SOFT-REJECTION INFORMED BAYESIAN OPTIMIZATION FOR EXPENSIVE SEQUENTIAL EXPERIMENTS

Anonymous authors
Paper under double-blind review

ABSTRACT

Bayesian optimization (BO) is widely used for expensive black-box optimization, but in many settings each evaluation is itself a long sequential experiment whose final outcome is the optimization target. Early stopping can further reduce cost; however, existing methods inherit the reliability of an underlying surrogate predictor, and can fail when trajectories cross, i.e., when a configuration that looks poor early overtakes others by the endpoint. We formally define this early stopping problem for sequential black-box optimization, and propose Soft-Rejection Informed Bayesian Optimization (SRBO) to free the stopping rule from this dependence. SRBO is built around an endpoint forecaster, a Direct-Forecasting Gaussian Process (DF-GP), one GP per decision step, that maps the partial trajectory and configuration directly to the endpoint. This forecaster is used in two ways. A Calibrated Composite Predictive Sequential Test (CCPST) turns the DF-GP forecast into a stopping rule calibrated by held-out cross-validation: for the cross-validated rule, under a per-epoch independence condition on the held-out stop indicators, the rate of falsely stopped winners has a finite-sample bound that holds even when trajectories cross and the forecast is mis-specified, and for the deployed rule it is audited on every benchmark. The same forecast is also fed back into the BO surrogate as a pseudo-observation when a trial is stopped to learn from rejected configurations. On eight cost-matched surrogate benchmarks built from real experimental and learning-curve data, SRBO attains the best average rank of the seven methods run on all eight and, against up to ten baselines, the lowest mean final regret on three of the five experimental benchmarks, significantly against every baseline on each: on in vitro permeation testing (IVPT) it is $3.3\times$ below the strongest baseline and $4\times$ to $959\times$ below the others ($p \leq 0.024$), on UCI Concrete compressive strength $7.1\times$ below the strongest ($p < 0.01$), and on a methane-coupling catalyst screen $2.7\times$ below it ($p \leq 0.024$). On the two remaining experimental benchmarks SRBO shows no significant difference from the leaders on one and trails single-fidelity BO on the other ($p = 0.02$), and on three learning-curve tasks BOHB leads.

1 INTRODUCTION

Bayesian optimization (BO) is a sample-efficient framework for optimizing expensive black-box functions (Jones et al., 1998; Shahriari et al., 2016; Garnett, 2023), with applications spanning drug design (Griffiths & Hernández-Lobato, 2020; Stanton et al., 2022), materials discovery (Frazier & Wang, 2016), and hyperparameter tuning (Snoek et al., 2012; 2015; Feurer et al., 2015). Many of these applications evaluate each configuration through a costly sequential experiment, such as multi-epoch neural network training, day-long permeation tests, or weeks-long curing of material specimens. Each experiment yields a time-indexed trajectory of intermediate observations whose endpoint y_T is the optimization target. Terminating unpromising experiments early can save substantial budget, but at $t < T$ only the partial trajectory $\mathbf{y}_{1:t}$ is observed. The central challenge is to do so without halting eventual high-y_T configurations, while preserving the information in stopped trials' partial trajectories.

Existing approaches to this trade-off fall into two families: bandit-style halving methods (Successive Halving (Jamieson & Talwalkar, 2016), Hyperband (Li et al., 2018), Bayesian Optimization HyperBand (BOHB) (Falkner et al., 2018)) that allocate budget across configurations and halt low-y_t

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.

trials, and curve-modeling methods (Freeze-Thaw BO (Swersky et al., 2014), learning-curve extrapolation (Domhan et al., 2015; Klein et al., 2017), and cost-sensitive freeze-thaw variants (Lee et al., 2025)) that model the trajectory along its time axis to guide pause-resume or termination. However, both families bind their stopping decisions to the proxy time axis: bandit halving treats y_t as a low-noise proxy for y_T, while curve-modeling methods extrapolate $\mathbf{y}_{1:t}$ forward to predict the endpoint. When curves cross (slow-gaining cement blends, late-accelerating permeation profiles, overfitting early validation) or trajectory shape mis-extrapolates, eventual high-y_T configurations may be stopped early.

Problem. In these settings, no existing early-stopping method has all three properties below. **Firstly**, the predictor inside both the stopping rule and the surrogate update should map $\mathbf{y}_{1:t}$ directly to y_T rather than extrapolate along the time axis (C1). **Secondly**, the false-stop rate (FSR) of eventual high-y_T winners should have a finite-sample bound that assumes nothing about that predictor's correctness, and its transfer to BO-selected trials should be audited (C2). **Lastly**, each halted experiment's partial trajectory should still inform the BO surrogate; otherwise iteration quality degrades relative to standard BO on full trajectories (C3).

Our approach. We propose **S**oft-**R**ejection Informed **B**ayesian **O**ptimization (**SRBO**), which couples two components. Its core principle is to stop unpromising trials reliably without falsely halting eventual high-y_T winners, treating each stopped trial as an informative *soft rejection* that reshapes the BO surrogate's posterior. First, a *Direct-Forecasting Gaussian Process* (DF-GP) predicts y_T directly from $\mathbf{y}_{1:t}$ and injects the prediction into the BO surrogate as a pseudo-observation, preserving stopped trials' information (addressing C1, C3). Second, a *Calibrated Composite Predictive Sequential Test* (CCPST) augments a per-step sequential stopping test with held-out cross-validation calibration, yielding, for its cross-validated calibration rule and under a per-epoch independence condition, a finite-sample trial-level FSR bound that holds even under a mis-specified DF-GP forecast, with the deployed rule audited empirically (addressing C2).

Our contributions are:

- **Problem formulation (Section 3).** We formalize this setting and give four pool-level diagnostics (Kendall τ_t, curve-crossing rate, bottom-K Jaccard, linear-fit RMSE) of partial-trajectory unreliability.
- **SRBO method with calibrated stopping (Section 4).** We couple DF-GP with CCPST and prove (Theorem 1), under a per-epoch independence condition, a per-epoch FSR bound $p_{\widehat{j}_m}^{\mathrm{CV},(m)} \leq \alpha_{\mathrm{fsr}} + \varepsilon_{n_{+,m}}$ for the cross-validated calibration rule that holds even under a mis-specified DF-GP forecast; the gap to the deployed rule is not bounded analytically and is audited on every benchmark.
- **Empirical validation (Section 5).** Under a 5-seed-tune / 20-seed-report protocol (no test-set selection) and a uniform cumulative cost budget $C = 1100$, SRBO attains the best aggregate rank across eight benchmarks and the lowest mean regret on three of the five experimental benchmarks (in vitro permeation test (IVPT), Concrete, oxidative coupling of methane (OCM)), significantly against every baseline on each. On the high-entropy-alloy oxygen-evolution benchmark (HEA-OER) it shows no significant difference from the leading methods, on Catechol single-fidelity BO edges it, and on the three Learning Curve Benchmark (LCBench) tasks BOHB wins. SRBO's mean runtime FSR is below its per-dataset tuned nominal level α_{fsr} on every benchmark, at most 0.005 on the five tasks of Table 12 and 0.022 on the three of Table 13.

2 Related Work

Bandit-Based Early Stopping. Bandit-based schemes accelerate hyperparameter optimization by evaluating a configuration to step $t < T$ and eliminating underperformers. Successive Halving (Jamieson & Talwalkar, 2016) and Hyperband (Li et al., 2018) use y_t as the halving signal across discrete rungs; Asynchronous Successive Halving Algorithm (ASHA) (Li et al., 2020) extends Successive Halving asynchronously, and BOHB (Falkner et al., 2018) couples Hyperband halving with a Tree-structured Parzen Estimator (TPE) candidate selector. These schemes rank configurations by y_t and therefore work best when y_t- and y_T-rankings agree, which fails in our setting (Section 3); Section 5 benchmarks SRBO against ASHA and a sync-bracket HyperbandEI with our protocol's EI candidate selector. DEHB (Awad et al., 2021), Population Based Training (Jaderberg et al., 2017)

and Optuna's median pruner (Akiba et al., 2019) also allocate budget by y_t rank, rely on the same agreement and are not compared here.

Learning-Curve Modeling. Freeze-thaw BO (Swersky et al., 2014) places a structured GP prior over learning curves; parametric families (Domhan et al., 2015) and Bayesian neural networks (Klein et al., 2017) extrapolate curves to the endpoint, and prior-data fitted networks (PFNs) (Müller et al., 2022) pre-trained on many synthetic curves from a parametric prior (Adriaensen et al., 2023; Rakotoarison et al., 2024) extend this paradigm. Cost-sensitive variants integrate cost and performance into the acquisition (Lee et al., 2025), and DyHPO (Wistuba et al., 2022) fits a deep-kernel GP over configuration, budget and a learned partial-curve embedding and advances the highest-expected-improvement configuration by one budget step; none provides finite-sample control on falsely halted winners. Their time-axis extrapolation fails when trajectory shape carries limited endpoint information (cumulative permeation profiles whose late-stage rate depends on formulation chemistry absent from $\mathbf{y}_{1:t}$), even when partial-trajectory rankings appear stable. Nothing restricts these formulations to machine-learning curves, so the separation from our setting is empirical rather than definitional. The LinExtrap and Freeze-Thaw baselines trail SRBO by more than two orders of magnitude on IVPT at $C = 1100$ and stall at regret 1229.95 on Concrete at every budget (Table 12, Appendix B.4). SRBO's DF-GP instead learns a *direct mapping* from partial trajectory to endpoint.

Sequential Tests and Optimal Stopping. Classical sequential analysis (Wald, 1945) delivers finite-sample error control under known likelihoods; Bayesian optimal stopping applies it per trial (BO-BOS (Dai et al., 2019)), Early Stopping BO (Stenger et al., 2024) truncates apparently suboptimal control episodes and feeds them back into BO heuristically, and campaign-level termination uses regret or budget bounds (Makarova et al., 2022; Wilson, 2024). All rest on correctly specified likelihoods, surrogate posteriors or application-specific heuristics; none bounds the per-trial, finite-sample rate of falsely halting a high-y_T winner under mis-specification, as CCPST's held-out Bernoulli calibration does for its cross-validated rule (Theorem 1).

Conformal Prediction and Held-out Hoeffding Calibration. Conformal prediction gives distribution-free coverage under exchangeability, with extensions to covariate shift, adaptive settings and quantile regression (Angelopoulos et al., 2021; Angelopoulos & Bates, 2021; Barber et al., 2023; Fannjiang et al., 2022; Gibbs & Candès, 2021; Lei et al., 2018; Romano et al., 2019; Tibshirani et al., 2019), and serves uncertainty-quantified candidate selection in BO (Stanton et al., 2023). Its quantile over exchangeable non-conformity scores does not compose with the per-step log-odds of the early-stopping Sequential Probability Ratio Test (SPRT). CCPST instead reinforces the SPRT with two safeguards estimated on held-out folds, an R^2-weighted stop threshold that makes stopping harder at low-R^2 time points and a distribution-free 5-fold held-out calibration that prunes early time points until the out-of-fold false-stop rate meets budget (Theorem 1). This calibration and its Hoeffding bound follow Xiao et al. (2026), who calibrate a Gaussian-process stopping rule for single experiments outside BO; CCPST embeds them in the BO loop with recalibration epochs and a union bound over candidate safe-time sets.

Handling Incomplete and Censored Observations. Classical censored-response regression (Hutter et al., 2013) treats stopped trials as censored, collapsing each trajectory into a scalar bound and discarding its shape, which matters when partial-trajectory and endpoint rankings disagree. SRBO instead injects DF-GP pseudo-observations into the EI surrogate, keeping an auxiliary censored module for diagnostic transparency only, outside the deployed acquisition (Appendix G, Section 5.3).

3 Problem Formulation

3.1 Standard Sequential Bayesian Optimization

We seek the global maximum of an expensive-to-evaluate black-box function $f : \mathcal{X} \to \mathbb{R}$ over a bounded domain $\mathcal{X} \subset \mathbb{R}^d$, i.e., $x^\star \in \arg\max_{x \in \mathcal{X}} f(x)$. All eight datasets are stated as maximization problems (the HEA-OER overpotential is negated). Standard BO sequentially selects configurations x_n by fitting a probabilistic surrogate, typically a Gaussian Process (GP) (Rasmussen & Williams, 2006), and maximizing an acquisition function $\alpha(x)$. The experiment at x_n runs to completion, yielding $Y_n = f(x_n) + \epsilon$, where $\epsilon \sim \mathcal{N}(0, \sigma^2)$ is Gaussian noise; the surrogate is updated with (x_n, Y_n) and the process repeats until a stopping condition, often the computational budget, is met.

3.2 Problem Setting

In many real-world applications, evaluating $f(x)$ entails a lengthy sequential experiment over a time or resource horizon T, which for configuration x generates a trajectory of intermediate proxy observations $\mathbf{y}_{1:T}(x) \coloneqq (y_1(x), y_2(x), \ldots, y_T(x))$ at discrete steps $t \in \{1, 2, \ldots, T\}$. The latent completed objective is a deterministic *functional* of the full trajectory, and the completed observation adds measurement noise:

$$f(x) \;=\; \Phi(\mathbf{y}_{1:T}(x))\,, \qquad Y(x) \;=\; f(x) + \epsilon, \qquad \epsilon \sim \mathcal{N}(0, \sigma^2), \tag{1}$$

where $\Phi : \mathbb{R}^T \to \mathbb{R}$ is an unknown, potentially nonlinear operator that summarizes the trajectory into the scalar quantity of interest (the Gaussian CDF used later in CCPST is denoted $\Phi_{\mathcal{N}}$). In the offline benchmarks, $Y(x)$ is the endpoint of the oracle fitted to the pool (Section 5.1), which defines regret and the FSR event; CCPST calibrates on the pool plus BO completions (Appendix F).

Scope of the present work. All eight datasets instantiate the endpoint-projection case $\Phi(\mathbf{y}_{1:T}) = y_T$ (Appendix C names each endpoint); the difficulty is not where the target sits along the trajectory but that the partial trajectory $\mathbf{y}_{1:t}$, for $t < T$, does not reliably determine y_T (Appendix A); when it does, BOHB leads (Section 5.2). SRBO and its FSR guarantee are designed and validated for this endpoint-target setting; the general Φ serves only the calibration argument's notation, and we make no empirical claim about non-endpoint targets.

No trajectory model is assumed. Neither the method nor its analysis posits a probabilistic law for $\mathbf{y}_{1:T}(x)$, such as independence across steps or an autoregressive structure in t. The DF-GP is a regression from $[x; \mathbf{y}_{1:t}]$ to the endpoint fitted per prefix length on the completed pool, so any dependence between steps is learned rather than modeled, and Theorem 1 rests on held-out Bernoulli stop indicators alone, so neither restricts the law of $\mathbf{y}_{1:T}(x)$ within a trajectory. Theorem 1's assumptions (Assumption 1) concern the calibration draw across trajectories, not the time structure of any one.

Problem statement. Given an evaluation horizon T and risk $\alpha_{\text{fsr}} \in (0, 1)$, we design a BO procedure that minimizes the simple regret $Y(x^\star) - \max_{i \le n} Y(x_i)$ within a cumulative cost budget C on proxy steps, that may STOP a trial after any $y_t(x_n)$, and that (i) bases stopping on a direct predictor of y_T from the partial trajectory $\mathbf{y}_{1:t}(x_n)$; (ii) targets $\text{FSR} \coloneqq \mathbb{P}[\text{STOP} \mid Y(x_n) \ge Y_{\text{th}}]$ at the nominal level α_{fsr}, where Y_{th} is the winner threshold, set to the 75th percentile of the completed targets in all our experiments (Section 4.2; theoretical per-epoch bound $\alpha_{\text{fsr}} + \varepsilon_{n_{+,m}}$ on the CV calibration-equivalent rule, Theorem 1; empirical audit on runtime $\le \alpha_{\text{fsr}}$ on every benchmark, Section 5.3); and (iii) reuses the partial trajectory $\mathbf{y}_{1:t}(x_n)$ of stopped trials to guide subsequent selections.

4 The SRBO Method

To address the challenges of Section 3 we propose SRBO, illustrated in Figure 1 and summarized in Algorithm 1 (Appendix G), where blue steps denote contributions specific to our method. SRBO aims to stop unpromising trials reliably without falsely halting eventual high-y_T winners, while treating each stopped trial as an informative *soft rejection* that reshapes the BO surrogate's posterior. Two components realize these aims: a Direct-Forecasting GP (DF-GP) that predicts y_T directly from the partial trajectory $\mathbf{y}_{1:t}$ and converts the prediction into a pseudo-observation (addressing C1, C3); and a Calibrated Composite Predictive Sequential Test (CCPST) that governs stopping by augmenting a per-step sequential test with held-out cross-validation calibration, yielding, for its cross-validated calibration rule and under Assumption 1 (A1), a finite-sample FSR bound that holds even when the DF-GP forecast is mis-specified (addressing C2). We present each, then Theorem 1, then the full algorithm.

4.1 Direct-Forecasting Gaussian Process (DF-GP) for Posterior Augmentation

Motivation. Hyperband-family early-stopping algorithms (Falkner et al., 2018; Li et al., 2020) do not use a halted trial to inform a model of the endpoint Y. ASHA (Li et al., 2020) treats the rung value y_t purely as an elimination signal and, as published, draws new configurations at random rather than from a fitted model; BOHB (Falkner et al., 2018) draws model-based candidates from the density model for the largest budget with enough observations, so once completed observations are available, which in our warm-started setting is from the start, a trial halted at a lower rung no longer informs

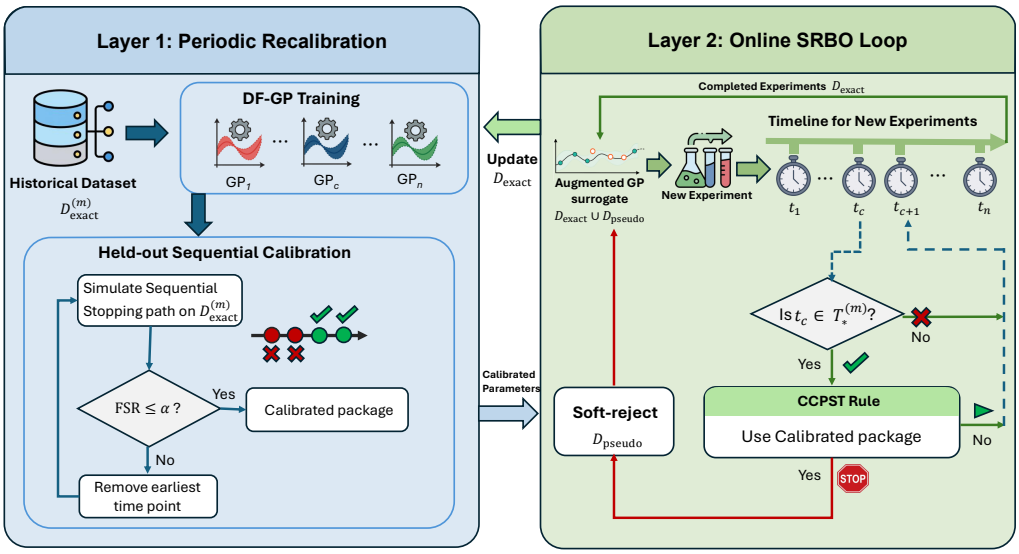


Figure 1: SRBO architecture: held-out calibration (Layer 1) and online soft-rejection BO loop (Layer 2).

candidate selection. Neither feeds the partial trajectory $\mathbf{y}_{1:t}(x_n)$ of a trial terminated at $t < T$ into a model of the endpoint, so what a halted trial reveals about Y is lost.

Definition. To recover this information, SRBO introduces the DF-GP, a secondary GP trained exclusively to map partial proxy trajectories to the final target. Let $\mathcal{D}_{\text{complete}} = \{(x_i, \mathbf{y}_{1:T}(x_i), Y(x_i))\}_{i=1}^{N}$ be the set of fully evaluated configurations, and define $z_t(x) := [x; \mathbf{y}_{1:t}(x)]$. For each decision step t, the DF-GP places a GP prior on the direct endpoint map,

$$g_t \sim \mathcal{GP}(m_t, k_t), \qquad Y(x) = g_t(z_t(x)) + \epsilon_{\text{DF}}, \qquad \epsilon_{\text{DF}} \sim \mathcal{N}(0, \sigma_{\text{DF}}^2), \tag{2}$$

so the DF-GP is a family of separate GPs, one per decision step t, each taking $z_t(x)$ of fixed dimension $d + t$ and labelled by the final endpoint $Y(x)$. Unlike curve-extrapolation models that fit a curve to $\mathbf{y}_{1:t}$ and project it forward to y_T, the DF-GP forecasts $Y(x)$ directly via a multivariate input-to-endpoint map, avoiding the error-accumulation of forward extrapolation and handling the $\mathbf{y}_{1:t} \to y_T$ mapping (C1).

In practice. When a new experiment x_{new} is early-stopped at step τ, for $\tau \in \{3, \ldots, T-1\}$, SRBO stores the stopped partial record $(x_{\text{new}}, \mathbf{y}_{1:\tau}(x_{\text{new}}))$. At each augmented-surrogate refit, the current DF-GP converts every stored partial record into a pseudo-label $\mu_\tau(x_{\text{new}}) = \mathbb{E}[g_\tau(z_\tau(x_{\text{new}})) \mid \mathcal{D}_{\text{complete}}]$, yielding a pseudo-observation $(x_{\text{new}}, \mu_\tau(x_{\text{new}}))$ for the BO acquisition surrogate $\mathcal{GP}_{\text{exact}}$. Pseudo-labels are thus refreshed after DF-GP recalibration rather than frozen at the stopping time. The augmented GP refits with inflated noise $\sigma_{\text{aug}}^2 = \gamma\sigma^2$ ($\gamma \geq 1$, default $\gamma = 3$) shared by exact and pseudo points, so γ acts as a global smoothing of the augmented surrogate rather than a selective discount on the DF-GP forecasts; it is applied only after at least one stopped trial exists. Implementation details and the $\text{SRBO}_{\text{no-DFGP}}$ ablation are in Appendix F and Section 5.3. The soft-rejection mechanism thus augments the acquisition surrogate's posterior (C3), steering the optimizer away from unpromising regions without the exact target value.

4.2 Calibrated Composite Predictive Sequential Test (CCPST)

FSR target. CCPST uses held-out calibration to obtain a finite-sample FSR bound for the CV calibration-equivalent rule; the deployed full-fit rule is validated empirically. We set $Y_{\text{th}}^{(m)}$ at each *recalibration epoch* m to the 75th percentile of the targets $\{Y(x_i)\}_{i \in \mathcal{D}_{\text{complete}}^{(m)}}$, where $\mathcal{D}_{\text{complete}}^{(m)}$ is the pool of fully-evaluated trajectories at the start of epoch m. The threshold is *fixed within the epoch* and recomputed at the next recalibration trigger (every 5 outer completions, or after 25 outer iterations

without them, in our deployed implementation). Theorem 1 therefore gives an epoch-wise FSR bound, made uniform over epochs by Bonferroni-adjusting the confidence (Appendix E). Under maximization $\{Y(x) \geq Y_{\text{th}}^{(m)}\}$ covers the top quartile (best 25%). For a trial at configuration x, we contrast $H_0 : Y(x) < Y_{\text{th}}^{(m)}$ vs. $H_1 : Y(x) \geq Y_{\text{th}}^{(m)}$, and define the trial-level false-stopping rate

$$\text{FSR}^{(m)} \;:=\; \mathbb{P}\Big[\textsc{Stop at any } s \in \mathcal{T}_\star^{(m)} \;\Big|\; H_1^{(m)}\Big], \tag{3}$$

which we *target* at a user-specified nominal level $\alpha_{\text{fsr}} \in (0,1)$ (Theorem 1 bounds it at $\alpha_{\text{fsr}} + \varepsilon_{n_{+,m}} + \Delta_m$, the first two terms being the calibration guarantee for the CV calibration-equivalent rule). Here $\mathcal{T}_\star^{(m)}$ is the safe-time set selected by held-out pruning within recalibration epoch m, and $H_1^{(m)} : Y(x) \geq Y_{\text{th}}^{(m)}$.

CCPST score. At step t, we query the DF-GP posterior $\mathcal{N}(\mu_t(x), \sigma_t^2(x))$ for $Y(x)$ given $\mathbf{y}_{1:t}(x)$, and define the per-step log-odds of promising versus unpromising as

$$\ell_t(x) \;=\; \log \frac{p_t(x)}{1 - p_t(x)}, \qquad p_t(x) \;:=\; 1 - \Phi_{\mathcal{N}}\left(\frac{Y_{\text{th}}^{(m)} - \mu_t(x)}{\sigma_t(x)}\right) \;=\; \Phi_{\mathcal{N}}\left(\frac{\mu_t(x) - Y_{\text{th}}^{(m)}}{\sigma_t(x)}\right), \tag{4}$$

where $\Phi_{\mathcal{N}}$ is the standard Gaussian CDF and $p_t(x) = \Pr[Y(x) \geq Y_{\text{th}}^{(m)} \mid \mathbf{y}_{1:t}(x)]$ is the posterior probability of being promising under the epoch-m threshold. Following the SPRT (Wald, 1945) bracket structure, we compare $\ell_t(x)$ against a raw threshold pair $(\log A, \log B)$ with $A = \beta/(1 - \alpha_{\text{fsr}})$, $B = (1-\beta)/\alpha_{\text{fsr}}$, where $\beta \in (0,1)$ is the SPRT-style lower-boundary shape parameter inherited from Wald's algebra ($\beta = 0.1$ in our deployed implementation; Appendix F): $\ell_t \leq \log A$ ($\Leftrightarrow$ p_t small, strong unpromising evidence) accepts H_0 (STOP); $\ell_t \geq \log B$ ($\Leftrightarrow$ p_t large, strong promising evidence) accepts H_1 (CONTINUE); otherwise OBSERVE another step. In the deployed loop CONTINUE and OBSERVE both run the trial to the next tested step, where the rule is re-applied, so only $\log A$ binds and $\log B$ merely labels the decision. The Wald-style thresholds only generate a family of candidate stopping rules; the FSR budget is enforced exclusively by held-out pruning, not by their nominal SPRT error interpretation. Unlike textbook SPRT we use the *instantaneous* per-step ratio rather than a running sum $\sum_{s \leq t} \ell_s$, because the DF-GP forecast at t uses $\mathbf{y}_{1:t}$ (so successive ℓ_s share strongly correlated inputs and a running sum would over-credit early evidence); the calibration steps below recover finite-sample FSR control of the cross-validated rule without that summation.

Reliability-weighted thresholds. Wald's bound assumes the per-step likelihood ratio is correctly specified. Since the DF-GP forecast at early t may be unreliable when the partial trajectory is, we down-weight evidence at low-R^2 time points. Within recalibration epoch m we estimate $\widehat{R}_t^{2,(m)}$ between the pooled out-of-fold DF-GP forecasts and ground-truth Y, then use the clipped reliability weight

$$\rho_t^{(m)} \;:=\; \max\Big\{\rho_{\min},\, \min\Big(1,\, \max(0, \widehat{R}_t^{2,(m)})\Big)\Big\}, \qquad \rho_{\min} = 0.01, \tag{5}$$

which matches the implementation safeguard against non-positive R^2. We replace the constant SPRT threshold by the time-varying

$$\log B_t^{(m)} \;=\; \log\Big(\tfrac{1-\beta}{\alpha_{\text{fsr}}}\,(\rho_t^{(m)})^{-1}\Big), \qquad \log A_t^{(m)} \;=\; \log\Big(\tfrac{\beta}{1-\alpha_{\text{fsr}}}\,\rho_t^{(m)}\Big), \tag{6}$$

so that small $\rho_t^{(m)}$ makes STOP harder; $\rho_t^{(m)} = 1$ recovers Wald's constant thresholds.

Held-out safe-time pruning. Even with R^2-weighting, testing the rule at every $t \in \{1, \ldots, T-1\}$ gives each promising trial up to $T-1$ chances to be falsely stopped, inflating the trial-level FSR. Within each epoch m we therefore keep the largest *nested* candidate $\mathcal{T}_{\widehat{j}_m}^{(m)} \subseteq \{3, \ldots, T-1\}$ whose simulated CV rule satisfies $\widehat{p}_{\widehat{j}_m}^{\text{CV},(m)} \leq \alpha_{\text{fsr}}$, starting from $\mathcal{T}_1^{(m)} = \{3, \ldots, T-1\}$ and dropping the earliest remaining t until the budget is met or the set is empty (stopping is thus never tested before three proxy observations or at the endpoint; Appendix E). The deployed rule stops the trial at the first $t \in \mathcal{T}_{\widehat{j}_m}^{(m)}$ with $\ell_t(x) \leq \log A_t^{(m)}$. The DF-GP mean for a held-out trajectory comes from a refit without it (5-fold CV on $\mathcal{D}_{\text{cal}}^{(m)}$); the choice of $\widehat{j}_m$ on the resulting $\widehat{p}_j^{\text{CV},(m)}$ is absorbed by the union bound over the K_m candidates in Theorem 1. Nesting gives at most $K_m = T-3$ candidates and a within-epoch union-bound penalty of $\log K_m$ in Theorem 1.

Assumption 1 (Per-epoch calibration condition)**.** *Within recalibration epoch m, consider two conditions: **(A1)** the held-out CV false-stop indicators used by safe-time pruning are i.i.d. Bernoulli variables for each candidate safe-time set, conditional on the epoch state, CV split, threshold $Y_{th}^{(m)}$, and fitted reliability weights $\{\rho_t^{(m)}\}$; **(A2)** the calibration trajectories are representative of the future BO-selected promising trajectories evaluated under the epoch-m rule. Theorem 1 uses (A1); (A2) keeps the population-shift part of its CV-to-runtime slack Δ_m small.*

The first clause is a finite-sample calibration condition, not a consequence of the DF-GP model; the second is the population-shift part of the CV-to-runtime transfer condition. Both are approximate here: 5-fold CV reduces leakage, but folds still share training data; the threshold $Y_{\text{th}}^{(m)}$, the pooled out-of-fold R^2 behind $\rho_t^{(m)}$ and the full-fit predictive standard deviation are all computed from the whole epoch pool, which couples the held-out indicators, and BO's acquisition can reshape the calibration pool relative to the population (Appendix E). Section 5.3 audits the deployed rule's runtime FSR directly, testing the conclusion rather than the conditions.

Theorem 1 (Per-epoch held-out Hoeffding FSR, explicit-slack form)**.** *Fix epoch m and the nested candidate safe-time sets $\{\mathcal{T}_j^{(m)}\}_{j=1}^{K_m}$ with $K_m \le T-3$. Let $\widehat{j}_m$ be the pruning choice: the first nonempty candidate satisfying $\widehat{p}^{CV,(m)}_{\widehat{j}_m} \le \alpha_{fsr}$ on $n_{+,m}$ held-out promising trajectories, or the empty fallback $\mathcal{T}_{K_m+1}^{(m)} = \emptyset$ if no nonempty candidate satisfies the budget. Under Assumption 1 (A1), with probability at least $1-\delta_m$ over the epoch-m calibration draw, the deployed false-stopping rate of* (3) *satisfies*

$$\mathrm{FSR}^{(m)} \;\le\; \underbrace{\alpha_{fsr}}_{\text{pruning budget}} \;+\; \underbrace{\varepsilon_{n_{+,m}}}_{\text{finite-sample tail}} \;+\; \underbrace{\Delta_m}_{\text{CV-to-runtime slack}}, \qquad \varepsilon_{n_{+,m}} \;:=\; \sqrt{\frac{\log\!\left(K_m/\delta_m\right)}{2\,n_{+,m}}}. \tag{7}$$

Only the first two terms come from calibration and use (A1) alone: they give $p^{CV,(m)}_{\widehat{j}_m} \le \alpha_{fsr} + \varepsilon_{n_{+,m}}$ for the CV calibration-equivalent rule. The third, $\Delta_m := \max\{0,\, \mathrm{FSR}^{(m)} - p^{CV,(m)}_{\widehat{j}_m}\}$ (Appendix E, Step 5), is the CV-to-runtime residual: it vanishes for that rule, is what Assumption 1 (A2) asserts to be small for the population shift (its algorithmic part, Step 5, is checked only empirically), and, like the coverage gap of Barber et al. (2023), is not bounded distribution-free. Setting $\delta_m = \delta/M_{\max}$ with $M_{\max} = 1 + \lfloor C/\min(5T, 75) \rfloor$, a deterministic bound on the number of epochs a run of budget C can deploy (21 here; Step 4 of the proof), gives a Bonferroni-uniform statement over all epochs. The proof and the empirical-Bernstein refinement are in Appendix E; the first two terms are distribution-free in the DF-GP forecast because they only use held-out Bernoulli stop indicators. If the empty fallback is selected, no trial can stop within that epoch and the bound holds with the convention $p^{CV,(m)}_{K_m+1} := 0$.

Unlike Wald's classical SPRT bound, which requires a correctly specified per-step likelihood ratio between H_0 and H_1, the calibration layer of Theorem 1 holds with no correctness assumption on the DF-GP forecast. The deployed runtime rule uses the full DF-GP fit and a smaller predictive standard-deviation floor, which Δ_m absorbs; Section 5.3 is therefore the binding evidence that the runtime rule stays below the per-dataset nominal FSR on the evaluated benchmarks. At our $n_{+,m} \in [10, 75]$ the concentration tail is loose, so the theorem serves as a calibration sanity check rather than the sole safety claim.

Overall algorithm. SRBO maximizes plain $\mathrm{EI}(x \mid \mathcal{GP}_{\text{exact}})$ on the DF-GP-augmented surrogate; stopped trials enter as DF-GP pseudo-observations and completed trials as exact targets in $\mathcal{D}_{\text{complete}}$, and the whole augmented surrogate is refit at the shared noise variance $\gamma\sigma^2$. Pseudocode and the diagnostic-only censored-GP bookkeeping are in Appendix G.

5 Experiments

We evaluate SRBO on eight sequential-experiment benchmarks, asking whether soft rejection reduces endpoint regret at matched cost (Q1), whether the deployed CCPST rule holds runtime FSR below the tuned nominal level (Q2), and which components drive the behavior (Q3).

5.1 Experimental Setup

Benchmarks. We use eight $T = 11$ surrogate benchmarks: each fits an oracle to a pool of trajectories, and every method queries that oracle over a continuous box around the pool (Appendix F), so the optimum and the regret scale are the oracle's and can lie outside the measured range (Appendix D). IVPT skin permeation ($d = 3$, 135 replicates of 28 formulations), UCI Concrete ($d = 7$, 127 pool) and LCBench credit-g/vehicle/segment ($d = 7$, 200 pool each) carry the full method set of Table 12; OCM heterogeneous catalysis ($d = 9$, 150 pool), HEA-OER electrocatalysis ($d = 4$, 150 pool) and Catechol flow chemistry ($d = 6$, 40 pool) carry the seven methods of Table 13. Search spaces and targets are in Appendix C.

Baselines. We compare against ten 20-seed baselines: single-fidelity BO (*StandardBO*), a linear endpoint-extrapolation control after Domhan et al. (2015) (*LinExtrap*), two CCPST-stopping controls that discard a stopped trial or virtual-fill it (*EI_Discard*, *EI_Virtual*) (Hutter et al., 2013), calibrated once on the recorded pool at the default $\alpha_{\text{fsr}} = 0.05$, explicit curve modeling (*Freeze-Thaw* (Swersky et al., 2014)), successive halving under EI (*ASHA* (Li et al., 2018; 2020), *HyperbandEI*) or TPE (*BOHB* (Falkner et al., 2018)), and the two freeze-thaw surrogates *ifBO* (Rakotoarison et al., 2024) and *DyHPO* (Wistuba et al., 2022), both run by our own runners with their published models, adapted as listed in Appendix F. All EI-based methods except DyHPO share the same initial designs and RBF-kernel GP surrogate; per-method descriptions, full configurations and the relation of DF-GP-ranked halving to CCPST are in Appendices F and H.

Metrics. We report simple regret $r_n = Y(x^\star) - \max_{i \le n} Y(x_i)$ (lower is better), cumulative cost $C_n = \sum_{i \le n} \tau_i$ in proxy steps, and the empirical runtime FSR, the fraction of promising trials (those with $Y(x_i) \ge Y_{\text{th}}^{(m_i)}$ at their scoring epoch m_i) that the deployed rule stopped. Regret is measured against the oracle's optimum over the search box, which exceeds the largest measured endpoint on every pool except OCM (Appendix D). A trial a method stops early is still credited with its oracle endpoint; the convention (Appendix F) is generous to methods that stop trials early, and counting only completed configurations leaves the leading method unchanged on every pool (Appendix C.5).

Protocol. All methods run under a fixed cumulative-cost budget $C = 1100$ proxy steps (100 full StandardBO iterations at $T = 11$); simple regret at this shared C is *cost-matched regret*, and methods that stop more trials evaluate more configurations within it (Appendix B.4 repeats the cap at $C \in \{550, 825\}$). We tune all four tunable methods identically on a 5-seed tuning pool (IDs 100 to 104) and report once on the disjoint 20-seed reporting pool (IDs 0 to 19), without test-set selection. Aggregates are mean $\pm$ SE; pairwise tests use two-sided Wilcoxon signed-rank.

Partial-trajectory reliability diagnostics. Four pool-level statistics (Appendix A) quantify partial-trajectory unreliability on every pool; of these, normalized linear-fit RMSE separates the five experimental pools (*high-mismatch*, 0.49 or above) from the three LCBench tasks (*low-mismatch*, 0.34 or below) without overlap.

5.2 Main Results (Q1)

Figure 2 and Table 12 report cost-matched regret at $C = 1100$. SRBO attains the lowest mean regret on the two high-mismatch tasks of Table 12, IVPT (linear-fit RMSE 0.66, mean 0.17) and Concrete (0.51, mean 4.02), significantly against all ten baselines on both ($p \le 0.024$ and $p < 0.01$). The strongest baseline on both is HyperbandEI, 3.3$\times$ and 7.1$\times$ above SRBO; BOHB has the highest regret of the eleven methods on IVPT (mean 166) and on Concrete ends 8.4$\times$ above SRBO, lower on only 6 of the 20 seeds. BOHB wins the three low-mismatch LCBench tasks (RMSE 0.34 or below), each at $p < 0.001$ against SRBO. The freeze-thaw surrogates also lead SRBO on LCBench (DyHPO on vehicle and segment, ifBO on vehicle).

On the three further benchmarks SRBO attains the lowest mean regret on OCM (0.024, significant against every method, $p \le 0.024$) but not on HEA-OER or Catechol (Table 13). On HEA-OER SRBO shows no significant difference from HyperbandEI, BOHB and StandardBO ($p \ge 0.08$), SRBO's median (0.0025) lying among the leaders' (0.0020 to 0.0041) while one seed at 0.074 raises its mean; on Catechol StandardBO's 0.0037 edges SRBO's 0.0044 ($p = 0.02$), with no significant difference between SRBO and ASHA, ifBO or HyperbandEI ($p \ge 0.08$). Across all eight benchmarks SRBO has the best average per-seed rank (2.56 against 3.44 for BOHB; Figure 3).

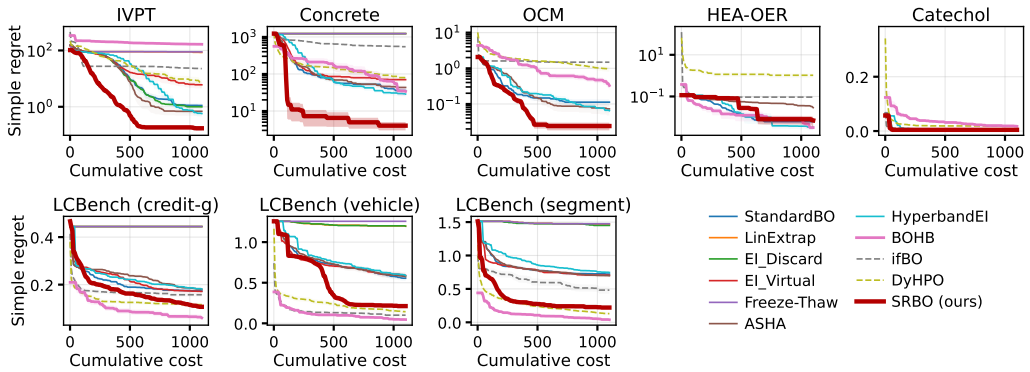


Figure 2: Mean simple regret vs. cumulative cost on all eight benchmarks (20 seeds, ± 1 SE; $C = 1100$, stepwise carry-forward to each run's endpoint cost, Table 9). Top row, the five experimental benchmarks; bottom row, the three LCBench tasks. The four stopping-rule controls were not run on OCM, HEA-OER and Catechol, whose regret is measured against one common optimum per pool (Table 13).

On Concrete BOHB's median and worst seed are 9.81 and 100.4 against SRBO's 2.15 and 21.0 ($p = 0.007$). The controlled BOHB variants of Appendix B.1 do not pin these losses on halving. HyperbandEI, which halves in Hyperband brackets as BOHB does and selects by EI but, unlike SRBO, keeps no stopped trial in its surrogate, trails SRBO significantly on all five pools of Table 12 and on OCM, by $1.7\times$ to $7.1\times$.

5.3 Calibration and Mechanism (Q2/Q3)

The deployed rule's 20-seed mean runtime FSR is below the nominal level (0.025 to 0.10) on every benchmark, at most 0.005 on the five tasks of Table 12 and 0.022 on the other three (Catechol, nominal 0.025; Figure 5, Table 1). Theorem 1 bounds the CV calibration-equivalent rule; the deployed full-fit rule is audited empirically (Appendix F). Ablations under the deployed protocol (Table 2) isolate the mechanism. Removing the DF-GP pseudo-observations degrades regret 306-fold on Concrete, where every seed ends at regret 1229.95, and 4.2- to 6.5-fold on the LCBench tasks ($p < 10^{-4}$ on all four), where the variant stops over 99% of its trials; on IVPT it changes regret by only $1.2\times$ ($p = 0.11$), stopping 67% of its trials. Stopping without injecting evidence therefore breaks BO wherever the stopped trials carry the search. Removing the auxiliary censored GP, which under the deployed rule enters neither the EI score nor the stopping decision, reproduces full SRBO seed by seed on 99 of the 100 pool and seed pairs and differs significantly from it on no pool ($p \geq 0.32$); the reportable mechanism is direct-forecast injection, not censored likelihood.

6 Limitations and Conclusion

Limitations. Appendix D discusses four: the empirical claim holds in the measured mismatch regime, not universally; the tail of Theorem 1 is loose at our calibration sizes; SRBO depends on the DF-GP, with safeguards that keep stopping conservative when its held-out R^2 is poor or promising completions are few; and every experiment replays a pre-collected pool through a fitted oracle whose optimum lies outside the measured range on seven of the eight pools.

Conclusion. We presented **SRBO**, which couples a direct-forecasting GP from partial trajectories to the endpoint with a calibrated composite predictive sequential test on the trial-level false-stop rate. Under a cost-matched protocol it has the best aggregate rank across eight benchmarks and the lowest mean regret on IVPT, Concrete and OCM, significantly against every baseline; it shows no significant difference from the leaders on HEA-OER, trails single-fidelity BO on Catechol, and BOHB leads the three LCBench tasks. Where early observations predict the endpoint well BOHB remains a strong default; where they do not, soft rejection keeps a stopped trial's information while held-out calibration bounds, under (A1), the cross-validated rate at which winners are stopped.

AI USE STATEMENT

In this work, we used generative AI tools (large language model assistants) for research execution, for corrections to the experimental methodology, for drafting and editing the paper, and for assistance with the theory. For research execution, the assistants wrote and ran the aggregation, diagnostic and reproduction scripts that produce the tables from the raw per-seed results, wrote the pool-construction and subsampling code for OCM, HEA-OER and Catechol, whose data are the published experimental records cited in Appendix C.4, wrote the ifBO and DyHPO runners, extended the diagnostics script to all eight pools, launched and monitored the cluster jobs, and carried out the data analyses behind the coverage comparison in Appendix A. The assistants also proposed and implemented corrections to the experimental methodology and to the scope of the claims, which the authors reviewed and approved. These were fixes to the shared GP surrogate and to the budget and regret accounting of the bracket-based baselines, with every affected run repeated; the completed-only regret of Appendix C.5; the Gaussian-process oracle for Catechol; the normalized linear-fit RMSE as the regime diagnostic; and the restriction of the finite-sample guarantee to the cross-validated rule under Assumption 1 (A1). For drafting, the assistants wrote drafts of sections of the paper, including table captions, appendix paragraphs and the supplement README, which the authors edited. For the theory, the assistants helped derive Theorem 1 and its empirical-Bernstein corollary. The problem formulation, the SRBO method and its central hypothesis are the authors'. Generative AI tools were not used to generate synthetic data; the controlled synthetic suite of Appendix B.2 is computed by a script from standard test functions and a Gaussian per-configuration noise proxy. Translation and qualitative or thematic data analysis are not applicable to this work. We also used generative AI tools to polish writing throughout the paper and for retrieval and discovery, including locating candidate benchmark datasets. Verification was AI-assisted, and its findings were reviewed by the authors. Every number in the result tables was regenerated from the shipped data and raw per-seed results, and the scripts behind Tables 12, 13, 2, 3, 5, 9, 10, 14 and 15 and behind the ablation block of Table 4 assert the published values and fail otherwise; every cited publication was resolved through its DOI or landing page and its title and author list checked against the record; and the method descriptions were checked against the shipped source code, and the statements and proofs of the theory were checked for consistency, in repeated audit passes. We take responsibility for the final content of this work, including text, claims or artifacts produced with the aid of generative AI.

REPRODUCIBILITY STATEMENT

The supplement contains the source of SRBO and of all ten baselines, the eight benchmark pools with the builders of the OCM, HEA-OER and Catechol pools, the SLURM runners that produced every reported run together with the configurations the tuning pools selected, and the per-seed result files behind every table and figure. Every result table and plotted figure has a driver that regenerates it from those files, and the drivers of Tables 12, 13, 2, 3, 5, 9, 10, 14 and 15, of the ablation block of Table 4 and of Figures 2 and 5 also assert the values printed here; Appendix F lists them under *Reproducibility commitment*, and Appendix C describes every pool and its construction.

REFERENCES

Steven Adriaensen, Herilalaina Rakotoarison, Samuel Müller, and Frank Hutter. Efficient Bayesian learning curve extrapolation using prior-data fitted networks. In *Advances in Neural Information Processing Systems*, volume 36, pp. 19858–19886, 2023. doi: 10.52202/075280-0872.

Takuya Akiba, Shotaro Sano, Toshihiko Yanase, Takeru Ohta, and Masanori Koyama. Optuna: A next-generation hyperparameter optimization framework. In *Proceedings of the 25th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining*, pp. 2623–2631. ACM, 2019. doi: 10.1145/3292500.3330701.

Anastasios N Angelopoulos and Stephen Bates. A gentle introduction to conformal prediction and distribution-free uncertainty quantification. *arXiv preprint arXiv:2107.07511*, 2021. doi: 10.48550/arXiv.2107.07511.

Anastasios N Angelopoulos, Stephen Bates, Michael I Jordan, and Jitendra Malik. Uncertainty sets for image classifiers using conformal prediction. In *International Conference on Learning Representations*, 2021. URL https://openreview.net/forum?id=eNdiU_DbM9.

Noor Awad, Neeratyoy Mallik, and Frank Hutter. DEHB: Evolutionary hyberband for scalable, robust and efficient hyperparameter optimization. In *Proceedings of the Thirtieth International Joint Conference on Artificial Intelligence*, pp. 2147–2153, 2021. doi: 10.24963/ijcai.2021/296.

Rina Foygel Barber, Emmanuel J Candès, Aaditya Ramdas, and Ryan J Tibshirani. Conformal prediction beyond exchangeability. *The Annals of Statistics*, 51(2):816–845, 2023. doi: 10.1214/23-AOS2276.

Toby Boyne, Juan Campos, Rebecca Langdon, Jixiang Qing, Yilin Xie, Shiqiang Zhang, Calvin Tsay, Ruth Misener, Daniel Davies, Kim Jelfs, Sarah Boyall, Thomas Dixon, Linden Schrecker, and Jose Pablo Folch. The Catechol benchmark: Time-series solvent selection data for few-shot machine learning. In *Advances in Neural Information Processing Systems*, volume 38, pp. 132632–132660, 2025. doi: 10.52202/085713-3994.

Jan Lukas Bürgel, Felix Thelen, Rico Zehl, Ridha Zerdoumi, Olga Krysiak, Wolfgang Schuhmann, and Alfred Ludwig. Dataset - exploration of nanostructured high-entropy alloys for key electrochemical reactions: a comparative study for the solid solution systems Cu-Pd-Pt-Ru, Ir-Pd-Pt-Ru and Ni-Pd-Pt-Ru. Zenodo, 2025. Version 1.0.

Jan Lukas Bürgel, Rico Zehl, Felix Thelen, Ridha Zerdoumi, Olga A. Krysiak, Benedikt Kohnen, Ellen Suhr, Wolfgang Schuhmann, and Alfred Ludwig. Exploration of nanostructured high-entropy alloys for key electrochemical reactions: a comparative study for the solid solution systems Cu–Pd–Pt–Ru, Ir–Pd–Pt–Ru and Ni–Pd–Pt–Ru. *Faraday Discussions*, 264:64–82, 2026. doi: 10.1039/d5fd00082c.

Zhongxiang Dai, Haibin Yu, Bryan Kian Hsiang Low, and Patrick Jaillet. Bayesian optimization meets Bayesian optimal stopping. In *International Conference on Machine Learning*, volume 97, pp. 1496–1506, 2019. URL https://proceedings.mlr.press/v97/dai19a.html.

Tobias Domhan, Jost Tobias Springenberg, and Frank Hutter. Speeding up automatic hyperparameter optimization of deep neural networks by extrapolation of learning curves. In *International Joint Conference on Artificial Intelligence*, pp. 3460–3468, 2015. URL https://www.ijcai.org/Proceedings/15/Papers/487.pdf.

Stefan Falkner, Aaron Klein, and Frank Hutter. BOHB: Robust and efficient hyperparameter optimization at scale. In *International Conference on Machine Learning*, volume 80, pp. 1437–1446, 2018. URL https://proceedings.mlr.press/v80/falkner18a.html.

Clara Fannjiang, Stephen Bates, Anastasios N Angelopoulos, Jennifer Listgarten, and Michael I Jordan. Conformal prediction under feedback covariate shift for biomolecular design. *Proceedings of the National Academy of Sciences*, 119(43):e2204569119, 2022. doi: 10.1073/pnas.2204569119.

Matthias Feurer, Aaron Klein, Katharina Eggensperger, Jost Springenberg, Manuel Blum, and Frank Hutter. Efficient and robust automated machine learning. In *Advances in Neural Information Processing Systems*, volume 28, pp. 2962–2970, 2015. URL https://papers.nips.cc/paper_files/paper/2015/hash/11d0e6287202fced83f79975ec59a3a6-Abstract.html.

Peter I Frazier and Jialei Wang. Bayesian optimization for materials design. In *Information Science for Materials Discovery and Design*, Springer Series in Materials Science, pp. 45–75. Springer International Publishing, 2016. doi: 10.1007/978-3-319-23871-5_3.

Roman Garnett. *Bayesian Optimization*. Cambridge University Press, 2023. doi: 10.1017/9781108348973.

Isaac Gibbs and Emmanuel Candès. Adaptive conformal inference under distribution shift. In *Advances in Neural Information Processing Systems*, volume 34, pp. 1660–1672, 2021. URL https://papers.nips.cc/paper_files/paper/2021/hash/0d441de75945e5acbc865406fc9a2559-Abstract.html.

Ryan-Rhys Griffiths and José Miguel Hernández-Lobato. Constrained Bayesian optimization for automatic chemical design using variational autoencoders. *Chemical Science*, 11(2):577–586, 2020. doi: 10.1039/C9SC04026A.

Frank Hutter, Holger H. Hoos, and Kevin Leyton-Brown. Bayesian optimization with censored response data. *arXiv preprint arXiv:1310.1947*, 2013. doi: 10.48550/arXiv.1310.1947.

Max Jaderberg, Valentin Dalibard, Simon Osindero, Wojciech M. Czarnecki, Jeff Donahue, Ali Razavi, Oriol Vinyals, Tim Green, Iain Dunning, Karen Simonyan, Chrisantha Fernando, and Koray Kavukcuoglu. Population based training of neural networks. arXiv preprint arXiv:1711.09846, 2017.

Kevin Jamieson and Ameet Talwalkar. Non-stochastic best arm identification and hyperparameter optimization. In *Artificial Intelligence and Statistics*, volume 51, pp. 240–248, 2016. URL https://proceedings.mlr.press/v51/jamieson16.html.

Donald R Jones, Matthias Schonlau, and William J Welch. Efficient global optimization of expensive black-box functions. *Journal of Global Optimization*, 13(4):455–492, 1998. doi: 10.1023/A:1008306431147.

Aaron Klein, Stefan Falkner, Jost Tobias Springenberg, and Frank Hutter. Learning curve prediction with Bayesian neural networks. In *International Conference on Learning Representations*, 2017. URL https://openreview.net/forum?id=S11KBYclx.

Dong Bok Lee, Aoxuan Silvia Zhang, Byungjoo Kim, Junhyeon Park, Steven Adriaensen, Juho Lee, Sung Ju Hwang, and Hae Beom Lee. Cost-sensitive freeze-thaw Bayesian optimization for efficient hyperparameter tuning. In *Advances in Neural Information Processing Systems*, volume 38, pp. 169531–169564, 2025. doi: 10.52202/085713-5644.

Jing Lei, Max G'Sell, Alessandro Rinaldo, Ryan J Tibshirani, and Larry Wasserman. Distribution-free predictive inference for regression. *Journal of the American Statistical Association*, 113(523):1094–1111, 2018. doi: 10.1080/01621459.2017.1307116.

Liam Li, Kevin Jamieson, Afshin Rostamizadeh, Ekaterina Gonina, Jonathan Ben-tzur, Moritz Hardt, Benjamin Recht, and Ameet Talwalkar. A system for massively parallel hyperparameter tuning. In *Proceedings of Machine Learning and Systems*, volume 2, pp. 230–246, 2020. URL https://proceedings.mlsys.org/paper_files/paper/2020/hash/a06f20b349c6cf09a6b171c71b88bbfc-Abstract.html.

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. URL https://jmlr.org/papers/v18/16-558.html.

Anastasia Makarova, Huibin Shen, Valerio Perrone, Aaron Klein, Jean Baptiste Faddoul, Andreas Krause, Matthias Seeger, and Cédric Archambeau. Automatic termination for hyperparameter optimization. In *International Conference on Automated Machine Learning*, volume 188, pp. 7/1–21, 2022. URL https://proceedings.mlr.press/v188/makarova22a.html.

Andreas Maurer and Massimiliano Pontil. Empirical Bernstein bounds and sample variance penalization. In *Proceedings of the 22nd Conference on Learning Theory (COLT)*, 2009. doi: 10.48550/arXiv.0907.3740. URL https://www.cs.mcgill.ca/~colt2009/papers/012.pdf.

Samuel Müller, Noah Hollmann, Sebastian Pineda Arango, Josif Grabocka, and Frank Hutter. Transformers can do Bayesian inference. In *International Conference on Learning Representations*, 2022. URL https://openreview.net/forum?id=KSugKcbNf9.

Thanh Nhat Nguyen, Thuy Tran Phuong Nhat, Ken Takimoto, Ashutosh Thakur, Shun Nishimura, Junya Ohyama, Itsuki Miyazato, Lauren Takahashi, Jun Fujima, Keisuke Takahashi, and Toshiaki Taniike. High-throughput experimentation and catalyst informatics for oxidative coupling of methane. *ACS Catalysis*, 10(2):921–932, 2020. doi: 10.1021/acscatal.9b04293.

Herilalaina Rakotoarison, Steven Adriaensen, Neeratyoy Mallik, Samir Garibov, Edward Bergman, and Frank Hutter. In-context freeze-thaw Bayesian optimization for hyperparameter optimization. In *International Conference on Machine Learning*, volume 235, pp. 41982–42008, 2024. URL https://proceedings.mlr.press/v235/rakotoarison24a.html.

Carl Edward Rasmussen and Christopher KI Williams. *Gaussian processes for machine learning*. MIT Press, 2006. doi: 10.7551/mitpress/3206.001.0001.

Yaniv Romano, Evan Patterson, and Emmanuel Candès. Conformalized quantile regression. In *Advances in Neural Information Processing Systems*, volume 32, pp. 3538–3548, 2019. URL https://papers.nips.cc/paper_files/paper/2019/hash/5103c3584b063c431bd1268e9b5e76fb-Abstract.html.

Bobak Shahriari, Kevin Swersky, Ziyu Wang, Ryan P Adams, and Nando de Freitas. Taking the human out of the loop: A review of Bayesian optimization. *Proceedings of the IEEE*, 104(1): 148–175, 2016. doi: 10.1109/JPROC.2015.2494218.

Jasper Snoek, Hugo Larochelle, and Ryan P Adams. Practical Bayesian optimization of machine learning algorithms. In *Advances in Neural Information Processing Systems*, volume 25, pp. 2951–2959, 2012. URL https://papers.nips.cc/paper_files/paper/2012/hash/05311655a15b75fab86956663e1819cd-Abstract.html.

Jasper Snoek, Oren Rippel, Kevin Swersky, Ryan Kiros, Nadathur Satish, Narayanan Sundaram, Mostofa Patwary, Prabhat, and Ryan Adams. Scalable Bayesian optimization using deep neural networks. In *International Conference on Machine Learning*, volume 37, pp. 2171–2180, 2015. URL https://proceedings.mlr.press/v37/snoek15.html.

Samuel Stanton, Wesley Maddox, Nate Gruver, Phillip Maffettone, Emily Delaney, Peyton Greenside, and Andrew Gordon Wilson. Accelerating Bayesian optimization for biological sequence design with denoising autoencoders. In *Proceedings of the 39th International Conference on Machine Learning*, volume 162, pp. 20459–20478, 2022. URL https://proceedings.mlr.press/v162/stanton22a.html.

Samuel Stanton, Wesley Maddox, and Andrew Gordon Wilson. Bayesian optimization with conformal prediction sets. In *Proceedings of The 26th International Conference on Artificial Intelligence and Statistics*, volume 206 of *PMLR*, pp. 959–986, 2023. URL https://proceedings.mlr.press/v206/stanton23a.html. arXiv:2210.12496.

David Stenger, Dominik Scheurenberg, Heike Vallery, and Sebastian Trimpe. Early stopping Bayesian optimization for controller tuning. In *Proceedings of the 63rd IEEE Conference on Decision and Control (CDC)*, pp. 3747–3753, 2024. doi: 10.1109/CDC56724.2024.10886051. arXiv:2501.11532.

Kevin Swersky, Jasper Snoek, and Ryan Prescott Adams. Freeze-thaw Bayesian optimization. *arXiv preprint arXiv:1406.3896*, 2014. doi: 10.48550/arXiv.1406.3896.

Ryan J Tibshirani, Rina Foygel Barber, Emmanuel Candès, and Aaditya Ramdas. Conformal prediction under covariate shift. In *Advances in Neural Information Processing Systems*, volume 32, 2019. URL https://proceedings.neurips.cc/paper/2019/hash/8fb21ee7a2207526da55a679f0332de2-Abstract.html.

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. doi: 10.1145/2641190.2641198.

Abraham Wald. Sequential tests of statistical hypotheses. *The Annals of Mathematical Statistics*, 16 (2):117–186, 1945. doi: 10.1214/aoms/1177731118.

James T. Wilson. Stopping Bayesian optimization with probabilistic regret bounds. In *Advances in Neural Information Processing Systems*, volume 37, pp. 98264–98296, 2024. doi: 10.52202/079017-3118. arXiv:2402.16811.

Martin Wistuba, Arlind Kadra, and Josif Grabocka. Supervising the multi-fidelity race of hyperparameter configurations. In *Advances in Neural Information Processing Systems*, volume 35, pp. 13470–13484, 2022. doi: 10.52202/068431-0979.

Yongrui Xiao, Mengjia Zhu, Yu Zhang, Dimitrios Tsaoulidis, and Tao Chen. Confidence-bound early stopping of experiments with sequential calibration. *Chemical Engineering Research and Design*, 230:756–766, 2026. doi: 10.1016/j.cherd.2026.05.013.

I-Cheng Yeh. Modeling of strength of high-performance concrete using artificial neural networks. *Cement and Concrete Research*, 28(12):1797–1808, 1998. doi: 10.1016/S0008-8846(98)00165-3.

Lucas Zimmer, Marius Lindauer, and Frank Hutter. Auto-PyTorch: Multi-fidelity metalearning for efficient and robust AutoDL. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 43 (9):3079–3090, 2021. doi: 10.1109/TPAMI.2021.3067763.

Klea Ziu, Ruben Solozabal, Srinivas Rangarajan, and Martin Takáč. A deep neural network for oxidative coupling of methane trained on high-throughput experimental data. *Journal of Physics: Energy*, 5(1):014009, 2023. doi: 10.1088/2515-7655/aca797.

A Partial-Trajectory Reliability Diagnostics

To measure how reliable partial trajectories are on the benchmarks introduced in Section 5.1 (Appendix C), we compute four pool-level diagnostics on each dataset (the same table also records SRBO's runtime false-stop rate against its nominal level and which method wins each pool). Let $\{(\mathbf{y}_{1:T}(x_i), Y(x_i))\}_{i=1}^{N}$ denote the pool of configurations with their full trajectories and targets. We evaluate: (i) the *Kendall rank correlation* $\tau_t = \tau(\mathbf{y}_t, Y)$ between proxy ranking at step t and target ranking; (ii) the *curve-crossing rate* at step $t = T/4$, defined as the fraction of configuration pairs (i, j) ranked in strictly opposite order, $(y_t^i - y_t^j)(Y^i - Y^j) < 0$ (a pair tied under either ranking counts as not crossed); (iii) the *bottom-K Jaccard index* between the K lowest-ranked configurations under $y_{T/4}$ and under Y, with $K = 10$ on every pool, which is the set an early-stopping rule discards first; and (iv) the *normalized linear-fit RMSE* $\mathrm{RMSE}_{\mathrm{lin}}/\mathrm{std}(Y)$, the residual RMSE of the least-squares line that predicts Y from $y_{T/4}$ across the pool divided by the standard deviation of Y (equivalently $\sqrt{1 - r^2}$ for the Pearson correlation r between $y_{T/4}$ and Y). For $T = 11$, $t = T/4$ denotes the third of the 11 checkpoints (0-based index $\lfloor T/4 \rfloor = 2$ in `compute_diagnostics.py`). A pure multi-fidelity regime would exhibit $\tau_{T/4} \to 1$, crossing rate $\to 0$, Jaccard $\to 1$, and $\mathrm{RMSE}_{\mathrm{lin}}$ small.

Table 1 quantifies partial-trajectory unreliability on every benchmark. Early-trajectory Kendall τ ranges from 0.58 (IVPT) to 0.84 (Catechol, just above the 0.83 of LCBench vehicle and segment), so between 7% and 21% of configuration pairs are ranked in the wrong order at $t = T/4$. Bottom-K Jaccard runs from 0.33 on IVPT to 0.82 on LCBench (credit-g) and Catechol, so on the hardest benchmark only five of the ten configurations that look worst at $t = T/4$ are still among the ten worst

Table 1: Partial-trajectory reliability diagnostics at $t = T/4$ of the trajectory (all four columns produced by `compute_diagnostics.py` in the supplement, with the bottom-K sets taken under a stable sort so that ties break by index and the column is platform-independent). The last two columns give SRBO's runtime false-stop rate (mean over the 20 seeds) against its nominal level, and the method with the lowest mean regret on that pool; the nominal is each pool's tuned α_{fsr}. Low Kendall τ, high crossing rate, low bottom-K Jaccard and large linear-fit RMSE indicate an early proxy that is mis-aligned with the target.

Dataset	$\tau_{T/4}$	Crossing	Jaccard	$\text{RMSE}_{\text{lin}}/\text{std}(Y)$	FSR / α_{fsr}	Best
IVPT	0.58	0.21	0.33	0.66	0.003/0.10	SRBO
HEA-OER	0.66	0.17	0.43	0.58	0.005/0.025	HyperbandEI
Concrete	0.69	0.16	0.54	0.51	0.001/0.025	SRBO
OCM	0.66	0.17	0.43	0.51	0.000/0.025	SRBO
Catechol	0.84	0.07	0.82	0.49	0.022/0.025	StandardBO
LCBench (credit-g)	0.65	0.15	0.82	0.34	0.004/0.05	BOHB
LCBench (vehicle)	0.83	0.08	0.67	0.32	0.004/0.05	BOHB
LCBench (segment)	0.83	0.08	0.54	0.30	0.005/0.025	BOHB

at the endpoint (a Jaccard index of 0.33 between two sets of ten). Normalized linear-fit RMSE places the three LCBench tasks, where BOHB leads, below every experimental pool, at 0.30 to 0.34 against 0.49 to 0.66; among the experimental pools it does not separate the three on which SRBO wins from the two on which it does not (Section 5.2, Figure 4).

Coverage and the two halving methods. Under the same cost-matched budget SRBO's and BOHB's stopping rules buy different things. BOHB halts most of what it starts, so its budget touches 172 to 201 configurations per seed on the three LCBench tasks against SRBO's 106 to 121; on IVPT and Concrete the same comparison gives $1.26\times$ and $1.63\times$. Coverage alone does not decide the outcome. HyperbandEI halves in Hyperband brackets as BOHB does, at its own tuned η and with StandardBO's EI selector, and touches 172 to 242 configurations per seed on every pool, yet it ends above SRBO on all three LCBench tasks, as on IVPT, Concrete and OCM (Tables 4 and 13), and at the shared default $\eta = 2$ it trails BOHB on all three LCBench tasks (Table 6). BOHB's lead on LCBench therefore belongs to its TPE selector on pools whose early observations predict the endpoint well (linear-fit RMSE 0.34 or below), and Appendix B.1 places its IVPT gap mainly on the same selector. Where the RMSE is 0.49 or above, BOHB trails SRBO on IVPT, Concrete, OCM and Catechol and shows no significant difference from it only on HEA-OER.

B Full Wilcoxon Significance Results, Per-Ablation FSR, and Sensitivity Sweeps

Figure 3 visualizes the average per-seed rank of each method aggregated across all eight benchmarks. It is restricted to the seven methods that were run on every one of them, which is the widest apples-to-apples comparison the data supports. Lower is better; SRBO achieves the best aggregate rank (mean $\pm$ SE, 2.56 ± 0.11), followed by BOHB (3.44 ± 0.18), HyperbandEI (4.09 ± 0.16) and StandardBO (4.18 ± 0.13). The aggregate favors SRBO because by mean regret it is never below fourth of the seven on any pool, whereas every other method is last or second to last on at least one (Tables 12 and 13). Figure 5 summarizes empirical FSR on all eight benchmarks against each benchmark's own nominal α_{fsr}. Table 2 reports the cell-level ablation numbers referenced from Section 5.3. Table 4 reports the pairwise Wilcoxon signed-rank p-values between SRBO and each baseline on per-seed final regret for the five pools of Table 12; those for OCM, HEA-OER and Catechol are in the caption of Table 13. Table 3 reports the empirical FSR of each ablation variant; full SRBO and no-cens stay below 0.01 everywhere, and on the four pools where no-DFGP stops essentially every trial its zero FSR reflects a failed search rather than calibration. Table 5 reports the hyperparameter sensitivity sweeps for the two principal knobs (ξ held fixed at its default value 0.1). Table 6 contrasts each tunable method's default configuration with the one its tuning pool selected, and Table 7 reports the three controlled BOHB variants discussed in Appendix B.1. Sweeping $\alpha_{\text{fsr}} \in \{0.01, 0.05, 0.10\}$ leaves the default 0.05 lowest on all three benchmarks (0.18, 15.80 and 0.120); the largest relative change is at 0.01 on LCBench (cg) (0.145 against 0.120). Sweeping $\gamma \in \{1, 3, 5\}$ gives the highest

regret at $\gamma = 5$ on IVPT and LCBench (cg) (0.22 and 0.150), while $\gamma = 1$ is lower in mean than the default on IVPT (0.17 against 0.18) and Concrete (12.26 against 15.80) and within noise of it on all three.

Table 2: Ablation study: final simple regret under the deployed protocol (each pool's tuned $(\alpha_{\text{fsr}}, \gamma, \xi)$ from Table 12, pure EI, $C = 1100$, 20 seeds, mean $\pm$ SE). SRBO (full) is run in the same job as the two variants, at the configuration and seeds of the Table 12 cell. Removing the DF-GP pseudo-observations (no-DFGP) degrades regret 306-fold on Concrete and 4.2- to 6.5-fold on the LCBench tasks, but only 1.2-fold on IVPT ($p = 0.11$); removing the censored GP (no-cens) changes no column significantly (Table 4). **Bold** marks the lowest value in each column.

Variant	IVPT	Concrete	LCBench (cg)	LCBench (veh)	LCBench (seg)
SRBO (full)	$\mathbf{0.21 \pm 0.03}$	$\mathbf{4.02 \pm 1.02}$	0.109 ± 0.007	$\mathbf{0.213 \pm 0.012}$	$\mathbf{0.234 \pm 0.013}$
$\text{SRBO}_{\text{no-cens}}$	$\mathbf{0.21 \pm 0.03}$	$\mathbf{4.02 \pm 1.02}$	$\mathbf{0.108 \pm 0.007}$	$\mathbf{0.213 \pm 0.012}$	$\mathbf{0.234 \pm 0.013}$
$\text{SRBO}_{\text{no-DFGP}}$	0.25 ± 0.03	1229.95 ± 0.00	0.455 ± 0.003	1.244 ± 0.007	1.512 ± 0.000

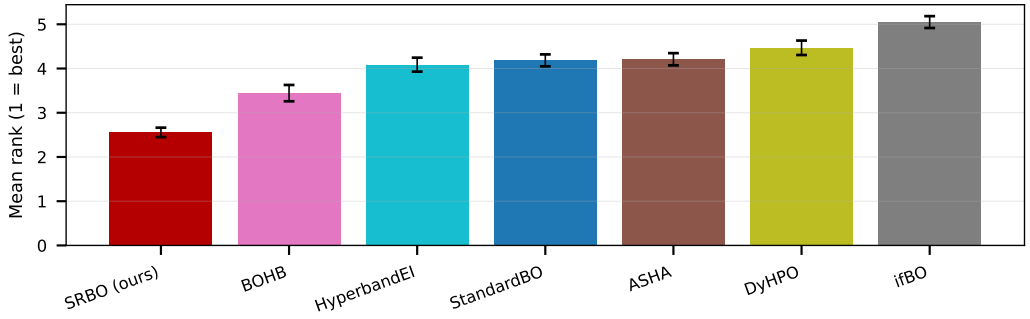


Figure 3: Average per-seed final-regret rank across all eight benchmarks (20 seeds $\times$ 8 datasets, the seven methods run on every benchmark); lower is better. Error bars: ± 1 SE over the 160 per-(dataset, seed) ranks. SRBO achieves the best rank (2.56), followed by BOHB (3.44), HyperbandEI (4.09) and StandardBO (4.18).

Table 3: Empirical runtime FSR of the ablation runs of Table 2 (same seeds and configurations; nominal α_{fsr} as tuned per pool, 0.10 on IVPT, 0.05 on LCBench (cg) and (veh), and 0.025 on Concrete and LCBench (seg)). Full SRBO and no-cens stay below 0.01 on every task. No-DFGP stops 67% of its trials on IVPT at a false-stop rate of 0.12, above its nominal 0.10, and over 99% of them on the other four pools, where none of the trials it stops would have been a winner, so its zero FSR there reflects a failed search (Table 2), not calibration.

Variant	IVPT	Concrete	LCBench (cg)	LCBench (veh)	LCBench (seg)
SRBO (full)	0.004	0.001	0.003	0.005	0.005
$\text{SRBO}_{\text{no-cens}}$	0.004	0.001	0.002	0.005	0.005
$\text{SRBO}_{\text{no-DFGP}}$	0.120	0.000	0.000	0.000	0.000

*Starred single-value cells mark cases where the tuning pool selected the default configuration; the selected default run is the value reported in Table 12.

B.1 BOHB Failure-Mode Diagnostics on IVPT and Concrete

To isolate *why* BOHB underperforms on the two pharmaceutical/materials benchmarks, we run three controlled diagnostic variants alongside vanilla BOHB, all at BOHB's default configuration ($\eta = 2$, $\gamma_{\text{tpe}} = 0.15$) rather than at the tuned configurations of Table 12 ($n = 20$ seeds each, identical TPE / random-init schedule, identical cost budget 1100). These controlled variants implicate halving only on IVPT. Disabling halving while keeping TPE lowers regret from 180.7 to 167.6 ($p < 10^{-4}$),

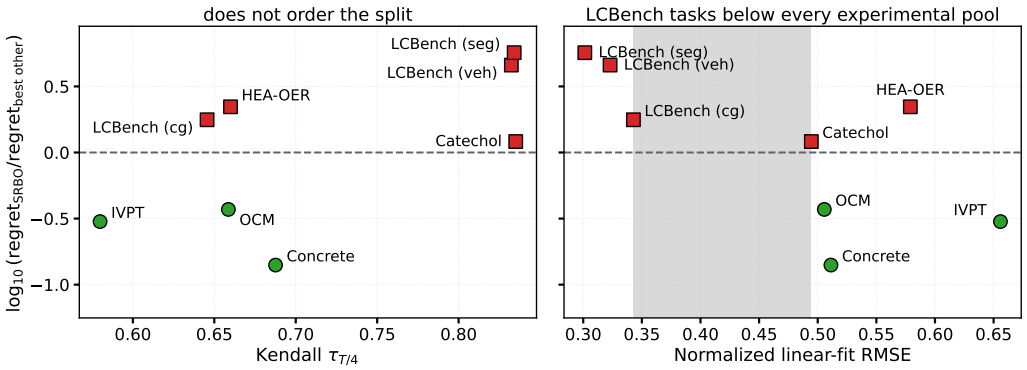


Figure 4: SRBO regret-ratio against two partial-trajectory reliability diagnostics, over all eight benchmarks. Each point is one benchmark; y-axis is $\log_{10}(\text{regret}_{\text{SRBO}}/\text{regret}_{\text{best baseline}})$. Green circles (negative y): SRBO has the lowest mean regret; red squares (positive y): another method does. *Left:* $\tau_{T/4}$ does not order the split. OCM ($\tau = 0.66$) is within 0.01 of HEA-OER, yet SRBO wins on OCM and not on HEA-OER. *Right:* normalized linear-fit RMSE places the three LCBench tasks, where BOHB leads by a factor of 1.8 to 5.7, below every experimental pool (0.34 or below against 0.49 or above; shaded gap), but it does not separate HEA-OER and Catechol from the three experimental pools on which SRBO wins.

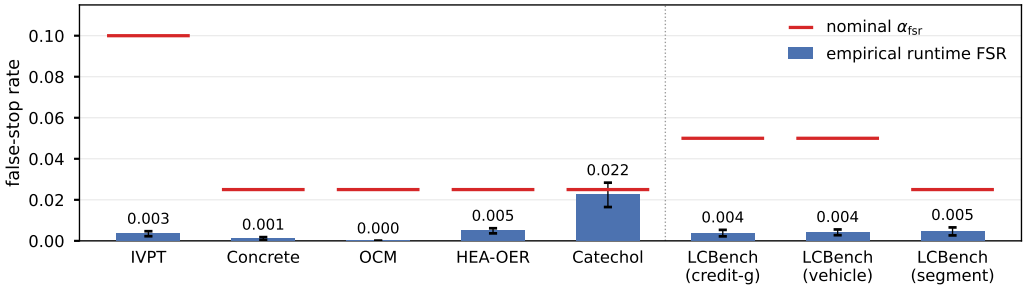


Figure 5: Empirical SRBO runtime FSR on all eight benchmarks. Blue bars are the mean per-seed runtime FSR over the 20 reporting seeds (± 1 SE), and each red segment is the benchmark's own nominal α_{fsr}, the tuned α_{fsr} of the runs behind the tables, 0.10 on IVPT, 0.05 on LCBench (cg) and (veh), and 0.025 on the other five. Every bar falls below its own nominal level; the smallest margin is Catechol's, 0.022 against 0.025 (within one SE), and every other bar is at least five times below its level.

whereas on Concrete disabling halving raises it from 72.2 to 146.5 ($p = 0.04$) and oracle halving does not separate from vanilla BOHB ($p = 0.39$), so the attribution stays mixed.

TPE-only (no halving).
Strip the Hyperband bracket entirely: each iteration, TPE suggests 1 candidate evaluated to step T at full cost. Tests TPE in isolation from halving on small continuous pools.

Oracle halving.
Keep the sync Hyperband brackets and TPE candidate selection, but rank survivors by the *true endpoint* y_T (oracle access via the pool simulator) instead of by the proxy y_t. Upper bound on what halving could achieve under perfect proxy ranking.

False-discard audit.
Vanilla BOHB instrumented to log, per bracket, whether the bracket's eventual best-y_T candidate is eliminated at any halving step.

Table 4: Two-sided Wilcoxon signed-rank p-values on per-seed final regret. Top block: SRBO vs each external baseline (cost-matched protocol $C = 1100$, $n = 20$ paired seeds per cell, tunable methods at their tuned configs; non-tunable at default). Bottom block: SRBO (full) vs each ablation variant of Table 2 (deployed protocol), $n = 20$ paired seeds on every dataset. Bold marks significant at the Wilcoxon test level $\alpha_{\text{sig}} = 0.05$ (distinct from the FSR risk α_{fsr}); values below 10^{-4} reported as "$< 10^{-4}$"; "identical" marks a pool on which the variant reproduces full SRBO seed by seed, so no test applies. An asterisk (*) marks cells where the baseline (not SRBO) achieves the lower mean regret; these are BOHB on all three LCBench tasks, DyHPO on LCBench (veh) and LCBench (seg), and ifBO on LCBench (veh).

SRBO vs.	IVPT	Concrete	LCBench (cg)	LCBench (veh)	LCBench (seg)
External baselines					
StandardBO	$\mathbf{< 10^{-4}}$	$\mathbf{< 10^{-4}}$	$\mathbf{< 10^{-4}}$	$\mathbf{< 10^{-4}}$	$\mathbf{< 10^{-4}}$
LinExtrap	$\mathbf{< 10^{-4}}$	$\mathbf{< 10^{-4}}$	$\mathbf{< 10^{-4}}$	$\mathbf{< 10^{-4}}$	$\mathbf{< 10^{-4}}$
EI_Discard	$\mathbf{< 10^{-4}}$	$\mathbf{< 10^{-4}}$	$\mathbf{< 10^{-4}}$	$\mathbf{< 10^{-4}}$	$\mathbf{< 10^{-4}}$
EI_Virtual	$\mathbf{< 10^{-4}}$	$\mathbf{< 10^{-4}}$	$\mathbf{0.0005}$	$\mathbf{< 10^{-4}}$	$\mathbf{< 10^{-4}}$
Freeze-Thaw	$\mathbf{< 10^{-4}}$	$\mathbf{< 10^{-4}}$	$\mathbf{< 10^{-4}}$	$\mathbf{< 10^{-4}}$	$\mathbf{< 10^{-4}}$
ASHA	$\mathbf{0.001}$	$\mathbf{< 10^{-4}}$	$\mathbf{< 10^{-4}}$	$\mathbf{< 10^{-4}}$	$\mathbf{< 10^{-4}}$
HyperbandEI	$\mathbf{0.024}$	$\mathbf{< 10^{-4}}$	$\mathbf{< 10^{-4}}$	$\mathbf{< 10^{-4}}$	$\mathbf{< 10^{-4}}$
BOHB	$\mathbf{< 10^{-4}}$	$\mathbf{0.007}$	$\mathbf{0.0007}^\star$	$\mathbf{< 10^{-4}}^\star$	$\mathbf{< 10^{-4}}^\star$
ifBO	$\mathbf{< 10^{-4}}$	$\mathbf{< 10^{-4}}$	$\mathbf{0.0001}$	$\mathbf{0.0009}^\star$	$\mathbf{< 10^{-4}}$
DyHPO	$\mathbf{< 10^{-4}}$	$\mathbf{< 10^{-4}}$	0.430	$\mathbf{0.036}^\star$	$\mathbf{0.004}^\star$
Ablation variants (SRBO full vs variant)					
$\text{SRBO}_{\text{no-cens}}$	identical	identical	0.32	identical	identical
$\text{SRBO}_{\text{no-DFGP}}$	0.11	$\mathbf{< 10^{-4}}$	$\mathbf{< 10^{-4}}$	$\mathbf{< 10^{-4}}$	$\mathbf{< 10^{-4}}$

Table 5: Hyperparameter sensitivity of SRBO around its default configuration (deployed EI acquisition, cost-matched $C = 1100$, 20 seeds, mean $\pm$ SE; one hyperparameter varied at a time from the default $(\alpha_{\text{fsr}}, \gamma, \xi) = (0.05, 3, 0.10)$, shown in **bold**). SRBO is robust to both: α_{fsr} over $[0.01, 0.10]$ changes the regret by at most $1.2\times$ on any benchmark, and the noise inflation γ over $\{1, 3, 5\}$ by at most $1.3\times$, with $\gamma = 1$ within noise of the default on all three (lower in mean on IVPT and Concrete).

HP	Value	IVPT	Concrete	LCBench (cg)
α_{fsr}	0.01	0.20 ± 0.03	16.35 ± 1.82	0.145 ± 0.009
	$\mathbf{0.05}$	$\mathbf{0.18 \pm 0.02}$	$\mathbf{15.80 \pm 2.02}$	$\mathbf{0.120 \pm 0.005}$
	0.10	0.20 ± 0.02	18.24 ± 2.90	0.122 ± 0.010
γ	1	0.17 ± 0.02	12.26 ± 1.65	0.127 ± 0.008
	$\mathbf{3}$	$\mathbf{0.18 \pm 0.02}$	$\mathbf{15.80 \pm 2.02}$	$\mathbf{0.120 \pm 0.005}$
	5	0.22 ± 0.02	15.48 ± 1.67	0.150 ± 0.013

Reading the diagnostic. On IVPT, removing halving entirely (TPE-only) gives a small but significant mean gain over vanilla BOHB ($180.7 \to 167.6$, $p < 10^{-4}$); oracle halving does *not* help (181.3, $p = 0.39$). About one in five halving brackets (18.9%) eliminates the bracket's best-y_T candidate, yet oracle access to y_T for ranking yields no improvement, suggesting that TPE candidate selection on the IVPT continuous space, rather than the proxy-ranking step, is the main bottleneck. On Concrete TPE-only is worse than vanilla (146.5 against 72.2, $p = 0.04$), so halving helps BOHB there; oracle halving is lower in mean ($72.2 \to 56.4$) but not significantly ($p = 0.39$), and only 3.0% of halving brackets discard their best candidate.

In these three controlled variants TPE on small continuous pools is a substantial bottleneck on IVPT, while on Concrete halving helps BOHB and proxy-ranked halving is not by itself a statistically decisive explanation of its gap to SRBO. HyperbandEI, which keeps Hyperband's bracket halving and replaces TPE by EI, is the strongest baseline on both pools (Table 12).

Practical guardrail. Prefer BOHB when normalized linear-fit RMSE is low (0.34 or below here, the three LCBench tasks, on all of which BOHB leads; Table 1). Kendall $\tau_{T/4}$ and crossing rate are

Table 6: Default-vs-tuned sensitivity for the four tunable methods (cost-matched protocol $C = 1100$, 20-seed reporting pool, mean final simple regret, no test-set regret minimization; SRBO defaults are $(\alpha_{\text{fsr}}, \gamma, \xi) = (0.05, 3, 0.10)$ under the deployed EI acquisition, matching the protocol used in Table 12). Cells show "default / selected non-default rerun" when both are tabulated; a single starred value means the tuning pool selected the default configuration. On IVPT ASHA's selected $\eta = 3$ and its default $\eta = 2$ have the same mean regret to two decimals. **Bold** marks the selected value used in Table 12. Selection is made on the disjoint 5-seed tuning pool, so a default value can remain selected even when an alternate reporting-pool value is lower. Tuning improves SRBO on all five pools, whereas BOHB's default configuration would have scored lower than the configuration its own tuning pool selected on LCBench (cg) and (seg) (0.059 and 0.031 against 0.060 and 0.039), HyperbandEI's on LCBench (cg) (0.180 against 0.181), and ASHA's on LCBench (seg) (0.690 against 0.697). We report the selected configurations throughout, so on these cells Table 12 understates these baselines, and BOHB leads LCBench (cg) and (seg) either way.

Method	IVPT	Concrete	LCBench (cg)	LCBench (veh)	LCBench (seg)
SRBO ($\alpha_{\text{fsr}}, \gamma, \xi$)	$0.18/\mathbf{0.17}$	$15.80/\mathbf{4.02}$	$0.120/\mathbf{0.107}$	$0.324/\mathbf{0.212}$	$0.537/\mathbf{0.221}$
BOHB (η, γ_{tpe})	$180.75/\mathbf{165.55}$	$72.21/\mathbf{33.80}$	$0.059/\mathbf{0.060}$	$0.051/\mathbf{0.046}$	$0.031/\mathbf{0.039}$
ASHA (η)	$0.69/\mathbf{0.69}$	$\mathbf{43.29}^*$	$\mathbf{0.180}^*$	$0.635/\mathbf{0.574}$	$0.690/\mathbf{0.697}$
HyperbandEI (η)	$\mathbf{0.58}^*$	$35.77/\mathbf{28.61}$	$0.180/\mathbf{0.181}$	$0.658/\mathbf{0.589}$	$0.808/\mathbf{0.746}$

Table 7: BOHB failure-mode diagnostics on IVPT and Concrete ($n = 20$ seeds, mean $\pm$ SE final simple regret at the shared cost budget $C = 1100$; median in brackets). All variants share the same TPE schedule, random initialization, and cost budget; only the indicated component is changed. The right column reports the per-bracket false-discard rate (% of completed halving brackets where the initial bracket's best-y_T candidate is eliminated by at least one halving step) under vanilla BOHB. Wilcoxon p-values vs. vanilla in parentheses.

Dataset	Vanilla BOHB	TPE-only (no halving)	Oracle halving	False-discard
IVPT	180.7 ± 1.5 [183.0]	167.6 ± 1.2 [167.0] ($p < 10^{-4}$)	181.3 ± 1.6 [183.0] ($p = 0.39$)	18.9%
Concrete	72.2 ± 20.1 [62.8]	146.5 ± 32.8 [93.5] ($p = 0.04$)	56.4 ± 11.6 [61.5] ($p = 0.39$)	3.0%

not enough on their own. Catechol has the highest $\tau_{T/4}$ and the lowest crossing rate of the eight pools, yet BOHB's regret there is $4\times$ SRBO's. Prefer SRBO over BOHB where the RMSE is 0.49 or above; there BOHB is significantly worse than SRBO on IVPT, Concrete, OCM and Catechol and not significantly different from it only on HEA-OER.

B.2 Controlled Mismatch Validation (Synthetic Stress Tests)

We add a controlled synthetic suite to isolate the effect of partial-trajectory unreliability. The tasks test whether SRBO's advantage appears when rank reversals are induced by construction. Each task uses a standard test function (Branin $d = 2$ pool 144, Hartmann-3 $d = 3$ pool 216, Rosenbrock-3 $d = 3$ pool 216) as the pool endpoint, $y_T(x_i) = f(x_i)$ at each pool point x_i (a grid point with a fixed jitter of at most 0.01 per coordinate), and constructs the proxy curve as $y_t(x) = w(t)\, f(x) + (1 - w(t))\, h(x)$, where $h(x) \sim \mathcal{N}(\text{mean } f, \text{var } f)$ is an i.i.d. per-config noise proxy and $w(t) = ((t-1)/(T-1))^p$ is the curve-blend weight with p set from the Gaussian relation $\tau = (2/\pi) \arcsin \rho$ between Kendall's τ and the correlation ρ of $y_{T/4}$ and y_T, so that $\tau_{T/4}$ is close to the target value. Target $\tau_{T/4}$ values and final regrets are reported in Table 8. As on IVPT, Concrete and the three LCBench tasks, every method queries a thin-plate-spline oracle that interpolates the grid (smoothing 0; Appendix F), and regret is measured to that oracle's optimum over the box, 3.43, 3.44 and 52.64 against maxima of -0.40, 3.86 and 0 for the three functions themselves. Every method runs at its default configuration under the protocol of Table 12 (cost budget $C = 1100$; SRBO with the deployed EI acquisition at $(\alpha_{\text{fsr}}, \gamma, \xi) = (0.05, 3, 0.10)$; HyperbandEI and BOHB at $\eta = 2$, BOHB with $\gamma_{\text{tpe}} = 0.15$), with no per-task tuning. BOHB ends $51\times$ to $144\times$ above the EI methods on all three tasks, whereas HyperbandEI, which runs the same brackets at the same η with the EI selector, stays within the range of the other EI methods, so the failure follows BOHB's TPE selector rather than halving. Among the EI methods SRBO leads on Hartmann-3 (0.007 against 0.010 to 0.014), shows no significant difference on Branin and trails on Rosenbrock-3 (2.90 against 1.02 to 1.08), so the suite does not by itself favor soft rejection over full-length EI.

Table 8: Controlled-mismatch synthetic suite. Final simple regret, mean ± SE on 20 seeds at default configs. Pool on a jittered regular grid; proxy curves constructed so that $\tau_{T/4}$ is close to the target column. Regret is measured to the optimum of the spline oracle fitted to the grid (Appendix B.2). BOHB fails on all three tasks, HyperbandEI does not, and SRBO leads only on Hartmann-3.

Dataset (d, N)	Target $\tau_{T/4}$	SRBO	BOHB	HyperbandEI	StandardBO	EI_Virtual
Branin (2, 144)	0.50	0.023 ± 0.007	2.08 ± 0.30	0.024 ± 0.001	0.029 ± 0.004	$\mathbf{0.021 \pm 0.004}$
Hartmann-3 (3, 216)	0.55	$\mathbf{0.007 \pm 0.001}$	0.87 ± 0.04	0.013 ± 0.001	0.014 ± 0.001	0.010 ± 0.001
Rosenbrock-3 (3, 216)	0.60	2.90 ± 0.07	147.0 ± 5.5	1.08 ± 0.14	1.03 ± 0.11	$\mathbf{1.02 \pm 0.10}$

What this test answers. *BOHB fails on every synthetic task* (regret $51\times$ to $73\times$ that of the next-worst method; Wilcoxon $p < 10^{-4}$ on every dataset), and HyperbandEI, with the same brackets and the EI selector, does not, so BOHB's failure follows its TPE selector rather than halving as such. Against the halving methods SRBO significantly beats BOHB on all three tasks and HyperbandEI on Hartmann-3 ($p = 0.0009$), does not differ significantly from HyperbandEI on Branin ($p = 0.81$) and loses to it on Rosenbrock-3 ($p < 10^{-4}$). Against StandardBO and EI_Virtual it wins Hartmann-3 ($p = 0.048$ against EI_Virtual and $p < 10^{-4}$ against StandardBO, $1.5\times$ and $2.1\times$ better), does not differ significantly from them on Branin ($p \geq 0.31$) and loses to them on Rosenbrock-3, so its position is task-dependent rather than monotone in $\tau_{T/4}$. **What it does not answer.** Whether SRBO transfers to further experimental datasets, which requires experimental data we do not collect here.

B.3 Endpoint Cumulative Costs

Table 9: Endpoint cumulative cost of the Table 12 runs (mean over the 20 reporting seeds of each run's final cumulative cost, against the nominal budget $C = 1100$). These are the endpoints of the cost vs. regret trajectories in Figure 2. No run spends past C. SRBO stops once the remaining budget cannot pay for its shortest informative trial, so it can end a few units short of C (its means are 1099 to 1100); every other method, the two bracket-based ones included, ends at exactly C on every seed. The three benchmarks of Table 13 follow the same pattern (SRBO 1097 to 1100, every other method 1100).

Method	IVPT	Concrete	LCBench (cg)	LCBench (veh)	LCBench (seg)
StandardBO	1100	1100	1100	1100	1100
LinExtrap	1100	1100	1100	1100	1100
EI_Discard	1100	1100	1100	1100	1100
EI_Virtual	1100	1100	1100	1100	1100
Freeze-Thaw	1100	1100	1100	1100	1100
ASHA	1100	1100	1100	1100	1100
HyperbandEI	1100	1100	1100	1100	1100
BOHB	1100	1100	1100	1100	1100
ifBO	1100	1100	1100	1100	1100
DyHPO	1100	1100	1100	1100	1100
SRBO (ours)	1100	1099	1099	1100	1100

B.4 Cost-Budget Comparison

To make the cost vs. performance trade-off explicit, we rerun the main experiment at two tighter hard cost budgets; each method runs until its cumulative proxy-step cost reaches C or its next step no longer fits (the two bracket-based methods stop at 817 under $C = 825$, where the next rung does not fit), at which point the best-so-far Y is recorded as the final regret. We test two budgets: 550 (half of StandardBO's 100-iteration cost of 1100) and 825 (75% of that cost). Table 10 reports final simple regret across IVPT / Concrete / LCBench credit-g at both budgets.

Reading the table. The cost-budget pattern reproduces the main-results split along the partial-trajectory reliability axis. LinExtrap, EI_Discard and Freeze-Thaw stall at regret 1229.95 on Concrete at both budgets, as in Table 12, and SRBO is the strongest of the methods that avoid it (18.94 and 16.29 against 47.95 to 156.89 for the others). On IVPT, SRBO at its default configuration leads

Table 10: Cost-budget comparison: final simple regret (mean $\pm$ SE, $n = 20$ per cell) when each method is run under a hard cumulative cost cap C, every method at its default configuration (SRBO $(\alpha_{\text{fsr}}, \gamma, \xi) = (0.05, 3, 0.10)$ with the deployed EI acquisition; ASHA, HyperbandEI and BOHB at $\eta = 2$; BOHB with $\gamma_{\text{tpe}} = 0.15$). Column headings: Std BO = StandardBO, LinEx = LinExtrap, Disc = EI_Discard, Virt = EI_Virtual, FT = Freeze-Thaw, HB-EI = HyperbandEI, BOHB = sync-bracket Hyperband-family with TPE candidate selector. † = every seed ends at the largest per-seed regret of that pool in Table 12 (the method stalls). ‡ = the stopping rule fires on essentially every trial, so the method degenerates to its warm-start set. **Bold** = best per row.

Budget	Dataset	Std BO	LinEx	Disc	Virt	FT	ASHA	HB-EI	BOHB	SRBO
$C = 550$	IVPT	6.14 ± 1.93	$90.31 \pm 0.00^{\ddagger}$	5.11 ± 1.59	18.72 ± 5.57	$90.31 \pm 0.00^{\ddagger}$	1.73 ± 0.34	16.29 ± 2.26	185.51 ± 1.81	$\mathbf{0.31 \pm 0.04}$
	Concrete	55.97 ± 9.51	$1229.95^{\dagger}$	$1229.95^{\dagger}$	85.53 ± 10.79	$1229.95^{\dagger}$	68.80 ± 9.81	57.46 ± 7.28	156.89 ± 44.38	$\mathbf{18.94 \pm 2.09}$
	LCBench (cg)	0.198 ± 0.008	$0.444 \pm 0.000^{\ddagger}$	$0.444 \pm 0.000^{\ddagger}$	0.206 ± 0.010	$0.444 \pm 0.000^{\ddagger}$	0.229 ± 0.008	0.213 ± 0.005	$\mathbf{0.065 \pm 0.008}$	0.143 ± 0.007
$C = 825$	IVPT	1.64 ± 0.67	$89.10 \pm 1.20^{\ddagger}$	1.35 ± 0.30	8.21 ± 2.66	$89.10 \pm 1.20^{\ddagger}$	0.61 ± 0.13	1.79 ± 0.73	183.39 ± 1.13	$\mathbf{0.21 \pm 0.02}$
	Concrete	52.44 ± 9.42	$1229.95^{\dagger}$	$1229.95^{\dagger}$	74.93 ± 10.45	$1229.95^{\dagger}$	55.03 ± 7.88	47.95 ± 6.14	132.75 ± 41.01	$\mathbf{16.29 \pm 1.97}$
	LCBench (cg)	0.178 ± 0.009	$0.444 \pm 0.000^{\ddagger}$	$0.444 \pm 0.000^{\ddagger}$	0.186 ± 0.010	$0.444 \pm 0.000^{\ddagger}$	0.209 ± 0.007	0.203 ± 0.005	$\mathbf{0.057 \pm 0.008}$	0.130 ± 0.006

at both budgets (0.31 and 0.21), ahead of ASHA (1.73 and 0.61) and EI_Discard (5.11 and 1.35); LinExtrap and Freeze-Thaw stop essentially every trial and end near 90, and BOHB ends at 183 or more.

BOHB pattern. BOHB fails on IVPT at both budgets (186 and 183), as in Table 12, and is heavy-tailed on Concrete (157 and 133 in mean); tuning at $C = 1100$ brings its Concrete mean to 33.8, but the default configuration used here does not recover. This is consistent with the mixed BOHB failure-mode diagnostics in Appendix B.1. However, BOHB *wins LCBench-cg at both budgets* (0.065 at $C = 550$, 0.057 at $C = 825$), beating every other method including SRBO, which is second at both budgets (0.143 and 0.130), ahead of every method other than BOHB, consistent with the regime-specific advantage discussed in Section 6. SRBO's gains concentrate in settings with pronounced partial-trajectory unreliability.

C Dataset Details and Partial-Trajectory Dynamics

Table 11: Sequential-experiment benchmark datasets. Pool size is the number of recorded trajectories (on IVPT, 135 physical replicates of 28 unique formulations).

Dataset	Domain	Dim. (d)	Pool size	Proxy steps (T)
IVPT	Pharmaceutical permeation	3	135	11
Concrete	Civil engineering materials	7	127	11
LCBench (credit-g)	AutoML HPO	7	200	11
LCBench (vehicle)	AutoML HPO	7	200	11
LCBench (segment)	AutoML HPO	7	200	11

We evaluate SRBO on eight benchmark datasets (Table 11 and Appendix C.4), five experimental pools on which early observations predict the endpoint poorly (normalized linear-fit RMSE 0.49 or above) and three LCBench tasks on which they predict it well (0.34 or below; Table 1). For each dataset, we outline the search space, the physical or computational meaning of the intermediate proxy y_t, and the final target objective Y.

C.1 In Vitro Permeation Test (IVPT)

The IVPT dataset is an internal set of in vitro permeation measurements on topical formulations: the goal is to discover an optimal formulation that maximizes skin permeation. The anonymized numerical data (formulation composition + 11 time-course permeation readings, $N = 135$ replicates) is shipped in the supplementary material at `data/ivpt/DATA1.csv` so that all reported IVPT numbers are independently reproducible from this submission; the full collection protocol (instrument configuration, batch provenance, quality-control criteria) will be released with a public DOI upon acceptance.

- **Search Space ($d = 3$):** the concentrations of three excipients, searched over the smallest box containing the 28 unique formulations of the pool (135 physical replicates) and 500 uniform draws from their range widened by 10% on each side.
- **Sequential Steps ($T = 11$):** Permeation is measured at 11 time points spanning a 30-hour experimental horizon ($\{1, 2, 3, 4, 6, 8, 22, 24, 26, 28, 30\}$ h after application), matching the time-grid stored in `data/ivpt/DATA1.csv`.
- **Proxy vs. Target:** The intermediate proxy y_t is the cumulative drug release at time step t; the target Y is the final trajectory endpoint y_T at $t = 30$ h. Because release dynamics vary drastically between formulations, extrapolating early y_t directly to y_T yields highly inaccurate rankings (Table 1).

C.2 Concrete Compressive Strength

This dataset tracks the curing process of high-performance concrete (Yeh, 1998) and is publicly available from the UCI Machine Learning Repository.

- **Search Space ($d = 7$):** the amounts of 7 continuous ingredients (cement, water, superplasticizer, etc.), searched over the smallest box containing the 127 pool mixtures and 500 uniform draws from their range widened by 5% on each side.
- **Sequential Steps ($T = 11$):** The UCI Concrete dataset stores one $(\text{age}, \text{strength})$ measurement per record rather than a curing trajectory; we group the 1030 records by their seven composition columns (428 distinct mixtures) and keep the 127 mixtures measured at four or more distinct ages with an earliest age of at most 7 days and a latest of at least 56 days, giving one trajectory per mixture at canonical ages $\{3, 7, 10, 14, 21, 28, 42, 56, 70, 90, 180\}$ days. Where a mixture reports several rows at one age, the entry at that age is their mean. Canonical ages a mixture does not report are filled by piecewise-linear interpolation (`np.interp`) over that mixture's own measured rows, with the nearest measured value held flat outside the measured range rather than extrapolated; for the three mixtures with non-identical replicate rows at an age, the interpolation runs over the replicate rows themselves, so a filled age takes a single replicate at each neighbouring age. Of the $127 \times 11 = 1397$ resulting entries, 522 (37.4%) are measured; in particular only 18 of the 127 mixtures report a 180-day strength, so for the other 109 the target Y is that mixture's last measured strength held flat (for two of them a single replicate at the last age, 0.98 and 0.44 MPa below the mean entry there). This yields a $T = 11$-step strength trajectory for each of the 127 real mixtures, and the construction is applied identically to every method.
- **Proxy vs. Target:** The proxy y_t is the partially cured compressive strength at step t. The target Y is the last column of the curve, y_T (matching `run_experiments.py`'s `Y_final_real = Y_all[:,-1]`), which is the measured 180-day strength for the 18 mixtures that report one and the last measured strength held flat for the other 109. Different chemical mixtures cure at different rates (mixtures high in fly ash gain strength slowly initially but reach high strength much later), so early-age strength is a poor proxy for the long-cure target.

C.3 LCBench (AutoML Hyperparameter Optimization)

LCBench (Zimmer et al., 2021) provides learning curves for hyperparameter optimization of neural networks on classification tasks indexed by the OpenML repository (Vanschoren et al., 2014). We use three of the LCBench tasks, identified by their OpenML dataset names `credit-g`, `vehicle`, and `segment`. The curves are taken from the LCBench authors' own `data_2k_lw` release rather than from OpenML. Each of these three tasks holds 2000 configurations in that release; our pool for a task is a fixed random subsample of 200 of them, drawn once per task with `np.random.default_rng(42)` and `rng.choice(2000, size=200, replace=False)` and then sorted, with the 50 recorded training epochs subsampled to 11 checkpoints (every fifth epoch plus the last) so that the trajectory length matches the other benchmarks.

- **Search Space ($d = 7$):** batch size, learning rate, momentum, weight decay, number of layers, maximum units per layer and maximum dropout rate, searched over the smallest box containing the 200 pool configurations and 500 uniform draws from their range widened by 5% on each side.
- **Sequential Steps ($T = 11$):** Each LCBench training trajectory is discretized into 11 epoch checkpoints (subsampled from the original 50-epoch curves for uniformity with IVPT/Concrete).

- **Proxy vs. Target:** The proxy y_t is the validation accuracy at epoch checkpoint t; the target Y is the validation accuracy at the *last* subsampled checkpoint (i.e., y_T, the endpoint of the same trajectory the trial was subsampled from); this matches our pool implementation in `src/run_experiments.py`, not the `final_val_accuracy` field that LCBench stores separately. Some learning curves cross (Table 1).

C.4 Result tables and the three further experimental benchmarks

Table 12: Main results: final simple regret at cost-matched budget $C = 1100$ (20 seeds, mean $\pm$ SE; lower is better; **bold**=best per column; †=every seed ends at the largest per-seed regret of that pool (the method stalls); ‡=stopping rule fires on essentially every trial, so the method degenerates to its warm-start set). SRBO achieves the lowest regret on the two high-mismatch tasks (IVPT, Concrete), significantly against every baseline (Table 4); BOHB leads on the three low-mismatch LCBench tasks, where the freeze-thaw surrogates also place ahead of SRBO (DyHPO on vehicle and segment, ifBO on vehicle). Endpoint costs are reported in Table 9 (Appendix B.3).

Method	IVPT	Concrete	LCBench (cg)	LCBench (veh)	LCBench (seg)
StandardBO	1.11 ± 0.22	35.74 ± 5.94	0.173 ± 0.007	0.555 ± 0.032	0.722 ± 0.014
LinExtrap	$84.98 \pm 3.83^{\ddagger}$	$1229.95^{\dagger}$	$0.444 \pm 0.000^{\ddagger}$	$1.196 \pm 0.008^{\ddagger}$	$1.469 \pm 0.013^{\ddagger}$
EI_Discard	0.99 ± 0.17	$1229.95^{\dagger}$	$0.444 \pm 0.000^{\ddagger}$	$1.196 \pm 0.008^{\ddagger}$	$1.450 \pm 0.014^{\ddagger}$
EI_Virtual	6.04 ± 1.59	70.30 ± 9.14	0.172 ± 0.011	0.589 ± 0.040	0.704 ± 0.031
Freeze-Thaw	$89.10 \pm 1.20^{\ddagger}$	$1229.95^{\dagger}$	$0.444 \pm 0.000^{\ddagger}$	$1.256 \pm 0.000^{\ddagger}$	$1.475 \pm 0.013^{\ddagger}$
ASHA	0.69 ± 0.21	43.29 ± 5.90	0.180 ± 0.008	0.574 ± 0.040	0.697 ± 0.020
HyperbandEI	0.58 ± 0.14	28.61 ± 4.47	0.181 ± 0.004	0.589 ± 0.017	0.746 ± 0.016
BOHB	165.55 ± 1.30	33.80 ± 8.84	$\mathbf{0.060 \pm 0.010}$	$\mathbf{0.046 \pm 0.008}$	$\mathbf{0.039 \pm 0.007}$
ifBO	22.66 ± 1.35	545.47 ± 24.52	0.157 ± 0.009	0.101 ± 0.017	0.482 ± 0.056
DyHPO	6.95 ± 2.42	78.56 ± 12.11	0.116 ± 0.008	0.143 ± 0.028	0.130 ± 0.021
SRBO (ours)	$\mathbf{0.17 \pm 0.03}$	$\mathbf{4.02 \pm 1.02}$	0.107 ± 0.007	0.212 ± 0.012	0.221 ± 0.016

Table 13: Three further experimental benchmarks (same cost-matched protocol as Table 12, $C = 1100$, 20 seeds, mean $\pm$ SE; lower is better; **bold**=best per column). The four stopping-rule controls of Table 12 (LinExtrap, EI_Discard, EI_Virtual, Freeze-Thaw) were not run on these pools and are omitted rather than left blank. Against SRBO the paired two-sided Wilcoxon is significant for every method on OCM (p from 2×10^{-6} to 0.024); on HEA-OER for ASHA, ifBO and DyHPO ($p < 0.007$) but not for BOHB, HyperbandEI or StandardBO ($p = 0.14$, 0.41 and 0.08), whose means and SRBO's lie within 0.004 of one another; and on Catechol for BOHB and DyHPO ($p < 10^{-4}$) and for StandardBO, which is lower than SRBO ($p = 0.02$), but not for ASHA, ifBO or HyperbandEI ($p \geq 0.08$).

Method	OCM	HEA-OER	Catechol
StandardBO	0.112 ± 0.024	0.0062 ± 0.0016	$\mathbf{0.0037 \pm 0.0012}$
ASHA	0.073 ± 0.026	0.0290 ± 0.0053	0.0039 ± 0.0010
HyperbandEI	0.065 ± 0.010	$\mathbf{0.0032 \pm 0.0008}$	0.0052 ± 0.0012
BOHB	0.329 ± 0.077	0.0032 ± 0.0007	0.0175 ± 0.0026
ifBO	1.483 ± 0.041	0.0929 ± 0.0090	0.0042 ± 0.0006
DyHPO	0.966 ± 0.166	1.0426 ± 0.2471	0.0188 ± 0.0027
SRBO (ours)	$\mathbf{0.024 \pm 0.006}$	0.0071 ± 0.0036	0.0044 ± 0.0003

These three pools cover domains the other five do not. In all three the sequential axis is a physical variable that costs more to extend, not a training budget. The OCM pool uses the high-throughput screening data of Nguyen et al. (2020) in the processed form that Ziu et al. (2023) released as `klea-ziu/OCM` under the MIT license; the HEA-OER pool uses the Ni-Pd-Pt-Ru oxygen-evolution measurements of the high-entropy-alloy dataset that Bürgel et al. (2025) deposited on Zenodo under CC BY to accompany Bürgel et al. (2026); and the Catechol pool comes from the Catechol benchmark of Boyne et al. (2025), released as `jpfolch/catechol_solvent_selection` under the MIT license.

- **OCM (heterogeneous catalysis).** Oxidative coupling of methane. Search space $d = 9$: three metal identities, a support identity, three loadings, temperature, and the CH_4:O_2 ratio. The three metal identities and the support enter as integer codes and are searched as continuous coordinates; the smoothed oracle's optimum, a C_2 yield of 13.46, lies at codes outside the pool, against a best measured yield of 18.19. The sequential axis is the contact time (residence time), which the published screen ran at nominal total flows of 20, 15 and 10 mL/min. Grouping the published measurements by catalyst, support, loadings, temperature and CH_4:O_2 ratio, and keeping the groups with at least three distinct recorded total flows, gives a grid of 1000 compositions. The screen's CH_4+O_2 partial-pressure levels are pooled within each recorded flow, so each native point is the mean of one to three measurements; runs recorded at 15.1 or 20.1 mL/min enter as points separate from those at 15 or 20 mL/min, so 68 of the 150 pool compositions carry four native points and 3 span only the 15 and 20 mL/min levels. We interpolate each composition's native points, on its own normalized contact-time axis, to $T = 11$ checkpoints (the Kendall τ between the shortest and the longest native point is 0.65 on the full grid), and the target is the C_2 yield at that composition's longest recorded contact time. We use a subsample of 150 so that the pool size matches IVPT, Concrete and HEA-OER (127 to 150) and the cost-matched budget $C = 1100$ remains meaningful. The subsample keeps all six compositions whose endpoint is at least 90% of the grid's best and 144 others, and its $\tau_{T/4}$ is 0.66 against 0.68 on the full grid. It tracks the marginals of the full grid: across the nine factors the largest shift in any level share is 0.073 (temperature) and six of the nine shift by 0.03 or less. The supplement ships the selected indices together with a script that rebuilds the pool from that grid and recomputes this comparison.
- **HEA-OER (electrocatalysis).** Oxygen-evolution activity of high-entropy alloys. Search space $d = 4$: the atomic fractions of Ni, Pd, Pt and Ru, each searched over the smallest interval containing its measured values and 500 uniform draws from that range widened by 5%, with no sum constraint. The 150 pool compositions lie on the quaternary simplex; the GP oracle's optimum lies off it, at fractions summing to 174 at%, and regret on this pool is measured to that point. The published materials library (library ML5 of the Zenodo record) holds 342 scanning-droplet-cell spots; the benchmark uses 150 of them, drawn by keeping the global optimum and stratifying over the ten endpoint deciles, with the draw chosen from a search over 400 seeds and sizes from 148 to 152 for a $\tau_{T/4}$ near 0.66 and at least four spots in the top tenth of the endpoint range (the full library has $\tau_{T/4} = 0.61$). The code of that draw was not kept, so the supplement ships the list of selected spots. The sequential axis is the current-density sweep, $T = 11$ log-spaced current levels up to 98% of the smallest maximum current over the 342 spots, and the target is the overpotential at the last level, entered negated so that the maximization convention of the pipeline minimizes it.
- **Catechol (flow chemistry).** A transient-flow rearrangement, 40 flow ramps. Search space $d = 6$: the Reichardt $E_T(30)$ polarity parameter and the three Kamlet-Taft parameters (α, β, π^*) of the ramp's solvent, the co-solvent fraction and temperature. The ramps that vary the solvent ratio do so at a near-constant residence time and do not span the residence-time grid, so every kept ramp runs a single entry of the benchmark's descriptor table (the pool uses 20 of its pure solvents and its three premixed binaries) and the co-solvent fraction is 0 or 1 throughout the pool. The sequential axis is residence time from 2 to 15 minutes, $T = 11$, so a later checkpoint means holding the reactor longer and therefore costs more, and the target is the Product-2 yield at the final residence time.

All seven methods that were run on these pools use the protocol of Table 12: cost-matched budget $C = 1100$, seeds $\{0, \ldots, 19\}$, tunable methods at the configuration their disjoint five-seed tuning pool selected. The four stopping-rule controls of Table 12 were not run here.

SRBO's runtime false-stop rate on these pools is 0.000 on OCM, 0.005 on HEA-OER and 0.022 on Catechol, below the nominal level of each ($\alpha_{\text{fsr}} = 0.025$) in every case. The mean number of trials carried to completion is 40.2, 96.1 and 37.3 respectively, so the stopping rule is active on OCM and Catechol and nearly inactive on HEA-OER.

The optimum used for the regret of each pool is recomputed per run, and on HEA-OER the recorded optima differ by up to 1.7×10^{-3}, which is a quarter of SRBO's regret there. We therefore reconstruct each run's incumbent exactly from the pair (recorded optimum, recorded regret) and report every

method against one common optimum per pool, so that no cell is advantaged by which optimum its run happened to locate.

C.5 COMPLETED-ONLY REGRET

Tables 12 and 13 credit a trial a method stops early with its oracle endpoint (Appendix F). Table 14 counts only the configurations a method ran to T, whose endpoint it observed; this is the set it can recommend from among its own trials. SRBO's best configuration is a completed trial on 159 of its 160 pool and seed pairs, so its column hardly moves, and the method with the lowest mean regret is the same as in Tables 12 and 13 on every pool. On IVPT, Concrete and OCM SRBO remains significantly lower than every baseline that completes a configuration on more than half of the seeds (paired two-sided Wilcoxon, $p \leq 0.024$). The regret of ifBO and DyHPO rises, by a factor of up to 491, because they advance configurations one step at a time and their incumbent is often an unfinished run, and LinExtrap and Freeze-Thaw, which stop essentially every trial, and EI_Discard, which does so on every pool except IVPT, complete no configuration after the warm start on most seeds of most pools. `regenerate_completed_only.py` regenerates the table from the per-trial logs in `run_logs/` for SRBO, StandardBO, the four stopping-rule controls and ASHA, and from the result files behind Tables 12 and 13 for BOHB, HyperbandEI, ifBO and DyHPO.

Table 14: Completed-only regret (20 seeds, mean final simple regret, $C = 1100$; the runs of Tables 12 and 13): only configurations a method ran to T count toward its incumbent. On OCM, HEA-OER and Catechol regret is measured from one common optimum per pool, as in Table 13. A superscript gives the number of seeds on which the method completed no configuration after the warm start, over which the mean is not taken; "none" marks a method that completed none on half of the seeds or more; "–" marks a method not run on that pool. **Bold** marks the lowest value in each column.

Method	IVPT	Concrete	LCB (cg)	LCB (veh)	LCB (seg)	OCM	HEA-OER	Catechol
StandardBO	1.11	35.74	0.173	0.555	0.722	0.112	0.0062	**0.0037**
LinExtrap	none	none	none	1.196	none	–	–	–
EI_Discard	0.99	none	none	1.196	none	–	–	–
EI_Virtual	6.04	70.30	0.172	0.589	0.704	–	–	–
Freeze-Thaw	none	none	none	none	none	–	–	–
ASHA	0.69	43.29	0.180	0.574	0.697	0.073	0.0290	0.0039
HyperbandEI	0.58	28.61	0.185	0.595	0.748	0.065	**0.0043**	0.0052
BOHB	166.43	34.71	**0.071**	**0.046**	**0.041**	0.372	0.0077	0.0229
ifBO	23.31	545.47	0.168	0.136	0.517	1.800	45.6578	none
DyHPO	52.11	167.32	0.128^{7}	0.176	0.141^{2}	1.110^{4}	none	none
SRBO (ours)	**0.17**	**4.02**	0.107	0.212	0.221	**0.024**	0.0094	0.0044

D LIMITATIONS IN FULL

Empirical scope. Our empirical claim is conditional rather than universal. SRBO targets the measurable regime in which early observations do not predict the endpoint reliably (normalized linear-fit RMSE of 0.49 or above, Table 1) and falsely eliminating an eventual winner is costly. It accordingly wins on three of the five experimental benchmarks (IVPT, Concrete, OCM), shows no significant difference from the leading methods on HEA-OER, is edged by StandardBO on Catechol, and is outperformed on the three LCBench tasks by BOHB throughout, and additionally by DyHPO on vehicle and segment and by ifBO on vehicle. Normalized linear-fit RMSE places the three LCBench tasks below every experimental pool, so on these eight pools the boundary with BOHB's regime falls in the gap between 0.34 and 0.49, a gap read off after the results were known; the diagnostics do not, however, separate the two experimental pools on which SRBO does not win from the three on which it does, and we have no diagnostic that does. Prospective validation in further experimental systems remains future work, as does evaluation on larger deep-learning learning-curve suites; our only neural-network training curves are the three LCBench tasks, all in the low-mismatch regime. *Calibration scale.* Theorem 1 bounds the CV calibration-equivalent rule under the per-epoch Bernoulli condition, Assumption 1 (A1); the concentration tail $\varepsilon_{n_{+,m}}$ is loose

at our $n_{+,m} \in [10, 75]$, and runtime FSR transfer is not bounded analytically; it rests on (A2) and on the empirical audit (Section 5.3). *Forecaster reliance.* SRBO depends on the DF-GP, which can overfit when few completed trajectories are available; two safeguards keep that failure conservative. CCPST's reliability weighting (Section 4.2) makes STOP harder when the DF-GP's held-out R^2 is poor, and below the promising-calibration floor of Appendix E.4 safe-time pruning selects the empty set, so nothing stops. A forecast whose unreliability shows in its held-out R^2 or in too few promising completions therefore costs forgone savings rather than falsely rejected endpoint winners, and SRBO degrades toward full-budget StandardBO. Neither safeguard catches a forecast that fits the calibration pool but misjudges the configurations BO goes on to select; under the no-DFGP ablation on IVPT the same stopping rule exceeds its nominal FSR (0.12 against 0.10, Table 3). These safeguards act on stopping only. The pseudo-label of a stopped trial is the DF-GP posterior mean, so a biased forecast shifts the augmented surrogate near that configuration, and no result bounds the effect on later selections; every pseudo-label is recomputed from the current DF-GP at each refit (Section 4), so an early error is not frozen in. *Offline replay.* All experiments are conducted on pre-collected pools rather than live experiments; the cost-matched protocol approximates online deployment but does not reproduce wall-clock or queue dynamics. Every method searches the same continuous box against the same oracle fitted to the pool (Appendix F) rather than looking up recorded points, and regret is measured against that oracle's optimum over the box, which can lie outside the measured range. On IVPT, Concrete and the three LCBench tasks, whose oracles interpolate the pool exactly (on IVPT, its 28 formulation means), that optimum is 504.16 on IVPT against a best formulation mean of 357.29 and a best single replicate of 434.98, 1282.51 against a largest measured endpoint of 82.60 on Concrete, and 1.05, 1.89 and 1.96 against largest measured endpoint accuracies of 0.74, 0.80 and 0.87 on LCBench (cg), (veh) and (seg). On Concrete and LCBench it lies on the boundary of the box, with five or six of the seven coordinates at a bound and therefore outside the measured ranges. Apart from LinExtrap, EI_Discard and Freeze-Thaw and from BOHB on IVPT, whose incumbent stays below the best formulation mean on every seed, every method's incumbent lies above the largest measured endpoint on every seed of these five pools, so regret on them is on the oracle's scale rather than in physical units, and the headline ratios on IVPT and Concrete are ratios of such distances. On the smoothed oracles the optimum is a C_2 yield of 13.46 on OCM against a best measured 18.19, at identity codes outside the pool; an overpotential of 426.8 mV on HEA-OER against a lowest measured 437.7 mV, at a composition off the simplex; and a yield of 0.448 on Catechol against a best measured 0.445, at an interior point of the box whose co-solvent fraction (1.004) exceeds one and whose Kamlet-Taft β (-0.038) is negative.

E THEORETICAL ANALYSIS: PROOF OF THEOREM 1

E.1 FORMAL SETUP

We restate the setup of Section 4.2 in precise notation, indexed by recalibration epoch m. Let $\mathcal{D}^{(m)}_{\text{complete}} = \{(x_i, \mathbf{y}_{1:T}(x_i), Y(x_i))\}$ be the pool of fully-evaluated points at the start of epoch m, $Y^{(m)}_{\text{th}}$ the corresponding 75th-percentile threshold, and $\mathcal{D}^{+(m)}_{\text{cal}} = \{i : Y(x_i) \geq Y^{(m)}_{\text{th}}\}$ its promising subset of cardinality $n_{+,m}$. For every $i \in \mathcal{D}^{+(m)}_{\text{cal}}$ let $\ell^{(m)}_t(x_i)$ be the CCPST per-step log-odds of (4) (using $Y^{(m)}_{\text{th}}$), and let $\{\log B^{(m)}_t, \log A^{(m)}_t\}$ be the R^2-weighted thresholds of (6) computed within epoch m.

Candidate safe-time sets. The monotone front-stripping step produces the nested family of candidate safe-time sets

$$\mathcal{T}^{(m)}_1 = \{3, \ldots, T-1\} \supset \mathcal{T}^{(m)}_2 = \{4, \ldots, T-1\} \supset \ldots \supset \mathcal{T}^{(m)}_{K_m} = \{T-1\}, \tag{8}$$

with $K_m = T-3$ candidates indexed by $j \in \{1, \ldots, K_m\}$ (and the trivial $\mathcal{T}^{(m)}_{K_m+1} = \emptyset$ at which no early stopping occurs).

Per-candidate stopping indicators. For a promising trial $i \in \mathcal{D}^{+(m)}_{\text{cal}}$ and candidate j, define two binary indicators that differ in whether the underlying posterior comes from 5-fold CV out-of-fold or

from the full DF-GP:

$$Z_{i,j}^{\mathrm{CV},(m)} \;:=\; \mathbf{1}\Big[\,\exists\, t\in\mathcal{T}_j^{(m)} : \ell_t^{\mathrm{CV},(m)}(x_i) \le \log A_t^{(m)}\,\Big], \tag{9}$$

$$Z^{\mathrm{run},(m)}(x) \;:=\; \mathbf{1}\Big[\,\exists\, t\in\mathcal{T}_{\widehat{j}_m}^{(m)} : \ell_t^{\mathrm{run},(m)}(x) \le \log A_t^{(m)}\,\Big] \quad \text{(deployed runtime, on novel } x\text{)}. \tag{10}$$

Here $\ell_t^{\mathrm{CV},(m)}$ uses the 5-fold CV out-of-fold posterior mean (computed on a DF-GP refit on $4/5$ of $\mathcal{D}_{\mathrm{complete}}^{(m)}$ disjoint from i's slice), the full-DF-GP predictive standard deviation, and the calibration-side standard-deviation floor $0.1\cdot\mathrm{std}(Y)$; $\ell_t^{\mathrm{run},(m)}$ uses the full-data DF-GP posterior mean and predictive standard deviation, the latter floored only at the numerical value 10^{-10}.

Per-candidate population and sample FSR. For each candidate j, the CV-population false-stopping rate is

$$p_j^{\mathrm{CV},(m)} \;:=\; \mathbb{P}\Big[Z_{(i^*),j}^{\mathrm{CV},(m)}=1 \,\Big|\, H_1^{(m)}\Big] \tag{11}$$

(where i^* is a fresh promising trajectory drawn from the same population as $\mathcal{D}_{\mathrm{cal}}^{+(m)}$; for a fresh trajectory the CV rule scores it with one of the five fold models chosen uniformly at random, so $p_j^{\mathrm{CV},(m)}$ is the fold-averaged false-stop probability, and conditional on the split (A1) treats the $n_{+,m}$ held-out indicators as draws from this rule), and its $n_{+,m}$-sample estimator is

$$\widehat{p}_j^{\mathrm{CV},(m)} \;:=\; \frac{1}{n_{+,m}} \sum_{i\in\mathcal{D}_{\mathrm{cal}}^{+(m)}} Z_{i,j}^{\mathrm{CV},(m)}. \tag{12}$$

The safe-time pruning step selects $\widehat{j}_m = \min\{j : \widehat{p}_j^{\mathrm{CV},(m)} \le \alpha_{\mathrm{fsr}}\}$ if such a nonempty candidate exists, and otherwise selects the empty fallback $\widehat{j}_m = K_m+1$; this is the first candidate in the front-stripping order that meets the budget, hence the largest retained safe-time set. The deployed runtime FSR within epoch m is $p^{\mathrm{run},(m)} := \mathbb{P}[Z^{\mathrm{run},(m)}(x)=1 \mid H_1^{(m)}]$, evaluated on novel BO-selected promising trajectories.

E.2 Proof of Theorem 1

For a single calibration whose safe-time set is fixed in advance ($K_m=1$), the tail of Step 2 reduces to the Hoeffding bound of Xiao et al. (2026, Proposition 1); the union over the K_m candidate sets, which covers the data-dependent choice of $\widehat{j}_m$, the recalibration epochs and the CV-to-runtime slack are what Theorem 1 adds.

Proof. **Step 1: Safe-time pruning enforces $\widehat{p}_{\widehat{j}_m}^{\mathrm{CV},(m)} \le \alpha_{\mathrm{fsr}}$ by construction.** Starting from $j=1$ (the largest set $\mathcal{T}_1^{(m)} = \{3,\dots,T-1\}$), the pruning step evaluates (9) on every held-out CV fold and computes the cross-validated $\widehat{p}_j^{\mathrm{CV},(m)}$. If $\widehat{p}_j^{\mathrm{CV},(m)} > \alpha_{\mathrm{fsr}}$ the earliest remaining time index is stripped (incrementing $j\to j+1$); the loop terminates at the first $\widehat{j}_m$ with $\widehat{p}_{\widehat{j}_m}^{\mathrm{CV},(m)} \le \alpha_{\mathrm{fsr}}$ (equivalently, the largest retained safe-time set that meets the budget), or at $\mathcal{T}_{K_m+1}^{(m)} = \emptyset$, in which case we set $\widehat{p}_{K_m+1}^{\mathrm{CV},(m)} := 0$ by convention (no trial can be stopped, so the false-stop rate is 0 vacuously). Hence after pruning,

$$\widehat{p}_{\widehat{j}_m}^{\mathrm{CV},(m)} \;\le\; \alpha_{\mathrm{fsr}}. \tag{13}$$

Step 2: Hoeffding concentration, simultaneous over candidates. Under Assumption 1 (A1), for each fixed j the indicators $\{Z_{i,j}^{\mathrm{CV},(m)}\}_{i\in\mathcal{D}_{\mathrm{cal}}^{+(m)}}$ are i.i.d. Bernoulli$(p_j^{\mathrm{CV},(m)})$. Hoeffding's inequality applied to bounded $[0,1]$-valued indicators gives, for any fixed j,

$$\mathbb{P}\Big[\,p_j^{\mathrm{CV},(m)} > \widehat{p}_j^{\mathrm{CV},(m)} + \varepsilon\,\Big] \;\le\; e^{-2n_{+,m}\varepsilon^2}, \qquad \forall \varepsilon>0. \tag{14}$$

Setting $\varepsilon = \varepsilon_{n_{+,m}} := \sqrt{\log(K_m/\delta_m)/(2n_{+,m})}$ and union-bounding over the K_m candidates yields

$$\mathbb{P}\Big[\,\exists j\in\{1,\dots,K_m\} :\, p_j^{\mathrm{CV},(m)} > \widehat{p}_j^{\mathrm{CV},(m)} + \varepsilon_{n_{+,m}}\,\Big] \le K_m\, e^{-2n_{+,m}\varepsilon_{n_{+,m}}^2} = \delta_m, \tag{15}$$

so with probability at least $1-\delta_m$ the bound $p_j^{\mathrm{CV},(m)} \le \widehat{p}_j^{\mathrm{CV},(m)} + \varepsilon_{n_{+,m}}$ holds *simultaneously for all* j.

Step 3: combine within epoch. On the high-probability event of Step 2, applying (15) at $j=\widehat{j}_m$ and chaining with (13):

$$p_{\widehat{j}_m}^{\mathrm{CV},(m)} \;\le\; \widehat{p}_{\widehat{j}_m}^{\mathrm{CV},(m)} + \varepsilon_{n_{+,m}} \;\le\; \alpha_{\mathrm{fsr}} + \varepsilon_{n_{+,m}}, \tag{16}$$

with probability at least $1-\delta_m$ over the epoch-m calibration draw. The validity at $j=\widehat{j}_m$ is the key reason we union-bound over *all* j rather than only the (data-dependent) selected one; this absorbs the post-selection bias from safe-time pruning.

Step 4: Cross-epoch uniform bound (Bonferroni). The number of recalibration epochs M in a deployed run is data-dependent, because early stopping changes how many completions the budget buys, so the union bound is taken over a deterministic upper bound instead. Every recalibration requires at least 5 new completions (cost at least $5T$) or 25 iterations of at least three proxy steps each (cost at least 75), so a run of cost budget C triggers at most $\lfloor C/\min(5T,75)\rfloor$ recalibrations and, counting the initial calibration of Algorithm 1, deploys at most $M_{\max}=1+\lfloor C/\min(5T,75)\rfloor$ epochs (21 for $C=1100$, $T=11$, where $5T=55<75$). To obtain a bound that holds for *all* epochs simultaneously with confidence at least $1-\delta$, set $\delta_m=\delta/M_{\max}$ in (16) for every m and union-bound over the epochs $m\in\{0,\dots,M_{\max}-1\}$: with probability at least $1-\delta$,

$$\forall m\in\{0,\dots,M_{\max}-1\}:\quad p_{\widehat{j}_m}^{\mathrm{CV},(m)} \;\le\; \alpha_{\mathrm{fsr}} + \sqrt{\tfrac{\log(K_m M_{\max}/\delta)}{2n_{+,m}}}, \tag{17}$$

adding $\log M_{\max}$ inside the square root of the per-epoch tail. In our cost-matched runs ($M\approx 16$ epochs on average across the eight benchmarks against $M_{\max}=21$, $K_m=8$, $\delta=0.05$, $n_{+,m}\approx 50$ on average, equivalently $\delta_m=\delta/M_{\max}=0.0024$; `calib_scale.py`), the cross-epoch tail evaluates to ≈ 0.28, so the uniform bound is $\alpha_{\mathrm{fsr}}+0.28$. The empirical runtime FSR remains well below this (Section 5.3).

Step 5: Transfer to the deployed runtime rule. The remaining gap is between the CV calibration-equivalent rule (covered by Equation (17)) and the rule actually deployed at runtime on novel BO-selected trials. The runtime FSR $p^{\mathrm{run},(m)}$ uses the indicator $Z^{\mathrm{run},(m)}$ (10), which differs from $Z^{\mathrm{CV},(m)}_{(\cdot,\widehat{j}_m)}$ in two respects: (i) the posterior mean comes from the *full* DF-GP fit on $\mathcal{D}^{(m)}_{\mathrm{complete}}$ rather than 5-fold CV out-of-fold; (ii) the predictive standard-deviation floor is 10^{-10} rather than $0.1\cdot\mathrm{std}(Y)$. The full-fit posterior mean can shift the per-step log-odds ℓ_t in either direction relative to the CV mean, and the smaller standard-deviation floor sharpens the decision: for good-but-underpredicted points it can increase stopping risk, which is why no analytic CV-to-runtime transfer is claimed. Define

$$\Delta_m \;:=\; \max\Big\{0,\; p^{\mathrm{run},(m)} - p_{\widehat{j}_m}^{\mathrm{CV},(m)}\Big\}. \tag{18}$$

We make no analytic claim about the sign or magnitude of Δ_m in general; closeness ($\Delta_m\approx 0$) is *not* provable distribution-free. Assumption 1 (A2) speaks only to the population shift between calibration and BO-selected trajectories; the full-fit posterior mean and the smaller floor are algorithmic differences between the two rules that a representativeness assumption on trajectories does not cover, so that part of Δ_m is checked only by the empirical audit. Section 5.3 reports the empirical runtime FSR pooled over the epochs of each run and averaged over the 20 reporting seeds, not $p^{\mathrm{run},(m)}$ epoch by epoch; on every benchmark this mean is below the per-dataset nominal α_{fsr} for SRBO, suggesting Δ_m is small on average on these tasks. Individual runs can exceed the nominal level where promising trials are few: on Catechol, with 24 to 37 promising trials per run, a single false stop gives 0.027 to 0.042 against the nominal 0.025, and 11 of the 20 runs exceed it. When the selection policy changes, Δ_m can be positive. The no-DFGP ablation, whose BO evaluates a different population under the same stopping rule, reaches a runtime FSR of 0.12 on IVPT against the nominal 0.10 that bounds its CV estimate (Table 3), which is consistent with a positive Δ_m. □

Remark on why R^2 weighting needs no correctness assumption in the proof. The R^2-weighted stop threshold $\log A_t^{(m)}$ only enters through the per-candidate indicators $Z_{i,j}^{\mathrm{CV},(m)}$ (9) and the runtime indicator $Z^{\mathrm{run},(m)}$ (10), both deterministic functionals of the DF-GP posterior and the clipped reliability weights $\rho_t^{(m)}$ estimated on held-out folds. The safe-time budget (13) is enforced against the *calibration-equivalent* CCPST score plus reliability-weighted threshold rule, so the Hoeffding

bound (14) does not require $\rho_t^{(m)}$ to recover the true DF-GP accuracy. Under-estimated $\rho_t^{(m)}$ makes the deployed rule more conservative, not less; over-estimated $\rho_t^{(m)}$ is absorbed into $p_j^{\mathrm{CV},(m)}$ for every candidate j, which the simultaneous-over-j Hoeffding (Step 2) controls and safe-time pruning then bounds at $j=\widehat{j}_m$ to α_{fsr} via $\widehat{p}_{\widehat{j}_m}^{\mathrm{CV},(m)}$. This is why the first two terms of Theorem 1 are distribution-free in the DF-GP forecast, with no correctness assumption on $\rho_t^{(m)}$, rather than requiring weighted exchangeability of a non-conformity score. The CV-to-runtime transfer (Step 5 of the proof) is the only place where DF-GP accuracy enters *indirectly*, via the residual Δ_m between $p_{\widehat{j}_m}^{\mathrm{CV},(m)}$ and $p^{\mathrm{run},(m)}$; this gap is empirically small for SRBO (the 20-seed mean runtime $\widehat{\mathrm{FSR}}$ stays at or below the per-dataset nominal α_{fsr} on every benchmark; Section 5.3), while under the no-DFGP ablation on IVPT the runtime rate exceeds the pruning budget (Step 5).

E.3 Empirical-Bernstein Refinement and Conservatism

Corollary 1 (Empirical-Bernstein refinement of Theorem 1)**.** *Under Assumption 1 (A1) and the same calibration-equivalent rule as Theorem 1, fix epoch m and the selected safe-time candidate $\widehat{j}_m$. With probability at least $1-\delta_m$ over the epoch-m calibration draw,*

$$p_{\widehat{j}_m}^{CV,(m)} \;\le\; \alpha_{fsr} \;+\; \underbrace{\sqrt{\frac{2\,\widehat{V}_{n_{+,m}}\,\log(2K_m/\delta_m)}{n_{+,m}}}}_{\textit{variance-adaptive tail}} \;+\; \underbrace{\frac{7\,\log(2K_m/\delta_m)}{3(n_{+,m}-1)}}_{\textit{bias term}}, \tag{19}$$

where $\widehat{V}_{n_{+,m}} = \frac{n_{+,m}}{n_{+,m}-1}\,\widehat{p}_{\widehat{j}_m}^{CV,(m)}(1-\widehat{p}_{\widehat{j}_m}^{CV,(m)})$ is the sample variance of the per-trajectory CV stop indicators at the selected candidate. If the empty fallback $\mathcal{T}_{K_m+1}^{(m)}=\emptyset$ is selected, the bound holds vacuously with $p_{K_m+1}^{CV,(m)}:=0$. The cross-epoch uniform version follows by Bonferroni ($\delta_m=\delta/M_{\max}$, Step 4 of the proof), and any runtime-rule transfer inherits the same caveat as Step 5 of the Theorem 1 proof (the residual Δ_m, audited empirically in Section 5.3).

Residual conservatism in Theorem 1. Three sources of looseness remain. First, the union bound over $K_m \le T-3$ nested candidate safe-time sets adds $\log K_m$ inside the square root of $\varepsilon_{n_{+,m}}$; for $K_m=8$ and $\delta_m=0.05$ this raises $\log(1/\delta_m)\approx 3.0$ to $\log(K_m/\delta_m)\approx 5.1$, a factor of about 1.3 on the tail, modest. Second, the Hoeffding tail treats per-trial indicators as worst-case Bernoulli, ignoring the variance reduction available when $\widehat{p}_{(m)}^{\mathrm{CV}}$ is small; Corollary 1 substitutes the variance-adaptive tail and tightens this in the regime $\widehat{p}_{(m)}^{\mathrm{CV}} \ll \alpha_{\mathrm{fsr}}$. Third, the cross-epoch Bonferroni adds $\log M_{\max}$ inside the same square root; for $M_{\max}=21$ this raises $\log(K_m/\delta)\approx 5.1$ to $\log(K_m M_{\max}/\delta)\approx 8.1$, a factor of about 1.26 on the tail, also modest.

Proof of Corollary 1. Replace the Hoeffding inequality in Step 2 of the Theorem 1 proof with the empirical-Bernstein inequality of Maurer & Pontil (2009, Theorem 4). For i.i.d. $X_1,\dots,X_n\in[0,1]$ with mean μ and sample variance $\widehat{V}_n = \frac{1}{n(n-1)}\sum_{i<j}(X_i-X_j)^2$, with probability at least $1-\delta$,

$$\mu \;\le\; \bar{X} \;+\; \sqrt{\frac{2\,\widehat{V}_n\log(2/\delta)}{n}} \;+\; \frac{7\log(2/\delta)}{3(n-1)}. \tag{20}$$

For each fixed candidate $j\in\{1,\dots,K_m\}$ apply this to $\{Z_{i,j}^{\mathrm{CV},(m)}\}_{i=1}^{n_{+,m}}$ at $\delta=\delta_m/K_m$ (per-candidate confidence) to obtain $p_j^{\mathrm{CV},(m)} \le \widehat{p}_j^{\mathrm{CV},(m)} + \sqrt{2\widehat{V}_{n_{+,m},j}\log(2K_m/\delta_m)/n_{+,m}} + 7\log(2K_m/\delta_m)/(3(n_{+,m}-1))$ with probability $\ge 1-\delta_m/K_m$, where $\widehat{V}_{n_{+,m},j} = \frac{n_{+,m}}{n_{+,m}-1}\widehat{p}_j^{\mathrm{CV},(m)}(1-\widehat{p}_j^{\mathrm{CV},(m)})$. Union-bounding over j makes this hold simultaneously for all candidates with probability $\ge 1-\delta_m$, hence in particular at $j=\widehat{j}_m$. Combining with the pruning condition $\widehat{p}_{\widehat{j}_m}^{\mathrm{CV},(m)} \le \alpha_{\mathrm{fsr}}$ yields Equation (19).

Hoeffding vs. empirical-Bernstein cross-over. At the pruning boundary $\widehat{p}^{\mathrm{CV},(m)}_{\widehat{j}_m} = \alpha_{\mathrm{fsr}}$, the two tails of Theorem 1 and Corollary 1 evaluate (taking $\delta_m = \alpha_{\mathrm{fsr}}$ for parity) to

$$\varepsilon^{\mathrm{Hoef}}_{n_{+,m}} = \sqrt{\frac{\log(K_m/\alpha_{\mathrm{fsr}})}{2n_{+,m}}}, \tag{21}$$

$$\varepsilon^{\mathrm{EB}}_{n_{+,m}} = \sqrt{\frac{2\alpha_{\mathrm{fsr}}(1-\alpha_{\mathrm{fsr}})\,\log(2K_m/\alpha_{\mathrm{fsr}})}{n_{+,m}-1}} + \frac{7\log(2K_m/\alpha_{\mathrm{fsr}})}{3(n_{+,m}-1)}. \tag{22}$$

Setting $\varepsilon^{\mathrm{Hoef}}_{n_{+,m}} = \varepsilon^{\mathrm{EB}}_{n_{+,m}}$ and solving numerically at $\alpha_{\mathrm{fsr}} = 0.05$, $K_m = 8$ yields the cross-over $n^\star_{+,m} \approx 252$. For $n_{+,m} < n^\star_{+,m}$ the Hoeffding tail of Theorem 1 is tighter because the $O(1/n_{+,m})$ bias term of empirical-Bernstein dominates; for $n_{+,m} \geq n^\star_{+,m}$ the variance-adaptive square-root term wins. In our experiments (Section 5.3) the 20-seed mean empirical *runtime* FSR stays below the per-dataset nominal α_{fsr} on every benchmark (closest on Catechol, 0.022 against 0.025), so the looseness of either theoretical bound is dominated by the empirical evidence.

E.4 Practical Requirements on $n_{+,m}$

For either bound to be *tight* (meaning the concentration tail is no larger than the calibration budget α_{fsr}, so the total stays below $2\alpha_{\mathrm{fsr}}$), we can solve $\varepsilon_{n_{+,m}} \leq \alpha_{\mathrm{fsr}}$ in closed form (taking $\delta_m = \alpha_{\mathrm{fsr}}$ for parity):

- **Hoeffding** (Theorem 1): $n_{+,m} \geq \log(K_m/\alpha_{\mathrm{fsr}})/(2\alpha^2_{\mathrm{fsr}})$. At $\alpha_{\mathrm{fsr}} = 0.05$, $K_m = 8$, $n_{+,m} \geq 1016$.
- **Empirical-Bernstein** (Corollary 1): at $\widehat{p}^{\mathrm{CV}}_{(m)} \leq \alpha_{\mathrm{fsr}}$, so that $\widehat{V}_{n_{+,m}}$ is at most about α_{fsr}, the tail is $\sqrt{2\alpha_{\mathrm{fsr}}\log(2K_m/\alpha_{\mathrm{fsr}})/n_{+,m}} + 7\log(2K_m/\alpha_{\mathrm{fsr}})/(3(n_{+,m}-1))$. The bias term cannot be dropped. At $n_{+,m} = 231$, the size the variance term alone would require, it equals 0.059 and exceeds the budget by itself. Solving the full inequality at $\alpha_{\mathrm{fsr}} = 0.05$, $K_m = 8$ gives $n_{+,m} \geq 660$, roughly $1.5\times$ fewer calibration samples than the Hoeffding requirement.

Both thresholds exceed the BO-sized calibration pools encountered in our benchmarks ($n_{+,m} \in [10, 75]$), so Theorem 1 and Corollary 1 are strictly *composite* statements whose informativeness depends on how much of the α_{fsr} budget is absorbed by the tail. The implementation enforces a minimum $n_{+,m} \geq \texttt{min_good_for_calibration} = 5$ safeguard (CCPST defaults to $\mathcal{T}^{(m)}_\star = \emptyset$, i.e., no early stopping, when fewer than 5 promising calibration trajectories are available), at which the bounds are loose but the 20-seed mean *empirical runtime* FSR we report in Section 5.3 is still at or below the per-dataset nominal α_{fsr} on every benchmark.

Further caveats.

- *Theorem conservatism (other axes).* The within-epoch union bound over $K_m \leq T-3$ candidate safe-time sets ignores positive correlation between adjacent prunings; the cross-epoch Bonferroni over $M_{\max}$ epochs ignores temporal dependence between consecutive recalibrations. The guarantee also requires the held-out CV indicators to be i.i.d. Bernoulli within each epoch (Assumption 1 (A1)), which holds only approximately here because the 5 folds share training data and because the threshold, reliability weights and predictive standard deviation are fitted on the whole epoch pool, and holds less well when BO's acquisition strongly reshapes the calibration pool; $\varepsilon_{n_{+,m}}$ then indicates the scale of the sampling error rather than a guaranteed bound on it. The CV-to-runtime transfer (Step 5 of the proof) introduces a residual Δ_m between $p^{\mathrm{CV},(m)}$ and $p^{\mathrm{run},(m)}$ from the in-sample full-DF-GP mean and the smaller standard-deviation floor; this residual is not bounded analytically and is empirically audited in Section 5.3.
- *Experimental scope.* All benchmarks are offline pool-based simulators rather than live experimental loops; deployment on real pharmaceutical or materials platforms will additionally face wet-lab measurement noise and non-stationary protocol drift that our simulators do not model.
- *Uninformative proxies.* If neither x nor $\mathbf{y}_{1:t}$ predicts Y on the held-out folds, CCPST's clipped reliability weight approaches the floor $\rho_{\min}$, the stopping threshold becomes highly conservative, and SRBO reverts toward StandardBO. If only $\mathbf{y}_{1:t}$ is uninformative, ρ_t instead measures how well the DF-GP predicts Y from x, and stopping rests on x rather than on the observed proxies; meta-learned priors and proxy-feature selection are natural extensions we leave to future work.

F Implementation Details

Surrogate and DF-GP. The two GPs differ in both inputs and kernel. The exact-GP surrogate takes the configuration x alone and uses an isotropic RBF kernel with a constant signal variance (`ConstantKernel`$\times$`RBF`), its length-scale bounded below at 0.1 in standardized input units: without the bound the marginal-likelihood fit can reach the optimizer floor of 10^{-5}, at which the posterior equals the prior at every unobserved point and EI is constant over the box. The same surrogate class, with the same bound, serves every EI-based method except DyHPO, whose deep-kernel GP is described in the ifBO and DyHPO paragraph below. It has no `WhiteKernel`, its observation noise being passed in as the fixed variance $\sigma^2 = 0.01$ on the standardized target (the surrogate z-scores Y before fitting, so this is 1% of the target variance, and $\sigma^2_{\text{aug}} = 0.03$ at the default $\gamma = 3$) rather than learned (see the soft-rejection noise paragraph below). The DF-GP uses `ConstantKernel`$\times$`RBF`$+$ `ConstantKernel`$\times$`DotProduct`$+$`WhiteKernel`, whose linear component captures monotone target trends (a global trend term, not a monotonicity constraint on the forecast) and whose `WhiteKernel` noise level is learned. Both are implemented on top of `sklearn.gaussian_process` with constant prior mean, Gaussian likelihood, and standard Type-II maximum-likelihood (MLE) kernel-hyperparameter fitting (no explicit hyperprior). The DF-GP takes the concatenated input $[x; \mathbf{y}_{1:t}]$ of dimension $d + t$, with x standardized feature-wise and each y_s standardized by the mean and standard deviation of checkpoint s in the DF-GP's current training set (after a $\log(1 + y)$ transform on IVPT), so the isotropic RBF acts on comparably scaled coordinates, and is re-fitted every 5 outer completions on the current $\mathcal{D}_{\text{complete}}$, a deliberate amortization of GP fitting cost over groups of new exact observations. A second trigger forces a refit after 25 iterations without enough completions, which folds in the completions that arrived since the last refit when the censoring rate is high. Each refit runs three restarts of the L-BFGS hyperparameter search against the five used for the initial fit.

DF-GP diagnostics. Three diagnostics on IVPT and Concrete test whether the DF-GP forecast reduces to ranking by the current value y_t. The first two use five-fold out-of-fold predictions of the $t = T/4$ model (the first three checkpoints), with folds split by pool row as in the calibration cross-validation, so on IVPT a held-out trajectory can have replicates of its formulation in the training folds. (a) Raising a held-out trajectory's whole prefix $\mathbf{y}_{1:t}$ by 0.25 standard deviations of the training-fold prefix values lowers the forecast for 7 of the 135 IVPT trajectories (5.2%) and 2 of the 127 Concrete configurations (1.6%), which no forecaster that is a non-decreasing function of y_t can do. (b) Refitting the same kernel and preprocessing on three input sets (Table 15) locates the ability to reorder crossing trajectories in the conditioning on x. From y_t alone the forecast repairs 6.9% and 3.1% of the pairs that y_t misorders, adding x raises this to 28.6% and 57.0%, and the rest of the prefix changes it by $+5.9$ and -1.6 points; on IVPT the full DF-GP inputs repair the most pairs without raising the rank agreement with Y. (c) Replacing the DF-GP inside SRBO by per-prefix isotonic regression of Y on y_t alone, with a constant predictive standard deviation equal to its in-sample residual standard deviation and the calibration, pseudo-observations and recalibration otherwise unchanged, raises the final regret from 4.02 to 7.97 on Concrete (worse on 16 of the 20 seeds; two-sided Wilcoxon $p = 4.2 \times 10^{-3}$), while on IVPT the two do not differ significantly (0.18 against 0.21; worse on 7, better on 8, tied on 5; $p = 0.33$), paired with the full SRBO row of Table 2 at the same configuration and seeds. On IVPT, then, the current value carries most of what the stopping rule needs, and the DF-GP's contribution shows in the pairwise diagnostics above rather than in final regret.

Table 15: Feature ablation of the DF-GP forecast at $t = T/4$ (five-fold out-of-fold predictions, the same kernel and preprocessing for every input set). Repair is the share of the pairs that y_t misorders relative to Y (1900 on IVPT and 1249 on Concrete, ties excluded) that the forecast orders correctly; Kendall τ is between forecast and Y, against 0.580 (IVPT) and 0.688 (Concrete) for ranking by y_t itself; RMSE is in the units of Y.

	IVPT			Concrete		
Inputs	Repair	τ	RMSE	Repair	τ	RMSE
y_t	6.9%	0.576	47.0	3.1%	0.678	6.89
$[x; y_t]$	28.6%	0.593	44.0	57.0%	0.784	5.02
$[x; \mathbf{y}_{1:t}]$ (DF-GP inputs)	34.5%	0.574	41.0	55.4%	0.758	6.39

CCPST hyperparameters. The nominal FSR level α_{fsr} is set per-dataset by tuning over $\{0.025, 0.05, 0.10\}$ on the disjoint 5-seed pool (the tuned values are $0.10, 0.025, 0.05, 0.05, 0.025$ on IVPT/Concrete/LCBench-cg/veh/seg and 0.025 on each of OCM, HEA-OER and Catechol), and the raw SPRT-style lower-boundary shape parameter is $\beta = 0.1$.

Implementation note on the deployed runtime rule. Every SRBO run reported here ran the shipped code, which passes the tuned α_{fsr} to every recalibration epoch; the empirical FSR audit in Section 5.3 measures the deployed rule. At each recalibration epoch m, the threshold $Y_{\text{th}}^{(m)}$ is set to the 75th percentile of the targets $\{Y(x_i)\}_{i \in \mathcal{D}_{\text{complete}}^{(m)}}$ then available, and is *held fixed within the epoch*; epoch-wise drift of $Y_{\text{th}}^{(m)}$ as $\mathcal{D}_{\text{complete}}^{(m)}$ grows is handled by the per-epoch / Bonferroni-uniform formulation of Assumption 1 and Theorem 1 (see Section 4.2 and Appendix E). The three-component stopping rule of Section 4.2 is calibrated by 5-fold cross-validation on the promising subset $\mathcal{D}_{\text{cal}}^{+(m)} = \{i : Y(x_i) \geq Y_{\text{th}}^{(m)}\}$: for each fold the DF-GP is refit on the remaining $4/5$ of $\mathcal{D}_{\text{complete}}^{(m)}$, every held-out promising trajectory is replayed through the score and reliability-weighted threshold components to count how many would be falsely stopped, and safe-time pruning iteratively drops the earliest remaining time point from $\mathcal{T}_\star^{(m)}$ until the cross-validated $\widehat{\text{FSR}}$ is at or below the calibration budget α_{fsr}. The reliability weight $\rho_t^{(m)}$ is obtained by clipping the pooled out-of-fold R^2 estimate across all CV folds as in Equation (5). **Recalibration cadence.** The full 5-fold CV calibration of $\{\log A_t^{(m)}, \log B_t^{(m)}, \mathcal{T}_\star^{(m)}\}$ is rerun whenever the DF-GP is retrained, i.e. every 5 outer completions or after 25 outer iterations without them; per-iteration recalibration would be ideal but would repeat these GP fits 5 to 14 times as often (the reported runs deploy 9 to 20 calibration epochs over 101 to 158 iterations on average), and the empirical FSR audit (Section 5.3, mean $\widehat{\text{FSR}} \leq \alpha_{\text{fsr}}$ on every benchmark) suggests this amortization is empirically sufficient. The implementation uses `min_good_for_calibration=5`: CCPST defaults to $\mathcal{T}_\star^{(m)} = \emptyset$ (no early stopping) whenever $n_{+,m} < 5$, which guards against pathological small-calibration regimes; the calibration pool is the historical pool $\mathcal{D}_{\text{exact}}^{(0)}$ plus the BO completions, and $\mathcal{D}_{\text{exact}}^{(0)}$ alone gives $n_{+,0} \geq 10$ on every benchmark, so the guard never delays the first stop and no campaign trials are spent on warm-up; the magnitude of the Hoeffding tail $\varepsilon_{n_{+,m}} = \sqrt{\log(K_m/\alpha_{\text{fsr}})/(2n_{+,m})}$ at our calibration sizes is discussed in Appendix E.4, while Section 5.3 reports the empirical runtime $\widehat{\text{FSR}}$ against each dataset's nominal α_{fsr}.

Acquisition. The deployed acquisition in all runs reported in Tables 12 and 13 is plain Expected Improvement $\text{EI}(x \mid \mathcal{GP}_{\text{exact}})$ on the DF-GP-augmented exact surrogate, with exploration parameter ξ selected per dataset on the tuning pool described below (default 0.1). The cost budget is enforced by the evaluation loop; it is not used inside the acquisition score. The acquisition is optimized over the continuous normalized search box $[0, 1]^d$, which maps each coordinate to its measured range widened by up to 5% on each side (up to 10% on IVPT; the box is the smallest one containing the pool and 500 uniform draws from the fully widened range): 500 uniformly sampled points are scored and the best 20 are refined by L-BFGS-B within the box. The benchmark pools are not enumerated as candidate sets; they are used to fit the oracle simulator and to warm-start the surrogate. Hyperparameters $(\alpha_{\text{fsr}}, \gamma, \xi)$ are selected from scratch under the deployed EI rule on the disjoint 5-seed tuning pool (seeds 100 to 104, grid $3 \times 3 \times 4 = 36$ configs per dataset), with reporting on the disjoint 20-seed pool (seeds 0 to 19); the tuning and reporting pools never overlap. The cost-aware variant, which is not deployed, is described in Appendix G; the ablations and sensitivity sweeps of Appendix B use the deployed rule.

Regret crediting. The simple regret of every method is computed from the best oracle endpoint among the configurations it credits, and a trial the method stops early is credited with its endpoint, although the method never observes it. SRBO, StandardBO, the four stopping-rule controls and ASHA do not credit a final trial cut short because the budget ran out; HyperbandEI and BOHB credit every configuration a bracket started, including survivors the budget left unpromoted, and ifBO and DyHPO credit a configuration when they select it. The convention favors methods that stop many trials. Algorithm 1 recommends the best completed configuration, and Table 14 (Appendix C.5) scores every method on its completed configurations only.

Baseline definitions. *LinExtrap* is a least-squares linear endpoint-extrapolation control in the spirit of Domhan et al. (2015) rather than an implementation of their method; it halts a trial when the fitted

line predicts an endpoint below the pool's 75th-percentile endpoint. *EI_Discard* applies CCPST-style stopping and discards halted trials, and *EI_Virtual* enters each CCPST-halted trial into the surrogate as an exact observation at the incumbent y_{best}; these correspond to the drop-censored and treat-as-uncensored baselines that Hutter et al. (2013) compare with their imputation. Both controls keep the DF-GP and CCPST calibration fitted on the recorded pool for the whole run, without recalibration, and run untuned at $\alpha_{\text{fsr}} = 0.05$; the SRBO$_{\text{no-DFGP}}$ ablation of Section 5.3, which keeps SRBO's recalibration and tuned configuration, is the comparison for the fate of a stopped trial. *Freeze-Thaw* (Swersky et al., 2014) places a GP over learning curves with the freeze-thaw exponential-decay mixture kernel, adapted here to a terminal upper-confidence halting rule rather than pause/resume.

Baseline tuning grids. The three tunable baselines are tuned on the same disjoint 5-seed pool (seeds 100 to 104) and selected by lowest mean final regret on that pool. ASHA and HyperbandEI sweep the elimination ratio $\eta \in \{2, 3, 4\}$ (3 configurations per dataset); BOHB sweeps $\eta \in \{2, 3, 4\}$ crossed with the TPE quantile $\gamma_{\text{tpe}} \in \{0.10, 0.15, 0.20, 0.25\}$ ($3 \times 4 = 12$ configurations per dataset). The defaults are $\eta = 2$ for all three and $\gamma_{\text{tpe}} = 0.15$ for BOHB; Table 6 contrasts each tunable method's default configuration with the one its tuning pool selected. Every sweep runs under the cost budget on all eight benchmarks through the tuning tasks of `hpc_reported_runs_task.sh`. ASHA's sweep ships as `asha_eta_tune/` and `asha_eta_sweep/`, from which `regenerate_asha_sweep.py` repeats the selection; it selects $\eta = 2$ on Concrete, LCBench (cg), OCM and Catechol, $\eta = 3$ on IVPT, LCBench (seg) and HEA-OER, and $\eta = 4$ on LCBench (veh). The tuning-pool runs of SRBO, BOHB and HyperbandEI ship in `tuning_results/`, and `regenerate_tuning.py` repeats every selection and checks it against the configuration each reported cell ran. Only grid points with all five tuning seeds take part, and exact ties go to the first grid point in enumeration order (for SRBO the smallest α_{fsr}, then γ, then ξ; for ASHA and HyperbandEI the smallest η; for BOHB the smallest η, then γ_{tpe}). The one exact tie is ASHA on Catechol, where $\eta = 2$ and $\eta = 3$ give identical tuning-pool regret on every seed; the rule takes $\eta = 2$, and $\eta = 3$ would report 0.0059 against 0.0039.

Soft-rejection noise. The augmented exact GP $\mathcal{GP}_{\text{exact}}$ is refit with a single noise variance $\sigma^2_{\text{aug}} = \gamma\sigma^2$ shared by exact and pseudo points, a coarsening of strict per-point inflation. The default is $\gamma = 3$. A larger γ smooths the whole augmented surrogate, exact and pseudo points alike; Table 5 (Appendix B) reports the sensitivity to γ. Note that under this shared-noise coarsening, $\gamma \to \infty$ inflates the noise on *both* exact and pseudo points, so the limit does not coincide with the SRBO$_{\text{no-DFGP}}$ ablation (which removes the pseudo stream while preserving exact-point precision); SRBO$_{\text{no-DFGP}}$ is reported as a separate ablation in Section 5.3.

Calibration vs. runtime target. Our default `target_mode=paper_realistic` mode calibrates $Y^{(m)}_{\text{th}}$ and the CV stop indicators on the raw historical replicate endpoints $\{Y_{\text{real}}(x_i)\}$ together with the endpoints of the trials it has run to completion, while runtime y_{true} at each BO query comes from the trained oracle simulator (one thin-plate-spline `RBFInterpolator` per time point with a degree-one polynomial tail, fit on the same historical data after averaging replicate formulations where a pool has them, with smoothing 0 on IVPT, Concrete and LCBench and 10^3 on OCM; HEA-OER and Catechol instead use a Matérn-$5/2$ GP oracle, because on HEA-OER the spline tail extrapolates unphysically off the composition simplex and on Catechol the smoothed spline placed its optimum at a corner of the box, 26% above the largest measured yield, which most methods approached at their first query, so that pool separated them little). On the IVPT benchmark (where each formulation has $\sim$5 noisy replicates), the raw-replicate endpoint and the aggregated-simulator endpoint are ≈ 0.74 Pearson-correlated with MAE ≈ 28.1 in the units of Y; the $Q_{0.75}$ thresholds align closely but individual "good/bad" classifications can disagree at the boundary. We therefore report empirical FSR using the runtime y_{true} definition and treat Theorem 1's CV-population FSR as a calibration-side proxy whose runtime transfer rests on Assumption 1 (A2) for the population shift and on the empirical audit for the rule differences (Appendix E, Step 5). An `oracle_aligned` calibration variant (which uses the same aggregated simulator for both calibration and runtime) is exposed in the supplement for users who prefer to remove this asymmetry; we did not rerun the Table 12 benchmarks under it.

ifBO and DyHPO. Both are run by our own runners (`src/run_ifbo.py`, `src/run_dyhpo.py`), which share the oracle, warm start, seeds and cost accounting of the other methods: each step advances one configuration by one checkpoint, a new configuration entering at the first, so $C = 1100$ is exactly 1100 steps, and each surrogate also receives, at no cost,

the historical pool trajectories on which SRBO's DF-GP is trained. ifBO combines the authors' pretrained FT-PFN (checkpoint v0.0.1 of the `ifbo` package) with the MFPI-random acquisition of Rakotoarison et al. (2024); DyHPO reimplements, in PyTorch and GPyTorch, the deep-kernel GP and multi-fidelity EI of the reference implementation of Wistuba et al. (2022). DyHPO keeps the reference implementation's defaults (Adam learning rate 10^{-3}, dense layers of 64 and 128 units), which differ from the learning rate 0.1 and the 128 and 256 units stated in the implementation details of Wistuba et al. (2022). Both score the unfinished runs together with fresh points of the unit box and a local refinement around the best candidate (the DyHPO reference implementation searches a finite tabular pool instead). Three changes each concern only one of the two methods: DyHPO's curve encoder sees the last three observed values rather than the whole observed prefix, and each DyHPO refit uses at most 800 rows (the last 400 rows in curve order, historical curves first and then configurations by start, and a random subsample of the rest); ifBO drops the oldest completed curves, then the oldest historical ones, from its context when the context would exceed the 1000 curve identities its checkpoint embeds, which happened on two LCBench (cg) seeds and four Catechol seeds. ifBO is deterministic given the seed. DyHPO's runner does not seed PyTorch, unlike the reference implementation, so a rerun reproduces its cells in distribution rather than seed by seed. Both run on CPU in a separate environment (`requirements_modern_baselines.txt`), at a median optimizer time of 127 and 105 minutes per run.

Infrastructure. All runs are 4-core CPU jobs on a shared SLURM cluster (no method uses a GPU), with 20 independent seeds per configuration. Across the eight benchmarks the mean optimizer time ranges from 8.7 to 39.2 seconds per iteration for SRBO and from 0.36 to 2.16 seconds for StandardBO; the difference is the cost of the repeated DF-GP fits and the 5-fold CCPST recalibration. This is overhead in optimizer time and not in trial cost, and it stays negligible beside the hours, days, or weeks that a single real evaluation takes in the target applications. Every run is seeded (apart from DyHPO's PyTorch generator, noted above), and every array task pins its BLAS threads; two runs of one seed can still diverge, so a rerun of a cell reproduces its distribution rather than its per-seed values: the full-SRBO row of Table 2, run separately from the Table 12 cell at the same configuration and seeds, gives 0.21 against 0.17 on IVPT (paired $p = 0.14$), the largest difference of the five pools.

Reproducibility commitment. We submit an anonymized supplement (`srbo_supplement.zip`, about 13 MB) containing: (i) the full SRBO source (DF-GP, CCPST, censored-GP, EI wrapper, plus ASHA, HyperbandEI, and BOHB baselines); (ii) the SLURM array runner `hpc_reported_runs_task.sh`, which produced every reported run of the nine methods implemented here under the deployed budget except the synthetic suite (`slurm/slurm_synth.sh`), the submission script that lists every array at its selected configuration, and the ifBO and DyHPO runners `src/run_ifbo.py` and `src/run_dyhpo.py` with their templates, which need the PyTorch environment of `requirements_modern_baselines.txt`; (iii) the seed list $\{0, \dots, 19\}$ plus the disjoint 5-seed tuning pool $\{100, \dots, 104\}$; (iv) full kernel/prior/optimizer configurations and the complete benchmark data for all eight benchmarks (UCI Concrete, LCBench credit-g/vehicle/segment, anonymized IVPT `data/ivpt/DATA1.csv`, OCM, HEA-OER and Catechol), with builders in `newbench_pools/` and `subsample_ocm.py` that rebuild the OCM, HEA-OER and Catechol pools from the published data, applying the shipped lists of selected rows for the two subsampled pools; (v) *the raw per-seed* `result_*.json` files behind every cell of Table 12, the external-baseline block of Table 4 and the Section 5.3 FSR of the five Table 12 benchmarks, in `table1_results/` (1100 per-seed runs, five pools by eleven methods by 20 seeds); the runs behind Tables 2, 5, 6, 7, 8 and 10, ASHA's η sweep and the isotonic counterfactual of Appendix F in `ablation_results/`, `sensitivity_results/`, `default_results/`, `diag_results/`, `synth_results/`, `costbudget_results/`, `asha_eta_tune/` and `asha_eta_sweep/`, and the tuning-pool runs behind every tuned configuration in `tuning_results/`; the 420 per-seed runs behind Table 13 and the OCM, HEA-OER and Catechol FSR of Section 5.3 ship in `newbench_results/` with the driver `aggregate_newbench.py`, which regenerates that table and asserts each cell against the published value, and the per-trial logs from which Table 14 reads the completed trials of SRBO, StandardBO, the four stopping-rule controls and ASHA ship in `run_logs/`; (vi) one-command drivers, all but the last of which read the raw JSONs without GP refitting: `regenerate_table1.py`, which prints all 55 cells of Table 12 and asserts each cell against the published value, `regenerate_tables.py`, which prints the external-baseline block of Table 4, Tables 7 and 8 and the Section 5.3 empirical FSR on the five benchmarks of Table 12, and `regenerate_ablation.py`, `regenerate_sensitivity.py`, `regenerate_default_vs_tuned.py`, `regenerate_costbudget.py`,

`regenerate_asha_sweep.py` and `regenerate_tuning.py`, which print Tables 2, 3, 5, 6 and 10 together with the ablation block of Table 4, ASHA's η selection and every other tuning-pool selection, and `regenerate_completed_only.py`, which prints and asserts Table 14, together with `calib_scale.py`, which replays the recalibration epochs and calibration-pool sizes quoted in Appendix E, and `regenerate_dfgp_diagnostics.py`, which refits the GPs of the DF-GP diagnostics in Appendix F in about half a minute and asserts their values; and (vii) plotting drivers `make_fig_regret.py` (Figure 2, asserting every panel endpoint against Tables 12 and 13, and with `--endpoint-table` Table 9, asserting each of its 55 cells), `make_fig_fsr.py` (Figure 5, asserting each bar against Table 1), `make_fig_rank.py` and `make_fig_regime.py`, plus the hand-drawn Figure 1 as the file included here (`fig1_srbo.pdf`). Upon acceptance we will replace the anonymized package with a permanent GitHub release (Apache 2.0 for code, CC-BY-4.0 for IVPT data) and publish the IVPT collection protocol with a public DOI.

G Pseudocode and Censored-GP Details

G.1 SRBO Pseudocode

In Algorithm 1 we use $\mathcal{D}_{\text{exact}}$ for the pool of trials with exactly observed targets and $\mathcal{D}_{\text{pseudo}}$ for the pool of stopped trials whose endpoint is replaced by a DF-GP forecast. $\mathcal{D}_{\text{exact}}$ is the same object as $\mathcal{D}_{\text{complete}}$ used in Section 4; we adopt the "exact" label here to contrast with "pseudo".

Complexity. Every BO step that ends in a completion or a soft rejection refits the augmented surrogate at $\mathcal{O}((|\mathcal{D}_{\text{exact}}| + |\mathcal{D}_{\text{pseudo}}|)^3)$, after $|\mathcal{D}_{\text{pseudo}}|$ DF-GP posterior queries of $\mathcal{O}(|\mathcal{D}_{\text{exact}}|^2)$ each. Each recalibration epoch is dominated by retraining the T per-checkpoint DF-GP models on $\mathcal{D}_{\text{exact}}$, once for the deployed forecaster and once inside each of the 5 CV folds, $\mathcal{O}(T\,|\mathcal{D}_{\text{exact}}|^3)$ in total; safe-time pruning then replays at most K_m nested candidate sets over the $n_{+,m}$ held-out promising trajectories at up to K_m time points, $\mathcal{O}(K_m^2\, n_{+,m}\, |\mathcal{D}_{\text{exact}}|^2)$. Each acquisition evaluation costs $\mathcal{O}((|\mathcal{D}_{\text{exact}}| + |\mathcal{D}_{\text{pseudo}}|)^2)$.

G.2 Auxiliary Censored-GP Bookkeeping

The released implementation maintains a secondary censored-response GP, following the censoring treatment of Hutter et al. (2013), that stores upper-censored bounds ("$Y(x) < Y_{\text{th}}^{(m)}$") for every stopped trial, retained for diagnostic transparency. The Expected Improvement value that ranks candidates is read off the DF-GP-augmented surrogate $\mathcal{GP}_{\text{exact}}$, so the censored GP does *not* enter the EI score, and CCPST operates on the DF-GP posterior. Under the deployed pure-EI rule the acquisition still evaluates a completion-probability proxy from the censored GP, but that quantity is discarded once the cost-aware division is disabled, so the censored GP affects neither the score nor the stopping decision. The no-cens ablation of Table 2 agrees. Run in the same job as full SRBO at the same configuration and seeds, the two coincide on 99 of the 100 pool and seed pairs; on the remaining pair the two runs diverge partway through the search, and the paired comparison in Table 4 finds no significant difference on any pool ($p \geq 0.32$). The implementation also supports a cost-aware ratio $\text{EI}/\mathbb{E}[\text{cost} \mid x]$ in which the censored GP feeds a completion-probability proxy $\phi_{\text{comp}}(x)$ to weight $\mathbb{E}[\text{cost} \mid x] = \phi_{\text{comp}} \cdot \text{full_cost} + (1-\phi_{\text{comp}}) \cdot \mathbb{E}[\text{stop_cost}]$; we disable that division because in our setting high $\mathbb{E}[\text{cost} \mid x]$ is itself a proxy for "CCPST keeps this configuration to completion", so the ratio penalized promising configurations rather than rewarding them.

H On the Treatment of Hyperband and BOHB

Hyperband (Li et al., 2018) and BOHB (Falkner et al., 2018) are the default bandit-style early-stopping baselines in AutoML. Their public implementations rank by y_t, which suits pools where y_t orders configurations as y_T does and is mis-calibrated when partial trajectories are unreliable, as documented in Table 1 (Kendall $\tau_{T/4} \in [0.58, 0.84]$, 7% to 21% pair inversions). Three Hyperband-family methods in the main comparison rank configurations by y_t: (i) pool-adapted ASHA (Li et al., 2018; 2020), async successive halving with y_t-quantile halting; (ii) HyperbandEI, sync-bracket Hyperband with the paper's EI candidate selector; and (iii) BOHB (Falkner et al., 2018), sync-bracket Hyperband with the canonical TPE density-ratio candidate selector. In both synchronous-bracket methods every

Algorithm 1 Soft-Rejection Informed Bayesian Optimization (SRBO). Blue steps are specific to our method.

Input: cost budget C, initial exact data $\mathcal{D}_{\text{exact}}^{(0)}$, FSR risk α_{fsr}, noise multiplier γ, EI exploration ξ, SPRT shape β, recalibration interval $K_{\text{recal}} = 5$.
Output: $x^\star \in \arg\max_{x_i \in \mathcal{D}_{\text{exact}}} Y(x_i)$.
Initialize exact surrogate $\mathcal{GP}_{\text{exact}}$, direct-forecasting $\mathcal{GP}_{\text{DF}}$, and $\mathcal{D}_{\text{pseudo}} \leftarrow \emptyset$. Set epoch counter $m \leftarrow 0$.
Calibrate CCPST (epoch $m = 0$): compute $Y_{\text{th}}^{(m)}$ as the 75th percentile of $\{Y(x_i)\}_{i \in \mathcal{D}_{\text{exact}}^{(m)}}$, the reliability weight $\rho_t^{(m)}$ via (5) and the R^2-weighted thresholds $\{\log B_t^{(m)}, \log A_t^{(m)}\}$ via (6), and the safe-time set $\mathcal{T}_\star^{(m)}$ via 5-fold held-out CV on the promising subset $\mathcal{D}_{\text{exact}}^{+(m)}$ (Section 4.2).
while budget C is not exhausted **do**
  $x_{\text{next}} \leftarrow \arg\max_{x \in \mathcal{X}} \text{EI}(x \mid \mathcal{GP}_{\text{exact}})$ ▷ Plain EI on the DF-GP-augmented surrogate
  Initiate experiment for x_{next}, observing $y_1, y_2, \ldots$ sequentially.
  for each observation step $t = 1, 2, \ldots, T-1$ as the trajectory unfolds **do**
    If $t \notin \mathcal{T}_\star^{(m)}$ **then** continue (observe-only step; no stopping test).
    Query $\mathcal{GP}_{\text{DF}}$ to obtain (μ_t, σ_t^2) from partial trajectory $\mathbf{y}_{1:t}(x_{\text{next}})$.
    Compute per-step log-odds ℓ_t via (4) using $Y_{\text{th}}^{(m)}$.
    if $\ell_t \le \log A_t^{(m)}$ **then**
      Soft-Rejection: store the partial record $(x_{\text{next}}, \mathbf{y}_{1:t}(x_{\text{next}}))$ in $\mathcal{D}_{\text{pseudo}}$ and terminate experiment; **break**.
    end if
  end for
  if experiment reached step T **then**
    Observe target $Y(x_{\text{next}})$ and add to $\mathcal{D}_{\text{exact}}$.
  end if
  Recompute pseudo-labels μ_τ for all records in $\mathcal{D}_{\text{pseudo}}$ using the current $\mathcal{GP}_{\text{DF}}$, then refit $\mathcal{GP}_{\text{exact}}$ on $\mathcal{D}_{\text{exact}} \cup \{(x, \mu_\tau)\}$ with a single inflated noise variance $\sigma_{\text{aug}}^2 = \gamma\sigma^2$ shared by exact and pseudo points (Section 4). Because this refit precedes the recalibration test below, the first candidate selected after a recalibration is scored by EI on pseudo-labels from the previous $\mathcal{GP}_{\text{DF}}$, while its stopping test already uses the retrained one; the pseudo-labels are refreshed at the next pass of this step.
  if K_{recal} new completions accumulated since the last recalibration, or 25 outer iterations elapsed since it **then**
    $m \leftarrow m + 1$; recompute $Y_{\text{th}}^{(m)}$, retrain $\mathcal{GP}_{\text{DF}}$ on $\mathcal{D}_{\text{exact}}$, and recalibrate $\rho_t^{(m)}, \{\log A_t^{(m)}, \log B_t^{(m)}\}$, $\mathcal{T}_\star^{(m)}$ via 5-fold CV on $\mathcal{D}_{\text{exact}}^{+(m)}$.
  end if
end while

configuration run to T enters the candidate model; with $T = 11$, which is not a power of η, a bracket's last survivor is promoted to T, and charged for it, before its endpoint is observed, and no rung is started or promoted past the budget. The main comparison also includes Freeze-Thaw (Swersky et al., 2014), a curve-modeling baseline, which is not a Hyperband-family method: it halts on its own learning-curve GP endpoint forecast rather than on y_t. ASHA, Freeze-Thaw and HyperbandEI all lose significantly to SRBO on IVPT and Concrete (Table 4), and so does BOHB: on IVPT with regret 165.55 against SRBO's 0.17 ($p < 10^{-4}$) and on Concrete with 33.80 ± 8.84 (median 9.81, max 100.4) against 4.02 ($p = 0.007$). Appendix B.1 reports three controlled BOHB variants (TPE-only without halving, oracle halving by y_T, false-discard audit): TPE-only modestly improves IVPT but remains poor (167.6 vs. 180.7, $p < 10^{-4}$), oracle halving does not improve IVPT (181.3, $p = 0.39$), and on Concrete removing halving hurts (146.5 vs. 72.2, $p = 0.04$) while oracle halving improves it directionally but not significantly (56.4 vs. 72.2, $p = 0.39$). HyperbandEI, which keeps Hyperband's bracket halving and replaces TPE by EI, is the strongest baseline on both pools. On IVPT it closes almost all of BOHB's gap to SRBO (0.58 against 165.55), so that gap comes mainly from the TPE selector; on Concrete it still ends $7.1\times$ above SRBO against BOHB's $8.4\times$, so most of that gap persists under EI. Halving under a common EI selector trails SRBO on every pool of Table 12. On the three LCBench tasks BOHB attains the lower mean regret, significantly on each ($p < 0.001$), while

HyperbandEI still trails SRBO; BOHB's TPE selector is the stronger tool where early observations predict the endpoint well, the regime SRBO was *not* designed for.

DF-GP-ranked halving. The proposition below compares successive halving on a DF-GP ranking signal with CCPST.

Proposition 1 (DF-GP-ranked halving and CCPST)**.** *Consider a Hyperband instance with a single-bracket successive-halving schedule that keeps the top $1/\eta$ fraction of survivors at each halving step (reduction factor $\eta > 1$), using as its ranking signal the DF-GP forecast $g(x) \coloneqq \hat{Y}(x \mid \mathbf{y}_{1:t})$ of Section 4. In a maximization setting, "keep the top $1/\eta$" means retain the x whose forecast lies* above *the $(1-1/\eta)$-th quantile of surviving-pool forecasts; equivalently, terminate those whose forecast lies below it. When surviving rankings are frozen between eliminations, this stopping event is analogous to the SRBO stopping rule of Section 4.2: both eliminate configurations whose endpoint forecast is sufficiently low, but DF-GP-ranked halving uses a* pool-relative *quantile cut that is recomputed per bracket, while CCPST uses an* absolute *per-step log-odds threshold against $Y_{th}^{(m)}$ with held-out FSR control. CCPST's safe-time pruning and FSR guarantee are, in addition, bracket-free.*

Proof sketch. The DF-GP produces a scalar ranking $g(x)$ monotone in predicted target. Under maximization, the halving operation retains x with $g(x)$ strictly above the $(1-1/\eta)$-th sample quantile of $\{g(\tilde{x}) : \tilde{x} \in \text{surviving pool}\}$, i.e., terminates x whenever $g(x)$ falls below that quantile. This rule parallels the SRBO stopping event $\{t \in \mathcal{T}_\star^{(m)},\ \ell_t(x) \le \log A_t^{(m)}\}$: both eliminate configurations whose endpoint forecast is sufficiently low, but the halving rule uses a pool-relative quantile that depends on the bracket composition, while CCPST uses an absolute log-odds threshold against $Y_{\text{th}}^{(m)}$ with a held-out FSR budget. The two rules also differ in scope: Hyperband recomputes the quantile once per bracket and inherits a union-bound penalty across brackets, whereas CCPST's safe-time pruning is bracket-free but is enforced only against $\widehat{\text{FSR}} \le \alpha_{\text{fsr}}$ on the held-out CV folds. □

Proposition 1 says that a "fair" Hyperband baseline (one that replaces y_t with a target-oriented forecast) is analogous to SRBO's per-trial stopping rule (configurations whose endpoint forecast is sufficiently low are eliminated either way) and differs from it in three ways. The Hyperband variant uses a pool-relative quantile recomputed per bracket and lacks CCPST's held-out Hoeffding FSR guarantee; it also ranks by the DF-GP point forecast alone, whereas the CCPST log-odds (4) standardize the forecast margin by the posterior standard deviation $\sigma_t(x)$, so configurations with equal point forecasts can be ordered differently by the two rules. Combining CCPST with Hyperband-style multi-bracket budget reallocation across trials is a promising direction we leave to future work.