcal{M}|$ budgets and fit four full-depth trees per budget. At tree count $b \in \{4, 8, 16\}$, the prefix means the first b full-depth trees. Let $\mathcal{E} = \{i : r_i(b) \geq 2\}$ be the rows with at least two OOB predictions, common to every active budget. When $|\mathcal{E}| > 0$, define the average risk estimate and the average raw OOB loss, respectively, by

$$\widehat{L}_B(m) = \frac{1}{|\mathcal{E}|} \sum_{i \in \mathcal{E}} \widehat{L}_{B,i}(m), \qquad \widehat{L}_{\mathrm{raw}}(m) = \frac{1}{|\mathcal{E}|} \sum_{i \in \mathcal{E}} \{Y_i - \bar{T}_i(m)\}^2.$$

Rank by $\widehat{L}_B$, retain the best $\lceil G/2 \rceil$ budgets, extend them to eight trees, and again retain half the active budgets, rounded up. Extend these survivors to 16 trees and compute the half-corrected score

$$\widehat{L}_{\mathrm{half}}(m) = \frac{\widehat{L}_B(m) + \widehat{L}_{\mathrm{raw}}(m)}{2} = \widehat{L}_{\mathrm{raw}}(m) - \frac{1}{2|\mathcal{E}|} \sum_{i \in \mathcal{E}} \left(\frac{1}{r_i} - \frac{1}{B} \right) S_i^2(m). \tag{38}$$

Every comparison uses the same eligible rows for all active budgets. Scoring requires at least two eligible rows; no completed case fails this requirement. Ties favor larger m.

Interpretation of half correction. The contribution of observation i to the half-corrected score is

$$\widehat{L}_{\mathrm{half},i}(m) := \tfrac{1}{2}\{\widehat{L}_{\mathrm{raw},i}(m) + \widehat{L}_{B,i}(m)\}.$$

Proof of the half-correction expectation identity. For fixed i, m, b and a supported count $r \geq 2$ under Proposition C.1,

$$\mathbb{E}\{\widehat{L}_{\mathrm{half},i}(m) \mid \mathcal{D}, r_i = r\} = L_{B,i}(m) + \tfrac{1}{2}(1/r - 1/B)v_i(m). \qquad \square$$

Thus each observation's contribution retains half of the excess variance term when $r < B$. The shrinkage reduces the coefficient on an estimated correction; it does not itself guarantee lower score variance or lower selection regret.

Shared computation and worker allocation. At each common node, budgets search nested prefixes of one uniform feature permutation: the size-m subset consists of the first m feature indices. The permutation is fresh given the preceding node history. Featurewise split fits are computed once within each active worker chunk, and budgets with the same winning split share child-node calculations. Separate chunks or diverging partitions are fitted separately. Worker chunks are balanced to reduce synchronization delays. This changes execution, not a budget's bootstrap samples, feature subsets, or predictions. Reuse requires identical ordered row multisets, candidate cutpoints, leaf-size constraints, and tie rules. A feature's best cutpoint is invariant to which other features are considered only at that same node. After budgets choose different splits, their descendants are processed as distinct states. One candidate–tree unit denotes one tree at one budget before crediting shared states. For G divisible by four, warm-started halving uses

$$8G + 8(G/2) + 16(G/4) = 16G. \tag{39}$$

The 4–8–16 screening schedule costs $8G$ candidate–tree units before sharing. Triggered refinement adds the nominated candidates' missing trees up to 50, including any midpoint reconstruction as specified in Appendix E.1. CPU, wall time, and feature–node work are measured separately because tree sizes and the amount of sharing vary with m.

D RECURSIVE PROOF AND COUPLED ESTIMATORS

D.1 RECURSIVE RESTRICTION, DIVERSITY, AND SNR

Proof of Theorem 3.1. For part (i), fix the background $\mathcal{A}$ from Section 3; all expectations in this part are conditional on $\mathcal{A}$. Draw the full sequence $\xi_1, \ldots, \xi_T$ before a run and ignore subsets after the path stops. This padding keeps the search indices common across completed hybrids. Set $\mathcal{F}_0 = \sigma(\mathcal{A})$, $\mathcal{F}_t = \sigma(\mathcal{A}, \xi_1, \ldots, \xi_t)$, and $\mathsf{M}_t = \mathbb{E}(H \mid \mathcal{F}_t)$. Conditional square integrability ensures these Hilbert-space means and second moments exist. The deterministic bound T gives $\mathsf{M}_T = H$ and $\mathsf{M}_0 = \mathbb{E}(H \mid \mathcal{A})$; the definition in Equation (11) is therefore $v_t = \mathbb{E}\,\|\mathsf{M}_t - \mathsf{M}_{t-1}\|^2$. Apply the ambiguity identity to the fully randomized terminal fit $H = H^{(T)}$:

$$\mathcal{R}(\bar{H}_B \mid \mathcal{A}) = \mathcal{R}(H \mid \mathcal{A}) - (1 - B^{-1})\mathbb{E}\,\|H - \mathbb{E}H\|^2 . \tag{40}$$

The hybrid risks telescope:

$$\mathcal{R}(H \mid \mathcal{A}) - \mathcal{R}(H_F \mid \mathcal{A}) = \sum_{t=1}^{T} \{\mathcal{R}(H^{(t)} \mid \mathcal{A}) - \mathcal{R}(H^{(t-1)} \mid \mathcal{A})\} = \sum_{t=1}^{T} \delta_t. \tag{41}$$

Moreover, $\mathsf{M}_t = \mathbb{E}(H \mid \mathcal{F}_t)$ is a Hilbert-space martingale. For $s < t$, its earlier increment is $\mathcal{F}_{t-1}$-measurable and $\mathbb{E}(\mathsf{M}_t - \mathsf{M}_{t-1} \mid \mathcal{F}_{t-1}) = 0$. Conditioning their inner product on $\mathcal{F}_{t-1}$ therefore gives zero. Expanding the squared norm of the sum of increments gives

$$\mathbb{E}\,\|H - \mathbb{E}H\|^2 = \sum_{t=1}^{T} \mathbb{E}\,\|\mathsf{M}_t - \mathsf{M}_{t-1}\|^2 = \sum_{t=1}^{T} v_t. \tag{42}$$

Combining Equations (40) and (42) proves Equation (13).

For part (ii), fix the design and common geometry, then condition only on Z. Use $\mathcal{F}_t = \sigma(Z, \xi_1, \ldots, \xi_t)$ and $\mathcal{F}_0 = \sigma(Z)$ for the diversity decomposition. The algorithm uses only Z and independent feature subsets, so the posterior mean remains κZ after those subsets are revealed. The posterior covariance of F contributes a common constant. For every t,

$$\begin{aligned}\mathcal{R}(H^{(t)} \mid Z) &= \mathbb{E}\left\|H^{(t)}\right\|^2 - 2\kappa\mathbb{E}\left\langle H^{(t)}, Z\right\rangle + \text{constant} \\ &= (1 - 2\kappa)\mathbb{E}\left\|H^{(t)}\right\|^2 + \text{constant},\end{aligned}$$

where the second equality uses the pathwise projection relation. Hence the posterior version of δ_t is $(2\kappa - 1)d_t$. Substitution into Equation (13) proves Equation (14). The sum of the d_t telescopes to $D_T = \|H_F\|^2 - \mathbb{E}\,\|H\|^2$, and Equation (42) gives $V_T = \mathbb{E}\,\|H - \mathbb{E}H\|^2$. The consequences following the theorem use $V_T \geq 0$: if $D_T \geq 0$ and $\kappa \leq 1/2$, the risk gap is nonpositive; if $D_T > 0$, dividing the strict risk inequality by $2D_T$ gives exactly Equation (15). □

Why response-adaptive CART still satisfies the projection relation.

Proof of the CART projection relation. Take one fixed set of n fitting observations with inner product $\langle u, v\rangle_n = n^{-1}\sum_i u_i v_i$. If a terminal partition Π places row i in leaf $A(i)$, let $n_{A(i)}$ denote the number of fitting rows in that leaf. The leaf-averaging matrix has entries

$$(P_\Pi)_{ij} = \frac{\mathbf{1}\{j \in A(i)\}}{n_{A(i)}}.$$

It is symmetric and idempotent, and $P_\Pi Z$ assigns each row the mean of Z in its leaf. Hence it is the orthogonal projection onto vectors constant within leaves. If the partition is selected adaptively from Z, the matrix $P_{\Pi(Z)}$ is random and nonlinear as a function of Z, but pathwise

$$\left\langle P_{\Pi(Z)}Z, Z\right\rangle = \left\|P_{\Pi(Z)}Z\right\|^2 . \qquad \square$$

The recursive SNR identity therefore does not require sample-split honesty. It does require every hybrid to fit the same score on a common geometry. Different row-bootstrap samples usually give different weighted leaf-averaging operators, which need not be orthogonal projections under one common inner product. Merely conditioning on their realized samples does not establish this common-projection assumption or the isotropic sequence model. The general recursive identity applies without that assumption; the posterior SNR formula requires it explicitly.

D.2 Stagewise estimators

Condition on the background $\mathcal{A}$ again. For population risk, let $C = \{(X_i, Y_i)\}_{i=1}^N$ be iid from the test law and independent of the hybrid randomizations and fitting sample. Define $\widehat{\mathcal{R}}_C(h) = N^{-1}\sum_i\{Y_i - h(X_i)\}^2$ and $\|h\|_C^2 = N^{-1}\sum_i h(X_i)^2$. For rollout $r = 1, \dots, R_0$, couple $H_r^{(t)}$ and $H_r^{(t-1)}$ by sharing their first $t-1$ feature subsets. Define

$$\widehat{\delta}_t = R_0^{-1}\sum_{r=1}^{R_0}\{\widehat{\mathcal{R}}_C(H_r^{(t)}) - \widehat{\mathcal{R}}_C(H_r^{(t-1)})\} \tag{43}$$

For the projection specialization, define

$$\widehat{d}_t = R_0^{-1}\sum_{r=1}^{R_0}\{\left\|H_r^{(t-1)}\right\|_C^2 - \left\|H_r^{(t)}\right\|_C^2\} \tag{44}$$

Proof of the expectation identities for Equations (43) *and* (44). The estimator $\widehat{\delta}_t$ satisfies $\mathbb{E}(\widehat{\delta}_t \mid \mathcal{A}) = \delta_t$, averaging over both fresh hybrid paths and C. Observation noise contributes the same constant to the two squared losses and cancels. If the realized hybrids are also conditioned on, averaging only over C instead targets their realized risk difference, not δ_t.

The estimator $\widehat{d}_t$ uses only predictions and evaluation covariates. It estimates the d_t defined in Equation (12) when the projection relation is posed in $L_2(P_X)$. For empirical least-squares projections, the same formula is evaluated on the common fitting grid that defines $\mathcal{H}$, rather than on an independent covariate sample. □

To estimate v_t, generate a pair $(H_{r,t}^{(1)}, H_{r,t}^{(2)})$ that shares the first t subsets and redraws remaining subsets $\xi_{t+1}, \dots, \xi_T$ independently. Define

$$\widehat{c}_t = \frac{1}{NR_0}\sum_{i=1}^{N}\sum_{r=1}^{R_0} H_{r,t}^{(1)}(X_i)H_{r,t}^{(2)}(X_i), \qquad \widehat{v}_t = \widehat{c}_t - \widehat{c}_{t-1} \tag{45}$$

Proof of the moment identities for Equation (45). Conditional on $\mathcal{A}$ and the first t subsets, the two completed fits are independent and both have conditional mean M_t. Their expected inner product is therefore $\|\mathsf{M}_t\|^2$. Consequently, $\widehat{v}_t$ satisfies $\mathbb{E}(\widehat{v}_t \mid \mathcal{A}) = v_t$ in $L_2(P_X)$, over the shared subsets, independent later subsets, and assessment covariates. A finite-sample difference $\widehat{v}_t$ need not be nonnegative, although $v_t \geq 0$. On a fixed empirical geometry, replace the average over C by the corresponding fixed-grid inner product. At $t = 0$, the two complete paths are independent; at $t = T$, they coincide. Use the same stored hybrid $H_r^{(t)}$ in the differences for stages t and $t+1$, and reuse each computed $\widehat{c}_t$ at both adjacent stages. With this consistency, the sums telescope exactly:

$$\sum_t \widehat{\delta}_t = R_0^{-1}\sum_r\{\widehat{\mathcal{R}}_C(H_r^{(T)}) - \widehat{\mathcal{R}}_C(H_r^{(0)})\}, \qquad \sum_t \widehat{v}_t = \widehat{c}_T - \widehat{c}_0.$$

For a deep tree, evaluating only at block endpoints estimates the accumulated contribution of all searches within each block. □

D.3 Why one-step ordering need not survive recursion

Proposition D.1 (A two-step reversal) In a three-feature null regression problem, there is a positive-definite covariance matrix and a positive-probability set of training samples on which an infinite feature-bagged ensemble exposing two features has lower conditional one-step prediction risk than full search, but higher conditional risk after a second forward-selection step and refitting. Here the risk integrates over a fresh test covariate, given the training sample.

Proof of Proposition D.1. Take

$$\Sigma = \begin{pmatrix} 1 & -.5 & .7 \\ -.5 & 1 & .1 \\ .7 & .1 & 1 \end{pmatrix}, \qquad Z = (-.7, .5, -.69)^\top.$$

The leading principal minors are 1, $3/4$, and $9/50$, so Σ is positive definite. In this example Z denotes the observed correlation vector $X^\top y/n$ for a fixed design with $X^\top X/n = \Sigma$, not the Hilbert-space response in Equation (5). Full search first selects feature 1 and has risk $.490000$. Under size-two exposure, the three equally likely subsets select features $1, 1, 3$; their infinite ensemble is $(-.466667, 0, -.230000)^\top$ and has risk $.420944$. After support S is selected, the residual score of feature j is

$$\frac{\{Z_j - \Sigma_{jS}\Sigma_{SS}^{-1}Z_S\}^2}{\Sigma_{jj} - \Sigma_{jS}\Sigma_{SS}^{-1}\Sigma_{Sj}}.$$

At step two, the randomized procedure has paths $(1, 3)$ with probability $2/3$ and $(3, 2)$ with probability $1/3$. Refitting and averaging gives risk $.615425$, whereas full search has risk $.568431$. The exact one-step risks are $7577/18000 < 49/100$; the exact two-step risks are $186437/302940 > 2899/5100$. All feature comparisons and risk inequalities are strict, so they persist in an open neighborhood of the displayed score. Under independent Gaussian null errors, $X^\top y/n \sim N(0, \sigma^2\Sigma/n)$ has a positive density everywhere; that neighborhood therefore has positive probability. This completes the counterexample. □

E Experimental details and supplementary results

E.1 Wide-grid benchmark and geometric refinement

Tasks and preprocessing. The panel contains all 12 public regression tasks in our development collection with at least 100 retained predictors (Vanschoren et al., 2014; Fischer et al., 2023; Grinsztajn et al., 2022; Rubachev et al., 2025). Table 3 identifies every task and reports the foldwise ranges of p and G. Eligibility depends on predictor count, not on the observed advantage of a selector. Each task uses a seeded shuffled five-fold assignment, rotating the test fold. Training and test samples are capped at 10,000 and 5,000 rows; smaller folds retain all rows. Numerical predictors are imputed by their training medians (zero for an entirely missing column), then transformed by the inverse hyperbolic sine, asinh. Categorical predictors use training-frequency encoding in one column each; missing values form a separate category and unseen test categories receive zero. Training-constant encoded columns are removed. The design is not subsequently standardized.

Acute Myeloid Leukemia predicts `MYC`, Cancer Drug Response predicts `Methotrexate_GDSC_1008`, and ECG5000 predicts `col_140`. Yolanda, Allstate Claims, Loan Default, and Sberbank Housing apply asinh to the response; the other eight tasks use no additional nonlinear response transform. All responses are then standardized by the outer training mean and standard deviation. Methods share the processed response and design, and risks are evaluated on this processed response scale. Trees use minimum leaf size five, at most 32 candidate cutpoints per feature, and grow until no legal split remains.

Candidate grid and comparators. The grid starts at $g = 1$ and repeatedly adds $\min\{p, \lfloor 1.1g \rfloor + 1\}$ until reaching p. We take its union with an auxiliary set initialized by the integer anchors 1, $\lceil \sqrt{p} \rceil$, $\lceil p/20 \rceil$, $\lceil p/10 \rceil$, $\lceil p/3 \rceil$, $\lceil p/2 \rceil$, $\lceil 3p/4 \rceil$, and p. After removing duplicates, enlarge this auxiliary set A to 24 values by repeatedly selecting the unused integer $m \in [p]$ that maximizes $\min_{a \in A} |\log(m/a)|$, breaking ties toward smaller m. The resulting grid has no fixed size cap and satisfies the following integer coverage property: every $m \in [p]$ has a grid value $g \le m$ with $m/g \le 1.1$. All methods use this grid and the same 256-tree target-risk curve. Full-grid OOB50 and OOB256 minimize raw OOB

Table 3: All 12 wide-feature tasks. Feature counts and grid widths show ranges over the five folds. Regret entries are five-fold means in percent. The last two columns are mean foldwise total-cost ratios of geometric refinement to standard halving (SH = 1), starting from prepared inputs. SH denotes standard successive halving.

				Mean regret (%)			Total cost / SH	
ID	Task	p	G	Ours	SH	OOB50	CPU	Wall
315	Communities and Crime	122	42	1.521	2.087	2.315	0.416	0.513
405	MTP	202	52	0.844	0.947	1.204	0.508	0.571
42705	Yolanda	100	40	0.912	1.072	0.259	0.519	0.593
44045	Allstate Claims	123–124	39–40	0.341	0.424	0.228	0.587	0.724
44046	Mercedes	351–357	54	0.403	0.399	0.913	0.421	0.532
44067	Loan Default	734	63	0.021	0.000	0.018	0.439	0.503
44793	ECG5000	140	45	3.410	2.657	1.799	0.360	0.422
46134	Acute Myeloid Leukemia	4251	85	1.036	1.348	1.142	0.395	0.421
46138	Cancer Drug Response	697	63	4.712	4.508	2.540	0.402	0.465
46300	CT Slice	378–379	54–56	1.497	2.138	3.172	0.387	0.449
46898	Sberbank Housing	289	54	0.345	0.144	0.285	0.464	0.511
46953	QSAR TID11	1024	66	0.993	0.900	0.715	0.446	0.512

loss after fitting 50 and 256 trees per budget. ECV25 uses 25 trees per budget to estimate 256-tree risk. Raw and corrected halving start every budget at eight trees, retain the best ceiling-half, extend them to 16 trees, prune again, and select among the remaining 32-tree prefixes. They differ only in whether each round uses raw OOB loss or the target-size-corrected score $\widehat{L}_B$ from Appendix C.2, with $B = 256$.

Scoring and two refinement triggers. The 4–8–16 allocation and half-corrected terminal score are those in Appendix C.2. Let h be the half-corrected leader among the 16-tree survivors. If the raw and half-corrected leaders differ, they are the two nominees. Otherwise, split the 16 tree indices into blocks 1:8 and 9:16. For block $k \in \{1, 2\}$, let r_{ik}, $\bar{T}_{ik}(m)$, and $S^2_{ik}(m)$ be row i's OOB count, prediction mean, and sample variance. Use the same rows $\mathcal{E}_{12} = \{i : r_{i1} \geq 2,\ r_{i2} \geq 2\}$ for both blocks and all surviving budgets. Define

$$c_{50} = \mathbb{E}(1/K \mid K > 0), \qquad K \sim \text{Binomial}\{50, (1 - 1/n)^n\},$$

where n is the outer training size, and compute

$$Q_k(m) = \frac{1}{|\mathcal{E}_{12}|} \sum_{i \in \mathcal{E}_{12}} \left[\{Y_i - \bar{T}_{ik}(m)\}^2 + \left(c_{50} - \frac{1}{r_{ik}} \right) S^2_{ik}(m) \right].$$

This coefficient targets the variance term of a raw 50-tree OOB score, whose number of predictions for a row is random. It differs from the $1/B$ coefficient for a fixed-size ensemble. Let c minimize Q_2 among survivors. The second trigger nominates $\{h, c\}$ only if $c \neq h$, $Q_1(c) - Q_1(h) > \tau$, and $Q_2(c) - Q_2(h) < -\tau$, where $\tau = 128\epsilon_{\text{mach}} \max\{1, \max_{k,m} |Q_k(m)|\}$ and ϵ_{mach} is double precision machine epsilon. If fewer than two common rows are available, or only one candidate remains, this second trigger is inactive. If neither trigger fires, return h at 16 trees. The block reversal is a training-side stability rule; the fixed-budget identities do not supply a confidence level for this adaptive comparison.

Geometric nomination and common terminal comparison. Order the two nominees as $a < b$. If the original grid has a value strictly between them, add the one minimizing

$$\left| \log m - \frac{\log a + \log b}{2} \right|, \qquad m \in \mathcal{M}, \quad a < m < b.$$

Equivalently, it is the grid value closest to the geometric mean $\sqrt{ab}$ on a logarithmic scale. The nomination uses only budget locations. It neither reads other candidates' scores nor assumes that the unknown risk minimum lies between the nominees. If there is no interior grid value, retain the original pair. All nominated candidates receive 50 trees from the same random streams, and raw OOB

loss over rows with at least one OOB prediction selects the winner. Score and midpoint ties favor the larger budget. A midpoint that left the earlier screening stage has its missing prefix reconstructed and charged. Existing models are retained, including the winning midpoint, and the selected 16- or 50-tree forest is completed to 256 trees.

Development sequence. The wider grid and comparator suite were fixed before their full five-fold evaluation. The refinement rule was developed on this same task library and then fixed before the paired resource execution. Its 60 training-only choices were sealed before reading their stored test-risk curves. The later complete-procedure runs verified these choices and the returned predictions; they supply new timing measurements, not a new accuracy test on unseen tasks. All 60 cases enter every main-panel summary.

E.2 Wide-grid resource measurements

Common work accounting. We count feature–node split searches, summed over all fitted trees: evaluating one feature at one node contributes one unit. A shared search is charged once even when several budgets use its result. This count records split-search work; it does not weight nodes by sample size or measure scoring, data preparation, prediction, or process overhead. For the baseline methods, total work is tuning work plus the stored 256-tree count of the selected budget minus its already fitted prefix (32 trees for halving, 25 for ECV25, and 50 for OOB50). OOB256 has already fitted every candidate to 256 trees; tuning-free methods incur the work of just their selected forest. The random-grid rule averages this count over all grid values. Geometric refinement uses the counts from its executed tuning and retained-model completion, including midpoint reconstruction. Both ways of counting halving agree exactly for tuning and completion in all 60 cases. Geometric refinement uses 47.87% of SH's total split-search work.

Provisional estimates for untimed methods. The times for OOB256, OOB50, corrected halving and ECV25 in Table 1 are temporary estimates, not complete-procedure measurements. Tuning-free methods are excluded from time comparisons. The same formulas are used separately for CPU and wall time in every task–fold case. Let a index procedures, and let subscripts t and c denote tuning and completion. Write H_t and H_{total} for measured SH tuning and total time, and $N_{a,t}$ for procedure a's historical native-kernel tuning time. For corrected halving, ECV25 and OOB50, set

$$\widehat{T}_{a,t} = H_t \frac{N_{a,t}}{N_{\text{SH},t}}.$$

For OOB256, multiply the OOB50 estimate by the ratio of their tuning work counts. This avoids using the evaluation-inclusive OOB256 reference clock. Let $W_{a,c}$ be the split-search work remaining after tuning. Estimate

$$\widehat{T}_{a,\text{total}} = \widehat{T}_{a,t} + (H_{\text{total}} - H_t)\frac{W_{a,c}}{W_{\text{SH},c}}.$$

SH has positive remaining work in all 60 cases. OOB256 requires no further tree fits; this approximation neglects its additional post-selection overhead.

The approximation assumes that relative tuning times transfer between kernels and that post-tuning time scales with remaining search work. Node sizes, startup, scoring, model transfer and prediction can violate these assumptions. Thus estimated ratios are rounded to two decimal places, carry no timing confidence intervals, and support only a provisional frontier. The complete measurements for ours and SH are retained exactly; the estimates are to be replaced by complete matched executions.

Matched timing. Both timed procedures use eight workers pinned to the same eight CPU cores, with one numerical-library thread per worker and the same optimized prefix-reduction split kernel. Standard halving retains its normal warm starts. Each method builds its own models; no fitted model is shared between methods. CPU sums parent and worker processor time. Wall time measures elapsed execution. Tuning includes score arithmetic, process-pool startup and shutdown, model transfer, and midpoint reconstruction. Total cost adds actual completion of the chosen forest and prediction. These are direct executions of selection and retained-model completion.

The timing panel contains 60 paired cases (120 method executions). Fifty-seven pairs alternate method order by task and fold; three retained pairs keep their original three-method measurement order. All selected budgets and final predictions matched the stored references.

Table 4: Directly measured cost ratios of geometric refinement to standard halving (SH = 1), with task-bootstrap 95% intervals. Both procedures use the same kernel, eight workers, and actual model reuse in all 60 cases. The second row starts from prepared inputs; the final row also includes input preparation and audit costs.

Measured scope	CPU / halving	Wall / halving
Tuning	0.2869 [0.2539, 0.3271]	0.3710 [0.3280, 0.4231]
Selected 256-tree forest, total	0.4412 [0.4110, 0.4775]	0.5123 [0.4743, 0.5585]
Including input preparation/audit	0.4429 [0.4125, 0.4791]	0.5221 [0.4832, 0.5687]

Table 5: Summed seconds over all 60 task–fold cases, with ratios of sums relative to halving. These ratios weight expensive cases differently from the task-level geometric summary in Table 4.

Scope	Resource	Geometric refinement	Halving	Ratio
Tuning	CPU	11133.3	35134.1	0.3169
Tuning	Wall	1749.9	4551.1	0.3845
Selected 256-tree forest, total	CPU	18835.7	43223.4	0.4358
Selected 256-tree forest, total	Wall	2775.1	5625.5	0.4933
Including input preparation/audit	CPU	18899.3	43287.1	0.4366
Including input preparation/audit	Wall	2838.5	5689.0	0.4990

For cost C, with proposed method A and halving H, let d index tasks and f index folds. The primary ratio is

$$\rho_C = \exp\left[\frac{1}{12}\sum_{d=1}^{12}\log\left\{\frac{1}{5}\sum_{f=1}^{5}\frac{C_{A,df}}{C_{H,df}}\right\}\right].$$

The same task-level quantities enter 20,000 paired bootstrap resamples (seed 20260916). Ratios of summed seconds are secondary summaries and need not equal ρ_C. Table 5 gives both methods' absolute totals. Energy was not measured, and component memory readings do not determine the simultaneous peak across all processes.

Including input preparation and validation gives task-level total CPU and wall ratios of 0.4429 [0.4125, 0.4791] and 0.5221 [0.4832, 0.5687]. The retained three cases audited a larger historical input manifest than the remaining 57; this additional scope is therefore reported separately from the main prepared-input measurements. Table 6 reports the latter 57 cases alone, averaging each task's available four or five folds. The full 60-case analysis remains primary.

E.3 Wide-grid refinement ablations

Table 8 holds the 4–8–16 screening, both triggers, and the raw OOB50 terminal score fixed. The endpoint-only row compares the two nominated budgets; the geometric-refinement row also includes an available geometric midpoint. The comparison isolates geometric nomination.

The midpoint is available in 18 of the 24 triggered cases. Seven already have its 16-tree prefix; eleven require reconstruction. It changes two choices, on MTP and CT Slice. Relative to the endpoint-only raw50 procedure, its mean regret reduction is 0.080 percentage points, with descriptive 95% interval [0.000, 0.235]. The task-specific prediction comparisons in Table 3 show the remaining heterogeneity relative to halving and OOB50.

E.4 One-search SNR mechanism

The linear experiment draws Gaussian features with identity, block correlation 0.7 in groups of five, AR(1) correlation 0.7, or global equicorrelation 0.5. At the fixed null, one-step least squares compares full search with a 100-member feature-bagged ensemble. Each combination of covariance and sample size uses 100 repetitions. Across the four covariance designs and three sample sizes, feature bagging improves all 1,200 trials. Figure 3(a) shows that the risk ratio remains below one as n grows, supporting the general-covariance null statement in Theorem 2.2.

Table 6: Sensitivity using only the 57 newly executed paired cases. Entries are ratios to halving (SH = 1) with task-bootstrap 95% intervals; this subset is not a complete five-fold panel.

Scope	CPU	Wall
Tuning	0.2824 [0.2467, 0.3248]	0.3643 [0.3185, 0.4193]
Selected 256-tree forest, total	0.4396 [0.4086, 0.4765]	0.5087 [0.4698, 0.5563]
Including input preparation/audit	0.4410 [0.4099, 0.4780]	0.5172 [0.4772, 0.5654]

Table 7: Paired accuracy contrasts on the 12-task, five-fold panel. The final column is geometric refinement minus the row's method, in percentage points, with a paired task-bootstrap 95% interval; negative values favor geometric refinement. Intervals describe the development panel and do not adjust for repeated method exploration.

Method	Mean regret (%)	Paired difference [95% CI]
Geometric refinement (ours)	1.336	–
Successive halving	1.385	-0.049 [-0.243, +0.154]
Corrected halving	1.563	-0.227 [-0.527, +0.038]
ECV25	1.716	-0.380 [-0.831, +0.055]
Full-grid OOB50	1.216	+0.120 [-0.407, +0.692]
Full-grid OOB256	0.976	+0.360 [+0.075, +0.676]
$p/3$	1.717	-0.381 [-0.918, +0.124]
$\sqrt{p}$	8.552	-7.216 [-14.118, -2.194]
All features	4.661	-3.325 [-6.990, -0.754]
Uniform grid	13.444	-12.108 [-26.157, -3.325]

The nonlinear experiment replaces each linear coordinate by a grouped CART stump search. Within an exposed feature, the stump searches a prespecified quantile grid; the outer search compares the featurewise winners. At zero signal, feature bagging improves all 2,400 trials. Figure 3(b) shows the absolute null-risk reduction as this within-feature search becomes richer. Panel (c) increases the amplitude of a step signal and shows how the advantage changes with signal strength. This fixed-signal example illustrates the restriction–diversity mechanism; the exact posterior boundary in Proposition 2.1 is derived under a Gaussian signal model.

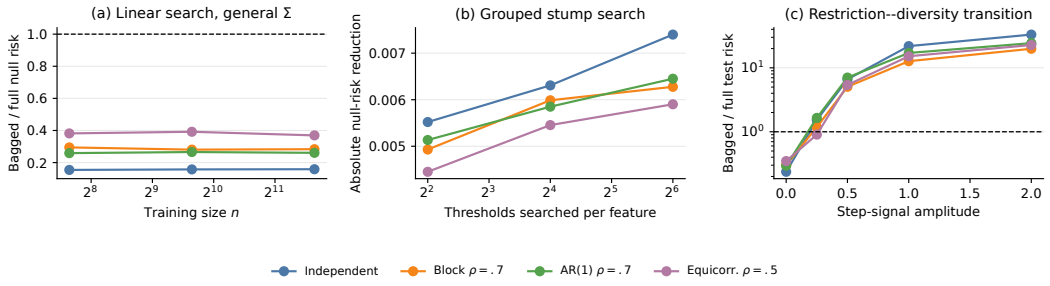


Figure 3: One-search mechanism experiments. (a) Bagged/full null-risk ratios under four Gaussian covariance structures. (b) Absolute null-risk reduction as a grouped stump search considers more candidate cutpoints. (c) Test risk as signal amplitude increases, showing the weak-to-strong SNR transition.

E.5 Recursive stagewise mechanism

Nine regression tasks use five seeded fold rotations. In each rotation, one fold is untouched test data, the next calibrates the stagewise estimators, and the remaining three fit the coupled trees. Depth-six trees use $m = \lceil \sqrt{p} \rceil$, minimum leaf size 20, and $B = 30$. Fifteen hybrid paths and 15 prefix-coupled pairs estimate Equation (43) and Equation (45) over six consecutive search blocks.

Table 8: Refinement ablations on the same 12 tasks and five folds. Only the last row is the main geometric-refinement procedure.

Procedure	Mean regret (%)
Endpoint-only, raw OOB50	1.416
Add geometric midpoint	1.336

Folds from the same dataset overlap in fitting observations, so results are averaged within dataset before interpretation.

The restriction and diversity sums telescope to numerical error below 5.6×10^{-17}. Calibration and untouched-test net contributions have correlation 0.751 across 270 related case–block values and 0.933 across the 54 dataset–block means; their accumulated signs agree on all nine datasets. Figure 4 connects these results to the main theory. Panel (a) shows the one-search SNR transition, while panel (b) shows that the recursive restriction-minus-diversity contributions estimated on calibration data track the same fitted couplings on untouched test data.

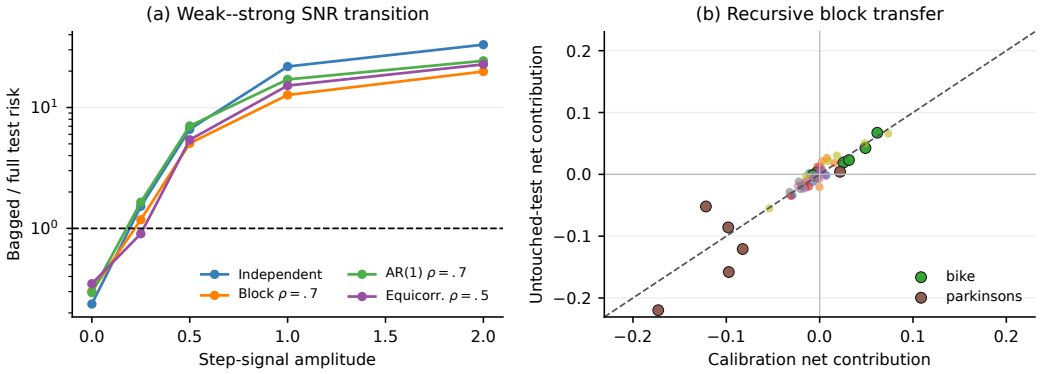


Figure 4: Mechanism checks. (a) One-search test risk as signal amplitude varies, supporting the weak-to-strong SNR transition. (b) Calibration and untouched- test net contributions across recursive search blocks; the dashed line is equality. The second panel supports transfer of the stagewise attribution.

Table 9 reports the six-block totals, and Figure 5 shows where the contributions arise for Bike and Parkinsons. Together they support the recursive interpretation: the final risk difference can be similar even when restriction and persistent diversity have different stagewise profiles.

Table 9: Stagewise attribution summed over six recursive search blocks. Restriction is $\sum_t \widehat{\delta}_t$, diversity credit is $(1 - B^{-1}) \sum_t \widehat{v}_t$, and net is their difference. Entries average five related rotations; test net evaluates the same couplings on the untouched fold.

Task	Cal. restriction	Cal. diversity	Cal. net	Test net
Airfoil	.062	.124	−.063	−.070
Auto MPG	.085	.061	.024	.051
Bike	.292	.135	.157	.154
Concrete	.118	.177	−.059	−.054
Energy	.025	.024	.002	.003
Parkinsons	−.341	.210	−.551	−.632
QSAR Fish	.008	.091	−.082	−.078
Wine	.006	.109	−.104	−.091
Yacht	.199	.118	.081	.110

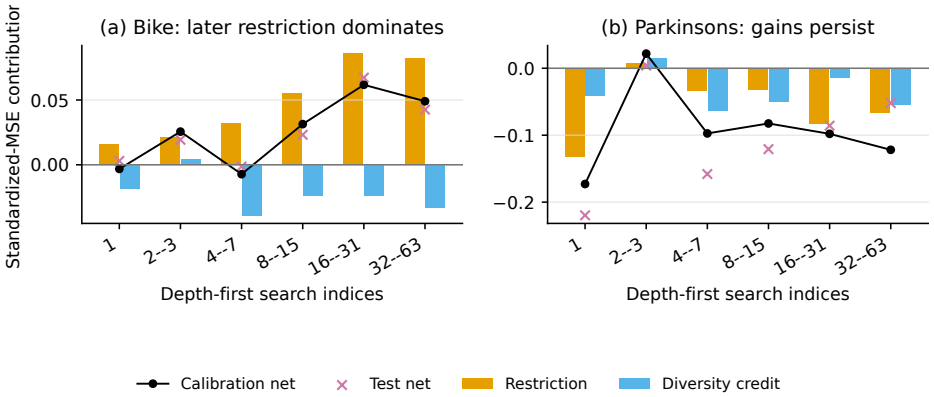


Figure 5: Recursive stagewise profiles for Bike and Parkinsons. Bars show calibration restriction and the negative diversity credit; the black line is their net contribution, and magenta crosses evaluate the same coupled fits on untouched test data.