

Beyond Fixed False Discovery Rates: Post-Hoc Conformal Selection with E-Variables

Meiyi Zhu and Osvaldo Simeone *Fellow, IEEE*

Abstract—Conformal selection (CS) uses calibration data to identify test inputs whose unobserved outcomes are likely to satisfy a pre-specified minimal quality requirement, while controlling the false discovery rate (FDR). Existing methods fix the target FDR level before observing data, which prevents the user from adapting the balance between number of selected test inputs and FDR to downstream needs and constraints based on the available data. For example, in genomics or neuroimaging, researchers often inspect the distribution of test statistics, and decide how aggressively to pursue candidates based on observed evidence strength and available follow-up resources. To address this limitation, we introduce post-hoc CS (PH-CS), which generates a path of candidate selection sets, each paired with a data-driven false discovery proportion (FDP) estimate. PH-CS lets the user select any operating point on this path by maximizing a user-specified utility, arbitrarily balancing selection size and FDR. Building on conformal e-variables and the e-Benjamini-Hochberg (e-BH) procedure, PH-CS is proved to provide a finite-sample post-hoc reliability guarantee whereby the ratio between estimated FDP level and true FDP is, on average, upper bounded by 1, so that the average estimated FDP is, to first order, a valid upper bound on the true FDR. PH-CS is extended to control quality defined in terms of a general risk. Experiments on synthetic and real-world datasets demonstrate that, unlike CS, PH-CS can consistently satisfy user-imposed utility constraints while producing reliable FDP estimates and maintaining competitive FDR control.

Index Terms—Conformal selection, false discovery rate, post-hoc, e-variables, multiple hypothesis testing

I. INTRODUCTION

A. Motivation

Conformal selection (CS) [1] provides a principled, distribution-free mechanism for choosing a subset of test inputs whose unobserved outcomes are likely to meet a given requirement. Examples of applications include early-stage drug discovery [2], [3], candidate gene selection in genomics [4], and feature selection in machine learning [5], [6]. CS is able to control the *false discovery rate* (FDR), i.e., the fraction of incorrectly selected candidates, at a pre-specified level $\alpha_{\max} \in [0, 1]$. A key limitation of this framework is that the target FDR level $\alpha_{\max}$ must be fixed *before* observing the test and calibration data. In practice, however, the “right” operating point is often dictated by the quality of the candidates and by downstream constraints such as budget caps or resource availability, all of which only become clear once the

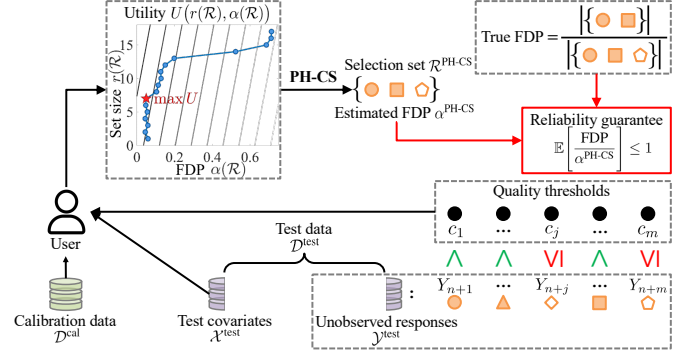


Fig. 1. Illustration of the PH-CS problem and framework. A user has access to labeled calibration data $\mathcal{D}^{\text{cal}}$ and unlabeled test covariates $\mathcal{X}^{\text{test}}$, while the test responses $\mathcal{Y}^{\text{test}}$ remain unobserved. Each test input X_{n+j} is associated with a threshold c_j representing the minimum quality requirement (2). Based on calibration and test data, PH-CS produces a selection set $\mathcal{R}^{\text{PH-CS}}$, together with an estimated FDP $\alpha^{\text{PH-CS}}$, by maximizing an arbitrary utility (26) that balances selection set size and reliability (i.e., FDP). Contour lines of the utility function $U(r(\mathcal{R}), \alpha(\mathcal{R}))$ are shown with darker lines indicating higher utility. PH-CS optimizes the utility over a candidate path, shown via blue dots, which is determined post-hoc based on calibration and test data. The red star marks the selected operating point. The reliability guarantee (1) ensures that the estimate $\alpha^{\text{PH-CS}}$ is, on average, an approximate upper bound on the true FDR (see (1) and (13)).

calibration and test data have been collected. Being forced to commit to a single FDR level a priori can lead to selection sets that are either too conservative, missing promising candidates that could have been useful for downstream applications, or too liberal, exceeding acceptable constraints for downstream users, with no mechanism for readjustment.

This tension between a priori guarantees and post-hoc flexibility arises across a wide range of applied settings. For instance, in early-stage drug discovery, after observing enrichment patterns, it may be desirable to tighten or relax the selection threshold so as to fill a fixed number of slots for downstream validation, rather than being locked to a nominal FDR level chosen before the screen [2], [3]. As another example, in genomics or neuroimaging, researchers may inspect the distribution of test statistics and decide how aggressively to pursue candidates based on observed evidence strength and available follow-up resources [4]. Finally, in feature selection for high-dimensional machine learning, users may wish to expand or contract the feature set after examining prediction accuracy or computational cost on a held-out sample [5], [6].

In each of these scenarios, as illustrated in Fig. 1, the user faces a *utility-driven* trade-off between the size of the selected set and the proportion of erroneous selections. Conventional CS cannot resolve this trade-off in a data-adaptive way, given that its FDR guarantee is tied to a single, pre-specified level.

To address this issue, this paper introduces *post-hoc CS* (PH-CS). As sketched in Fig. 1, PH-CS produces a path of candidate selection sets together with data-driven *false*

The work of M. Zhu and O. Simeone was supported by an Open Fellowship of the EPSRC (EP/W024101/1). The work of O. Simeone was also supported by EPSRC (EP/X011852/1) and ERC (No. 101198347).

Meiyi Zhu is with the Department of Engineering, King’s College London, WC2R 2LS, London, U.K. (e-mail: meiyi.l.zhu@kcl.ac.uk).

Osvaldo Simeone is with the Institute for Intelligent Networked Systems, Northeastern University London, E1 8PH London, U.K. (e-mail: o.simeone@northeastern.edu).

discovery proportion (FDP) estimates. The FDP is the fraction of incorrectly selected candidates, and its average is the FDR. Based on the candidate selection sets, PH-CS allows the user to select the operating point that maximizes an arbitrary non-negative utility of set size and estimated FDP. Importantly, the estimated FDP results in reliability guarantee that holds *uniformly* over the path, so the choice of operating point can be made *after* the data have been observed without violating statistical validity. In particular, the estimated FDP is, on average, an approximately valid upper bound on the FDR (see Sec. I-C for a precise statement).

B. Related Work

1) *Conformal Prediction and Conformal Selection*: Conformal prediction constructs distribution-free prediction sets with finite-sample coverage guarantees under exchangeability assumptions [7]–[11]. Conformal prediction builds on *conformal p-variables*, statistics that target the null hypothesis of exchangeability [12]. Reference [1] recast selection of test data units as the multiple testing of random null hypotheses, leveraging conformal p-variables and the Benjamini-Hochberg (BH) procedure [13], [14] to control the FDR at a fixed level. Other applications of conformal p-variables include anomaly detection [15].

Since the introduction of CS [1], numerous extensions have been proposed. In terms of the selection criterion, the work [16] extended CS to sensitivity analysis of individual treatment effects under covariate shift, reference [17] generalized CS to multivariate responses via a regional monotonicity condition, and the paper [18] handled conjunctive and disjunctive selection conditions. Targeting the statistical power, reference [19] improved power through data-reuse strategies for score optimization, and the work [20] achieved asymptotically optimal power via a likelihood-ratio-based rule inspired by the Neyman-Pearson paradigm. Addressing the decision protocol, CS has been adapted to online settings with irrevocable decisions [21] or with real-time feedback [22], and to interactive settings that support adaptive model updates during screening [23]. Other extensions include promoting diversity in the selected set [24] and applications to compound screening in drug discovery [3]. In all these studies, however, the FDR level must be fixed a priori.

2) *E-variables and E-BH*: E-variables [25]–[27] offer an alternative to p-variables for quantifying evidence against a null hypothesis, with large values indicating stronger evidence. Reference [28] developed the e-Benjamini-Hochberg (e-BH) procedure, which controls the FDR using e-variables under arbitrary dependence among the test statistics, a property that the classical BH procedure does not enjoy in general. In the conformal setting, the work [29] introduced conformal e-variables based on a score-ratio statistic that replaces the rank-based conformal p-values introduced in [7], and the paper [30] combined conformal e-values with e-BH to achieve derandomized novelty detection with provable FDR control. Further developments have focused on improving the power of e-BH, including conditional calibration strategies [31], compound and weighted e-values that concentrate the testing

budget on promising hypotheses [32], and randomization-based enhancements [33]. Our work also adopts conformal e-variables and e-BH, but targets post-hoc selection rather than prediction or novelty detection.

3) *Post-hoc and Post-selection Inference*: Post-hoc inference in multiple testing aims to provide valid error guarantees for rejected sets chosen after observing the data. In the p-value framework, reference [34] established a foundational approach by bounding the number of false discoveries uniformly over all possible rejected sets. This framework was further developed by [35] and [36], who provided tighter post-hoc bounds on the FDP via reference families and permutation-based methods, respectively. In the e-value framework, reference [37] demonstrated the flexibility of e-values for post-selection inference through e-value-based confidence intervals. References [38], [39] introduced backward conformal prediction, which fixes the prediction set size and adapts the coverage level accordingly, yielding a reliability condition that directly inspires the post-hoc guarantee pursued in this work. A related direction bounds the FDP with high probability, simultaneously over all rejection thresholds [40], [41], which we adopt as one of our baselines.

4) *Conformal Selection with E-variables*: Recent work has combined conformal e-variables with e-BH for selection and related tasks. Reference [42] applied this combination to selection from hierarchical data, constructing count-based e-variables from threshold exceedances after a data-dependent cutoff. More recently, the work [43] extended CS to settings where the quality of each selection is measured by a continuous loss rather than a binary indicator, using risk-adjusted e-variables to control a generalized notion of FDR. In both works, however, the e-variable construction depends on a thresholding level that must be fixed before the data are seen, so that changing it requires recomputing the e-variables. This prevents direct application to post-hoc selection, where the operating point is chosen after observing the data.

C. Main Contributions

The main contributions of this paper are summarized as follows.

- **Post-hoc conformal selection**: We propose PH-CS, a novel CS framework that generates a path of candidate selection sets with associated FDP estimates (see Fig. 1). Among the candidate selection sets, PH-CS selects the operating point that maximizes a user-specified utility function balancing set size and reliability (Algorithm 1). Like [42] and [43], PH-CS replaces conformal p-variables and the BH procedure with conformal e-variables [29] and the e-BH procedure [28]. However, unlike these prior works, whose e-variable constructions are tied to a specific nominal level, PH-CS exploits the level-uniform property of e-BH to ensure simultaneous validity across all nominal levels, enabling post-hoc selection of the operating point.
- **Finite-sample post-hoc reliability guarantee**: We prove that the FDP estimate $\alpha^{\text{PH-CS}}$ produced by PH-CS for any

test batch satisfies, on average, the reliability condition

$$\mathbb{E} \left[\frac{\text{FDP}}{\alpha^{\text{PH-CS}}} \right] \leq 1, \quad (1)$$

implying that the average of the FDP estimate $\alpha^{\text{PH-CS}}$, i.e., $\mathbb{E}[\alpha^{\text{PH-CS}}]$, is an approximate upper bound on the true FDR. In fact, this condition coincides, up to second-order terms, with the inequality $\mathbb{E}[\alpha^{\text{PH-CS}}] \geq \text{FDR}$ [27], [38]. This guarantee holds for *any* data-dependent choice of operating point along the e-BH solution path, under the only assumptions of exchangeability and score monotonicity.

- **Flexible utility designs:** PH-CS accommodates post-hoc selection based on an arbitrary utility function obtained from FDP and set size. This formulation supports, e.g., selection sets that control the size of the selected set, as well as additive trade-off designs, enabling the user to tailor the size-reliability balance to the application at hand.
- **Extensions to risk control and weighted selection:** We extend PH-CS to settings where the quality of each selection is measured by a continuous loss rather than a binary indicator [43], yielding *post-hoc risk-controlled selection* (PH-RCS) with the same reliability guarantee. We further incorporate priority weights that allow the user to assign higher importance to more promising test inputs while preserving post-hoc validity.
- **Empirical validation:** Through experiments on three synthetic settings and three real-world datasets, we demonstrate that PH-CS consistently satisfies user-imposed size and utility constraints that conventional CS cannot enforce, while providing reliable FDP estimates and maintaining competitive FDR performance.

The remainder of the paper is organized as follows. Sec. II formulates the post-hoc selection problem and the utility-based objective, and Sec. III reviews conventional CS. The proposed PH-CS framework and its theoretical guarantee are developed in Sec. IV, followed by extensions to risk control and weighted selection in Sec. V. Experimental results are presented in Sec. VI, and Sec. VII concludes this paper. Proofs and additional experiments are deferred to the appendices.

II. PROBLEM FORMULATION

As illustrated in Fig. 1, we study a selection problem in which a user leverages a labeled calibration sample to select a subset of unlabeled inputs with desirable properties from a batch of candidates [1], [37], [42]. The goal of this selection is to balance selection size and reliability: selecting more inputs is generally preferable, but it can increase the fraction of selected inputs that fail to meet the pre-specified requirements. The rest of this section formalizes setting and requirements.

A. Calibration and Test Data

The system has access to a labeled calibration sample $\mathcal{D}^{\text{cal}} = \{(X_i, Y_i)\}_{i=1}^n$, collected from past operations and used to calibrate selection decisions. At a given epoch, the system observes a batch of m unlabeled test covariates $\mathcal{X}^{\text{test}} = \{X_{n+j}\}_{j=1}^m$, generated together with unobserved

responses $\mathcal{Y}^{\text{test}} = \{Y_{n+j}\}_{j=1}^m$. We assume that the combined collection of calibration and test pairs $\mathcal{D}^{\text{cal}} \cup \mathcal{D}^{\text{test}}$, with $\mathcal{D}^{\text{test}} = \{(X_{n+j}, Y_{n+j})\}_{j=1}^m$, is exchangeable. This is the case when all samples $(X, Y) \in \mathcal{D}^{\text{cal}} \cup \mathcal{D}^{\text{test}}$ are independent.

Larger values of the response variable correspond to more desirable outcomes. Accordingly, for each test input $j = 1, \dots, m$, the user specifies a threshold c_j that represents a minimum target performance level. The goal is to identify test inputs X_{n+j} whose (unobserved) outcome Y_{n+j} strictly exceeds the minimum value c_j , i.e.,

$$Y_{n+j} > c_j. \quad (2)$$

A generalization of this requirement that allows for the control of a general loss $L(X_{n+j}, Y_{n+j})$ is presented in Sec. V following [43].

B. Selection Rule and False Discovery Rate

A selection rule R maps the observed data $(\mathcal{D}^{\text{cal}}, \mathcal{X}^{\text{test}})$ to a data-dependent selection subset

$$\mathcal{R} = R(\mathcal{D}^{\text{cal}}, \mathcal{X}^{\text{test}}) \subseteq \{1, \dots, m\}, \quad (3)$$

whose indices are interpreted as the test inputs in set $\mathcal{X}^{\text{test}}$ the user deems likely to satisfy the requirement in (2).

A *false discovery* occurs when a test point is selected even though it fails to satisfy the requirement in (2), i.e., when one has the inclusion $j \in \mathcal{R}$ and also the inequality $Y_{n+j} \leq c_j$. We quantify the realized error using the FDP, i.e.,

$$\text{FDP}(\mathcal{R}, \mathcal{Y}^{\text{test}}) = \frac{\sum_{j=1}^m \mathbb{1}\{j \in \mathcal{R}, Y_{n+j} \leq c_j\}}{\max\{1, |\mathcal{R}|\}}, \quad (4)$$

and define the FDR as its expectation, i.e.,

$$\text{FDR}(R) = \mathbb{E}[\text{FDP}(\mathcal{R}, \mathcal{Y}^{\text{test}})], \quad (5)$$

where the expectation is taken over the joint distribution of the selected set $\mathcal{R}$ and of the true test labels $\mathcal{Y}^{\text{test}}$.

C. Utility-Based Objective

The design of the selection rule R entails a trade-off between set size and reliability. On the one hand, selecting a larger set $\mathcal{R}$ is generally preferable, as it may produce a larger yield in a production system, give more options to downstream applications [16], [44], or require fewer queries to further evaluate unselected units using high-cost validation systems [45], [46]. On the other hand, a larger set $\mathcal{R}$ may also increase the fraction of selected inputs that fail to meet requirement (2). Conversely, a more conservative rule that produces a smaller set $\mathcal{R}$ may reduce the number of erroneous selections, but it may also exclude desirable inputs satisfying the condition (2).

CS does not provide any control for the set size $|\mathcal{R}|$, targeting only a maximum FDR value α_{max} via the inequality $\text{FDR}(\mathcal{R}) \leq \alpha_{\text{max}}$ (see Sec. III). In contrast, in this work we assume that, given the dataset $\mathcal{D}^{\text{cal}}$ and the unlabeled test data $\mathcal{X}^{\text{test}}$, the user wishes to balance size and reliability in the selection of set $\mathcal{R}$.

To elaborate, we write as

$$r = r(\mathcal{R}) = \sum_{j \in \mathcal{R}} \pi_j \quad (6)$$

the weighted size of set $\mathcal{R}$, where $0 \leq \pi_j \leq 1$ are arbitrary weights. With $\pi_j = 1$, we obtain the set cardinality $r(\mathcal{R}) = |\mathcal{R}|$. Alternatively, if the weight π_j is an estimate of the probability of the requirement (2), the size (6) becomes an estimate of the average number of correct selections that satisfy condition (2).

Let $U : [0, m] \times (0, 1) \mapsto [0, +\infty)$ be a user-specified non-negative utility $U(r, \alpha)$ that is non-decreasing in the set size r in (6) and non-increasing in the FDP α . The specific ways in which the utility $U(r, \alpha)$ increases with respect to set size r and decreases with the FDP α characterize the user-defined trade-off between these two criteria. For example, the utility function $U(r, \alpha)$ may be chosen in one of the following ways:

- *Constrained-size design*: Given a minimum selection size $r_{\min} \in \{0, 1, \dots, m\}$, the user can prioritize smaller FDP levels among rules with size $r \geq r_{\min}$ by adopting the utility

$$U(r, \alpha) = (1 - \alpha) \cdot \mathbb{1}\{r \geq r_{\min}\}. \quad (7)$$

Such a lower bound arises when a downstream stage must process a guaranteed number of candidates, as in filling a screening budget in drug discovery [2], [3] or shortlisting genes for follow-up [4].

- *Additive trade-off*: A general trade-off between size r and FDP values can be expressed via the additive utility

$$U(r, \alpha) = u(r) - \lambda v(\alpha) + C, \quad (8)$$

where $u : [0, m] \rightarrow [0, +\infty)$ is a non-decreasing function; $v : (0, 1) \rightarrow [0, +\infty)$ is a non-decreasing function; hyperparameter $\lambda > 0$ controls the relative importance of set size r and FDP α ; and the constraint $C \geq 0$ is chosen to ensure the non-negativity of utility (8). Varying the hyperparameter λ yields a continuum of possible operating points between size-oriented and reliability-oriented behavior.

- *Multiplicative trade-off*: Alternatively, the size r and reliability $1 - \alpha$ can be combined via the multiplicative utility

$$U(r, \alpha) = u(r) \cdot v(1 - \alpha), \quad (9)$$

where function $u(\cdot)$ and $v(\cdot)$ are non-decreasing as in (8). With the choices $u(r) = r$ and $v(x) = x$, together with weights $\pi_j = 1$ in (6), the utility (9) captures the expected number of correct selections among the r selected inputs, i.e., $U(r, \alpha) = r(1 - \alpha)$.

We are ideally interested in supporting any selection rule (3) that addresses the problem

$$\max_{\mathcal{R} \subseteq \{1, \dots, m\}} U(r(\mathcal{R}), \alpha(\mathcal{R})), \quad (10)$$

where $r(\mathcal{R})$ is the weighted set size (6) and $\alpha(\mathcal{R}) = \text{FDP}(\mathcal{R}, \mathcal{Y}^{\text{test}})$ is the FDP in (4) for set $\mathcal{R}$. However, the FDP $\alpha(\mathcal{R})$ in (4) depends on the true responses $\mathcal{Y}^{\text{test}}$, which are unknown. Therefore, we propose to study selection rules that use only data $\mathcal{D}^{\text{cal}}$ and $\mathcal{X}^{\text{test}}$, taking the form

$$\mathcal{R} = R(\mathcal{D}^{\text{cal}}, \mathcal{X}^{\text{test}}) = \arg \max_{\mathcal{R} \subseteq \{1, \dots, m\}} U(r(\mathcal{R}), \hat{\alpha}(\mathcal{R})), \quad (11)$$

where $\hat{\alpha}(\mathcal{R})$ is an estimate of the quantity $\text{FDP}(\mathcal{R}, \mathcal{Y}^{\text{test}})$ based on the available data. Note that, throughout this paper, the notation $\arg \max$ is used to denote any of the maximizers of the given function.

In order for problem (11) to be a useful approximation of the original optimization (10), we impose that the estimate $\hat{\alpha}(\mathcal{R})$ used in problem (11) satisfy the inequality

$$\mathbb{E} \left[\frac{\text{FDP}(\mathcal{R}, \mathcal{Y}^{\text{test}})}{\hat{\alpha}(\mathcal{R})} \right] \leq 1, \quad (12)$$

where the expectation is taken over the joint distribution of the selected set $\mathcal{R}$ and of the true test labels $\mathcal{Y}^{\text{test}}$. This inequality, inspired by [38], stipulates that, on average, the ratio between the true FDP and the estimate $\hat{\alpha}(\mathcal{R})$ does not exceed 1. The ratio (12) compares the true FDP with its estimate within each run, and it reduces to the standard requirement $\mathbb{E}[\text{FDP}] \leq \alpha$ when $\hat{\alpha}(\mathcal{R})$ is a fixed level α .

For a general data-dependent $\hat{\alpha}(\mathcal{R})$, a first-order Taylor expansion of $1/\hat{\alpha}(\mathcal{R})$ around the mean $\mathbb{E}[\hat{\alpha}(\mathcal{R})]$, as shown in Appendix A, implies that condition (12) is approximately, up to a quadratic error of order $O((\hat{\alpha}(\mathcal{R}) - \mathbb{E}[\hat{\alpha}(\mathcal{R})])^2)$, equivalent to the inequality [27], [38]

$$\text{FDR}(R) \leq \mathbb{E}[\hat{\alpha}(\mathcal{R})]. \quad (13)$$

Accordingly, the estimate $\hat{\alpha}(\mathcal{R})$ provides, on average, an approximately valid upper bound on the true FDR (5). An alternative criterion based on a probably approximately correct (PAC) requirement [40], [41] will be explored in Sec. VI-B.

Overall, as illustrated in Fig. 1, our goal is to design a selection rule that addresses problem (11), while leveraging an FDR estimate $\hat{\alpha}(\mathcal{R})$ satisfying the inequality (12). The resulting class of methods can adjust set size $r(\mathcal{R})$ and FDP $\alpha(\mathcal{R})$, in a post-hoc manner, as a function of the observed data $\mathcal{D}^{\text{cal}}$ and $\mathcal{X}^{\text{test}}$, while providing a reliable estimate of the FDR.

III. PRELIMINARIES: CONVENTIONAL CONFORMAL SELECTION

To provide the necessary background, in this section we briefly review the conventional CS framework based on conformal p-variables and the BH procedure introduced in [1]. CS is a selection rule of the form (3) that imposes a constraint on the FDR (5), i.e.,

$$\text{FDR}(R) \leq \alpha_{\max} \quad (14)$$

for a pre-specified target maximum level $\alpha_{\max} \in (0, 1)$. Unlike the general utility-based rule (11) studied in this work, CS cannot adjust set size $|\mathcal{R}|$ and FDP level α as a function of the observed data $\mathcal{D}^{\text{cal}}$ and $\mathcal{X}^{\text{test}}$. Rather, it can only ensure the average guarantee (14). CS serves as a baseline, highlighting the limitations in the state of the art that motivate the proposed class of selection methods.

A. Selection as Multiple Hypothesis Testing

For each test input $X_{n+j} \in \mathcal{X}^{\text{test}}$, CS considers the *random null hypothesis*

$$H_j : Y_{n+j} \leq c_j, \quad (15)$$

so that rejecting hypothesis H_j corresponds to selecting input X_{n+j} . With this formulation, false discoveries coincide with selected indices $j \in \mathcal{R}$ for which the null hypothesis H_j is true. Unless stated otherwise, we assume that the target levels are fixed constraints, not dependent on data, although our results can be generalized (see Sec. V).

To test the random nulls (15), conventional CS uses a predictive model $\mu : \mathcal{X} \rightarrow \mathbb{R}$, whose output $\mu(X)$ estimates the response for input X , to define a non-negative *conformity score* $S : \mathcal{X} \times \mathcal{Y} \rightarrow \mathbb{R}_+$. The purpose of the score $S(X, Y)$ is to quantify how likely it is for the true response associated with input X under the data distribution to exceed, i.e., to be better than, level Y . For instance, one may use the scores $S(X, Y) = \max\{\mu(X) - Y, 0\}$ or $S(X, Y) = \exp(\mu(X) - Y)$, which are non-negative and non-increasing in Y [1].

The score $S(X, Y)$ is assumed to be monotone non-increasing in the response Y , i.e., for every $X \in \mathcal{X}$, we have the inequality

$$S(X, Y) \geq S(X, Y') \quad (16)$$

for $Y \leq Y'$. This condition captures the desired property that the likelihood of exceeding a higher (better) value $Y' \geq Y$ cannot be larger than the corresponding likelihood for level Y .

On the calibration samples, CS computes the scores $S_i = S(X_i, Y_i)$ for all data points $i = 1, \dots, n$. Moreover, for each test input X_{n+j} , CS evaluates the score at the required output threshold c_j as $\hat{S}_{n+j} = S(X_{n+j}, c_j)$ for $j = 1, \dots, m$. Note that the score $\hat{S}_{n+j}$ is computable, since it depends only on the test input X_{n+j} . Furthermore, given the assumed monotonicity of the score function, it provides the highest score for responses compatible with the condition (2).

CS converts the calibration scores $\{S_i\}_{i=1}^n$ and the test scores $\hat{S}_{n+j}$ into a *conformal p-variable* P_j for each random null hypothesis H_j in (15), with $j = 1, \dots, m$. This is given by

$$P_j = \frac{1 + \sum_{i=1}^n \mathbb{1}\{S_i > \hat{S}_{n+j}\}}{n+1}, \quad j = 1, \dots, m. \quad (17)$$

Note that if the test score $\hat{S}_{n+j}$ coincides with some calibration scores S_i , ties are resolved by randomized tie-breaking: among the equal-score calibration points, a uniformly random subset is counted as exceeding $\hat{S}_{n+j}$ [1]. The quantity (17) effectively counts the fraction of calibration points with higher scores S_i than the score $\hat{S}_{n+j}$ for the test input at the target response c_j .

A key subtlety in CS is that the null hypothesis H_j is random because it involves the unobserved test outcome Y_{n+j} . Consequently, the quantity P_j is not a classical p-variable for the null hypothesis H_j . Instead, the quantity P_j is a conformal p-variable for the related null hypothesis that the multiset $\{S_i\}_{i=1}^n \cup \{\hat{S}_{n+j}\}$ is exchangeable [1]. Intuitively, if this exchangeability hypothesis does not hold, the threshold c_j is incompatible with test input X_{n+j} , and one should reject the null hypothesis H_j . Accordingly, a small value of the quantity P_j provides evidence against exchangeability and, in turn, against the random null hypothesis H_j [1].

B. BH Selection

Given the conformal p-variables $\{P_j\}_{j=1}^m$ and a pre-specified nominal level $\alpha_{\max} \in (0, 1)$, CS evaluates the rejection set $\mathcal{R}$ via the BH procedure [1], [13], [14]. Specifically, sorting the p-variables $P_{(1)} \leq \dots \leq P_{(m)}$, BH computes the maximum index k for which the p-variable $P_{(k)}$ does not exceed the corrected threshold $\alpha_{\max} k/m$, i.e.,

$$k(\alpha_{\max}) = \max \left\{ k \in \{1, \dots, m\} : P_{(k)} \leq \frac{\alpha_{\max} k}{m} \right\}, \quad (18)$$

with the convention $\max(\emptyset) = 0$. Then it outputs the selected set

$$R_{\alpha_{\max}}^{\text{CS}}(\mathcal{D}^{\text{cal}}, \mathcal{X}^{\text{test}}) = \left\{ j \in \{1, \dots, m\} : P_j \leq \frac{\alpha_{\max} k(\alpha_{\max})}{m} \right\} \quad (19)$$

if $k(\alpha_{\max}) > 0$ and $R_{\alpha_{\max}}^{\text{CS}}(\mathcal{D}^{\text{cal}}, \mathcal{X}^{\text{test}}) = \emptyset$ if $k(\alpha_{\max}) = 0$.

C. Theoretical Guarantees

CS satisfies the following guarantee on the FDR.

Theorem 1 ([1, Theorem 3]: Fixed-level FDR control). Assume that (i) the combined calibration and test pairs $\mathcal{D}^{\text{cal}} \cup \mathcal{D}^{\text{test}}$ are i.i.d; and (ii) the score function $S(X, Y)$ is monotone non-increasing in Y as in (16). Let the levels $\{c_j\}_{j=1}^m$ and the target FDR $\alpha_{\max} \in (0, 1)$ be fixed in advance, independently of the observed data. Then, the CS selection rule $R_{\alpha_{\max}}^{\text{CS}}$ defined in (19) satisfies the inequality

$$\text{FDR}(R_{\alpha_{\max}}^{\text{CS}}) \leq \alpha_{\max}. \quad (20)$$

IV. POST-HOC CONFORMAL SELECTION

We now develop PH-CS, a data-driven selection method capable of addressing the post-hoc utility optimization in (10) based on an FDR estimate $\hat{\alpha}(\mathcal{R})$ satisfying the requirement (12). As in the conventional CS reviewed in Sec. III, PH-CS views selection as multiple hypothesis testing of the random nulls $H_j : Y_{n+j} \leq c_j$ in (15). Unlike CS, however, PH-CS can address the post-hoc optimization (10) rather than being constrained to the FDR requirement (14). Technically, this is done by replacing conformal p-variables (17) and the BH procedure adopted by CS with conformal e-variables and e-BH [28].

A. Scores and Conformal E-variables

Like CS, in order to construct candidate selection sets, PH-CS relies on a predictive model $\mu(\cdot)$ to obtain conformity scores $S(X, Y)$, where larger values indicate stronger evidence that the outcome exceeds level Y for input X . As discussed in Sec. III-A, the score $S(X, Y)$ satisfies the monotonicity condition (16). Furthermore, as in CS, PH-CS computes the scores $S_i = S(X_i, Y_i)$ for the calibration data $i = 1, \dots, n$, and the scores $\hat{S}_{n+j} = S(X_{n+j}, c_j)$ for the test data $j = 1, \dots, m$. We recall that the score $\hat{S}_{n+j}$ can be interpreted as quantifying the evidence, for test input X_{n+j} , in favor of meeting the requirement $Y_{n+j} > c_j$.

While CS adopts the p-variables (17), PH-CS adopts e-variables targeting the random null hypothesis (15). An e-variable is a random variable whose average under the null does not exceed 1. E-variables provide a more robust way to measure evidence against a null hypothesis as compared to a p-variable [37]. In particular, as shown in [27], e-variables support the post-hoc selection of significance levels in hypothesis testing, while p-variables require significance levels to be specified in advance.

While different choices are possible [30], [42], [43], we focus here on the *conformal e-variable* introduced in [29]. Unlike the e-variables presented in [30], [42], [43], the adopted e-variable has the key advantage that it does not depend on a target FDR level, supporting the desired utility optimization step. For each test point X_{n+j} with $j = 1, \dots, m$, the conformal e-variable is defined as [29]

$$E_j = \frac{\hat{S}_{n+j}}{\frac{1}{n+1} \left(\sum_{i=1}^n S_i + \hat{S}_{n+j} \right)}. \quad (21)$$

This statistic is an e-variable for the null hypothesis that the multiset $\{S_i\}_{i=1}^n \cup \{\hat{S}_{n+j}\}$ is exchangeable [29]. In fact, under this assumption, the average of the random variable E_j equals 1 (see Appendix B). Accordingly, a large value of the statistic (21) provides evidence that exchangeability does not hold, which, in turn, supports rejecting the random null hypothesis $H_j : Y_{n+j} \leq c_j$ in (15) by following the same arguments given in Sec. III-A.

B. Utility-Driven Selection

PH-CS aims to output a data-dependent set $\mathcal{R} \subseteq \{1, \dots, m\}$ that maximizes the utility $U(r(\mathcal{R}), \hat{\alpha}(\mathcal{R}))$ as per problem (11), where the FDP estimate $\hat{\alpha}(\mathcal{R})$, computed from data $(\mathcal{D}^{\text{cal}}, \mathcal{X}^{\text{test}})$, must serve, on average, as a reliable upper bound on the FDR as formalized by (12). To address problem (11), PH-CS starts by restricting the optimization to a discrete set of possible subsets $\mathcal{R}_k \subseteq \{1, \dots, m\}$ with $k \in \mathcal{K}$, where $\mathcal{K}$ is a discrete set of integers. To motivate our choice of candidate subsets $\{\mathcal{R}_k\}_{k \in \mathcal{K}}$, we review next the operation of e-BH [28], an FDR-controlling testing procedure based on e-variables.

For a given nominal FDR level $\alpha \in (0, 1)$, e-BH selects indices $j \in \{1, \dots, m\}$ whose e-variable E_j exceeds a threshold $t(\alpha)$, i.e.,

$$\mathcal{R}(\alpha) = \{j \in \{1, \dots, m\} : E_j \geq t(\alpha)\}. \quad (22)$$

In a manner similar to BH (see (19)), the threshold $t(\alpha)$ is obtained by imposing the self-consistency condition that, if k points are selected, i.e., if $|\mathcal{R}(\alpha)| = k$, the cutoff must be given by [28]

$$t(\alpha) = \frac{m}{\alpha k}. \quad (23)$$

Let $E_{(1)} \geq \dots \geq E_{(m)}$ be the ordered e-variables. The e-BH set (22) can change only when the cutoff $t(\alpha)$ crosses one of the order statistics $\{E_{(k)}\}_{k=1}^m$. Therefore, when adopting e-BH with possibly varying FDR levels α , one can focus without loss of generality on the discrete family of candidate subsets

$$\mathcal{R}_k = \{j \in \{1, \dots, m\} : E_j \geq E_{(k)}\}, \quad k = 0, 1, \dots, m. \quad (24)$$

Algorithm 1 PH-CS

Input: Calibration data $\mathcal{D}^{\text{cal}} = \{(X_i, Y_i)\}_{i=1}^n$, unlabeled test covariates $\mathcal{X}^{\text{test}} = \{X_{n+j}\}_{j=1}^m$, thresholds $\{c_j\}_{j=1}^m$, and utility $U(\cdot, \cdot)$

- 1: Compute calibration scores $\{S_i\}_{i=1}^n$, and test scores $\{\hat{S}_{n+j}\}_{j=1}^m$
- 2: Construct conformal e-variables $\{E_j\}_{j=1}^m$ via (21)
- 3: Sort $\{E_j\}_{j=1}^m$ in non-increasing order as $E_{(1)} \geq E_{(2)} \geq \dots \geq E_{(m)}$
- 4: For each $k = 1, \dots, m$, define set $\mathcal{R}_k = \{j : E_j \geq E_{(k)}\}$ and FDP estimate $\hat{\alpha}(\mathcal{R}_k) = \min\{1, m/(kE_{(k)})\}$ with the convention $(\alpha_0, \mathcal{R}_0) = (0, \emptyset)$
- 5: Among all candidate sets $\{\mathcal{R}_k\}_{k=0}^m$, choose $\mathcal{R}^{\text{PH-CS}} = \arg \max_{\mathcal{R} \in \{\mathcal{R}_0, \dots, \mathcal{R}_m\}} U(r(\mathcal{R}), \hat{\alpha}(\mathcal{R}))$.
- 6: Set $\alpha^{\text{PH-CS}} = \hat{\alpha}(\mathcal{R}^{\text{PH-CS}})$

Output: Selected set $\mathcal{R}^{\text{PH-CS}}$ and error level $\alpha^{\text{PH-CS}}$

Note that the size of the set (24) increases in the index $k = 0, 1, \dots, m$, i.e., we have the inclusion $\mathcal{R}_0 \subseteq \mathcal{R}_1 \subseteq \dots \subseteq \mathcal{R}_m$.

PH-CS optimizes problem (11) over the sets $\mathcal{R}_k$ in (24). For each candidate $\mathcal{R}_k$, the FDP estimate $\hat{\alpha}(\mathcal{R}_k)$ is obtained by inverting the equality (23) as

$$\hat{\alpha}(\mathcal{R}_k) = \min \left\{ 1, \frac{m}{kE_{(k)}} \right\}, \quad (25)$$

where the first term in the minimum ensures that the estimate $\hat{\alpha}(\mathcal{R}_k)$ does not exceed 1. The resulting post-hoc optimization problem solved by PH-CS is then

$$\begin{aligned} \mathcal{R}^{\text{PH-CS}} &= R^{\text{PH-CS}}(\mathcal{D}^{\text{cal}}, \mathcal{X}^{\text{test}}) \\ &= \arg \max_{\mathcal{R} \in \{\mathcal{R}_0, \dots, \mathcal{R}_m\}} U(r(\mathcal{R}), \hat{\alpha}(\mathcal{R})), \end{aligned} \quad (26)$$

and reports the associated FDP level as

$$\alpha^{\text{PH-CS}} = \hat{\alpha}(\mathcal{R}^{\text{PH-CS}}). \quad (27)$$

C. Summary of PH-CS

Algorithm 1 summarizes the proposed PH-CS procedure. Starting from the calibration sample and an unlabeled test batch, PH-CS computes conformity scores $S(X_i, Y_i)$ on the calibration pairs, and evaluates the scores $S(X_{n+j}, c_j)$ at the requirement thresholds c_j on the test covariates. These quantities are then converted into conformal e-variables $\{E_j\}_{j=1}^m$ using (21). Next, the e-variables are sorted in non-increasing order to generate the finite collection of candidate sets $\mathcal{R}_k$ in (24). Finally, PH-CS selects the set $\mathcal{R}_k$ by maximizing the utility (26), producing the selected set $\mathcal{R}^{\text{PH-CS}}$ in (26) together with the chosen nominal level $\alpha^{\text{PH-CS}}$ in (27). In the next subsection, we will show that this construction yields a valid post-hoc reliability estimate in the sense of inequality (12).

D. Reliability Guarantee

The following theorem shows that PH-CS, which is defined in Algorithm 1, returns an FDP estimate $\alpha^{\text{PH-CS}}$ for the selected set $\mathcal{R}^{\text{PH-CS}}$ that satisfies the desired reliability property (12).

Theorem 2 (Post-hoc reliability guarantee). Assume that (i) the combined calibration and test pairs $\mathcal{D}^{\text{cal}} \cup \mathcal{D}^{\text{test}}$ are exchangeable, and (ii) the score $S(X, Y)$ is monotone non-increasing in Y as in (16). Then, the PH-CS output $(\mathcal{R}^{\text{PH-CS}}, \alpha^{\text{PH-CS}})$ defined in (26) satisfies the average requirement

$$\mathbb{E} \left[\frac{\text{FDP}(\mathcal{R}^{\text{PH-CS}}, \mathcal{Y}^{\text{test}})}{\alpha^{\text{PH-CS}}} \right] \leq 1, \quad (28)$$

where the average is evaluated with respect to the joint distribution of the selected set $\mathcal{R}^{\text{PH-CS}}$ and of the true test labels $\mathcal{Y}^{\text{test}}$.

Proof: See Appendix B.

As explained in Sec. II, the inequality (28) implies, up to a first-order approximation, the condition (13), showing that the estimate $\alpha^{\text{PH-CS}}$ is, on average, a valid upper bound on the FDR. Following [38], this property can also be leveraged to estimate the FDR from the calibration sample via batch resampling.

As a final remark, Theorem 2 can be shown to hold even if the target levels $\{c_j\}_{j=1}^m$ depend on the available data $(\mathcal{D}^{\text{cal}}, \mathcal{X}^{\text{test}})$ (see Appendix B). In contrast, extending Theorem 1 to data-dependent thresholds requires the additional positive regression dependence on a subset condition [1].

Theorem 2 guarantees the reliability of the estimate $\alpha^{\text{PH-CS}}$ in terms of the average ratio in (28). As explained in Sec. II, the condition (28) is related to (13), i.e., to the inequality $\mathbb{E}[\text{FDP}(\mathcal{R}^{\text{PH-CS}}, \mathcal{Y}^{\text{test}})] \leq \mathbb{E}[\alpha^{\text{PH-CS}}]$. A natural question is whether one can establish the desired inequality $\text{FDP}(\mathcal{R}^{\text{PH-CS}}, \mathcal{Y}^{\text{test}}) \leq \alpha^{\text{PH-CS}}$ pointwise with high probability, rather than on average. The following result shows that, as the calibration and test sets grow, this is indeed the case.

Proposition 1 (Asymptotic reliability). Assume that (i) the calibration and test pairs $\mathcal{D}^{\text{cal}} \cup \mathcal{D}^{\text{test}}$ are i.i.d.; (ii) the score $S(X, Y)$ is monotone non-increasing in Y as per inequality (16); and (iii) the score has a finite second moment $\mathbb{E}[S(X, Y)^2] < \infty$ and a positive mean $\mathbb{E}[S(X, Y)] > 0$. Then, for every $\varepsilon > 0$, the PH-CS output $(\mathcal{R}^{\text{PH-CS}}, \alpha^{\text{PH-CS}})$ in (26)–(27) satisfies the limit

$$\lim_{n, m \rightarrow \infty} \mathbb{P}(\text{FDP}(\mathcal{R}^{\text{PH-CS}}, \mathcal{Y}^{\text{test}}) \leq \alpha^{\text{PH-CS}} + \varepsilon) = 1. \quad (29)$$

Proof: See Appendix C.

Sec. VI-B will discuss an alternative baseline scheme that guarantees a finite-sample counterpart of condition (29) formulated as a PAC requirement [40], [41].

V. POST-HOC RISK-CONTROLLED SELECTION

The PH-CS framework developed in Sec. IV measures the quality of each selection through the binary cost $\mathbb{1}\{Y_{n+j} \leq c_j\}$ appearing in the numerator of the FDP (4): a selected input either meets the requirement (2), incurring a zero cost, or not, incurring a cost equal to 1. In many applications, however, the consequence of selecting an input is more naturally quantified by a continuous loss $\mathcal{L}(X, Y) \in [0, 1]$, such as a squared prediction error or a semantic distance to a reference output [43]. We show below that the post-hoc mechanism in Algorithm 1

extends directly to this setting, yielding a generalized strategy that we refer to as PH-RCS.

1) *Generalized Risk and Error Metric:* For each test input X_{n+j} , the cost of selection is captured by the loss

$$L_j = \mathcal{L}(X_{n+j}, Y_{n+j}) \in [0, 1]. \quad (30)$$

Replacing the binary indicator in (4) with the continuous risk L_j in (30), the generalized FDP, defined as

$$\text{FDP}^g(\mathcal{R}, \mathcal{Y}^{\text{test}}) = \frac{\sum_{j=1}^m L_j \mathbb{1}\{j \in \mathcal{R}\}}{\max\{1, |\mathcal{R}|\}}, \quad (31)$$

measures the per-selected unit risk, and its average yields the generalized FDR

$$\text{FDR}^g(R) = \mathbb{E}[\text{FDP}^g(\mathcal{R}, \mathcal{Y}^{\text{test}})]. \quad (32)$$

When we set $L_j = \mathbb{1}\{Y_{n+j} \leq c_j\}$, the quantities FDP^g and FDR^g reduce to the FDP in (4) and the FDR in (5), respectively.

2) *Risk-Adjusted E-Variables:* In Sec. IV, the conformal e-variable E_j in (21) is defined by the inequality $\mathbb{E}[\mathbb{1}\{H_j\} \cdot E_j] \leq 1$, where H_j is the binary null event defined in (15). For a more general risk function, reference [43] introduces risk-adjusted e-variables as random variables satisfying the inequality

$$\mathbb{E}[L_j E_j^g] \leq 1. \quad (33)$$

Intuitively, a large value of the generalized e-variable E_j^g provides evidence that the risk L_j is small, since a simultaneously large risk would violate the mean constraint (33). Concrete constructions of risk-adjusted e-variables are provided in [43].

3) *Algorithm and Guarantee:* Given risk-adjusted e-variables $\{E_j^g\}_{j=1}^m$ satisfying condition (33), PH-RCS applies Algorithm 1 with the e-variable E_j replaced by the generalized e-variable E_j^g throughout. The nested candidate sets $\mathcal{R}_k$ in (24), the FDP estimate $\hat{\alpha}(\mathcal{R}_k)$ in (25), and the utility optimization in (26) all retain the same form. We denote the resulting output by $(\mathcal{R}^{\text{PH-RCS}}, \alpha^{\text{PH-RCS}})$ to distinguish it from the binary setting in Sec. IV. The reported level $\alpha^{\text{PH-RCS}}$ now serves as a reliability certificate for the generalized FDP (31) rather than the standard FDP (4). The following result establishes the post-hoc reliability of this generalized procedure.

Theorem 3 (Post-hoc risk-control guarantee). Assume that, for each $j = 1, \dots, m$, the statistic $E_j^g \geq 0$ satisfies (33). Then the PH-RCS output $(\mathcal{R}^{\text{PH-RCS}}, \alpha^{\text{PH-RCS}})$ of Algorithm 1, with E_j replaced by E_j^g , satisfies the average requirement

$$\mathbb{E} \left[\frac{\text{FDP}^g(\mathcal{R}^{\text{PH-RCS}}, \mathcal{Y}^{\text{test}})}{\alpha^{\text{PH-RCS}}} \right] \leq 1, \quad (34)$$

where the average is evaluated with respect to the joint distribution of the selected set $\mathcal{R}^{\text{PH-RCS}}$ and of the true test labels $\mathcal{Y}^{\text{test}}$.

Proof: See Appendix D.

VI. EXPERIMENTS

In this section, we evaluate PH-CS under the constrained-size utility (7), the additive trade-off utility (8), and the multiplicative trade-off utility (9). We compare PH-CS against conventional CS [1] and three data-driven baselines that instead estimate the FDP from p-variables, from a data split, or from a PAC bound in lieu of the average ratio condition (13). Further results on additional datasets and regressors are collected in Appendix G, and Appendix H empirically validates the large-sample behavior of Proposition 1.

A. Settings

We consider two different experimental settings, including both synthetic and real data.

1) *Synthetic Data*: Following reference [1], in each run, in the synthetic-data setting, we independently generate a training sample, a calibration sample, and a test batch with sizes 1000, $n = 1000$, and $m = 100$, respectively, using the same data-generating mechanism. According to this mechanism, the covariate vector $X = [X^{(1)}, \dots, X^{(20)}]$ has i.i.d. entries with uniform distribution $U([-1, 1])$, and the response follows the relationship $Y = f(X) + \varepsilon$ with function $f(X) = 5(X^{(1)}X^{(2)} + e^{X^{(4)}-1})$ and additive noise satisfying one of the following models: (i) homoscedastic noise: $\varepsilon \sim \mathcal{N}(0, (0.15)^2)$; and (ii) heteroscedastic noise: $\varepsilon | X \sim \mathcal{N}(0, \tilde{\sigma}(X)^2)$, where $\tilde{\sigma}(X) = 0.1(5.5 - |f(X)|)/2$. All results are averaged over 100 independent trials. As in [1], we use gradient boosting as the regressor to obtain the prediction function $\mu(X)$. We set the target outcome to $c_j = c = 0$ for all test inputs $j = 1, \dots, m$.

2) *Real Data*: We also consider three real-world datasets, which are randomly split into training, calibration, and test subsets at each run:

- **Recruitment**: The campus recruitment dataset [1] has 215 samples, where $Y = 1$ indicates a successful placement, and the covariates $X \in \mathbb{R}^{21}$ include features such as exam percentages, degree type, and specialization. In each run, the data are split into 45 training samples, 85 calibration samples, and 85 test samples. We use a gradient boosting classifier as the prediction model $\mu(X)$.
- **Musk**: The Musk dataset [30] contains 6598 molecular conformations with covariates $X \in \mathbb{R}^{166}$ describing spatial shape features, and a binary label Y indicating musk activity ($Y = 1$) or not ($Y = 0$). In each run, the data are split into training, calibration, and test subsets with proportions (0.8, 0.1, 0.1). We use a support vector machine as the prediction model $\mu(X)$.
- **Shuttle**: The Shuttle dataset [30] contains 58000 samples with covariates $X \in \mathbb{R}^9$ representing sensor readings, and label Y binarized by setting $Y = 1$ for all classes other than class 1, and $Y = 0$ for class 1. In each run, the data are split into training, calibration, and test subsets with proportions (0.5, 0.25, 0.25). We use logistic regression as the prediction model $\mu(X)$.

For all three real-data classification datasets, we use the same threshold $c_j = c = 0$, so that the requirement $Y_{n+j} > c_j$ identifies the positive class $Y_{n+j} = 1$.

3) *Score Function*: For the real-data classification tasks, the prediction $\mu(X)$ represents the predicted probability for class $Y = 1$, while for the synthetic regression setting, the model output is first mapped to the interval $(0, 1)$ via min-max normalization using the range observed on the training set. As discussed in Sec. III-A, the conformity score $S(X, Y)$ should capture the likelihood that the outcome exceeds Y , and must be non-negative and non-increasing in Y (see (16)). Following [1], given the target c , we adopt the piece-wise constant function

$$S(X, Y) = \begin{cases} \left(\frac{\mu(X)}{1 - \mu(X)} \right)^\gamma, & \text{if } Y \leq c, \\ \delta, & \text{if } Y > c, \end{cases} \quad (35)$$

where $\gamma > 0$ is a tuning parameter and $\delta = 10^{-6}$ is a small constant. When $Y \leq c$, the score increases with the predicted quality $\mu(X)$, and a larger hyperparameter γ amplifies this contrast, improving the discriminative power of the e-BH procedure for well-trained predictors. When $Y > c$, the score is clipped to δ , reducing the denominator of the e-variable (21) and boosting the e-values for promising inputs. A larger parameter γ sharpens the score (35) toward the hard threshold used by indicator-type e-variables [30], [42], [43], which yield a single selection set, whereas the continuous score generates the entire path (24). We use $\gamma = 3$ on the well-separated synthetic data and $\gamma = 50$ on the harder real data.

4) *Performance Metrics*: We report the selected set size $|\mathcal{R}|$, the realized FDP (4), its declared estimate $\alpha^{\text{PH-CS}}$ (27), and the utility (10), all introduced earlier, together with the *power*, i.e., the fraction of test points satisfying the requirement (2) that are selected:

$$\text{Power}(\mathcal{R}, \mathcal{Y}^{\text{test}}) = \frac{\sum_{j=1}^m \mathbb{1}\{j \in \mathcal{R}, Y_{n+j} > c_j\}}{\max\{1, \sum_{j=1}^m \mathbb{1}\{Y_{n+j} > c_j\}\}}. \quad (36)$$

Also known as the true discovery rate, the power complements the FDP, with larger values preferable.

B. Baseline Methods

Beside conventional CS [1], which targets a pre-specified FDP level $\alpha_{\max}$, we compare PH-CS against three data-driven baselines that also address the post-hoc objective (10). All three build on the BH counterpart of the e-BH path (24) used by PH-CS, spanning the same range of set sizes, and differ only in how the FDP estimate of the selected set is constructed.

- 1) *Naïve PH-CS* takes the nominal BH level of each candidate set as its FDP estimate and keeps the most useful operating point. Since this level is chosen after inspecting the data, it carries no validity guarantee.
- 2) *Split PH-CS* restores a guarantee through data splitting. It selects the level on one half of the calibration data, and runs conventional CS at that fixed level on the other half. Thus, the guarantee of Theorem 1 applies, but at the cost of calibrating on only part of the data.
- 3) *PAC PH-CS* replaces the in-expectation reliability of PH-CS with a high-probability bound that upper bounds the realized FDP of all candidate sets simultaneously [40], [41]. The bound holds with probability $1 - \eta$, but requires

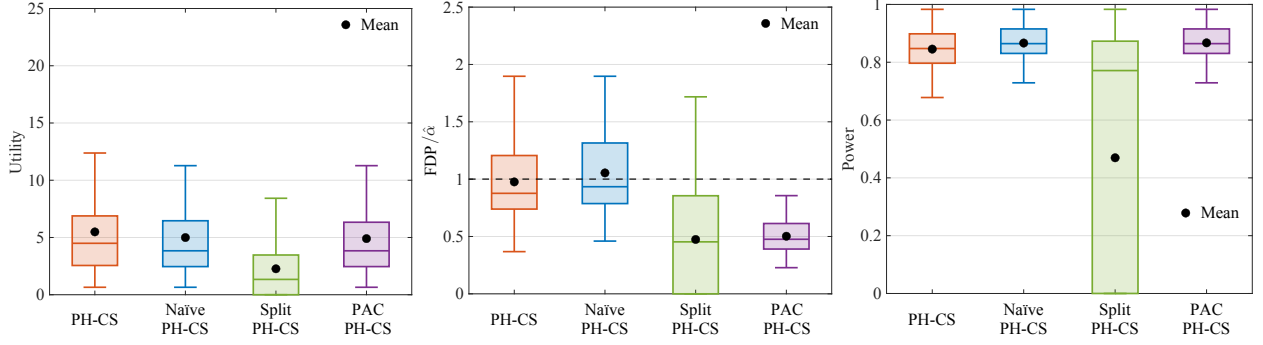


Fig. 2. Box plots of the realized utility (left), the ratio between the realized FDP (4) and the declared level $\hat{\alpha}$ (middle), and the power (right) under the additive trade-off utility (8) on the Recruitment dataset, for PH-CS and the three baselines of Sec. VI-B.

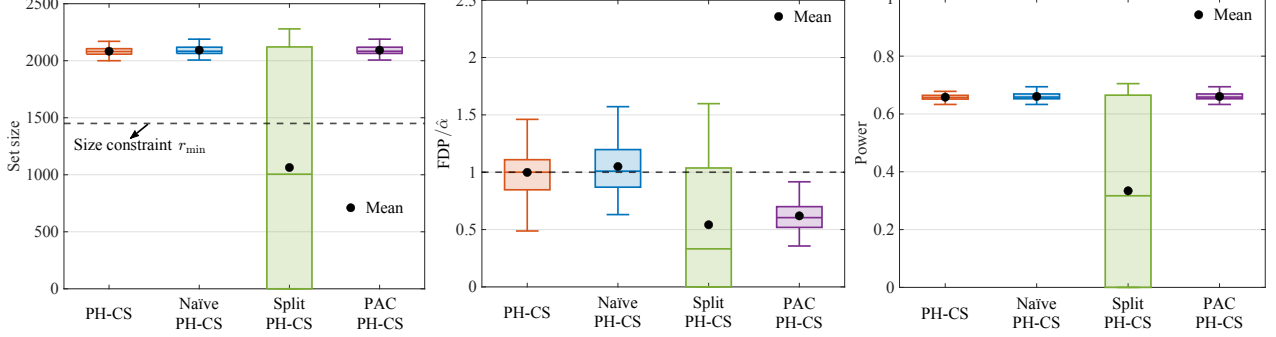


Fig. 3. Box plots of the selected set size (left), the ratio between the realized FDP (4) and the declared level $\hat{\alpha}$ (middle), and the power (right) under the constrained-size utility (7) on the Shuttle dataset, for PH-CS and the three baselines of Sec. VI-B.

the extra confidence level η and, being simultaneous over all sets, may be loose at the selected operating point.

The constructions of the three FDP estimates are detailed in Appendix F.

C. Comparison with Baselines

We compare PH-CS against the three post-hoc baselines of Sec. VI-B on the Recruitment dataset under the additive trade-off utility (8) with $U(r, \alpha) = \log r - \lambda \log(1/(1-\alpha))$, $r = |\mathcal{R}|$, and $\lambda = 15$, and on the Shuttle dataset under the constrained-size utility (7) with $r_{\min} = 0.1m$. The PAC PH-CS approach uses confidence level $\eta = 0.1$.

Fig. 2 and Fig. 3 report the achieved utility or selected set size, the ratio between the realized FDP (4) and the declared level $\hat{\alpha}$, and the power (36). PH-CS attains the highest utility, meets the size constraint in every run, and remains competitive in power. Its average ratio between the realized FDP and the declared level stays closest to one, so its FDP estimate satisfies the post-hoc reliability (12) and is also the tightest among all methods.

Naïve PH-CS instead reports the nominal level of the selected set as its FDP estimate, which is chosen by maximizing the utility and is no longer valid in advance. Its average ratio exceeds one, so the post-hoc reliability (12) no longer holds. Split PH-CS and PAC PH-CS keep the average ratio below one and remain reliable, but are overly conservative. Split PH-CS uses only half of the data for calibration, which weakens its selection, yielding the lowest utility on the Recruitment dataset and frequently violating the size constraint on the Shuttle dataset. PAC PH-CS upper bounds the FDP of all candidate sets at once, so its estimate is loose at the selected set. Overall,

PH-CS is the only method that is both reliable and effective for the selection task.

As PH-CS outperforms all three post-hoc baselines, we adopt it as our method and compare it with conventional CS [1] in the remaining experiments.

D. Constrained-Size Utility

Under the constrained-size utility in (7) with $r = |\mathcal{R}|$, the goal is to select, for each test batch, a set of test samples of size at least $r_{\min}$, while keeping the realized FDP as small as possible. In our experiments, we set $r_{\min} = 0.5m$ for synthetic data and $r_{\min} = 0.1m$ for real data, corresponding to selecting at least 50% and 10% of the test batch, respectively. The resulting histograms of the selected set size, realized FDP, and power are shown in Fig. 4 for the synthetic data with homoscedastic noise and the Musk dataset. To enable a fair comparison with CS [1], we set the target FDR level $\alpha_{\max}$ in (18)–(19) to the empirical average of the declared levels $\alpha^{\text{PH-CS}}$ produced by PH-CS over the 100 trials. This way, both PH-CS and CS guarantee the same FDR level, while PH-CS also controls the selected set size for test run.

In the left panels of Fig. 4, the selected sets produced by PH-CS are seen to always satisfy the minimum-size requirement, whereas the selected sets produced by CS fall below this threshold in a non-negligible fraction of realizations. At the same time, the middle panels show that PH-CS is generally competitive in terms of FDP, although the capacity for size control may come at the cost of a larger FDP on test batches, most notably on the Musk dataset. The right panels further show that PH-CS attains higher power, as it reliably selects a set of the required size while CS, tied to a single pre-specified level, often selects far fewer points. Overall, these results

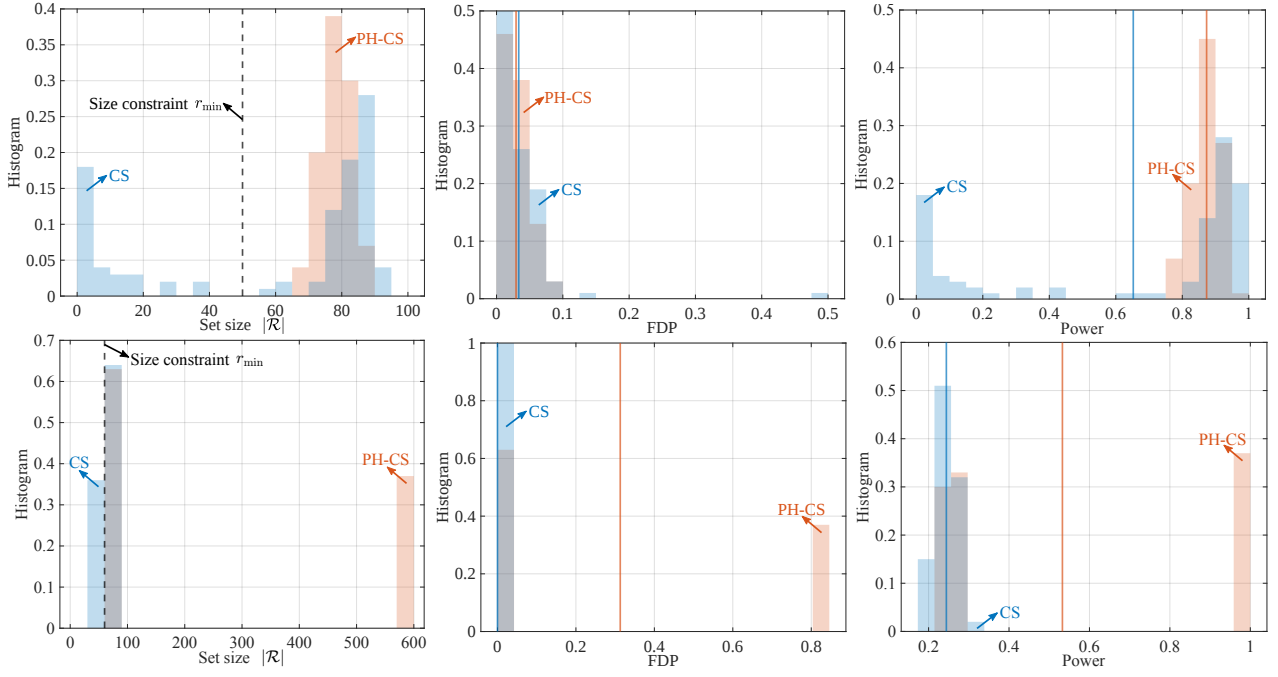


Fig. 4. Histograms of the selected set size (left), realized FDP (middle), and power (right) under the constrained-size utility (7) on the synthetic data with homoscedastic noise (top row) and the Musk dataset (bottom row). In the middle and right panels, the blue and red vertical lines indicate the average values for CS and PH-CS, respectively.

highlight the practical advantage of the post-hoc construction in (26). Unlike CS, which is tied to a single pre-specified FDR level, PH-CS can adaptively select an operating point so as to satisfy the size requirement $r_{\min}$ in each realization, while obtaining competitive FDR performance as compared to CS and offering reliable estimates of the FDP. This latter property is illustrated in Sec. VI-F.

E. Power-Aware Utility

We now consider two utilities that target the correct selections rather than the raw set size: (i) the multiplicative utility (9) with $U(r, \alpha) = r(1 - \alpha)$ and $r = |\mathcal{R}|$; and (ii) the additive utility (8) with $U(r, \alpha) = \log r - \lambda \log(1/(1 - \alpha))$, $\lambda = 15$, and the weighted size r given by (6), where π_j estimates the probability of the event $\{Y_{n+j} > c_j\}$ based on the model. Specifically, for classification, we set $\pi_j = \mu(X_{n+j})$, and for regression, we model $Y_{n+j} | X_{n+j} \sim \mathcal{N}(\mu(X_{n+j}), \sigma(X_{n+j})^2)$, with $\sigma(X)$ estimated from the training residuals, and set

$$\pi_j = 1 - \Phi\left(\frac{c_j - \mu(X_{n+j})}{\sigma(X_{n+j})}\right), \quad (37)$$

where Φ is the standard normal cumulative distribution function.

Fig. 5 reports the realized utility, power, and FDP for the multiplicative utility on the Shuttle dataset and the weighted utility on the Recruitment dataset, with CS matched to the average declared level as before. Under both utilities, PH-CS attains a higher average utility and a higher power than CS at a comparable FDP. This shows that, by rewarding the estimated correct selections, the utility can be steered toward power while the FDP remains controlled.

F. Empirical Evaluation of the FDP Estimate

A key property of PH-CS is its capacity to provide reliable conservative estimates $\alpha^{\text{PH-CS}}$ of the FDP, as formalized by (28)–(13). To illustrate this, Fig. 6 reports the relationship between the realized FDP ($\mathcal{R}^{\text{PH-CS}}, \mathcal{Y}^{\text{test}}$) and the declared level $\alpha^{\text{PH-CS}}$ over 100 trials via a scatter plot evaluated on the Shuttle dataset under the constrained-size utility (7) and on the Recruitment dataset under the additive trade-off utility (8).

The main observation is that, in both cases, the joint distribution of true and estimated FDP is concentrated around the diagonal, indicating that the declared level $\alpha^{\text{PH-CS}}$ tracks the realized FDP reasonably well across realizations. Some points lie above the diagonal, meaning that the realized FDP can occasionally exceed the declared level, but the average estimated FDP $\mathbb{E}[\alpha^{\text{PH-CS}}]$ is larger than the true FDR $\text{FDR}(\mathcal{R}^{\text{PH-CS}})$, confirming the theoretical results in (28) and (13).

VII. CONCLUSION

This paper has introduced post-hoc conformal selection (PH-CS), a framework that extends conventional CS by allowing the user to choose the operating point after observing the data while retaining a finite-sample reliability guarantee. PH-CS replaces conformal p-variables and the BH procedure with conformal e-variables and the e-BH procedure, exploiting the level-uniform property of e-BH to generate a path of candidate selection sets, each paired with a data-driven FDP estimate. The user then selects the operating point that maximizes an arbitrary utility balancing set size and reliability. We proved that the reported FDP estimate $\alpha^{\text{PH-CS}}$ satisfies the post-hoc reliability condition $\mathbb{E}[\text{FDP}/\alpha^{\text{PH-CS}}] \leq 1$ under exchangeability and score monotonicity, and extended the framework to continuous risk control (PH-RCS) and priority-weighted selection. Experiments on synthetic and real-world datasets

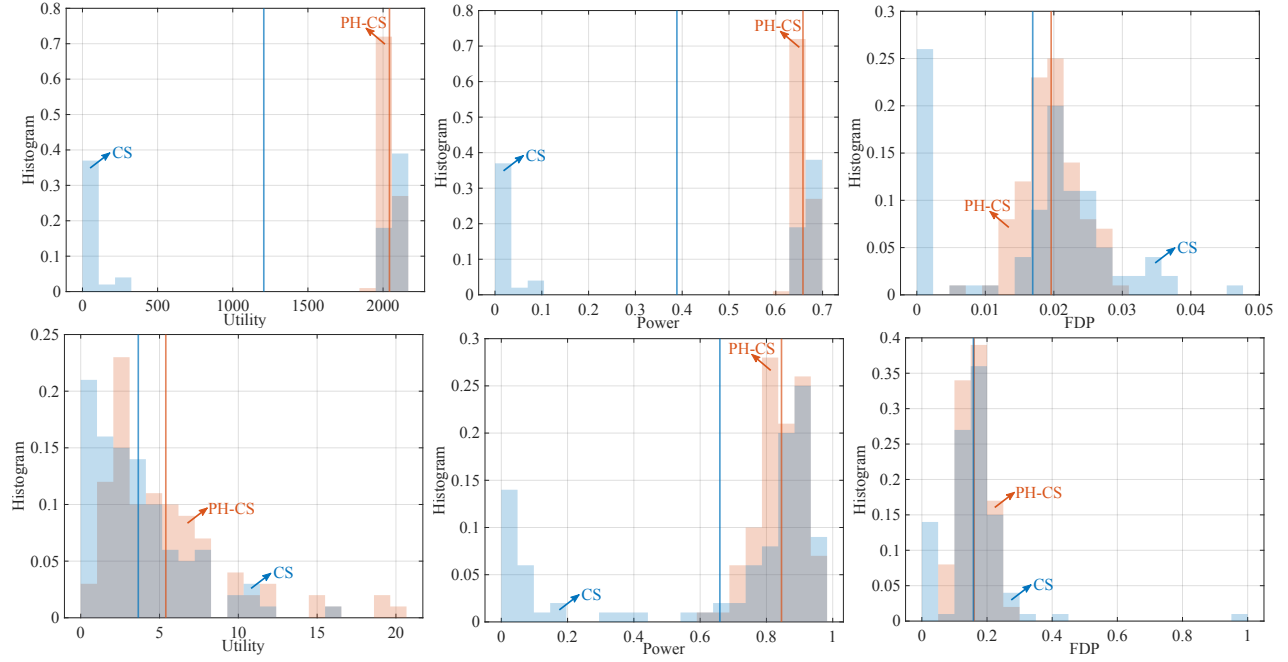


Fig. 5. Histograms of the realized utility (left), power (middle), and FDP (right) under the multiplicative utility (9) on the Shuttle dataset (top row) and the additive utility (8) with weighted set size (6) on the Recruitment dataset (bottom row). In each panel, the blue and red vertical lines indicate the average values for CS and PH-CS, respectively.

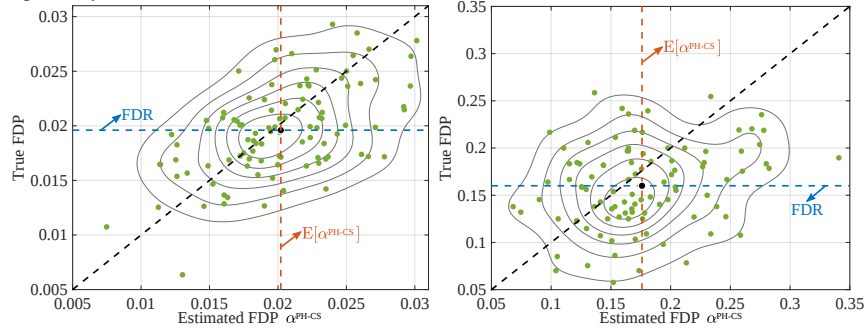


Fig. 6. Scatter plots of the realized FDP, $FDP(\mathcal{R}^{PH-CS}, \mathcal{Y}^{test})$ in (4), versus the estimated level α^{PH-CS} (27) over 100 random seeds under the constrained-size utility (7) on the Shuttle dataset (left) and the additive trade-off utility (8) on the Recruitment dataset (right). Contour lines show the kernel density of the joint distribution. The black dashed line is the reference at which estimated and true FDP coincide, while the vertical red and horizontal blue dashed lines mark the average estimate $\mathbb{E}[\alpha^{PH-CS}]$ and the true FDR, $FDR(R^{PH-CS})$, respectively.

confirmed that PH-CS consistently satisfies user-imposed size and utility constraints that CS cannot enforce, while maintaining competitive FDR control and providing reliable FDP estimates across different regressors. Overall, PH-CS presents a principled solution for utility-driven selection with post-hoc flexibility, establishing a foundation for further research in data-adaptive conformal inference.

Future work could explore extensions of PH-CS to settings with distribution shift between calibration and test data [12], [16], [43], the use of optimized e-variable designs [47] and improved e-BH procedures [48], and the design of level-independent boosting strategies for conformal e-variables that improve power while preserving post-hoc validity.

REFERENCES

- [1] Y. Jin and E. J. Candès, “Selection by prediction with conformal p-values,” *J. Mach. Learn. Res.*, vol. 24, no. 244, pp. 1–41, 2023.
- [2] D. Boldini, L. Friedrich *et al.*, “Machine learning assisted hit prioritization for high throughput screening in drug discovery,” *ACS Cent. Sci.*, vol. 10, no. 4, pp. 823–832, 2024.
- [3] T. Bai, P. Tang *et al.*, “Conformal selection for efficient and accurate compound screening in drug discovery,” *J. Chem. Inf. Model.*, vol. 65, no. 24, pp. 13 070–13 085, 2025.
- [4] K. Korthauer, P. K. Kimes *et al.*, “A practical guide to methods controlling false discoveries in computational biology,” *Genome Biol.*, vol. 20, no. 1, 2019.
- [5] C. Dai, B. Lin *et al.*, “False discovery rate control via data splitting,” *J. Am. Stat. Assoc.*, vol. 118, no. 544, pp. 2503–2520, 2023.
- [6] P. Stoica and P. Babu, “False discovery rate (FDR) and familywise error rate (FER) rules for model selection in signal processing applications,” *IEEE Open J. Signal Process.*, vol. 3, pp. 403–416, 2022.
- [7] V. Vovk, A. Gammerman *et al.*, *Algorithmic learning in a random world*. Springer, 2005, vol. 29.
- [8] A. N. Angelopoulos and S. Bates, “Conformal prediction: A gentle introduction,” *Found. Trends Mach. Learn.*, vol. 16, no. 4, pp. 494–591, 2023.
- [9] C. Xu and Y. Xie, “Conformal prediction for time series,” *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 45, no. 10, pp. 11 575–11 587, 2023.
- [10] S. Park, K. M. Cohen *et al.*, “Few-shot calibration of set predictors via meta-learned cross-validation-based conformal prediction,” *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 46, no. 1, pp. 280–291, 2024.
- [11] V. Jensen, F. M. Bianchi *et al.*, “Ensemble conformalized quantile regression for probabilistic time series forecasting,” *IEEE Trans. Neural Netw. Learn. Syst.*, vol. 35, no. 7, pp. 9014–9025, 2024.
- [12] R. F. Barber and R. J. Tibshirani, “Unifying different theories of conformal prediction,” *arXiv preprint arXiv:2504.02292*, 2025.

- [13] Y. Benjamini and Y. Hochberg, “Controlling the false discovery rate: a practical and powerful approach to multiple testing,” *J. R. Stat. Soc. Ser. B Methodol.*, vol. 57, no. 1, pp. 289–300, 1995.
- [14] Y. Benjamini and D. Yekutieli, “The control of the false discovery rate in multiple testing under dependency,” *Ann. Stat.*, vol. 29, no. 4, pp. 1165–1188, 2001.
- [15] S. Bates, E. Candès *et al.*, “Testing for outliers with conformal p-values,” *Ann. Stat.*, vol. 51, no. 1, pp. 149–178, 2023.
- [16] Y. Jin, Z. Ren *et al.*, “Sensitivity analysis of individual treatment effects: A robust conformal inference approach,” *Proc. Natl. Acad. Sci. U.S.A.*, vol. 120, no. 6, p. e2214889120, 2023.
- [17] T. Bai, Y. Zhao *et al.*, “Multivariate conformal selection,” in *Proc. Int. Conf. Mach. Learn.*, vol. 267, 2025, pp. 2535–2559.
- [18] Q. Hao, W. Liao *et al.*, “Multi-condition conformal selection,” *arXiv preprint arXiv:2510.08075*, 2025.
- [19] T. Bai and Y. Jin, “Optimized conformal selection: Powerful selective inference after conformity score optimization,” *arXiv preprint arXiv:2411.17983*, 2024.
- [20] J. Qin, Y. Liu *et al.*, “Revamping conformal selection with optimal power: A Neyman–Pearson perspective,” *arXiv preprint arXiv:2502.16513*, 2025.
- [21] K. Liu, H. Xi *et al.*, “Online conformal selection with accept-to-reject changes,” in *Proc. AAAI Conf. Artif. Intell.*, vol. 40, no. 28, 2026, pp. 23 765–23 773.
- [22] L. Lu, Y. Huo *et al.*, “Feedback-enhanced online multiple testing with applications to conformal selection,” *arXiv preprint arXiv:2509.03297*, 2025.
- [23] Y. Gui, Y. Jin *et al.*, “ACS: An interactive framework for conformal selection,” *arXiv preprint arXiv:2507.15825*, 2025.
- [24] Y. Nair, Y. Jin *et al.*, “Diversifying conformal selections,” *arXiv preprint arXiv:2506.16229*, 2025.
- [25] V. Vovk and R. Wang, “E-values: Calibration, combination and applications,” *Ann. Stat.*, vol. 49, no. 3, pp. 1736–1754, 2021.
- [26] P. Grünwald, R. de Heide *et al.*, “Safe testing,” *J. R. Stat. Soc. Ser. B Stat. Methodol.*, vol. 86, no. 5, pp. 1091–1128, 2024.
- [27] N. W. Koning, “Post-hoc α hypothesis testing and the post-hoc p -value,” *arXiv preprint arXiv:2312.08040*, 2025.
- [28] R. Wang and A. Ramdas, “False discovery rate control with e-values,” *J. R. Stat. Soc. Ser. B Stat. Methodol.*, vol. 84, no. 3, pp. 822–852, 2022.
- [29] A. A. Balinsky and A. D. Balinsky, “Enhancing conformal prediction using e-test statistics,” *arXiv preprint arXiv:2403.19082*, 2024.
- [30] M. Bashari, A. Epstein *et al.*, “Derandomized novelty detection with FDR control via conformal e-values,” *Adv. Neural Inf. Process. Syst.*, vol. 36, pp. 65 585–65 596, 2023.
- [31] J. Lee and Z. Ren, “Boosting e-BH via conditional calibration,” *arXiv preprint arXiv:2404.17562*, 2024.
- [32] N. Ignatiadis, R. Wang *et al.*, “Asymptotic and compound e-values: multiple testing and empirical Bayes,” *arXiv preprint arXiv:2409.19812*, 2024.
- [33] Z. Xu and A. Ramdas, “More powerful multiple testing under dependence via randomization,” *arXiv preprint arXiv:2305.11126*, 2023.
- [34] J. J. Goeman and A. Solari, “Multiple testing for exploratory research,” *Statist. Sci.*, vol. 26, no. 4, pp. 584–597, 2011.
- [35] G. Blanchard, P. Neuvial *et al.*, “Post hoc confidence bounds on false positives using reference families,” *Ann. Stat.*, vol. 48, no. 3, pp. 1281–1303, 2020.
- [36] J. Hemerik, A. Solari *et al.*, “Permutation-based simultaneous confidence bounds for the false discovery proportion,” *Biometrika*, vol. 106, no. 3, pp. 635–649, 2019.
- [37] Z. Xu, R. Wang *et al.*, “Post-selection inference for e-value based confidence intervals,” *Electron. J. Stat.*, vol. 18, no. 1, pp. 2292–2338, 2024.
- [38] E. Gauthier, F. Bach *et al.*, “Backward conformal prediction,” *arXiv preprint arXiv:2505.13732*, 2025.
- [39] E. Gauthier, F. Bach *et al.*, “E-values expand the scope of conformal prediction,” *arXiv preprint arXiv:2503.13050*, 2025.
- [40] U. Gazin, G. Blanchard *et al.*, “Transductive conformal inference with adaptive scores,” in *Proc. Int. Conf. Artif. Intell. Stat.*, 2024, pp. 1504–1512.
- [41] Z. Song, Y. Jin *et al.*, “Everywhere valid bounds on false discovery proportions in conformal inference,” *arXiv preprint arXiv:2605.20726*, 2026.
- [42] Y. Lee and Z. Ren, “Selection from hierarchical data with conformal e-values,” *arXiv preprint arXiv:2501.02514*, 2025.
- [43] T. Bai and Y. Jin, “Conformal selective prediction with general risk control,” *arXiv preprint arXiv:2603.24704*, 2026.
- [44] G. De Toni, N. Okati *et al.*, “Towards human-AI complementarity with prediction sets,” *Adv. Neural Inf. Process. Syst.*, vol. 37, pp. 31 380–31 409, 2024.
- [45] J. G. Richens, C. M. Lee *et al.*, “Improving the accuracy of medical diagnosis with causal machine learning,” *Nat. Commun.*, vol. 11, no. 1, p. 3923, 2020.
- [46] P. Carracedo-Reboredo, J. Liñares-Blanco *et al.*, “A review on machine learning approaches and trends in drug discovery,” *Comput. Struct. Biotechnol. J.*, vol. 19, pp. 4538–4558, 2021.
- [47] N. W. Koning and S. van Meer, “Fuzzy prediction sets: Conformal prediction with e-values,” *arXiv preprint arXiv:2509.13130*, 2025.
- [48] Z. Xu, A. Solari *et al.*, “Bringing closure to false discovery rate control: A general principle for multiple testing,” *arXiv preprint arXiv:2509.02517*, 2025.

APPENDIX

A. Link Between Post-hoc Reliability and FDR Control

For completeness, we provide a short derivation supporting the relation (13), which links the post-hoc reliability condition (12) to an approximate upper bound on the FDR in terms of $\mathbb{E}[\hat{\alpha}(\mathcal{R})]$. The approximation is most informative when the estimate $\hat{\alpha}(\mathcal{R})$ has a sufficiently small variance.

A first-order Taylor expansion of $1/\hat{\alpha}(\mathcal{R})$ around $\mathbb{E}[\hat{\alpha}(\mathcal{R})]$ yields

$$\frac{1}{\hat{\alpha}(\mathcal{R})} = \frac{1}{\mathbb{E}[\hat{\alpha}(\mathcal{R})]} - \frac{\hat{\alpha}(\mathcal{R}) - \mathbb{E}[\hat{\alpha}(\mathcal{R})]}{\mathbb{E}[\hat{\alpha}(\mathcal{R})]^2} + O((\hat{\alpha}(\mathcal{R}) - \mathbb{E}[\hat{\alpha}(\mathcal{R})])^2). \quad (38)$$

Substituting (38) into (12) and neglecting the higher-order term yields

$$\begin{aligned} \mathbb{E}\left[\frac{\text{FDP}(\mathcal{R}, \mathcal{Y}^{\text{test}})}{\hat{\alpha}(\mathcal{R})}\right] &\approx \frac{\mathbb{E}[\text{FDP}(\mathcal{R}, \mathcal{Y}^{\text{test}})]}{\mathbb{E}[\hat{\alpha}(\mathcal{R})]} \\ &\quad - \frac{\mathbb{E}[\text{FDP}(\mathcal{R}, \mathcal{Y}^{\text{test}})(\hat{\alpha}(\mathcal{R}) - \mathbb{E}[\hat{\alpha}(\mathcal{R})])]}{\mathbb{E}[\hat{\alpha}(\mathcal{R})]^2} \\ &\approx \frac{\mathbb{E}[\text{FDP}(\mathcal{R}, \mathcal{Y}^{\text{test}})]}{\mathbb{E}[\hat{\alpha}(\mathcal{R})]} = \frac{\text{FDR}(R)}{\mathbb{E}[\hat{\alpha}(\mathcal{R})]} \leq 1, \quad (39) \end{aligned}$$

which yields the approximate bound (13).

B. Proof of Theorem 2

We first introduce an *oracle* counterpart of the conformal e-variable in (21), obtained by evaluating the score at the unobserved test label Y_{n+j} , i.e.,

$$E_j^* = \frac{S_{n+j}}{\frac{1}{n+1}(\sum_{i=1}^n S_i + S_{n+j})}, \quad j = 1, \dots, m, \quad (40)$$

where $S_{n+j} = S(X_{n+j}, Y_{n+j})$ for $j = 1, \dots, m$, and we assume $\sum_{i=1}^n S_i + S_{n+j} > 0$ almost surely since the score is non-negative as defined in Sec. III-A.

Step 1: $\mathbb{E}[E_j^*] = 1$

By assumption (i) in Theorem 2, for each fixed j , the collection $\{S_i\}_{i=1}^n \cup \{S_{n+j}\}$ is exchangeable. Hence

$$\begin{aligned} \mathbb{E}[E_j^*] &= \mathbb{E}\left[\frac{(n+1)S_{n+j}}{\sum_{i=1}^n S_i + S_{n+j}}\right] \\ &= \frac{1}{n+1} \mathbb{E}\left[\sum_{k \in \{1, \dots, n, n+j\}} \frac{(n+1)S_k}{\sum_{i=1}^n S_i + S_{n+j}}\right] \\ &= \frac{1}{n+1} \mathbb{E}[n+1] = 1, \quad (41) \end{aligned}$$

where the second equality uses exchangeability to replace the expectation of the single term involving S_{n+j} by the average of all $n+1$ symmetric terms $S_1, \dots, S_n, S_{n+j}$. Based on this, we apply the following steps.

Step 2: Under the null H_j , $E_j \leq E_j^*$

To compare E_j in (21) with its oracle counterpart E_j^* in (40), we study how the e-variable changes with its score input while keeping the calibration sum fixed.

Fix $a = \sum_{i=1}^n S_i$ and define the scalar mapping

$$f(S) = \frac{(n+1)S}{a+S}, \quad (42)$$

which is the functional form of the e-variable as a function of the score input S . A direct calculation gives

$$f'(S) = \frac{a(n+1)}{(a+S)^2} \geq 0, \quad (43)$$

so $f(S)$ is non-decreasing in S . Consequently, for each j , we have

$$E_j = f(\hat{S}_{n+j}) \text{ and } E_j^* = f(S_{n+j}). \quad (44)$$

By assumption (ii) in Theorem 2, the score $S(X, Y)$ is monotone non-increasing in Y . Hence, under the null hypothesis H_j in (15), i.e., on the event $\{Y_{n+j} \leq c_j\}$, we have

$$S_{n+j} = S(X_{n+j}, Y_{n+j}) \geq S(X_{n+j}, c_j) = \hat{S}_{n+j}. \quad (45)$$

Since function $f(S)$ is non-decreasing in score S , this implies

$$E_j \leq E_j^* \text{ under } H_j \text{ for } j = 1, \dots, m. \quad (46)$$

Step 3: Post-hoc reliability via e-BH consistency

We now adapt the proof strategy of [1, Theorem 3] to the e-variable and e-BH setting. Fix any $\alpha \in (0, 1)$ and let $\mathcal{R} = \mathcal{R}(\alpha)$ denote the rejection set obtained by the e-BH procedure (22)–(23). The selection set $\mathcal{R}$ is then self-consistent with threshold $m/(\alpha|\mathcal{R}|)$, in the sense that

$$j \in \mathcal{R} \iff E_j \geq \frac{m}{\alpha|\mathcal{R}|}. \quad (47)$$

Expanding the FDP (4) using (47), and noting that $E_j \leq E_j^*$ on the event $\{Y_{n+j} \leq c_j\}$ by (46) and that $\mathbb{1}\{Y_{n+j} \leq c_j\} \leq 1$, we obtain

$$\begin{aligned} \text{FDP}(\mathcal{R}, \mathcal{Y}^{\text{test}}) &= \frac{\sum_{j=1}^m \mathbb{1}\{Y_{n+j} \leq c_j\} \mathbb{1}\{j \in \mathcal{R}\}}{\max\{1, |\mathcal{R}|\}} \\ &\leq \frac{1}{|\mathcal{R}|} \sum_{j=1}^m \mathbb{1}\{Y_{n+j} \leq c_j\} \mathbb{1}\left\{E_j^* \geq \frac{m}{\alpha|\mathcal{R}|}\right\} \\ &\leq \frac{1}{|\mathcal{R}|} \sum_{j=1}^m \mathbb{1}\left\{E_j^* \geq \frac{m}{\alpha|\mathcal{R}|}\right\}. \end{aligned} \quad (48)$$

Applying the elementary bound $\mathbb{1}\{E \geq t\} \leq E/t$ for $E \geq 0$ and $t > 0$ with $E = E_j^*$ and $t = m/(\alpha|\mathcal{R}|)$, and noting that $|\mathcal{R}|$ cancels, gives

$$\text{FDP}(\mathcal{R}, \mathcal{Y}^{\text{test}}) \leq \frac{1}{|\mathcal{R}|} \sum_{j=1}^m \frac{\alpha|\mathcal{R}|}{m} E_j^* = \frac{\alpha}{m} \sum_{j=1}^m E_j^*. \quad (49)$$

Dividing both sides by α yields the level-uniform bound

$$\frac{\text{FDP}(\mathcal{R}, \mathcal{Y}^{\text{test}})}{\alpha} \leq \frac{1}{m} \sum_{j=1}^m E_j^* \quad (50)$$

for all $\alpha \in (0, 1)$. Taking expectations and using $\mathbb{E}[E_j^*] = 1$ from (41) gives

$$\mathbb{E}\left[\frac{\text{FDP}(\mathcal{R}, \mathcal{Y}^{\text{test}})}{\alpha}\right] \leq 1. \quad (51)$$

Since the right-hand side of (50) depends on neither α nor the thresholds $\{c_j\}_{j=1}^m$, the bound holds when α is replaced by any data-dependent choice measurable with respect to $(\mathcal{D}^{\text{cal}}, \mathcal{X}^{\text{test}})$, and remains valid even if the thresholds $\{c_j\}_{j=1}^m$ are themselves functions of $(\mathcal{D}^{\text{cal}}, \mathcal{X}^{\text{test}})$. In particular, taking $\alpha = \alpha^{\text{PH-CS}}$ and $\mathcal{R} = \mathcal{R}(\alpha^{\text{PH-CS}}) = \mathcal{R}^{\text{PH-CS}}$ yields (28).

The key step enabling post-hoc validity is the inequality $\mathbb{1}\{E \geq t\} \leq E/t$, which yields a bound free of α (see (50)). In contrast, conventional CS relies on the tail calibration $\mathbb{P}(P \leq t) \leq t$, whose analogue would require $\mathbb{E}[1/P] < \infty$, which fails even for $P \sim \text{U}(0, 1)$. This is why CS guarantees are tied to a pre-specified level and cannot extend to post-hoc selection.

C. Proof of Proposition 1

By the definition of the FDP (4), the claim (29) requires the realized FDP of the selected set $\mathcal{R}^{\text{PH-CS}}$ to stay below the reported level $\alpha^{\text{PH-CS}}$ in large samples. We obtain this from the level-uniform bound (50) in Appendix B, which holds for every level $\alpha \in (0, 1)$ and for the data-dependent level $\alpha^{\text{PH-CS}}$. Evaluated at $\alpha = \alpha^{\text{PH-CS}}$ and $\mathcal{R} = \mathcal{R}^{\text{PH-CS}}$, it reads

$$\text{FDP}(\mathcal{R}^{\text{PH-CS}}, \mathcal{Y}^{\text{test}}) \leq \alpha^{\text{PH-CS}} \cdot \frac{1}{m} \sum_{j=1}^m E_j^*, \quad (52)$$

where E_j^* is the oracle e-variable (40). Since $\alpha^{\text{PH-CS}} \leq 1$ by (25), this yields the deterministic inequality

$$\text{FDP}(\mathcal{R}^{\text{PH-CS}}, \mathcal{Y}^{\text{test}}) \leq \alpha^{\text{PH-CS}} + \left| \frac{1}{m} \sum_{j=1}^m E_j^* - 1 \right|. \quad (53)$$

By (53), the event $\{|\sum_{j=1}^m E_j^*/m - 1| \leq \varepsilon\}$ is contained in $\{\text{FDP}(\mathcal{R}^{\text{PH-CS}}, \mathcal{Y}^{\text{test}}) \leq \alpha^{\text{PH-CS}} + \varepsilon\}$, so for every $\varepsilon > 0$,

$$\begin{aligned} &\mathbb{P}\left(\text{FDP}(\mathcal{R}^{\text{PH-CS}}, \mathcal{Y}^{\text{test}}) \leq \alpha^{\text{PH-CS}} + \varepsilon\right) \\ &\geq \mathbb{P}\left(\left|\frac{1}{m} \sum_{j=1}^m E_j^* - 1\right| \leq \varepsilon\right). \end{aligned} \quad (54)$$

Since this probability is also at most one, it therefore suffices to prove that, for every $\varepsilon > 0$,

$$\lim_{n, m \rightarrow \infty} \mathbb{P}\left(\left|\frac{1}{m} \sum_{j=1}^m E_j^* - 1\right| \leq \varepsilon\right) = 1. \quad (55)$$

By definition, the score is non-negative, so dropping the term S_{n+j} from the denominator of the oracle e-variable (40) only enlarges it. This yields, for each j ,

$$0 \leq \frac{S_{n+j}}{\frac{1}{n+1} \sum_{i=1}^n S_i} - E_j^* = \frac{(n+1)S_{n+j}^2}{\left(\sum_{i=1}^n S_i\right)\left(\sum_{i=1}^n S_i + S_{n+j}\right)} \leq \frac{(n+1)S_{n+j}^2}{\left(\sum_{i=1}^n S_i\right)^2}, \quad (56)$$

where the last step again drops S_{n+j} from a denominator. Averaging (56) over $j = 1, \dots, m$, the gap

$$R_{n,m} = \frac{1}{m} \cdot \frac{\sum_{j=1}^m S_{n+j}}{\frac{1}{n+1} \sum_{i=1}^n S_i} - \frac{1}{m} \sum_{j=1}^m E_j^* \quad (57)$$

satisfies

$$0 \leq R_{n,m} \leq \frac{n+1}{\left(\sum_{i=1}^n S_i\right)^2} \cdot \frac{1}{m} \sum_{j=1}^m S_{n+j}^2. \quad (58)$$

By assumption (i) in Proposition 1, the scores are i.i.d, so for every $\varepsilon > 0$ the law of large numbers gives

$$\lim_{n \rightarrow \infty} \mathbb{P}\left(\left|\frac{1}{n} \sum_{i=1}^n S_i - \mathbb{E}[S(X, Y)]\right| > \frac{\varepsilon}{2}\right) = 0, \quad (59a)$$

$$\lim_{m \rightarrow \infty} \mathbb{P}\left(\left|\frac{1}{m} \sum_{j=1}^m S_{n+j} - \mathbb{E}[S(X, Y)]\right| > \frac{\varepsilon}{2}\right) = 0, \quad (59b)$$

with $\mathbb{E}[S(X, Y)] > 0$ by assumption (iii). Since the first term in (57) is the ratio of these two averages, (59) gives

$$\lim_{n, m \rightarrow \infty} \mathbb{P}\left(\left|\frac{1}{m} \cdot \frac{\sum_{j=1}^m S_{n+j}}{\frac{1}{n+1} \sum_{i=1}^n S_i} - 1\right| > \frac{\varepsilon}{2}\right) = 0. \quad (60)$$

For the gap, the finite second moment of assumption (iii) keeps $\sum_{j=1}^m S_{n+j}^2/m$ bounded, while (59) makes $\sum_{i=1}^n S_i$ grow as $n \mathbb{E}[S(X, Y)]$, so the upper bound in (58) gives

$$\lim_{n, m \rightarrow \infty} \mathbb{P}\left(R_{n,m} > \frac{\varepsilon}{2}\right) = 0. \quad (61)$$

By (57) and the triangle inequality, $|\sum_{j=1}^m E_j^*/m - 1|$ is at most the sum of the two deviations bounded in (60) and (61). These two limits therefore establish (55), which with (54) proves (29).

D. Proof of Theorem 3

The proof follows the same strategy as Step 3 of the proof of Theorem 2 in Appendix B, but replaces the oracle argument in Steps 1–2 with the risk-adjusted e-variable condition (33).

Fix $\alpha \in (0, 1)$ and let $\mathcal{R} = \mathcal{R}(\alpha)$ denote the e-BH output with E_j replaced by E_j^g . As in (47), the self-consistency of e-BH gives

$$j \in \mathcal{R} \iff E_j^g \geq \frac{m}{\alpha|\mathcal{R}|}. \quad (62)$$

Applying (62) and the elementary bound $1\{E \geq t\} \leq E/t$ to the generalized FDP (31), we obtain

$$\begin{aligned} \text{FDP}^g(\mathcal{R}, \mathcal{Y}^{\text{test}}) &= \frac{\sum_{j=1}^m L_j 1\{j \in \mathcal{R}\}}{\max\{1, |\mathcal{R}|\}} \\ &\leq \frac{1}{|\mathcal{R}|} \sum_{j=1}^m L_j \frac{\alpha|\mathcal{R}|}{m} E_j^g = \frac{\alpha}{m} \sum_{j=1}^m L_j E_j^g. \end{aligned} \quad (63)$$

Dividing both sides by α yields the level-uniform bound

$$\frac{\text{FDP}^g(\mathcal{R}, \mathcal{Y}^{\text{test}})}{\alpha} \leq \frac{1}{m} \sum_{j=1}^m L_j E_j^g. \quad (64)$$

Taking expectations and applying (33) gives

$$\mathbb{E}\left[\frac{\text{FDP}^g(\mathcal{R}, \mathcal{Y}^{\text{test}})}{\alpha}\right] \leq \frac{1}{m} \sum_{j=1}^m \mathbb{E}[L_j E_j^g] \leq 1. \quad (65)$$

Since the right-hand side of (64) does not depend on α , substituting $\alpha = \alpha^{\text{PH-RCS}}$ and $\mathcal{R} = \mathcal{R}^{\text{PH-RCS}}$ yields (34).

E. Priority-Weighted Selection

In many applications, test inputs are not equally important. For instance, in drug discovery certain molecular scaffolds may warrant closer examination, or in clinical settings high-risk patients may deserve priority [28], [32]. Assigning higher weights to more promising inputs also serves as a power-boosting mechanism by concentrating the testing budget where it is most needed [28], [32]. To incorporate such preferences while preserving the post-hoc reliability guarantee, we allow the user to assign a nonnegative weight w_j to each test input X_{n+j} , subject to the budget constraint

$$\sum_{j=1}^m w_j \leq m. \quad (66)$$

The weights are applied to risk-adjusted e-variables $\{E_j^g\}_{j=1}^m$ (see (33)), yielding weighted e-variables [28], [32]

$$\tilde{E}_j^g = w_j E_j^g, \quad j = 1, \dots, m. \quad (67)$$

Algorithm 1 is then applied with the e-variables $\{E_j^g\}_{j=1}^m$ replaced by the quantities $\{\tilde{E}_j^g\}_{j=1}^m$ in (67), producing a weighted candidate path and a utility-selected operating point $(\mathcal{R}^{\text{PH-RCS}}, \alpha^{\text{PH-RCS}})$. The following result confirms that the post-hoc guarantee of Theorem 3 is preserved under weighting.

Corollary 1 (Weighted post-hoc guarantee). *Let $\{w_j\}_{j=1}^m$ be nonnegative weights satisfying (66). Then, under the same assumption of Theorem 3, the output of Algorithm 1 with E_j^g replaced by $\tilde{E}_j^g$ in (67) satisfies the inequality*

$$\mathbb{E}\left[\frac{\text{FDP}^g(\mathcal{R}^{\text{PH-RCS}}, \mathcal{Y}^{\text{test}})}{\alpha^{\text{PH-RCS}}}\right] \leq 1. \quad (68)$$

Proof: By the proof of Theorem 3, the level-uniform bound (64) holds under (33). Replacing E_j^g by $\tilde{E}_j^g = w_j E_j^g$ in Algorithm 1 replaces each summand $L_j E_j^g$ in (64) by $w_j L_j E_j^g$, giving

$$\frac{\text{FDP}^g(\mathcal{R}, \mathcal{Y}^{\text{test}})}{\alpha} \leq \frac{1}{m} \sum_{j=1}^m w_j L_j E_j^g. \quad (69)$$

Taking expectations and applying the budget constraint (66) yields

$$\mathbb{E}\left[\frac{\text{FDP}^g(\mathcal{R}, \mathcal{Y}^{\text{test}})}{\alpha}\right] \leq \frac{1}{m} \sum_{j=1}^m w_j \mathbb{E}[L_j E_j^g] \leq \frac{1}{m} \sum_{j=1}^m w_j \leq 1. \quad (70)$$

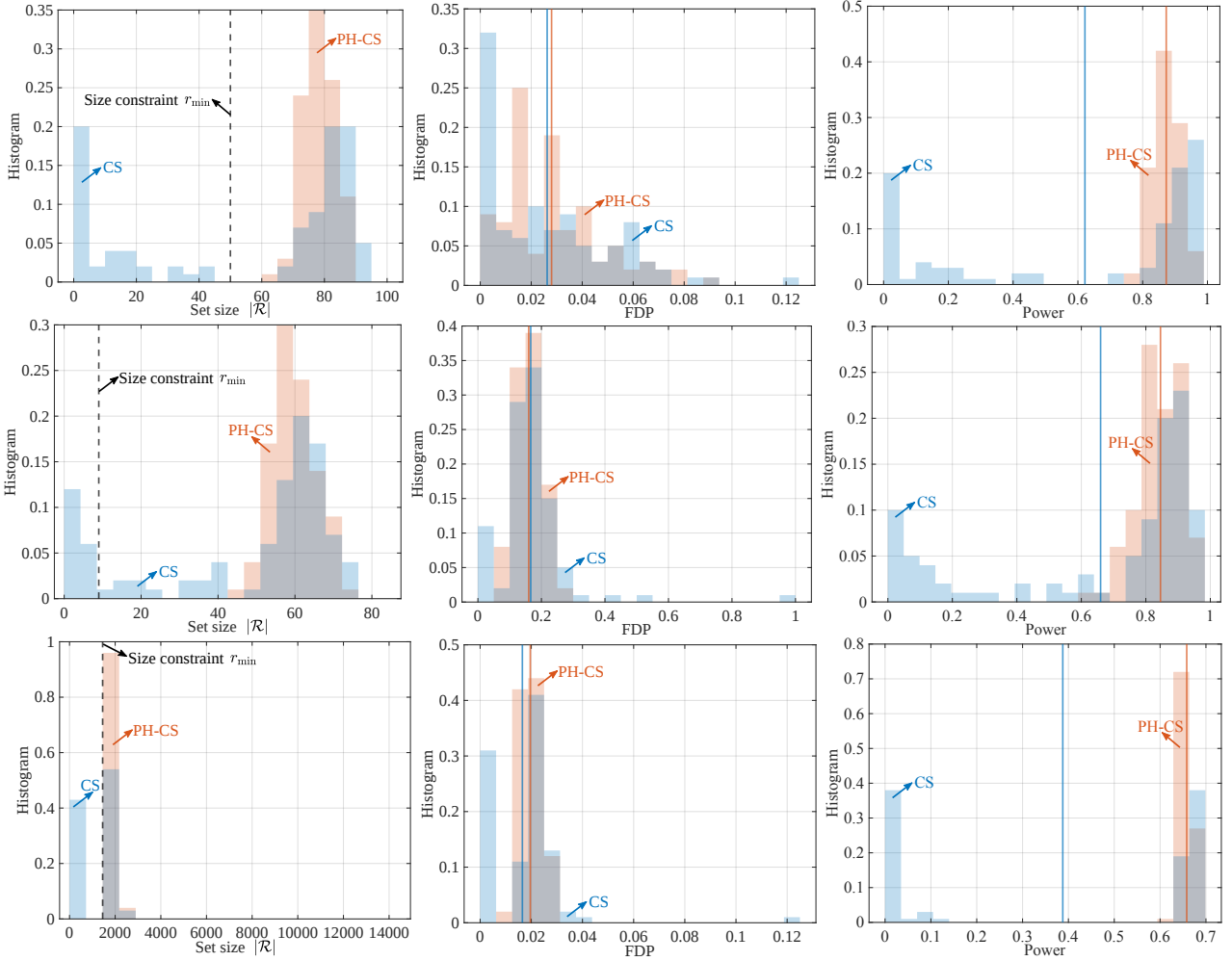


Fig. 7. Histograms of the selected set size (left), realized FDP (middle), and power (right) under the constrained-size utility (7) on the synthetic data with heteroscedastic noise using gradient boosting (top row), the Recruitment dataset (middle row), and the Shuttle dataset (bottom row). In the middle and right panels, the blue and red vertical lines indicate the average values for CS and PH-CS, respectively.

Since the right-hand side of (69) is free of α , substituting $\alpha = \alpha^{\text{PH-RCS}}$ and $\mathcal{R} = \mathcal{R}^{\text{PH-RCS}}$ yields (68).

When the loss is selected as $L_j = \mathbb{1}\{Y_{n+j} \leq c_j\}$, the guarantee (68) recovers Theorem 2.

F. Details of the Baseline Methods

This appendix details the three baselines introduced in Sec. VI-B. All three baselines build on the nested family of selected sets obtained from the conformal p-variables (17) via the BH procedure [13], [14], i.e.,

$$\mathcal{R}_k^{\text{BH}} = \{j \in \{1, \dots, m\} : P_j \leq P_{(k)}\}, \quad k = 0, 1, \dots, m, \quad (71)$$

where $P_{(1)} \leq \dots \leq P_{(m)}$ are the sorted conformal p-variables (17) and $\mathcal{R}_0^{\text{BH}} = \emptyset$. This family is the BH counterpart of the e-BH path (24) used by PH-CS, and spans the same range of set sizes. The three baselines report an FDP estimate for the selected set, and differ only in how this estimate is constructed.

1) *Naïve PH-CS*: This baseline applies conventional CS across a range of nominal levels and retains the most useful operating point. For each candidate set $\mathcal{R}_k^{\text{BH}}$, inverting the

BH self-consistency condition (18) gives the nominal level associated with the set, which we adopt as its FDP estimate

$$\hat{\alpha}^n(\mathcal{R}_k^{\text{BH}}) = \min \left\{ 1, \frac{m P_{(k)}}{k} \right\}, \quad (72)$$

where, as in (25), the minimum ensures that the estimate does not exceed 1. Naïve PH-CS then solves the problem

$$\mathcal{R}^n = \arg \max_{\mathcal{R} \in \{\mathcal{R}_0^{\text{BH}}, \dots, \mathcal{R}_m^{\text{BH}}\}} U(r(\mathcal{R}), \hat{\alpha}^n(\mathcal{R})), \quad (73)$$

and reports the associated level $\alpha^n = \hat{\alpha}^n(\mathcal{R}^n)$. Since this level is selected after inspecting the data, it is no longer fixed in advance, and the fixed-level guarantee of Theorem 1 does not apply.

2) *Split PH-CS*: This baseline restores a fixed-level guarantee through data splitting. The calibration set is partitioned into two disjoint halves $\mathcal{D}^{\text{cal}} = \mathcal{D}_1 \cup \mathcal{D}_2$, which are used in two steps: (i) the Naïve PH-CS procedure (73) is applied with calibration restricted to dataset $\mathcal{D}_1$, guaranteeing an FDR level α^s ; (ii) conventional CS (19) is run on dataset $\mathcal{D}_2$ with α^s held fixed as the target level and as the reported FDP estimate. Because the level α^s is selected from $\mathcal{D}_1$ and is therefore independent of $\mathcal{D}_2$, Theorem 1 applies conditionally

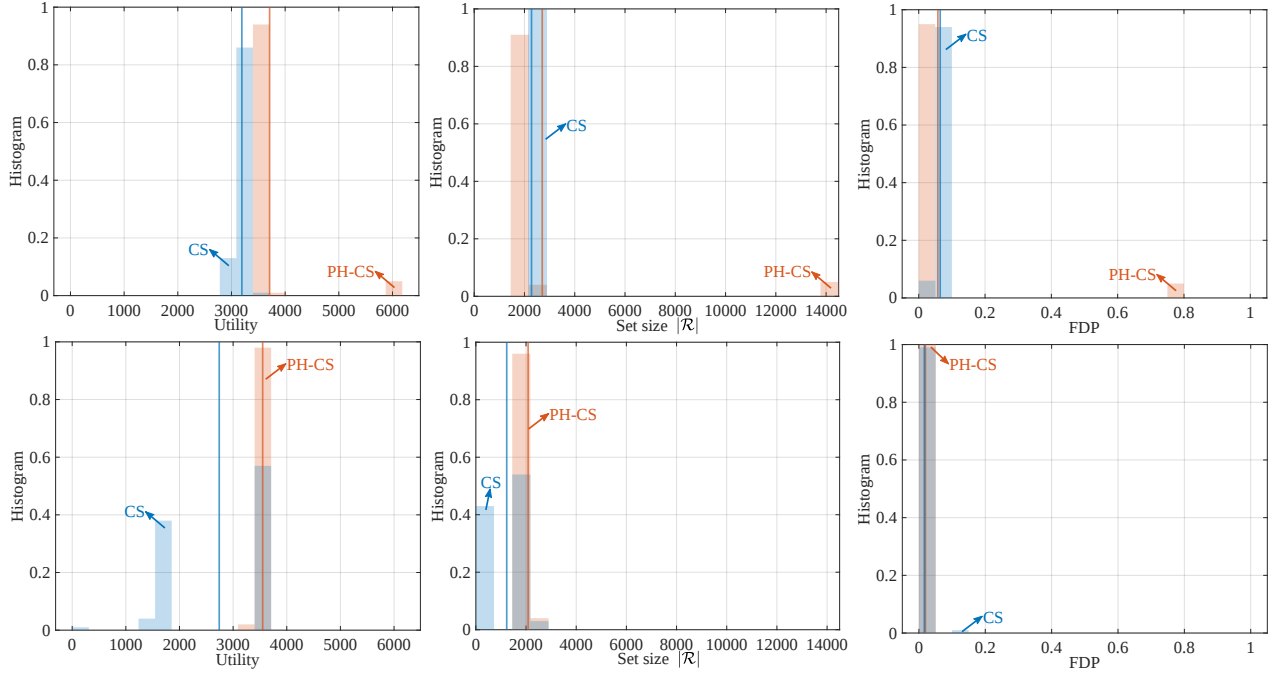


Fig. 8. Histograms of the realized utility (left), selected set size (middle), and FDP (right) under the additive trade-off utility (8) on the Shuttle dataset, with $\lambda = 12800$ (top row) and $\lambda = 14000$ (bottom row). In each panel, the blue and red vertical lines indicate the corresponding average values of CS and PH-CS, respectively.

on the selected level. This reliability is obtained at the cost of calibrating on only a fraction of the available data.

3) *PAC PH-CS*: Rather than reporting a level whose reliability holds only on average, this approach upper bounds, with high probability, the realized FDP (4) of every candidate set at once [40], [41]. Specifically, it provides an FDP estimate $\hat{\alpha}^{\text{PAC}}(\mathcal{R}_k^{\text{BH}})$ that, by construction, upper bounds the realized FDP of all candidate sets $\mathcal{R}_k^{\text{BH}}$ in (71) simultaneously as per the PAC requirement

$$\mathbb{P}\left(\text{FDP}(\mathcal{R}_k^{\text{BH}}, \mathcal{Y}^{\text{test}}) \leq \hat{\alpha}^{\text{PAC}}(\mathcal{R}_k^{\text{BH}}), \forall k\right) \geq 1 - \eta. \quad (74)$$

Because this bound holds for all k at once, the operating point can be chosen after inspecting the data. Accordingly, this baseline selects the set

$$\mathcal{R}^{\text{PAC}} = \arg \max_{\mathcal{R} \in \{\mathcal{R}_0^{\text{BH}}, \dots, \mathcal{R}_m^{\text{BH}}\}} U(r(\mathcal{R}), \hat{\alpha}^{\text{PAC}}(\mathcal{R})), \quad (75)$$

and reports the associated bound $\alpha^{\text{PAC}} = \hat{\alpha}^{\text{PAC}}(\mathcal{R}^{\text{PAC}})$. We emphasize that the guarantee (74) is of a different nature from the post-hoc reliability (12) targeted by PH-CS. It controls the realized FDP with high probability rather than in expectation, requiring the specification of the additional confidence level η . Furthermore, being valid simultaneously over all sets, the estimate $\hat{\alpha}^{\text{PAC}}(\mathcal{R}^{\text{PAC}})$ may be loose at the single selected operating point.

The estimate $\hat{\alpha}^{\text{PAC}}$ is constructed as follows, following [40], [41]. The realized FDP (4) of the set $\mathcal{R}_k^{\text{BH}}$ can be written as

$$\text{FDP}(\mathcal{R}_k^{\text{BH}}, \mathcal{Y}^{\text{test}}) = \frac{m\hat{F}_m^0(P_{(k)})}{\max\{1, |\mathcal{R}_k^{\text{BH}}|\}}, \quad (76)$$

where $\hat{F}_m^0(t) = (\sum_{j=1}^m \mathbb{1}\{Y_{n+j} \leq c_j\} \mathbb{1}\{P_j \leq t\})/m$ is the empirical distribution of the null p-variables. Only the

numerator is unknown, as it depends on the unobserved test labels.

The distribution of $\hat{F}_m^0$, however, is known. A conformal p-variable (17) enters only through the ranks of the conformity scores, and under exchangeability all orderings of the scores are equally likely. The distribution of $\hat{F}_m^0$ is therefore the same as if the scores were i.i.d. uniform variables, independently of the data [40]. Repeatedly sampling the p-variables from this distribution shows how far $\hat{F}_m^0$ can rise, and an envelope $G : [0, 1] \rightarrow [0, 1]$ exceeded with probability at most η then bounds it with probability at least $1 - \eta$ [41],

$$\mathbb{P}(\hat{F}_m^0(t) \leq G(t), \forall t \in [0, 1]) \geq 1 - \eta. \quad (77)$$

The shape of G depends on the deviation statistic used, ranging from the linear Dvoretzky-Kiefer-Wolfowitz (DKW) envelope [40] to the tighter, variance-adaptive higher-criticism (HC) envelope [41].

Substituting the bound (77) into (76) replaces the unknown numerator $m\hat{F}_m^0(P_{(k)})$ by $mG(P_{(k)})$, giving the FDP estimate

$$\hat{\alpha}^{\text{PAC}}(\mathcal{R}_k^{\text{BH}}) = \frac{mG(P_{(k)})}{\max\{1, |\mathcal{R}_k^{\text{BH}}|\}}. \quad (78)$$

Since (77) holds for all $t \in [0, 1]$ simultaneously, evaluating it at $t = P_{(k)}$ for every k and substituting into (76) yields the guarantee (74), where the same probability $1 - \eta$ is inherited from (77).

G. Additional Experimental Results

This appendix collects additional experimental results supplementing Sec. VI, under the constrained-size utility (7) and the additive trade-off utility (8), and across alternative regressors.

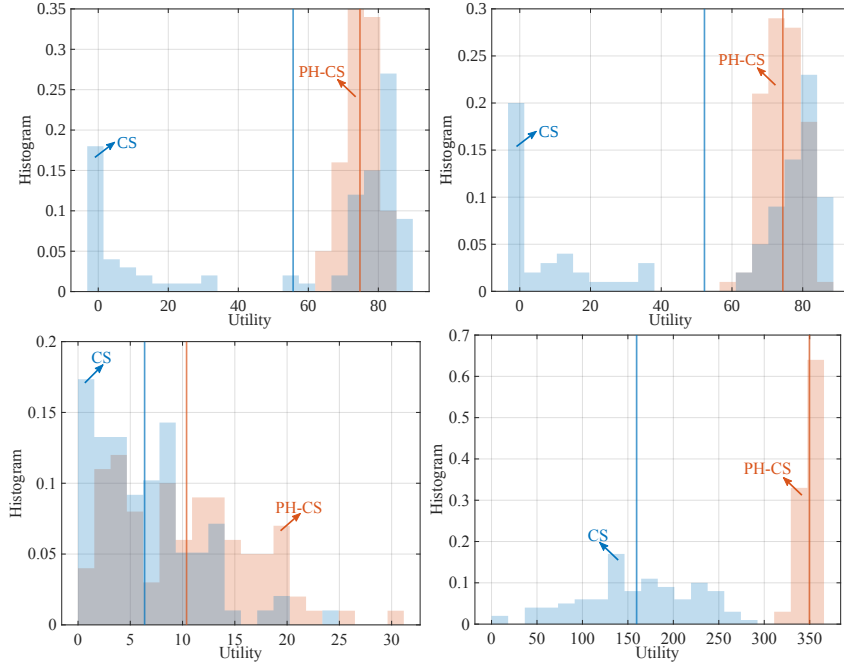


Fig. 9. Histograms of the realized utility under the additive trade-off utility (8) on the synthetic data with homoscedastic (top left) and heteroscedastic (top right) noise using gradient boosting, and on the Recruitment (bottom left) and Musk (bottom right) datasets. The blue and red vertical lines indicate the average utility of CS and PH-CS, respectively.

1) *Constrained-Size Utility*: We provide additional results under the constrained-size utility (7), supplementing Sec. VI-D with the heteroscedastic synthetic data, the Recruitment dataset, and the Shuttle dataset. As shown in Fig. 7, the same conclusion holds across all three settings: PH-CS always satisfies the minimum-size requirement, whereas CS falls below it in a non-negligible fraction of runs, while PH-CS also attains a higher power at a comparable FDP.

2) *Additive Trade-off Utility*: We consider the additive trade-off utility in (8), instantiated as $U(r, \alpha) = r - \lambda\alpha$ with $r = |\mathcal{R}|$. PH-CS selects, for each test batch, the candidate set that maximizes this utility, and CS is matched to its average declared level as in Sec. VI-D. To illustrate the effect of the trade-off parameter λ , Fig. 8 reports the realized utility, selected set size, and FDP on the Shuttle dataset for $\lambda = 12800$ and $\lambda = 14000$.

PH-CS achieves a larger average realized utility than CS, although not necessarily for every individual realization. This is expected because PH-CS maximizes the utility in (11) using the declared level $\alpha^{\text{PH-CS}}$, whereas the plotted utility in (10) is evaluated using the true realized FDP, so finite-sample deviations between the two can occasionally favor CS. As λ increases from the top row to the bottom row, the utility places more emphasis on reducing the FDP. Accordingly, PH-CS becomes more conservative, yielding a smaller selected set and a smaller realized FDP, while the change for CS is induced only through the matched target level. The realized utility remains comparable across the two choices of λ , confirming that λ provides a direct way to adjust the size-reliability trade-off within the same utility framework.

We further report results on additional datasets. For the synthetic data, we use $U(r, \alpha) = r - \lambda\alpha$ with $\lambda = 500$. For the Recruitment dataset, we use $U(r, \alpha) = \log r - \lambda \log(1/(1-\alpha))$

with $\lambda = 15$, and for the Musk dataset, we use $U(r, \alpha) = r - \lambda\alpha$ with $\lambda = 1690$. On both real datasets we set the score hyperparameter to $\gamma = 5$ rather than the value used in Sec. VI. As shown in Fig. 9, PH-CS achieves a larger average realized utility than CS across all four settings, consistent with the Shuttle results above, so its advantage carries over to a different γ .

3) *Alternative Regressors*: We present additional synthetic-data results obtained using random forest and support vector regression, supplementing the results in Sec. VI based on gradient boosting. Under the constrained-size utility, Fig. 10 shows that the same qualitative conclusion holds for random forest: PH-CS consistently satisfies the minimum-size requirement in each realization, whereas CS violates this constraint in a non-negligible fraction of runs. PH-CS also attains a higher power while keeping the FDP comparable. Under a power-aware weighted utility, where each test input is weighted by its estimated true-discovery probability π_j via (37), Fig. 11 confirms that PH-CS achieves higher average utility and higher power than CS at a comparable FDP when using support vector regression. Overall, these results indicate that the empirical behavior of PH-CS is robust to the choice of regressor.

H. Empirical Validation of the Asymptotic Behavior

This appendix illustrates the large-sample behavior of PH-CS on the synthetic data of Sec. VI with homoscedastic noise. In both experiments below, the calibration and test sets grow jointly, with $n = m$ ranging over $\{100, 200, 500, 1000, 2000, 5000\}$ and 200 trials per size, so that the joint growth matches the asymptotic regime being illustrated.

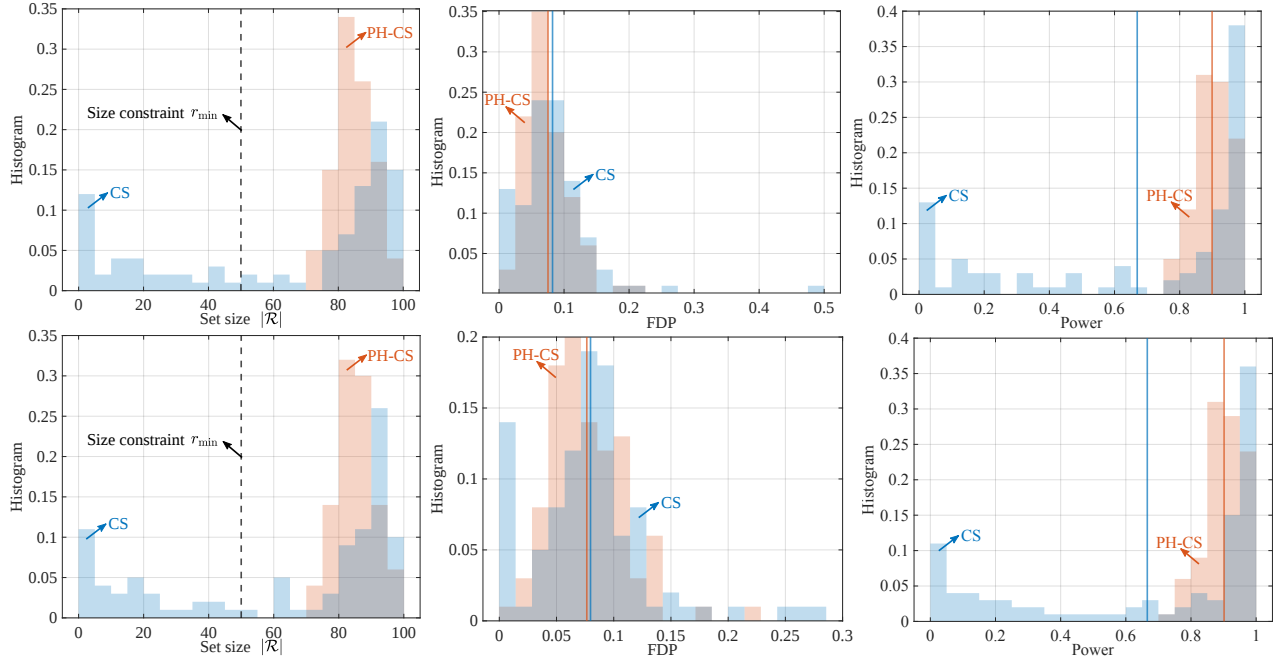


Fig. 10. Histograms of the selected set size (left), realized FDP (middle), and power (right) under the constrained-size utility (7) on synthetic data with homoscedastic (top row) and heteroscedastic (bottom row) noise using random forest. In the middle and right panels, the blue and red vertical lines indicate the average values for CS and PH-CS, respectively.

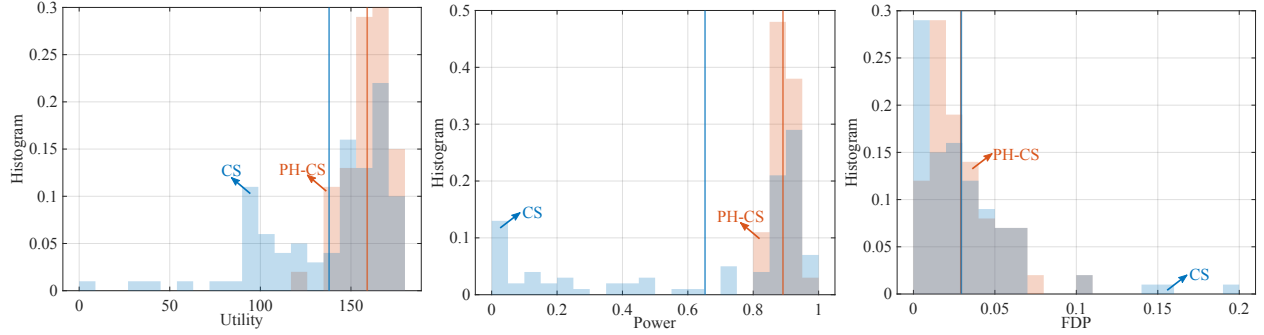


Fig. 11. Histograms of the realized utility (left), power (middle), and FDP (right) under the additive utility (8) with weighted set size (6) on the synthetic data with homoscedastic noise using support vector regression. In each panel, the blue and red vertical lines indicate the average values for CS and PH-CS, respectively.

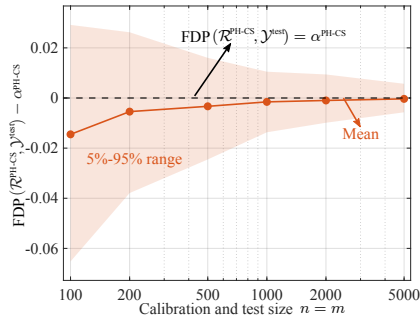


Fig. 12. Gap $\text{FDP}(\mathcal{R}^{\text{PH-CS}}, \mathcal{Y}^{\text{test}}) - \alpha^{\text{PH-CS}}$ versus the calibration and test size $n = m$, on the synthetic data with homoscedastic noise under the constrained-size utility (7). The solid line is the mean over 200 trials and the shaded band is the 5%-95% range.

1) Asymptotic Reliability: This experiment illustrates Proposition 1 under the constrained-size utility (7), with the size floor set to $r_{\min} = 0.5m$ to keep the operating point comparable across sizes. We track the gap $\text{FDP}(\mathcal{R}^{\text{PH-CS}}, \mathcal{Y}^{\text{test}}) - \alpha^{\text{PH-CS}}$ between the realized FDP and the reported level. As shown in Fig. 12, the 5%-95% range narrows toward a point

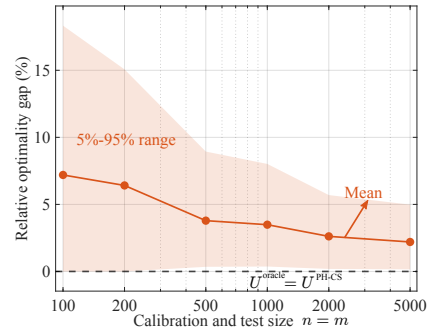


Fig. 13. Relative optimality gap $(U^{\text{oracle}} - U^{\text{PH-CS}})/U^{\text{oracle}}$ between PH-CS and an oracle that selects on the same candidate path using the true FDP, versus the calibration and test size $n = m$, on the synthetic data with homoscedastic noise under the additive utility $U = r - \lambda\alpha$ with $\lambda = 5m$. The solid line is the mean over 200 trials and the shaded band is the 5%-95% range.

as $n = m$ grows, so the scatter of the realized FDP around $\alpha^{\text{PH-CS}}$ in Fig. 6 is a finite-sample effect. Its upper edge, positive at small samples, falls below any fixed margin as the sample grows, in agreement with the asymptotic bound of

Proposition 1. The range concentrates around a small negative mean gap, showing that $\alpha^{\text{PH-CS}}$ is on average a slightly conservative estimate of the realized FDP, consistent with the average guarantee of Theorem 2.

2) *Optimality Gap*: This experiment compares PH-CS with an oracle that searches the same score-ordered candidate path (24) but selects the operating point using the true FDP rather than the estimate $\alpha^{\text{PH-CS}}$, with both utilities evaluated at the true FDP. We use the additive utility $U(r, \alpha) = r - \lambda\alpha$ with $\lambda = 5m$, so that the weight scales with the problem size. Fig. 13 reports the relative gap $(U^{\text{oracle}} - U^{\text{PH-CS}})/U^{\text{oracle}}$. As $n = m$ grows, both the mean gap and its 5%–95% range shrink, indicating that the operating point selected by PH-CS through $\alpha^{\text{PH-CS}}$ approaches the one the oracle selects with the true FDP. On this well-separated data, where the score keeps the estimate close to the realized FDP, the utility attained by PH-CS is therefore close to the best achievable on the candidate path.