28 May 2026

Beyond Go/No-Go Decisions: A Regional Selection Framework for Uncertainty-Aware Molecule Screening

Weihua Shi[1,2], Yixuan Li[1,2], Tian Bai[3], Yijie Zhang[2,4], Kaiqiong Zhao, Marc-André Legault[2], Hui Peng[5], Yue Zhao[6], Eric D. Kolaczyk[1,2], Xiang Yu, Archer Y. Yang[1,2]

1. Department of Mathematics and Statistics, McGill University
2. Mila-Quebec AI Institute
3. Department of Statistics, Stanford University
4. School of Computer Science, McGill University
5. Department of Chemistry, University of Toronto
6. Department of Mathematics, University of York

Abstract

In drug discovery, quantitative structure–activity relationship (QSAR) models are widely used to guide Go/No-Go decisions within the Design–Make–Test–Analyze (DMTA) cycle. However, conventional decision heuristics typically rely on a single cutoff, leading to a rigid binary select/discard paradigm. This approach is particularly ill-suited for borderline compounds near the decision boundary, where screening decisions are especially sensitive to prediction uncertainty and premature choices may either discard viable leads or advance likely failures, thereby increasing downstream assay costs. To address this limitation, we propose Regional Selection (RS), an uncertainty-aware three-way decision framework that partitions compounds into Predicted Pass, Predicted Fail, and Predicted Indeterminate regions. By explicitly reserving high-uncertainty compounds for targeted follow-up, RS avoids the pitfalls of premature binary classification. We formalize this framework through Regional Selection Inference (RSI), which casts region assignment as a multiple-hypothesis testing problem. We develop two imple- mentations of RSI: an empirical calibration-based method (RSI-EC), which thresholds uncertainty-normalized scores via empirical calibration, and a conformal selectionbased method (RSI-CS), which constructs conformal p-values for region assignment. RSI-EC is supported by large-sample calibration arguments, whereas RSI-CS provides finite-sample, distribution-free guarantees under exchangeability. Extensive evaluations across 15 high-dimensional QSAR benchmarks show that both RSI procedures reliably control the

false discovery rate while maintaining high screening power. In limited-data regimes, RSI-CS yields particularly stable FDR control, whereas RSI-EC can be slightly less conservative; both perform strongly as sample sizes increase. We further study a cost-aware extension that incorporates asymmetric downstream costs through the score construction while keeping the nominal FDR target fixed. This extension introduces a tuning parameter that can reduce realized downstream cost, with dataset-dependent trade-offs against screening power. Overall, RSI offers a mathematically grounded and resource-aware alternative to single-threshold screening, allowing discovery teams to better balance decision confidence with assay budgets.

Beyond Go/No-Go Decisions: A Regional Selection Framework for Uncertainty-Aware Molecule Screening

Weihua Shi,[†,‡] Yixuan Li,[*,†,‡] Tian Bai,[¶] Yijie Zhang,[§,‡] Kaiqiong Zhao,[‖] Marc-André Legault,[⊥,‡] Hui Peng,[#] Yue Zhao,[@] Eric D. Kolaczyk,[†,‡] Xiang Yu,[*,△] and Archer Y. Yang[*,†,‡]

†*Department of Mathematics and Statistics, McGill University, Montreal, Quebec, Canada*

‡*Mila - Quebec AI Institute, Montreal, Quebec, Canada*

¶*Department of Statistics, Stanford University, USA*

§*School of Computer Science, McGill University, Montreal, Quebec, Canada*

‖*Department of Mathematics and Statistics, York University, Toronto, Canada*

⊥*Faculty of Pharmacy, Université de Montréal, Montreal, Quebec, Canada*

#*Department of Chemistry, University of Toronto, Toronto, Canada*

@*Department of Mathematics, University of York, York, United Kingdom*

△*Lilly Research Laboratories, USA*

E-mail: yixuan.li2@mail.mcgill.ca; sean.yu2@lilly.com; archer.yang.yi@gmail.com

Abstract

In drug discovery, quantitative structure–activity relationship (QSAR) models are widely used to guide Go/No-Go decisions within the Design–Make–Test–Analyze (DMTA) cycle. However, conventional decision heuristics typically rely on a single cutoff, leading to a rigid binary select/discard paradigm. This approach is particularly ill-suited

for borderline compounds near the decision boundary, where screening decisions are especially sensitive to prediction uncertainty and premature choices may either discard viable leads or advance likely failures, thereby increasing downstream assay costs. To address this limitation, we propose Regional Selection (RS), an uncertainty-aware three-way decision framework that partitions compounds into Predicted Pass, Predicted Fail, and Predicted Indeterminate regions. By explicitly reserving high-uncertainty compounds for targeted follow-up, RS avoids the pitfalls of premature binary classification. We formalize this framework through Regional Selection Inference (RSI), which casts region assignment as a multiple-hypothesis testing problem. We develop two implementations of RSI: an empirical calibration-based method (RSI-EC), which thresholds uncertainty-normalized scores via empirical calibration, and a conformal selection-based method (RSI-CS), which constructs conformal p-values for region assignment. RSI-EC is supported by large-sample calibration arguments, whereas RSI-CS provides finite-sample, distribution-free guarantees under exchangeability. Extensive evaluations across 15 high-dimensional QSAR benchmarks show that both RSI procedures reliably control the false discovery rate while maintaining high screening power. In limited-data regimes, RSI-CS yields particularly stable FDR control, whereas RSI-EC can be slightly less conservative; both perform strongly as sample sizes increase. We further study a cost-aware extension that incorporates asymmetric downstream costs through the score construction while keeping the nominal FDR target fixed. This extension introduces a tuning parameter that can reduce realized downstream cost, with dataset-dependent trade-offs against screening power. Overall, RSI offers a mathematically grounded and resource-aware alternative to single-threshold screening, allowing discovery teams to better balance decision confidence with assay budgets.

Introduction

"Confirmations should count only if they are the result of risky predictions" - Karl Popper

Early-stage drug discovery requires selecting promising candidate compounds from vast chemical libraries under severe experimental resource constraints[1–5]. Because assaying every compound is costly and time-consuming, computational screening methods are routinely used to decide which compounds should be advanced, excluded, or reserved for further evaluation. Among these methods, quantitative structure–activity relationship (QSAR) models[6–12] provide a widely used framework for predicting biological or chemical activity from molecular structure.

In many virtual screening applications[13], however, the goal is not merely accurate prediction but reliable decision-making under uncertainty. In particular, screening pipelines aim to identify truly promising compounds while limiting false selections, since each false selection may trigger costly downstream experiments. Standard QSAR pipelines typically output point predictions of activity[6,7], but these predictions alone do not provide calibrated uncertainty quantification or direct control of selection error rates such as the false discovery rate (FDR)[14]. The Sheridan method[15–17] and conformal selection (CS)[18–22] are two approaches that support screening decisions under FDR control. The Sheridan method relies on asymptotic large-sample arguments and therefore provides asymptotic FDR control, whereas CS provides finite-sample FDR control under exchangeability through a distribution-free construction. Despite these advances, existing FDR-based screening approaches still typically culminate in a single threshold, and hence in a binary select/discard rule.

However, this binary formulation is poorly suited to compounds whose predicted activities lie close to the decision boundary. For these compounds, the margin between the prediction and the cutoff is small, so minor estimation errors or model uncertainty can change the assigned pass/fail label. For example, in virtual screening for inhibition of the human *ether-à-go-go*-related gene (hERG) potassium channel, a major driver of drug-induced cardiotoxicity, risk is often operationalized using an IC_{50} cutoff on the order of 10 μM. Two structurally similar compounds may have predicted IC_{50} values of 9.9 and 10.1 μM. Under a strict binary rule, these two compounds would receive opposite decisions. Given typical *in vitro* assay

variability and QSAR model uncertainty, however, the two predictions may be practically indistinguishable. Forcing an immediate pass/fail decision in this regime may either advance a likely failure or discard a viable lead.

The distinction is important in the Design–Make–Test–Analyze (DMTA) cycle, where different stages carry different operational consequences. At the *Make* stage, decisions often concern synthesis prioritization, and a binary Go/No-Go rule may remain operationally acceptable when the immediate cost of retaining an uncertain candidate is relatively low. At the *Test* stage, however, the downstream consequences of a decision change substantially: advancing a borderline compound may waste assay capacity and resources, whereas discarding it too early may eliminate a useful lead. In this higher-stakes regime, a three-way triage strategy becomes more appropriate, separating

- **Predicted Pass**: compounds with strong evidence for advancement, which may skip the current assay and proceed to the next stage, thereby saving the cost of the current assay; the risk is that, if the prediction is incorrect, the compound may fail at a later and more expensive downstream stage;
- **Predicted Fail**: compounds with strong evidence for exclusion, which are not advanced to the current assay, thereby saving current assay resources; the corresponding risk is the opportunity cost of missing a genuinely promising candidate;
- **Predicted Indeterminate**: compounds whose status remains ambiguous for the current assay decision, and therefore warrants a lower-cost confirmatory evaluation before a clear advancement or exclusion decision is made.

Importantly, binary and trinary decision strategies should not be viewed as fundamentally disjoint problems. Rather, they reflect different operational cost structures. A trinary decision strategy becomes useful when compounds assigned to the Predicted Indeterminate region are sent to a distinct, lower-cost follow-up workflow. In this setting, the Predicted Indeterminate label has its own operational meaning: it does not make a definitive Pass or Fail decision,

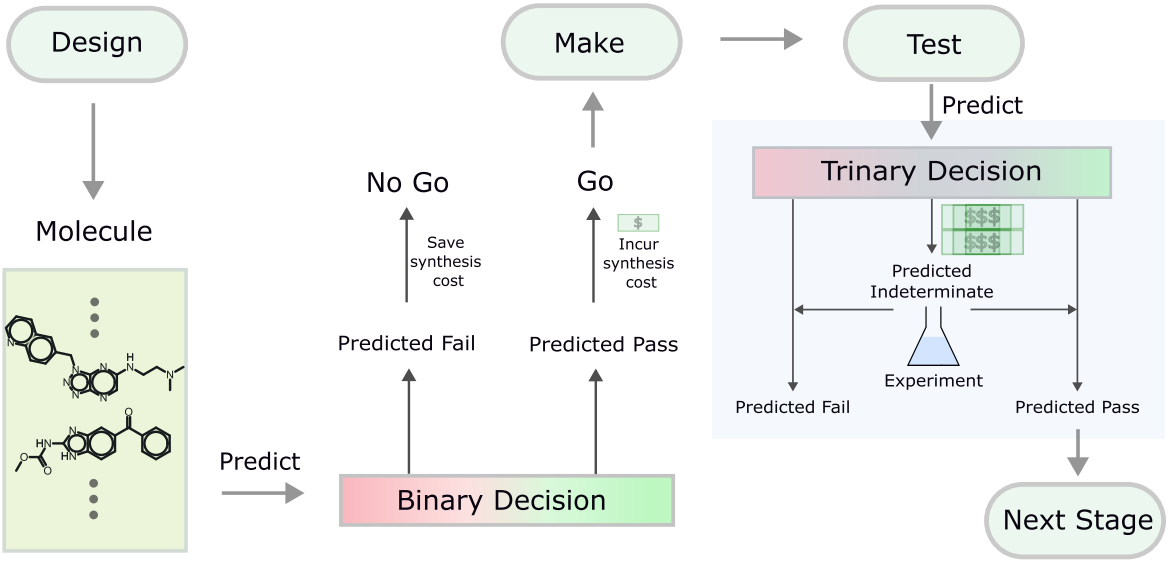


Figure 1: Schematic illustration of how binary and trinary decision strategies arise within the DMTA cycle. Binary decisions are associated with the *Make* stage, whereas the *Test* stage more naturally motivates a trinary decision structure with an explicit Indeterminate pathway.

but instead defers borderline compounds to additional validation. Therefore, the Predicted Indeterminate region should not be collapsed into either Predicted Pass or Predicted Fail. If no such lower-cost intermediate workflow exists, then the Predicted Indeterminate label provides little practical benefit, and the procedure effectively reduces to a binary decision strategy.

Figure 1 places the distinction between binary and trinary decision strategies within the DMTA cycle, while Figure 2 gives a more detailed view of their operational consequences. At the *Make* stage, binary Go/No-Go decisions may remain adequate, since the opportunity cost of missing a good candidate is high whereas the cost of high-throughput screening is relatively low. As a discovery program moves to downstream stages, assay costs gradually increase, for example, from in vitro to in vivo assays and from rodent to primate studies. This change in operational stakes naturally motivates the trinary RS structure at the *Test* stage, with an explicit Indeterminate pathway.

Motivated by this operational perspective, we propose Regional Selection (RS), a frame-

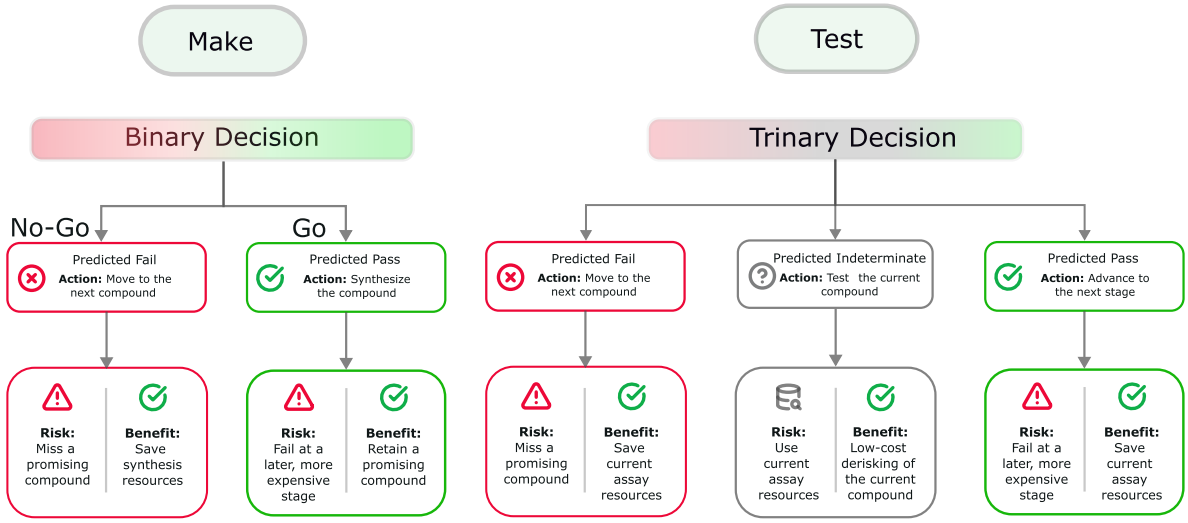


Figure 2: The binary and trinary decision frameworks across the *Make* and *Test* stages.

work that partitions candidate compounds into three regions: (i) Predicted Pass, containing compounds with strong statistical evidence for advancement; (ii) Predicted Fail, containing compounds with strong evidence for exclusion; and (iii) Predicted Indeterminate, containing near-threshold compounds for which deferral to follow-up evaluation is preferable to an immediate binary decision. For such compounds, an explicit deferral option is valuable because a premature binary classification may trigger unnecessary downstream assay costs.

To formalize this idea, we introduce *Regional Selection Inference* (RSI), which casts region assignment as a multiple-hypothesis testing problem. In this formulation, the hypotheses are defined with respect to the true but unobserved response category of each compound, as determined by the activity cutoffs, whereas the selected sets determine the predicted decision labels. We consider two complementary inferential targets: identifying compounds whose responses lie in the True Indeterminate region, and identifying compounds whose responses lie in the True Pass or True Fail regions. For these targets, we develop two implementations: an empirical calibration-based method (RSI-EC), which thresholds uncertainty-normalized scores via empirical calibration, and a conformal selection-based method (RSI-CS), which constructs conformal p-values for region assignment[23,24]. RSI-EC is supported by large-sample

calibration arguments, whereas RSI-CS provides finite-sample, distribution-free guarantees under exchangeability. Across high-dimensional QSAR benchmarks, we show that both procedures achieve reliable error control and strong screening performance. We further develop cost-aware extensions for both RSI-EC and RSI-CS, allowing asymmetric downstream costs to be incorporated through method-specific score constructions while preserving the original FDR target.

Methodology

In this section, we develop RSI for assigning test compounds to the Predicted Fail, Predicted Indeterminate, or Predicted Pass region. The section first formulates RS region assignment as a multiple-hypothesis testing problem on the test set. We then introduce two complementary testing settings, which differ in their null hypotheses and error targets, and implement each setting using both RSI-EC and RSI-CS. Figure 3 provides an overview of RSI-EC and RSI-CS workflows within the proposed RS framework.

Problem Formulation

Let $x_i \in \mathcal{X}$ denote the molecular descriptors for compound i, and let $y_i \in \mathbb{R}$ represent its measured chemical activity. We consider a labeled dataset $D_{\text{obs}} = \{(x_i, y_i)\}_{i=1}^{n}$ and an unlabeled test set $D_{\text{test}} = \{x_{n+j}\}_{j=1}^{m}$, for which the true activity levels $\{y_{n+j}\}_{j=1}^{m}$ are unobserved. We assume that samples in D_{obs} and D_{test} are exchangeable.

Motivated by the goal of precise regional identification, we formulate the screening task as a multiple-hypothesis testing problem on D_{test}[20,25]. For each test compound $j \in D_{\text{test}}$, we construct a hypothesis test whose outcome induces an RS region assignment. Thus, each test compound is assigned to one of three predicted decision regions: Predicted Fail, Predicted Indeterminate, or Predicted Pass. We consider two complementary settings that differ in their null formulations and error targets.

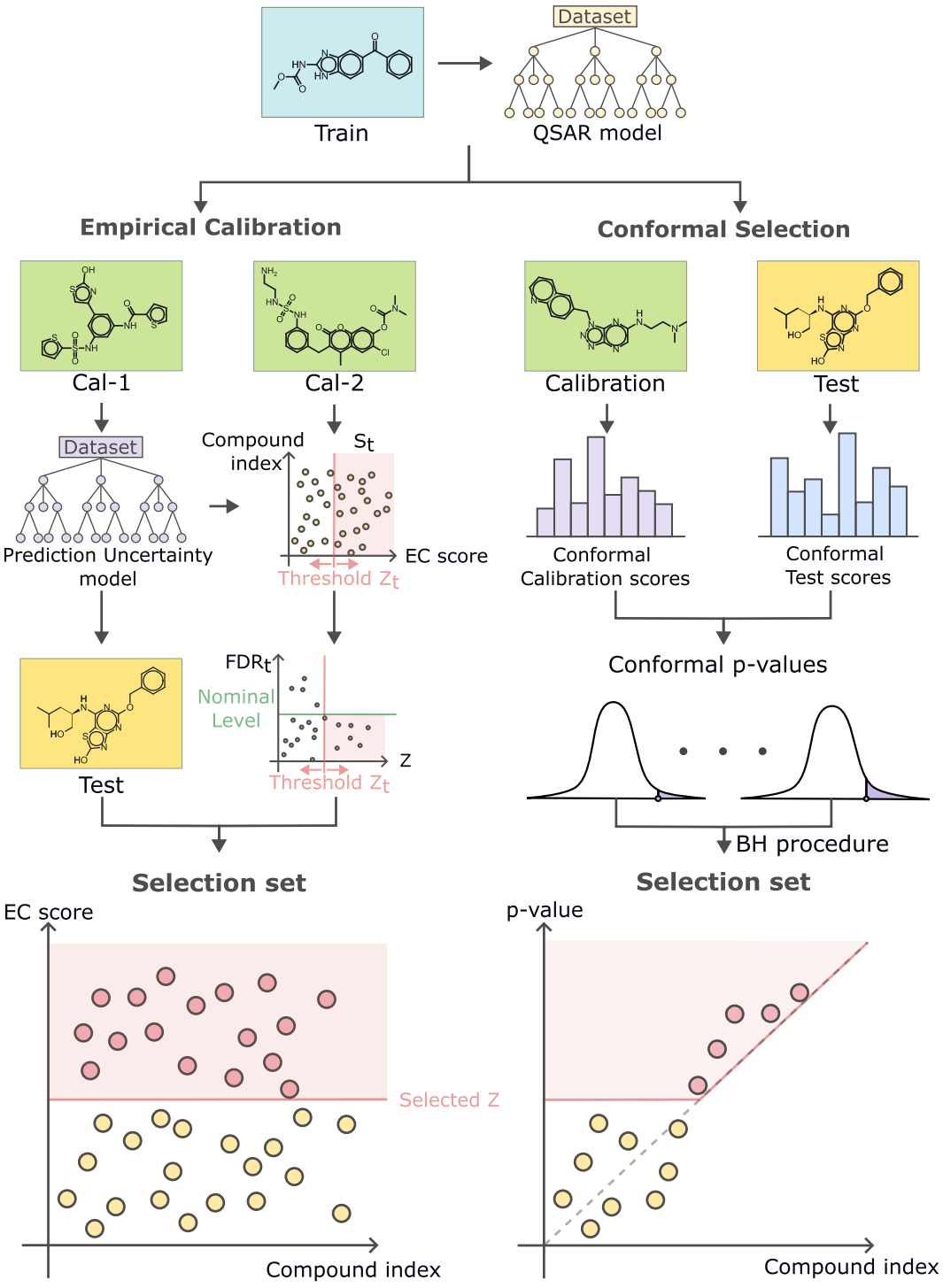


Figure 3: Overview of RSI-EC and RSI-CS applied to Regional Selection.

- **Setting I (True Indeterminate):** The goal is to identify borderline compounds and assign them to the Predicted Indeterminate region. Accordingly, the alternative hypothesis corresponds to True Indeterminate region, and the rejection set $\mathcal{S}$ defines the Predicted Indeterminate region. We control the FDR among Predicted Indeterminate assignments, i.e., the expected proportion of compounds in the True Fail and True Pass regions mistakenly assigned to Predicted Indeterminate region.

- **Setting II (True Fail and True Pass):** The goal is to certify compounds as Predicted Fail and Predicted Pass regions when evidence is strong, with the Predicted Indeterminate region defined by exclusion. Accordingly, the alternative hypothesis corresponds to the True Fail and True Pass regions, and the rejection set $\mathcal{S}$ defines the Predicted Fail and Predicted Pass regions. We control the FDR among assignments in the Predicted Fail and Predicted Pass regions, i.e., the expected proportion of compounds in the True Indeterminate mistakenly assigned to the Predicted Fail and Predicted Pass regions.

Table 1 shows a schematic comparison of the two settings. The remainder of this section develops the two hypothesis-testing formulations corresponding to Setting I and Setting II.

Table 1: Two Regional Selection screening settings.

Setting	Target region	Default decision
I	True Indeterminate	True Fail and True Pass regions
II	True Fail and True Pass regions	True Indeterminate

Setting I

We first describe a baseline two-test construction for Setting I. We then present our refined formulation and the corresponding statistical guarantees, which will later be instantiated using RSI-CS and RSI-EC.

Baseline approach. As a conceptual baseline for Setting I, we consider an independent two-threshold screening procedure based on thresholds $c_1 < c_2$. For each compound $j \in \{1, \ldots, m\}$ in D_{test} (with activity y_{n+j}), the procedure applies two one-sided hypothesis tests independently:

$$\textbf{Stage 1:} \quad H_{0j}^{(1)} : y_{n+j} \leq c_1 \quad \text{vs.} \quad H_{1j}^{(1)} : y_{n+j} > c_1,$$

$$\textbf{Stage 2:} \quad H_{0j}^{(2)} : y_{n+j} \geq c_2 \quad \text{vs.} \quad H_{1j}^{(2)} : y_{n+j} < c_2.$$

Let $\mathcal{S}^{(1)}$ and $\mathcal{S}^{(2)}$ denote the rejection sets (that is, the sets of rejected candidates) of Stage 1 and Stage 2, respectively. In Stage 1, rejection of $H_{0j}^{(1)}$ provides evidence that $y_{n+j} > c_1$. As a decision rule, compounds outside $\mathcal{S}^{(1)}$ are assigned to the Predicted Fail region and $\mathcal{S}_{\text{red}} = (\mathcal{S}^{(1)})^{\text{c}}$. In Stage 2, rejection of $H_{0j}^{(2)}$ provides evidence that $y_{n+j} < c_2$. As a decision rule, compounds outside $\mathcal{S}^{(2)}$ are assigned to the Predicted Pass region and $\mathcal{S}_{\text{green}} = (\mathcal{S}^{(2)})^{\text{c}}$. Compounds rejected in both stages satisfy both one-sided evidence conditions and are therefore assigned to the Predicted Indeterminate region ($\mathcal{S}_{\text{gray}} = \mathcal{S}^{(1)} \cap \mathcal{S}^{(2)}$) as illustrated in the top panel of Figure 4.

However, this construction implicitly relies on a coherent overlap between the two rejection sets. When the two one-sided tests are applied separately, no structural constraint ensures that $\mathcal{S}^{(1)}$ and $\mathcal{S}^{(2)}$ overlap in a stable or meaningful way. Each test is subject to statistical error, and its rejection set can deviate from the corresponding activity-based region in different directions. Consequently, $\mathcal{S}_{\text{gray}}$ can be arbitrarily small and may even be empty, yielding a degenerate procedure that fails to produce a meaningful Predicted Indeterminate region (in the bottom of Figure 4). This illustrates that the independent two-threshold baseline does not provide a principled mechanism or guarantees for identifying borderline compounds.

Our approach. Motivated by this observation, we reformulate the original multiple testing problem as a unified multiple-testing formulation with compound-wise hypotheses. For each compound $j \in \{1, \ldots, m\}$ in D_{test}, we directly test whether y_{n+j} lies within the True

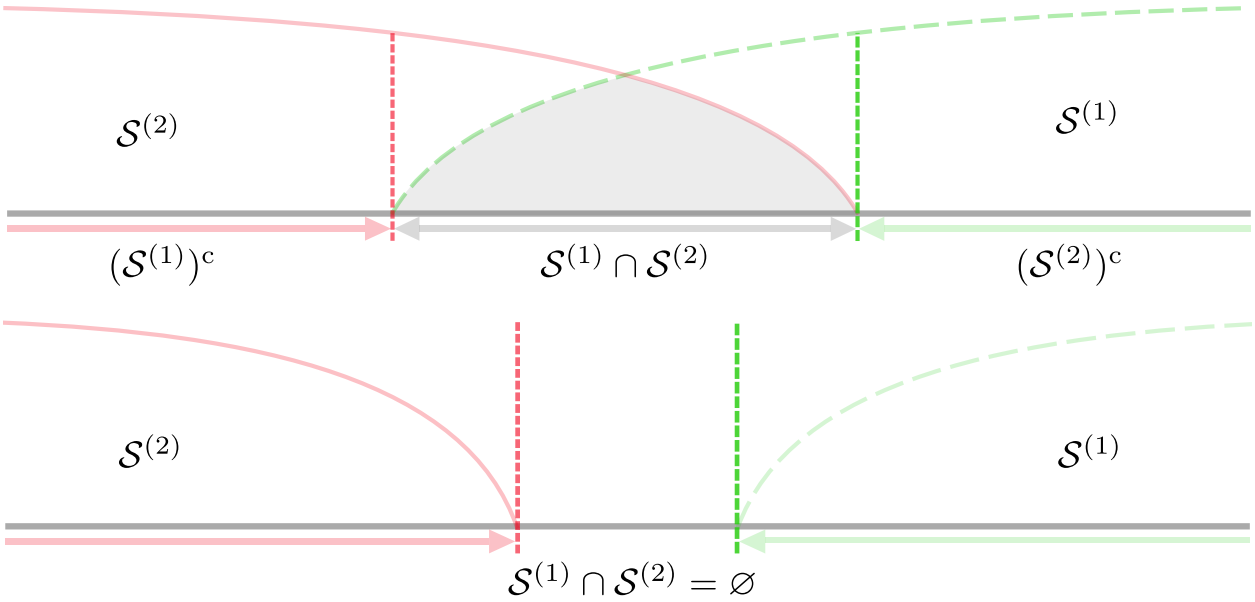


Figure 4: Illustration of the regions induced by the independent two-threshold procedure (Setting I). The upper diagram shows a non-empty intersection $\mathcal{S}^{(1)} \cap \mathcal{S}^{(2)}$, whereas the lower diagram shows the case $\mathcal{S}^{(1)} \cap \mathcal{S}^{(2)} = \varnothing$. The green, red, and gray regions correspond to Predicted Pass, Predicted Fail, and Predicted Indeterminate, respectively.

Indeterminate region. The resulting hypothesis test is formulated as follows:

$$H_{0j} : y_{n+j} \in \mathcal{R}_0 \quad \text{vs.} \quad H_{1j} : y_{n+j} \in \mathcal{R}_1.$$

where the True Fail and True Pass regions (True regions) are combined as $\mathcal{R}_0 = (-\infty, c_1] \cup [c_2, +\infty)$, and the True Indeterminate region is $\mathcal{R}_1 = (c_1, c_2)$. Under this hypothesis test, compounds are assigned to the Predicted Fail and Predicted Pass regions by default unless the null hypothesis is rejected. Let $\mathcal{S}$ denote the index set of test compounds for which H_{0j} is rejected, i.e.,

$$\mathcal{S} := \{j : H_{0j} \text{ is rejected}\}.$$

In Setting I, the rejection set $\mathcal{S}$ corresponds to the Predicted Indeterminate region and $\mathcal{S}_{\text{gray}} = \mathcal{S}$. Therefore, its complement $\mathcal{S}^{\mathrm{c}}$ represents the Predicted regions ($\mathcal{S}_{\text{red}} \cup \mathcal{S}_{\text{green}} = \mathcal{S}^{\mathrm{c}}$); see Figure 5. This formulation adopts a conservative screening strategy, under which compounds are assigned to the Predicted Indeterminate region only after rejecting the null

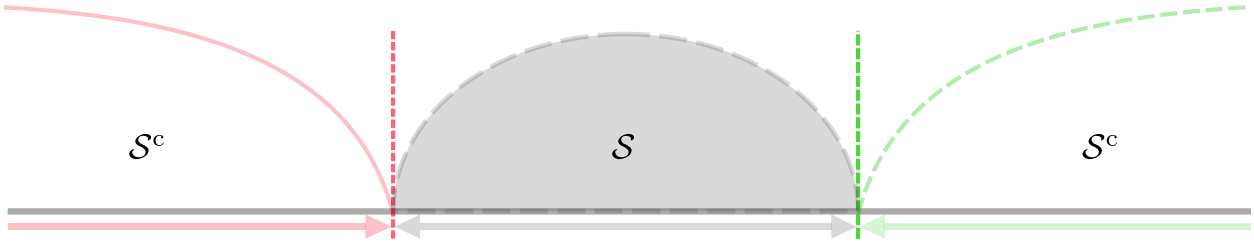


Figure 5: Illustration of the three regions induced by the proposed RS framework (Setting I). The rejection set $\mathcal{S}$ corresponds to the Predicted Indeterminate region, while its complement $\mathcal{S}^c$ defines the Predicted Fail and Predicted Pass regions (Predicted regions).

hypothesis.

Within this multiple testing framework, we focus on controlling FDR, defined as the expected proportion of compounds in the True Fail and True Pass regions incorrectly assigned to the Predicted Indeterminate region among all selected compounds. All expectations are taken with respect to the joint distribution of the unobserved test activities $\{y_{n+j}\}_{j=1}^{m}$,

$$\text{FDR} = \mathbb{E}\left[\frac{|\mathcal{S} \cap \mathcal{H}_0|}{|\mathcal{S}|}\right], \tag{1}$$

where $\mathcal{H}_0 = \{j : y_{n+j} \in \mathcal{R}_0\}$, with the convention that $0/0 = 0$. Our goal is to ensure FDR $\leq q$ where q is a user-defined threshold. While controlling FDR, we seek to maximize the screening power, which represents the ability of the procedure to identify truly Indeterminate compounds:

$$\text{Power} = \mathbb{E}\left[\frac{|\mathcal{S} \cap \mathcal{H}_1|}{|\mathcal{H}_1|}\right],$$

where $\mathcal{H}_1 = \{j : y_{n+j} \in \mathcal{R}_1\}$. This reformulation introduces the key structural element missing from the independent two-test baseline. Rather than defining the Predicted Indeterminate region as an ad hoc intersection of two separately generated rejection sets, our approach treats Predicted Indeterminate assignment as the rejection set of a single multiple-testing problem. This gives the target region a direct statistical definition and allows FDR control to be imposed explicitly on Predicted Indeterminate assignments, with power quantifying the

ability to identify truly borderline compounds.

Setting II

Setting II provides a complementary formulation to Setting I. The null hypothesis is that the compound belongs to the True Indeterminate region. We first describe a baseline two-test construction for Setting II, then present our refined formulation.

Baseline approach. As a conceptual baseline for Setting II, we first consider an independent two-threshold screening procedure analogous to that in Setting I. Specifically, we perform the following two one-sided tests:

$$\textbf{Stage 1:} \quad H_{0j}^{(1)} : y_{n+j} > c_1 \quad \text{vs.} \quad H_{1j}^{(1)} : y_{n+j} \leq c_1,$$

$$\textbf{Stage 2:} \quad H_{0j}^{(2)} : y_{n+j} < c_2 \quad \text{vs.} \quad H_{1j}^{(2)} : y_{n+j} \geq c_2.$$

Let $\mathcal{S}^{(1)}$ and $\mathcal{S}^{(2)}$ denote the rejection sets of Stage 1 and Stage 2, respectively. Under Stage 1, rejecting $H_{0j}^{(1)}$ provides evidence that $y_{n+j} \leq c_1$, corresponding to the Predicted Fail region and $\mathcal{S}_{\text{red}} = \mathcal{S}^{(1)}$. Under Stage 2, rejecting $H_{0j}^{(2)}$ provides evidence that $y_{n+j} \geq c_2$, corresponding to the Predicted Pass region and $\mathcal{S}_{\text{green}} = \mathcal{S}^{(2)}$. The Predicted Indeterminate region is then defined as the complement of their union ($\mathcal{S}_{\text{gray}} = (\mathcal{S}^{(1)} \cup \mathcal{S}^{(2)})^{\text{c}}$) as illustrated in the top panel of Figure 6. However, this two-stage construction again relies on an ad hoc combination of independent tests and does not provide a principled mechanism for controlling error rates over the resulting Predicted-region assignments. As a result, it does not yield rigorous guarantees for regional assignment.

Our approach. For each compound $j \in \{1, \ldots, m\}$, we directly test whether the unobserved activity falls outside the True Indeterminate region through a single compound-wise hypothesis test:

$$H_{0j} : y_{n+j} \in \mathcal{R}_0 \quad \text{vs.} \quad H_{1j} : y_{n+j} \in \mathcal{R}_1.$$

where $\mathcal{R}_0 = (c_1, c_2)$ is the True Indeterminate region and $\mathcal{R}_1 = (-\infty, c_1] \cup [c_2, \infty)$ is the

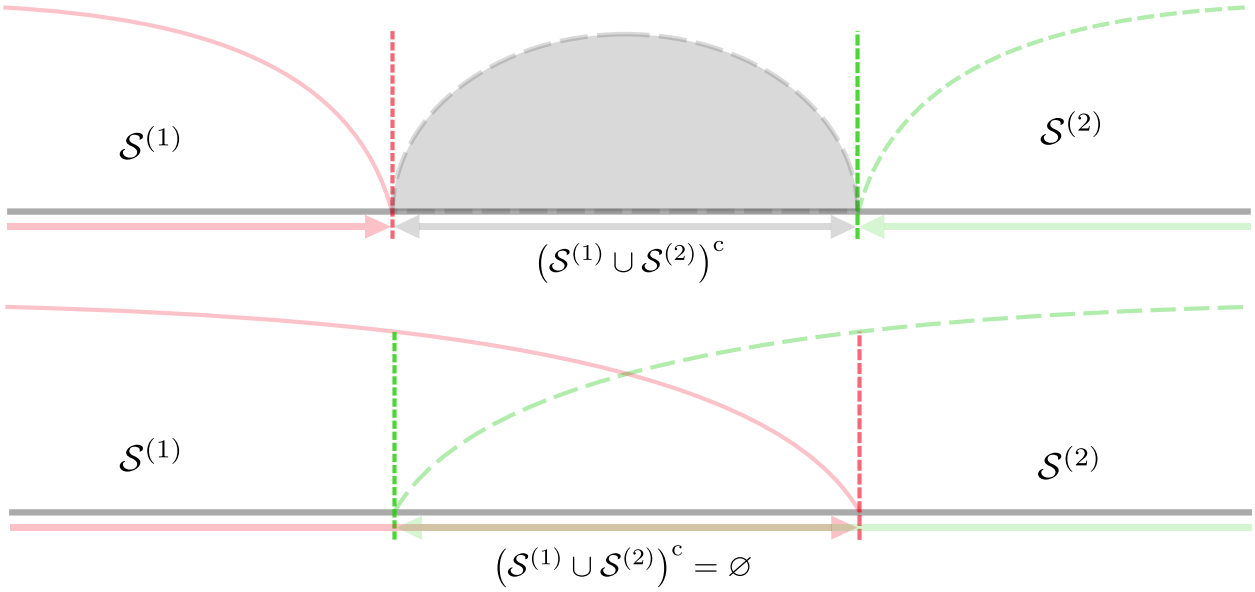


Figure 6: Illustration of the regions induced by the independent two-threshold procedure (Setting II). The upper diagram shows a non-empty Predicted Indeterminate region $(\mathcal{S}^{(1)} \cup \mathcal{S}^{(2)})^{\mathrm{c}}$, whereas the lower diagram illustrates a degenerate case in which $(\mathcal{S}^{(1)} \cup \mathcal{S}^{(2)})^{\mathrm{c}} = \varnothing$.

union of the True Fail and True Pass regions. Compared with Setting I, Setting II reverses the roles of the True Indeterminate and True regions in the null and alternative hypotheses.

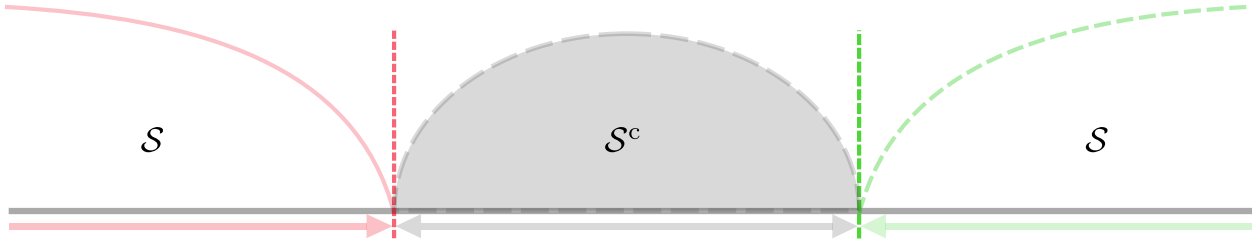


Figure 7: Illustration of the three regions induced by the proposed RS framework in Setting II. The rejection set $\mathcal{S}$ corresponds to the Predicted regions, while its complement $\mathcal{S}^{\mathrm{c}}$ defines the Predicted Indeterminate region.

Under this framework, let $\mathcal{S}$ denote the index set of test compounds for which H_{0j} is rejected, i.e.,

$$\mathcal{S} := \{j : \text{reject } H_{0j}\}.$$

In Setting II, the rejection set $\mathcal{S}$ corresponds to the Predicted regions ($\mathcal{S}_{\text{red}} \cup \mathcal{S}_{\text{green}} = \mathcal{S}$).

Accordingly, its complement $\mathcal{S}^{\mathrm{c}}$ corresponds to the Predicted Indeterminate region and $\mathcal{S}_{\mathrm{gray}} = \mathcal{S}^{\mathrm{c}}$ (see Figure 7). This formulation reflects a conservative design choice: compounds are assigned to the Predicted Indeterminate region by default, and are moved to the Predicted Pass or Predicted Fail regions only when the null hypothesis is rejected.

As in Setting I, all expectations are taken with respect to the joint distribution of the unobserved test activities $\{y_{n+j}\}_{j=1}^{m}$. Within this multiple testing framework, we control the FDR, defined as the expected proportion of True Indeterminate compounds incorrectly assigned to the Predicted regions among all selected compounds:

$$\mathrm{FDR} = \mathbb{E}\left[\frac{|\mathcal{S} \cap \mathcal{H}_0|}{|\mathcal{S}|}\right], \tag{2}$$

where $\mathcal{H}_0 = \{j : y_{n+j} \in \mathcal{R}_0\}$. At the same time, we define screening power to quantify how effectively the procedure selects the True-region compounds:

$$\mathrm{Power} = \mathbb{E}\left[\frac{|\mathcal{S} \cap \mathcal{H}_1|}{|\mathcal{H}_1|}\right],$$

where $\mathcal{H}_1 = \{j : y_{n+j} \in \mathcal{R}_1\}$. Under FDR control, higher power means that more True-region compounds are correctly assigned to the Predicted regions, while avoiding excessive erroneous selection of borderline compounds.

Relative to the baseline construction, this refined formulation aligns Setting II with Setting I in an important respect. The Predicted regions are now defined through a single coherent multiple-testing problem, rather than by taking an ad hoc union of rejection sets arising from separate tests for the two Predicted regions. The formulations above define the regional testing targets and their corresponding error criteria. We now show how to realize these targets in practice through two implementations: RSI-EC, which calibrates a score threshold empirically, and RSI-CS, which constructs conformal p-values and applies the Benjamini–Hochberg (BH) procedure[14,26].

Statistical Implementation

To implement the proposed RS framework, we consider two statistical approaches: the RSI-EC and RSI-CS. They differ in both construction and theoretical guarantees.

The RSI-EC is based on empirical calibration. It thresholds activity scores according to their empirical behavior in one or more calibration sets. Although this approach is widely used in practice, its guarantees rely on large-sample approximations. It can be combined with any predictive model that provides both a point prediction and an uncertainty estimate. To ensure that the standardized score is well defined, the cutoff, prediction, and uncertainty must refer to the same target quantity and be measured on the same scale.

The RSI-CS is based on the conformal selection framework. It constructs p-values from the exchangeability of the training, calibration, and test samples. Its FDR control guarantee does not depend on the choice of predictive model or on asymptotic approximations. Under exchangeability, it provides rigorous finite-sample validity without parametric modeling assumptions.

Table 2 summarizes the key methodological differences between the two approaches.

Table 2: Comparison of Statistical Properties of RSI-CS and RSI-EC.

Method	**Mechanism**	**Type of Guarantee**	**Data Splits**
RSI-EC	score thresholding	Asymptotic FDR control	$D_{\text{train}} + D_{\text{cal1}} + D_{\text{cal2}}$.
RSI-CS	p-value based	Finite-sample FDR control	$D_{\text{train}} + D_{\text{cal}}$.

Empirical Calibration-based method

RSI-EC provides a score that quantifies the distance of a prediction from a threshold relative to local uncertainty. We extend this construction to the RS framework by incorporating both decision boundaries c_1 and c_2.

In our implementation, the local uncertainty is represented by a pointwise prediction-error function $\hat{\sigma}(x)$ rather than a global MSE. Following the Sheridan method, we estimate $\hat{\sigma}(x)$

using a separate error model trained on cross-validated prediction errors from the original predictive model. For a new compound, this model yields a pointwise RMSE-type error estimate that reflects the expected uncertainty of the prediction[17]. We then use $\hat{\sigma}(x)$ to normalize the distance-to-boundary scores in RSI-EC. We randomly partition the labeled dataset D_{obs} into a proper training set D_{train} and two calibration sets $D_{\text{cal1}} = \{(x_i, y_i) : i \in I_{\text{cal1}}\}$ and $D_{\text{cal2}} = \{(x_i, y_i) : i \in I_{\text{cal2}}\}$, with sizes $n_1 := |D_{\text{cal1}}|$ and $n_2 := |D_{\text{cal2}}|$.

General procedure. The RSI-EC procedure consists of three main steps:

1. **Training:** Use D_{train} to fit a regression model $\hat{\mu}(\cdot)$ for the conditional mean activity $\mu(x) = \mathbb{E}(Y \mid X = x)$, which represents the expected activity of compound x.

2. **Calibration:** Use D_{cal1} to construct a pointwise uncertainty estimator $\hat{\sigma}(\cdot)$. Then, for each $(x_i, y_i) \in D_{\text{cal2}}$, define RSI-EC score by

$$z_i := z(\hat{\mu}(x_i),\, \hat{\sigma}(x_i))\,, \qquad i \in I_{\text{cal2}}.$$

   Here, the score function $z(\cdot, \cdot)$ is setting-specific, and its explicit form is given later. Assume $\hat{\sigma}(x) > 0$ for all x. Let $z_{\min} := \min_{i \in I_{\text{cal2}}} z_i$ and $z_{\max} := \max_{i \in I_{\text{cal2}}} z_i$ be the minimum and maximum calibration scores on D_{cal2}, respectively. For a prespecified grid size $N \geq 2$, we construct a cutoff grid on the interval $[z_{\min}, z_{\max}]$ by selecting N equally spaced points $\mathcal{T} := \{t_1, t_2, \cdots, t_N\}$, with $t_1 = z_{\min}$ and $t_N = z_{\max}$. For each candidate cutoff $t_k \in \mathcal{T}$, define the corresponding selection set

$$\mathcal{S}_k := \{\, i \in I_{\text{cal2}} : z_i \geq t_k \,\}.$$

   Using the setting-specific FDR estimator (cf. (1) in Setting I and (2) in Setting II), compute $\widehat{\text{FDR}}_k$ for each $\mathcal{S}_k$, and choose the smallest cutoff satisfying the target level q:

$$\hat{t} \;=\; \min\Big\{\, t_k \in \mathcal{T} : \widehat{\text{FDR}}_k \leq q\,,\; k = 1, 2, \cdots, N \Big\}.$$

If no candidate cutoff satisfies the constraint, set $\hat{t} = +\infty$.

3. **Test:** For each test point x_{n+j} ($j = 1, \ldots, m$, where $m := |D_{\text{test}}|$), compute $z_{n+j} := z(\hat{\mu}(x_{n+j}), \hat{\sigma}(x_{n+j}))$, and output the selected test compounds

$$\mathcal{S} := \{\, j \in \{1, \ldots, m\} : z_{n+j} \geq \hat{t} \,\}.$$

RSI-EC scores for different settings. RSI-EC adapts to different screening objectives through a setting-specific score function $z(\hat{\mu}(x_i), \hat{\sigma}(x_i))$, which ranks compounds according to a standardized distance relative to the two cutoffs $c_1 < c_2$.

Setting I. In Setting I, compounds are scored by their standardized proximity to the interior of the interval (c_1, c_2):

$$z_i \;=\; \frac{\min\{\hat{\mu}(x_i) - c_1,\; c_2 - \hat{\mu}(x_i)\}}{\hat{\sigma}(x_i)}. \tag{3}$$

Note that $z_i > 0$ if and only if $\hat{\mu}(x_i) \in (c_1, c_2)$, whereas $z_i \leq 0$ otherwise. Therefore, larger values of z_i correspond to predictions lying deeper inside (c_1, c_2), while more negative values indicate predictions lying farther outside the interval, in both cases relative to the local uncertainty $\hat{\sigma}(x_i)$.

Setting II. In Setting II, compounds are scored by their standardized deviation away from the center interval (c_1, c_2):

$$z_i \;=\; \frac{\max\{c_1 - \hat{\mu}(x_i),\; \hat{\mu}(x_i) - c_2\}}{\hat{\sigma}(x_i)}.$$

Here $z_i > 0$ if and only if $\hat{\mu}(x_i) \notin (c_1, c_2)$ and $z_i \leq 0$ otherwise. Thus, larger values of z_i indicate predictions lying farther into the exterior region $\mathcal{R}_1$, while non-positive values correspond to predictions within the center interval, again scaled by $\hat{\sigma}(x_i)$.

The RSI-EC procedure provides a practical approach to the RS framework through empirical calibration and thresholding of an uncertainty-normalized score. However, this approach does not provide exact finite-sample FDR control, since its threshold calibration relies on empirical estimates whose validity is only justified asymptotically. We therefore also consider the CS procedure, which provides finite-sample guarantees under exchangeability.

Conformal Selection-based method

The RSI-CS realizes the RS framework by converting region-based hypotheses into valid conformal p-values and then applying the BH procedure[14,26] to control the FDR.

We use the same test dataset notation $D_{\text{test}} = \{x_{n+j}\}_{j=1}^{m}$ as defined earlier, and randomly split $D_{\text{obs}} = \{(x_i, y_i)\}_{i=1}^{n}$ into disjoint training and calibration subsets by partitioning the index set $\{1, \ldots, n\}$ into I_{train} and I_{cal}. We define $D_{\text{train}} = \{(x_i, y_i) : i \in I_{\text{train}}\}$ and $D_{\text{cal}} = \{(x_i, y_i) : i \in I_{\text{cal}}\}$, with size $n_{\text{cal}} := |D_{\text{cal}}|$.

The core of CS is a nonconformity function $V(x, y)$, which measures how atypical a sample (x, y) is relative to the calibration data. For calibration samples $i \in I_{\text{cal}}$, the nonconformity scores are computed as $V_i = V(x_i, y_i)$. For each test compound j, since value y_{n+j} is unobserved, a candidate score $\widehat{V}_j = V(x_{n+j}, c)$ is computed by plugging in a predetermined value c specified by the null hypothesis H_{0j}. Score functions V are different for Settings I and II. A more detailed examination of V and the selection of the value c will be presented in a subsequent subsection. RSI-CS then constructs conformal p-values p_j for $j = 1, \ldots, m$ for the compounds in D_{test} by comparing $\widehat{V}_j$ to the calibration scores $\{V_i\}_{i \in I_{\text{cal}}}$ in a rank-based manner; smaller p_j indicates stronger evidence against corresponding null H_{0j}. The BH procedure is then applied to $p_1, \ldots, p_m$ to find the final selection set with FDR control. Here the only setting-dependent component is the score construction (i.e., the choice of V and c); once p-values are obtained, the same BH procedure is applied identically in both Settings I and II.

General procedure. The RSI-CS procedure consists of the following three steps:

1. **Training:** Fit the predictive model required for the score construction below using D_{train}. In our implementation, we use a probabilistic classifier $\hat{\pi}(\cdot)$ to estimate $\pi(x) = \mathbb{P}(Y \in \mathcal{R}_1 \mid X = x)$. Although the activity outcome is continuous, we convert it here into a binary region-membership label indicating whether $Y \in \mathcal{R}_1$. This reformulation is introduced to facilitate score construction in our two-sided regional setting, where it helps define a score function with the desired regional monotonicity property. We therefore train a binary probabilistic classifier on D_{train} to estimate $\pi(x)$.

   Specifically, we transform the continuous compound activity outcome into a binary label indicating membership in $\mathcal{R}_1$, and then train a binary probabilistic classifier on D_{train}. Additional details on this activity transformation, along with comparisons across alternative data transformations and model choices, are provided in Appendix B.3.

2. **Calibration & scoring:** Compute calibration scores V_i for $(x_i, y_i) \in D_{\text{cal}}$ and, for each test compound in D_{test}, compute a corresponding test score $\widehat{V}_j$. In our convention, stronger evidence against H_{0j} corresponds to smaller scores. The conformal p-value for test compound j is

$$p_j \;=\; \frac{\sum_{i \in I_{\text{cal}}} \mathbb{1}\{V_i < \widehat{V}_j\} + U_j \left(1 + \sum_{i \in I_{\text{cal}}} \mathbb{1}\{V_i = \widehat{V}_j\}\right)}{n_{\text{cal}} + 1},$$

   where $U_j \sim \text{Unif}(0,1)$ is used for breaking ties in scores[18,19]. Smaller values of p_j indicate stronger evidence against H_{0j}.

3. **Test (BH procedure):** Apply the BH procedure to the set of conformal p-values associated with the test compounds. Let $p_{(1)} \leq \cdots \leq p_{(m)}$ denote the ordered p-values (with $m := |D_{\text{test}}|$), and define $r = \max\{k : p_{(k)} \leq kq/m\}$ (with $r = 0$ if the set is

empty). Return the final selection set

$$\mathcal{S} \;=\; \{\, j \in \{1, \ldots, m\} : p_j \leq p_{(r)} \,\}.$$

To complete the RSI-CS construction, it remains to specify score functions that are compatible with the regional hypotheses and preserve valid FDR control.

Score functions for different settings. In standard conformal selection, to ensure statistical guarantees in FDR control, the one-sided testing formulation typically relies on a score function satisfying a monotonicity condition. In the RS framework, however, decisions are driven by regional (two-sided) hypotheses, so the usual monotonicity condition is no longer directly applicable. Instead, we require the score function to satisfy the following *regional monotonicity* condition[23,24] to obtain valid FDR control.

Definition 1 (Regional monotonicity)**.** *A score function $V : \mathcal{X} \times \mathcal{Y} \to \mathbb{R}$ satisfies the regional monotonicity property if*

$$V(x, y') \leq V(x, y) \quad \textit{for all } x \in \mathcal{X},\ y' \in \mathcal{R}_0,\ y \in \mathcal{R}_1.$$

For Settings I and II, we adopt the clipped score (defined below) as the default score construction because it is naturally aligned with the regional hypothesis structure, satisfies the regional monotonicity property required by the RSI-CS framework, and achieves strong empirical selection power. The proof of regional monotonicity is provided in Appendix A. For empirical reference, Appendix B.2 additionally compares the clipped score with two residual-based heuristic baselines, which do not satisfy regional monotonicity and are included only to illustrate the empirical advantage of the clipped construction under the RS framework.

Setting I. In Setting I, the target (alternative) region is $\mathcal{R}_1 = (c_1, c_2)$ and the null region is $\mathcal{R}_0 = (-\infty, c_1] \cup [c_2, +\infty)$. Train the predictive model $\hat{\pi}(x)$ to estimate the probability

$\mathbb{P}(y \in \mathcal{R}_1 \mid X = x)$ and define the scores as

$$\textbf{Calibration:} \quad V_i = V(x_i, y_i) = M\,\mathbb{1}\{y_i \in \mathcal{R}_1 \setminus \partial\mathcal{R}_1\} - \hat{\pi}(x_i),\ i \in I_{\text{cal}},$$

where $\partial\mathcal{R}_1$ is the boundary of $\mathcal{R}_1$, i.e. $\{c_1, c_2\}$, and M is chosen sufficiently large so that the target and non-target scores are separated, i.e.,

$$\inf_x V(x, y) \;\geq\; \sup_x V(x, y'), \qquad \text{for all } y' \in \mathcal{R}_0 \text{ and } y \in \mathcal{R}_1.$$

This holds whenever $M \geq \sup_x \hat{\pi}(x) - \inf_x \hat{\pi}(x)$; a convenient sufficient choice is $M \geq 2\sup_x |\hat{\pi}(x)|$[18].

For test compounds $j = 1, \ldots, m$, we define the test score as:

$$\textbf{Test:} \quad \widehat{V}_j := V(x_{n+j}, c) = -\hat{\pi}(x_{n+j}), \qquad c \in \partial\mathcal{R}_1 = \{c_1, c_2\},$$

where c can be either c_1 or c_2. Because the clipped score excludes the boundary, we have $V(x, c_1) = V(x, c_2) = -\hat{\pi}(x)$, so $\widehat{V}_j$ is well defined and independent of the choice of boundary point.

Setting II. In Setting II, the target (alternative) region is the exterior of the center interval, $\mathcal{R}_1 = (-\infty, c_1] \cup [c_2, +\infty)$, with null region $\mathcal{R}_0 = (c_1, c_2)$. We again train $\hat{\pi}(x) = \widehat{\mathbb{P}}(y \in \mathcal{R}_1 \mid X = x)$, and use the same functional form of the score:

$$\textbf{Calibration:} \quad V_i = M\,\mathbb{1}\{y_i \in \mathcal{R}_1 \setminus \partial\mathcal{R}_1\} - \hat{\pi}(x_i),\ i \in I_{\text{cal}}, \qquad \textbf{Test:} \quad \widehat{V}_j = -\hat{\pi}(x_{n+j}).$$

Thus, Setting II changes the screening target by redefining $\mathcal{R}_1$, while preserving the same conformal construction and the downstream BH procedure.

In the formulations above, false discoveries are treated symmetrically under the target FDR criterion. In practice, however, different types of erroneous regional assignments may

induce substantially different downstream costs. This motivates a cost-aware extension of RS, introduced below.

Extension: Cost-Aware Regional Selection

We next consider cost asymmetry under Setting I. In this setting, false discoveries may arise from Predicted regions. Specifically, a false discovery corresponds to selecting a compound whose true response lies in the True Fail and True Pass regions $\mathcal{R}_0 = (-\infty, c_1] \cup [c_2, \infty)$ as if it belonged to the True Indeterminate region $\mathcal{R}_1 = (c_1, c_2)$. The two components of $\mathcal{R}_0$ may have different practical consequences. In early-stage drug discovery, incorrectly advancing a truly failing compound ($Y \leq c_1$) to additional validation may be substantially more costly than temporarily delaying a truly passing compound ($Y \geq c_2$), which is already a promising candidate and may only require an additional confirmatory assay.

Let $\mathcal{S} \subseteq \{1, \ldots, m\}$ denote the selected set of test compounds, and let y_{n+j} denote the response of the j-th test compound, used only for evaluation after selection. To evaluate the downstream cost induced by these asymmetric false discoveries, we define the average cost

$$\text{AverageCost} = \frac{\sum_{j \in \mathcal{S}} \left[c_{\text{fail}} \mathbb{1}\{y_{n+j} \leq c_1\} + c_{\text{pass}} \mathbb{1}\{y_{n+j} \geq c_2\} \right]}{|\mathcal{S}| \vee 1}. \tag{4}$$

Here $c_{\text{fail}} > 0$ is the cost assigned to incorrectly selecting a truly failing compound, and $c_{\text{pass}} > 0$ is the cost assigned to incorrectly selecting a truly passing compound. When $|\mathcal{S}| = 0$, the average cost is defined to be zero. Thus, (4) measures the average realized downstream cost per selected compound.

In Setting I, we consider the case $c_{\text{fail}} \gg c_{\text{pass}}$. Since the selected set is intended to recover compounds in the True Indeterminate region, the cost adjustment should penalize compounds in the True Fail region most strongly, compounds in the True Pass region more mildly, and compounds in the True Indeterminate region least. The quantity in (4) is an evaluation criterion rather than a directly available objective at the time of selection, because the test

responses y_{n+j} are unobserved when the decision is made. Nevertheless, it suggests a natural cost-sensitive surrogate based on the conditional probabilities of the True Fail and True Pass regions. Let

$$\hat{p}_{\text{fail}}(x) = \widehat{\mathbb{P}}(Y \leq c_1 \mid X = x), \qquad \hat{p}_{\text{pass}}(x) = \widehat{\mathbb{P}}(Y \geq c_2 \mid X = x),$$

denote the estimated probabilities that a compound with feature vector x lies in the True Fail and True Pass regions, respectively. We define the cost-adjustment factor

$$C_\eta(x) = \exp\left[-\eta \left\{c_{\text{fail}}\hat{p}_{\text{fail}}(x) + c_{\text{pass}}\hat{p}_{\text{pass}}(x)\right\}\right],$$

where $\eta \geq 0$ is a cost-scaling parameter. Hence $0 < C_\eta(x) \leq 1$, and larger values of η produce stronger downweighting for compounds with larger predicted downstream costs. Since the cost term is kept on the original cost scale, η should be interpreted relative to the chosen values of c_{fail} and c_{pass}.

The form of $C_\eta(x)$ implements the intended ordering of penalties across regions. Compounds likely to belong to the True Indeterminate region have small $\hat{p}_{\text{fail}}(x)$ and small $\hat{p}_{\text{pass}}(x)$, so $C_\eta(x)$ remains close to one. Compounds likely to lie in the True Pass region ($Y \geq c_2$) have large $\hat{p}_{\text{pass}}(x)$ but small $\hat{p}_{\text{fail}}(x)$, and hence are downweighted mainly through the smaller cost c_{pass}. In contrast, compounds likely to lie in the True Fail region ($Y \leq c_1$) have large $\hat{p}_{\text{fail}}(x)$, which is multiplied by the larger cost c_{fail} and therefore induces stronger downweighting. Thus, under Setting I with $c_{\text{fail}} \gg c_{\text{pass}}$, the adjustment penalizes compounds likely to be True Indeterminate least, compounds likely to lie in the True Pass region moderately, and compounds likely to lie in the True Fail region most strongly.

Cost-aware scoring in RSI-EC. For RSI-EC, let z_i denote the baseline distance-to-boundary score in (3), where larger values provide stronger evidence for the True Indeterminate region. Since z_i is signed, directly multiplying it by $C_\eta(x_i) \in (0, 1]$ would incorrectly increase

negative scores toward zero. We therefore use the sign-aware cost-adjusted score

$$z_i^{(\eta)} = z_i \{C_\eta(x_i)\}^{\mathrm{sgn}(z_i)},$$

with the convention $z_i^{(\eta)} = 0$ when $z_i = 0$. Equivalently, positive scores are multiplied by $C_\eta(x_i)$, whereas negative scores are divided by $C_\eta(x_i)$. The empirical calibration step is otherwise unchanged, with $z_i^{(\eta)}$ replacing z_i.

Cost-aware scoring in RSI-CS. For RSI-CS, we define the cost-adjusted Indeterminate probability score as

$$\hat{\pi}_\eta(x) = \hat{\pi}(x) C_\eta(x),$$

where

$$\hat{\pi}(x) = \widehat{\mathbb{P}}(Y \in \mathcal{R}_1 \mid X = x) = \widehat{\mathbb{P}}(c_1 < Y < c_2 \mid X = x)$$

is the baseline Indeterminate probability score. Since $0 < C_\eta(x) \leq 1$, the adjusted score remains nonnegative and satisfies $\hat{\pi}_\eta(x) \leq \hat{\pi}(x)$. It therefore keeps the interpretation of a cost-downweighted probability of belonging to the True Indeterminate region.

The corresponding calibration and test scores are

$$V_i^{(\eta)} = M \mathbb{1}\{y_i \in \mathcal{R}_1 \setminus \partial\mathcal{R}_1\} - \hat{\pi}_\eta(x_i), \qquad i \in I_{\mathrm{cal}},$$

and

$$\widehat{V}_j^{(\eta)} = -\hat{\pi}_\eta(x_{n+j}), \qquad j = 1, \ldots, m.$$

Here I_{cal} denotes the calibration index set, and $M > 1$ is the constant used in the clipped-score construction. The conformal p-values and the BH selection step are then computed in the same way as in the baseline RSI-CS procedure, with $V_i^{(\eta)}$ and $\widehat{V}_j^{(\eta)}$ replacing the original scores.

This cost-aware modification preserves the *regional monotonicity* structure required by

RSI-CS. For any fixed x, the term $\hat{\pi}_\eta(x)$ depends only on the feature vector and not on the realized response y. Hence it acts as a constant shift in the score as a function of y. If $y' \in \mathcal{R}_0$ and $y \in \mathcal{R}_1 \setminus \partial\mathcal{R}_1$, then

$$V^{(\eta)}(x, y') = -\hat{\pi}_\eta(x), \qquad V^{(\eta)}(x, y) = M - \hat{\pi}_\eta(x).$$

Therefore,

$$V^{(\eta)}(x, y') \leq V^{(\eta)}(x, y)$$

whenever $M > 0$. In particular, the usual choice $M > 1$ used in the clipped-score construction remains sufficient. Thus, the cost-aware RSI-CS score changes the ranking across compounds through $C_\eta(x)$, but it does not reverse the required ordering between the Predicted Fail, Predicted Pass, and Predicted Indeterminate regions for a fixed x. Consequently, treating the fitted models used to construct $C_\eta(x)$ and $\hat{\pi}_\eta(x)$ as fixed, the finite-sample FDR control argument for RSI-CS continues to apply, provided the remaining assumptions of the baseline CS procedure hold.

Numerical Studies

In this section, we evaluate the numerical performance of the proposed Regional Selection framework on the following datasets. To maintain a concise presentation and emphasize the resource-allocation objective, we focus on Setting I as the primary screening scenario, while additional results for Setting II are deferred to Appendix C. Setting I captures the practically most relevant scenario, where the goal is to identify candidates that warrant experimental validation efficiently.

Dataset. We employed 15 activity prediction tasks from the 2012 Merck Molecular Activity Challenge (Kaggle)[27–29]. Each task provides experimentally measured activity values as continuous regression targets, together with fixed-length molecular descriptor representations

released with the competition. Across tasks, the number of compounds ranges from approximately 1,500 to 38,000, and the descriptor dimensionality ranges from roughly 4,100 to 9,200. These benchmarks have been widely used in QSAR evaluation and provide a standardized testbed spanning heterogeneous sample sizes and high-dimensional molecular representations. Although the underlying QSAR tasks are formulated as regression problems, our screening

Table 3: Summary of dataset sizes, descriptor dimensionality, activity cutoffs (c_1, c_2), and the resulting Indeterminate proportion for the 15 Kaggle QSAR datasets.

Dataset	**Number of Compounds**	**Number of Descriptors**	**Cutoffs** (c_1, c_2)	**Indeterminate Proportion (%)**
3A4	37241	9177	(4.5, 6.0)	21
CB1	8716	5555	(6.5, 8.0)	14
DPP4	6148	5025	(6.0, 7.0)	36
HIVINT	1815	4186	(6.0, 7.0)	61
HIVPROT	3212	5751	(7.0, 9.0)	28
LOGD	37388	8623	(3.0, 5.0)	12
METAB	1569	4372	(40.0, 70.0)	11
NK1	9965	5592	(8.5, 9.5)	18
OX1	5351	4601	(5.8, 7.5)	29
OX2	11151	5462	(6.0, 8.0)	41
PGP	6399	4731	(0.1, 0.8)	32
PPB	8651	4991	(1.0, 2.0)	49
RAT_F	6105	5525	(1.0, 1.9)	20
TDI	4165	5712	(0.0, 1.0)	65
THROMBIN	5059	5282	(6.0, 9.0)	14

framework operates through region-based decisions induced by two activity cutoffs (c_1, c_2), which define the True Fail, True Indeterminate, and True Pass regions. Because such cutoffs are application-dependent and often chosen heuristically in practice, we fix one cutoff pair per dataset to instantiate the RS task; the resulting values are reported in Table 3. These cutoff pairs are held fixed across all repetitions and across all competing methods within each dataset. They are chosen to yield a non-degenerate True Indeterminate region whose prevalence is neither vanishingly small nor overwhelmingly large. Across the 15 datasets, the resulting proportion of compounds in the True Indeterminate region, i.e., $y \in (c_1, c_2)$, ranges from approximately 11% to 65%. This variability produces a realistic range of borderline-set

sizes and facilitates comparison of screening behavior under different region prevalences. Table 3 summarizes the dataset sizes, descriptor dimensions, cutoff pairs, and resulting True Indeterminate proportions.

With the datasets and RS cutoffs specified, we next describe the experimental protocol used to apply RSI-EC, RSI-CS, and the baseline screening procedure and to evaluate their performance.

Procedure. Unless otherwise stated, all methods follow the data-splitting and calibration protocols described in Section Methodology. For each dataset, we evaluate performance under Setting I in two data regimes: (i) the full dataset, and (ii) a limited-data regime obtained by first uniformly sampling 10% of the compounds without replacement and then applying the same splitting procedure within that subset. Within each regime, we randomly partition the available data into D_{train} (50%), D_{cal} (35%), and D_{test} (15%).

We compare three screening procedures: (i) RSI-CS, (ii) RSI-EC, and (iii) the baseline two-stage construction introduced in Section Methodology. For RSI-CS and RSI-EC, the default predictive models are random forests. In RSI-EC, the regression model $\hat{\mu}$ is used to estimate activity, and an additional pointwise uncertainty model $\hat{\sigma}(\cdot)$ is used to normalize the distance-to-boundary score. Accordingly, the 35% calibration split is further divided into D_{cal1} (20% of the total data) and D_{cal2} (15%) for uncertainty estimation and score-threshold calibration, respectively. In RSI-CS, the calibration split D_{cal} is used directly to construct conformal scores and p-values.

For the Baseline method, we implement the independent two-stage construction described earlier using conformal selection separately at the two boundaries. More precisely, one one-sided conformal screening step is used to test against the Predicted Fail region, and another is used to test against the Predicted Pass region, and the final Predicted Indeterminate set is taken as the intersection of the two resulting rejection sets. This preserves the conceptual structure of the baseline while making its implementation comparable to the proposed

methods.

To account for variability due to random splitting, and due to the additional subsampling step in the 10% regime, we repeat the entire pipeline 100 times for each dataset and report averages over repetitions. Performance is evaluated on the held-out test split by comparing the induced regional assignments with the observed test activities. Throughout this section, the reported false discovery metric is the empirical false discovery proportion (FDP), averaged over repetitions.

Results. Figures 8 and 9 compare observed FDP and power for RSI-CS, RSI-EC, and Baseline across the 15 Kaggle QSAR datasets under Setting I. Here the Baseline refers to the independent two-stage construction introduced in Section Methodology. We vary the nominal target level over $q \in \{0.1, 0.2, \ldots, 1.0\}$ and plot either observed FDP or power on the y-axis against q on the x-axis. The gray dashed line ($y = x$) represents ideal FDP calibration.

Across most datasets and nominal levels, both RSI-CS and RSI-EC show observed FDP curves that track the target level reasonably closely in both data regimes. This pattern is especially clear on the full datasets (Figure 9, top), where the two methods exhibit similar behavior over a broad range of q. On the 10% subsets (Figure 8, top), RSI-CS is generally more stable, particularly at smaller nominal levels, whereas RSI-EC exhibits somewhat greater variability, which is consistent with its calibration-by-thresholding construction and large-sample motivation. By contrast, the baseline method shows less stable FDP control across datasets. In several cases, the observed FDP exceeds the nominal level over part of the q range, indicating inflation rather than conservativeness.

Power increases monotonically with q across datasets (Figures 8 and 9, bottom), as expected: allowing a larger false discovery budget permits more selections and therefore more true discoveries. On the 10% subsets, RSI-EC often attains slightly higher power than RSI-CS, consistent with less conservative behavior in smaller samples. This gap narrows on the full datasets, where the two methods achieve broadly similar power while maintaining comparable

observed FDP behavior. By contrast, the Baseline method often attains comparable or even higher power, but these gains are frequently accompanied by weaker FDP control, with observed FDP exceeding the nominal level in several datasets.

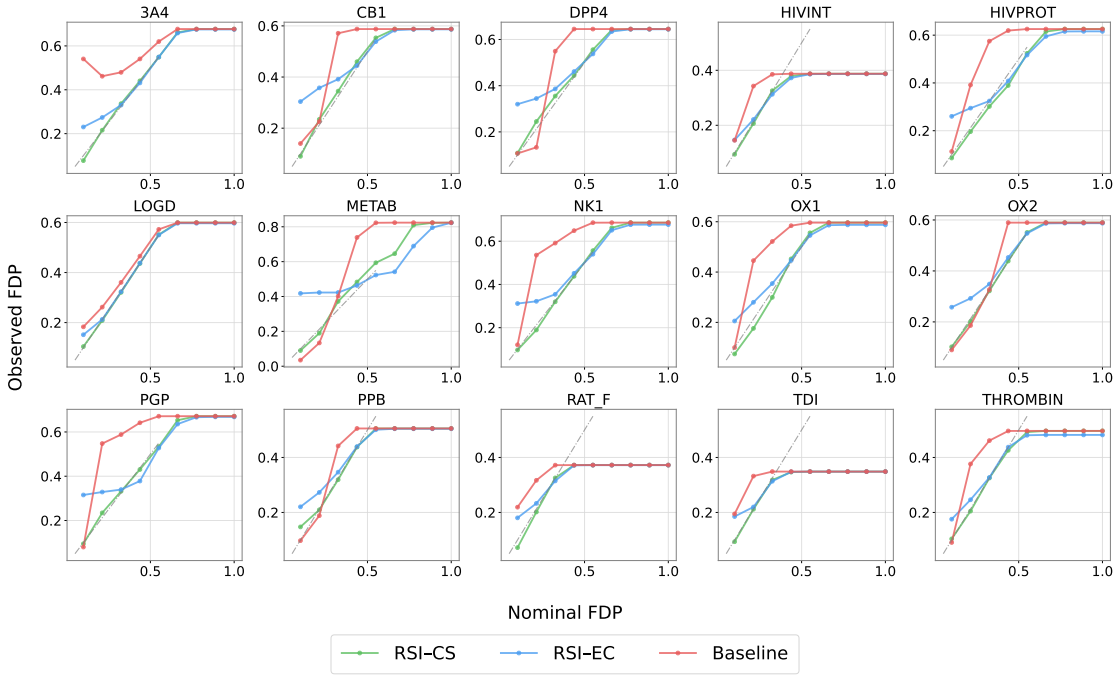
Several datasets also exhibit a plateau effect at larger nominal levels. Once the selected set has essentially saturated under a given score construction and cutoff specification, further increasing q no longer materially changes the selected compounds. As a result, both observed FDP and power flatten out.

Taken together, these results support the empirical validity of the proposed RSI framework under Setting I. Relative to the independent two-stage baseline, the proposed region-aware formulation yields a more effective FDP–power trade-off while maintaining a direct statistical interpretation of the final regional assignment. We further assess robustness to key design choices by varying (i) the prediction model, (ii) the nonconformity score construction in RSI-CS, and (iii) the data transformation used in the conformal pipeline; see Appendix B for details.

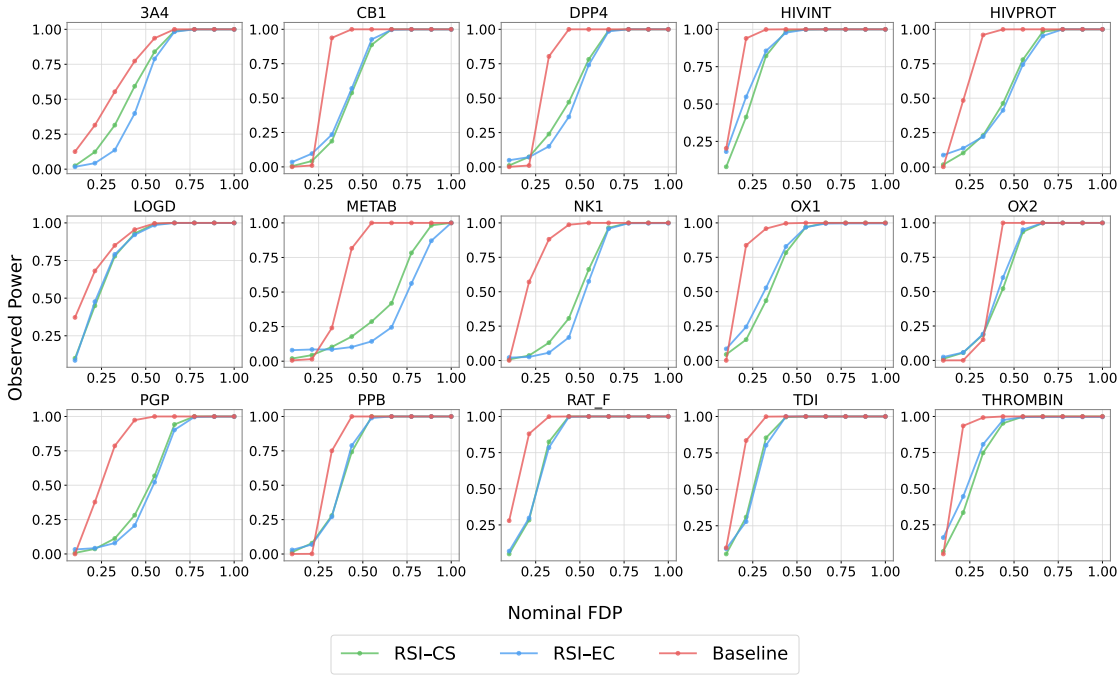
Cost-aware trade-off analysis. We next examine how the cost-aware tuning parameter affects average cost, power, observed FDP, and the composition of the selected set, with particular attention to the resulting cost–power trade-off. Throughout this analysis, q denotes the target nominal FDR level for the set of compounds selected for Indeterminate follow-up. We fix the asymmetric cost weights at $(c_{\text{pass}}, c_{\text{fail}}) = (1, 10)$, so that incorrectly carrying forward a truly failing compound is treated as substantially more costly than delaying a truly passing compound.

The baseline procedure corresponds to $\eta = 0$. Larger values of η increasingly downweight compounds with larger predicted cost exposure $C(x)$. In the diagnostic plots, we report the representative tuning values

$$\eta \in \{0, 0.01, 0.03, 0.05, 0.1, 0.2\}.$$

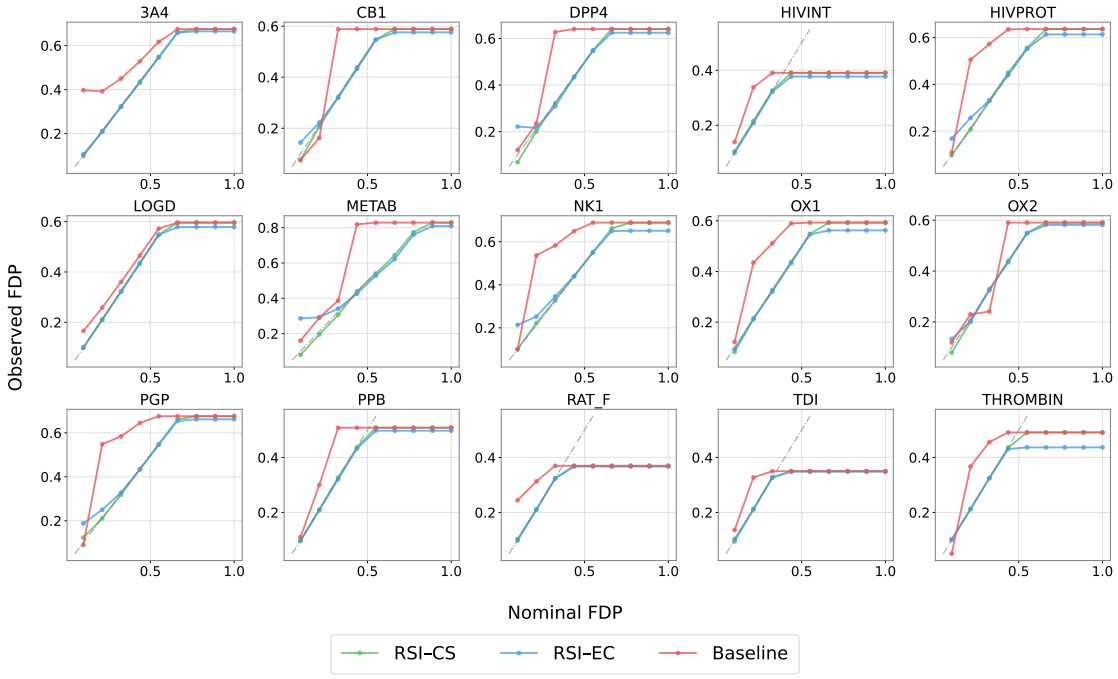


(a) FDP control on 10% subsets of the 15 Kaggle datasets.

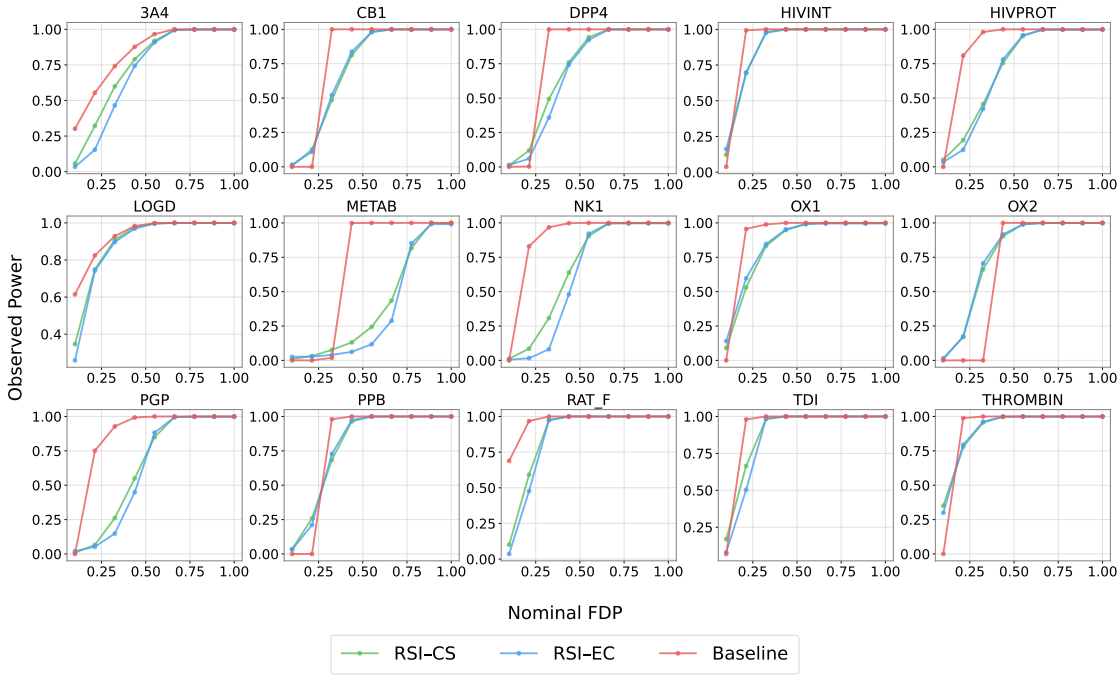


(b) Power on 10% subsets of the 15 Kaggle datasets.

Figure 8: Observed FDP (a) and Power (b) as functions of the nominal level $q \in \{0.1, 0.2, \ldots, 1.0\}$ on 10% random subsets of the 15 Kaggle datasets. The gray dashed line indicates ideal calibration (FDP $= q$).



(a) FDP control on the entirety of the 15 Kaggle datasets.



(b) Power on the entirety of the 15 Kaggle datasets.

Figure 9: Observed FDP (a) and Power (b) as functions of the nominal level $q \in \{0.1, 0.2, \ldots, 1.0\}$ on the full datasets. The gray dashed line indicates ideal calibration (FDP $= q$).

For each configuration, we record the average cost, empirical power, observed FDP, and the mean number of selected compounds whose true responses fall in the True Fail, True Indeterminate, and True Pass regions.

In the main text, we focus on the representative nominal level $q = 0.3$ and display two datasets, HIVPROT and OX2, for both RSI-EC and RSI-CS in Figure 10. These two datasets are selected because they clearly illustrate the empirical cost–power trade-off induced by the cost-aware adjustment. The purpose of this figure is diagnostic: it shows how increasing η can reduce average cost while changing power and the composition of the selected set. Results for other datasets and nominal levels are reported in Appendix D.

Each row corresponds to one dataset. The left panel reports the absolute operating characteristics: average cost, observed FDP, and empirical power as functions of η. The middle panel reports changes relative to the baseline $\eta = 0$. The right panel reports changes in the selected-set composition, again relative to $\eta = 0$.

More precisely, let $A(\eta)$, $P(\eta)$, and $F(\eta)$ denote the average cost, empirical power, and observed FDP at tuning value η, respectively, for a fixed dataset, method, and nominal level q. The middle panel uses the following quantities:

$$\Delta_{\text{cost}}(\eta) = 100\frac{A(0) - A(\eta)}{A(0)},$$

$$\Delta_{\text{power}}(\eta) = 100\{P(\eta) - P(0)\}, \qquad \Delta_{\text{FDP}}(\eta) = 100\{F(\eta) - F(0)\}.$$

Thus, cost is reported as a relative percentage reduction, whereas power and FDP are reported as percentage-point changes. A positive value of Δ_{cost} means that the cost-aware score reduces average cost relative to the baseline. A positive value of Δ_{power} means that empirical power increases relative to the baseline, while a positive value of Δ_{FDP} means that the observed FDP increases relative to the baseline.

The right panel decomposes the selected set. Let $n_{\text{fail}}(\eta)$, $n_{\text{ind}}(\eta)$, and $n_{\text{pass}}(\eta)$ denote the mean numbers of selected compounds whose true responses lie in the True Fail, True

Indeterminate, and True Pass regions, respectively. We plot

$$\Delta n_{\text{fail}}(\eta) = n_{\text{fail}}(\eta) - n_{\text{fail}}(0),$$

$$\Delta n_{\text{ind}}(\eta) = n_{\text{ind}}(\eta) - n_{\text{ind}}(0), \qquad \Delta n_{\text{pass}}(\eta) = n_{\text{pass}}(\eta) - n_{\text{pass}}(0).$$

These curves identify the source of the cost and power changes. A decrease in Δn_{fail} indicates that the cost-aware score selects fewer truly failing compounds than the baseline, which is desirable under $c_{\text{fail}} > c_{\text{pass}}$. An increase in Δn_{ind} contributes to power, since these are the target compounds for Indeterminate follow-up. Changes in Δn_{pass} reflect the extent to which truly passing compounds are assigned to Predicted Indeterminate follow-up rather than directly treated as Predicted Pass.

We first consider RSI-EC. In Figure 10(a), the observed FDP remains close to the baseline across the tuning path, indicating that the cost-aware modification mainly changes the ranking and composition of the selected set rather than visibly increasing the empirical false-discovery proportion. For HIVPROT, average cost decreases as η increases, while empirical power is preserved and slightly improved. The selected-count decomposition shows that this improvement is associated with an increase in selected True Indeterminate compounds and a decrease in selected True Fail compounds. Thus, for HIVPROT, RSI-EC reduces average cost without sacrificing power.

For OX2 under RSI-EC, the same pattern is more pronounced. Average cost decreases with η, while power increases relative to the baseline. The selected-count panel shows a substantial increase in selected True Indeterminate compounds, together with fewer selected True Fail compounds at larger values of η. This indicates that the cost-aware score enriches the selected set with target True Indeterminate compounds while reducing costly false selections. Hence, in these two examples, RSI-EC shows a favorable cost-aware pattern: average cost is reduced while empirical power is maintained or improved.

We next consider RSI-CS. In Figure 10(b), the cost reduction is stronger than under

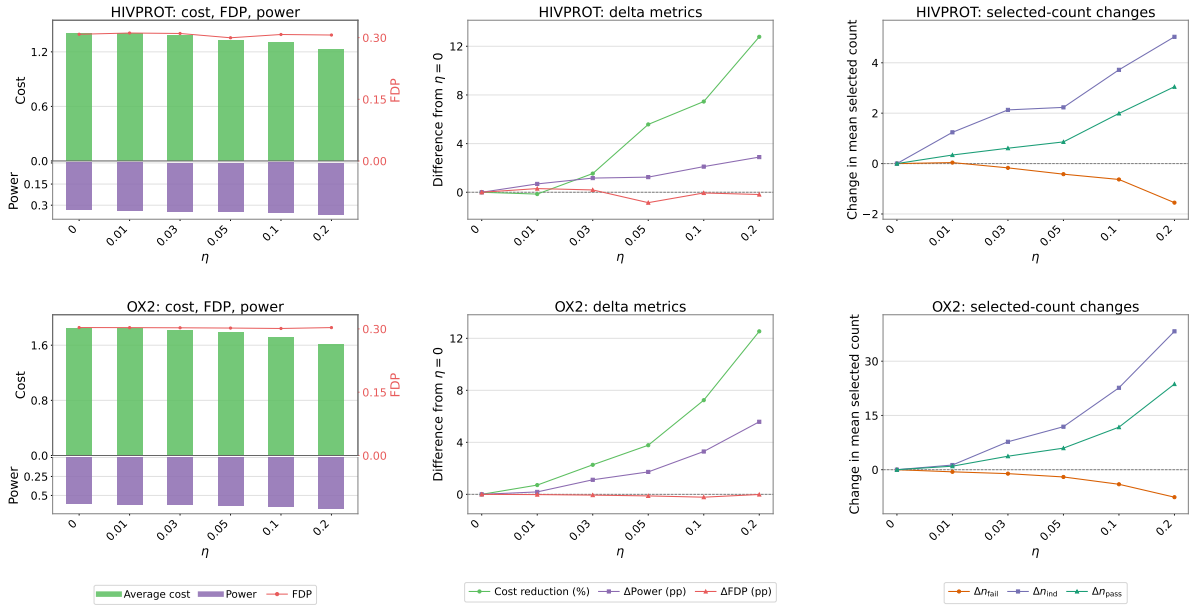
RSI-EC. For both HIVPROT and OX2, average cost decreases substantially as η increases, while the observed FDP remains broadly stable relative to the baseline. However, the power response displays a clearer cost–power trade-off. For small to moderate values of η, power is approximately stable, but at larger values of η power decreases, especially for OX2.

The selected-count decomposition explains this behavior. Under RSI-CS, increasing η removes many selected True Fail compounds, which directly reduces average cost under the asymmetric cost weights. At the same time, stronger cost downweighting eventually removes some selected True Indeterminate compounds, particularly for OX2. Since True Indeterminate compounds are the target discoveries, this reduction translates into lower empirical power. Thus, RSI-CS exhibits the intended cost-sensitive behavior, but also shows that overly aggressive cost downweighting can reduce power.

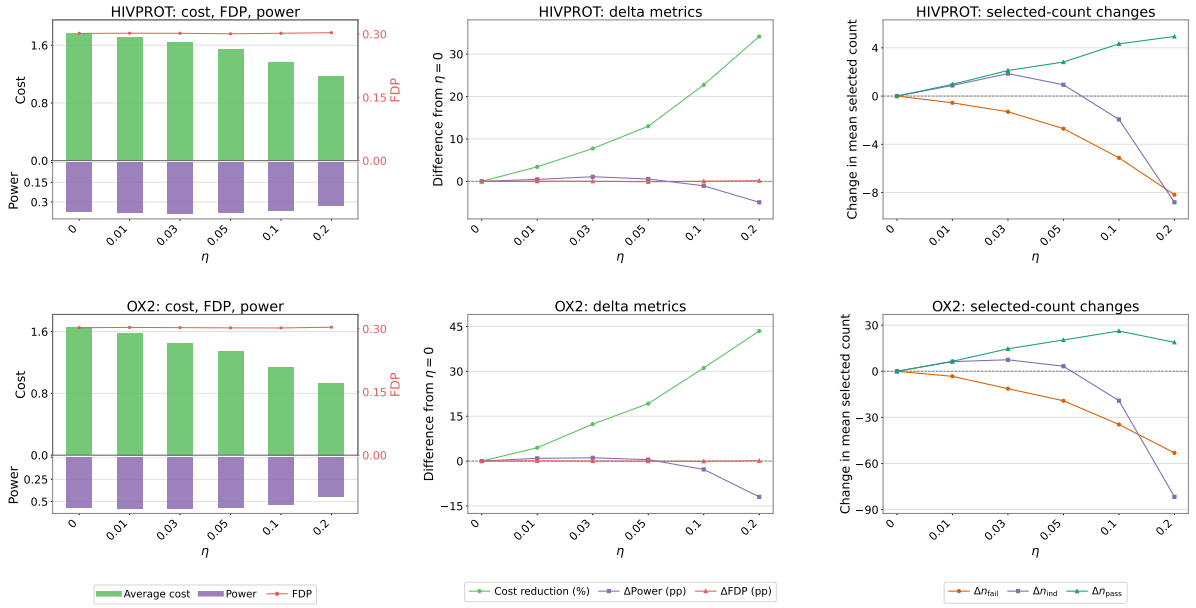
Overall, Figure 10 shows that η acts as a cost–power tuning parameter. RSI-EC gives a favorable pattern in these two datasets, reducing average cost while preserving or improving power. RSI-CS gives stronger cost reductions, but at large η this can come at the cost of selecting fewer true Indeterminate compounds. These results suggest that η should be chosen to balance cost reduction against power, rather than simply being made as large as possible.

Conclusion

We introduced Regional Selection (RS), a three-way decision framework for molecular screening that extends the standard binary paradigm. By partitioning candidate compounds into Predicted Pass, Predicted Fail, and Predicted Indeterminate regions, RS provides a principled way to defer borderline cases for secondary validation rather than forcing immediate binary decisions. To implement this framework, we developed Regional Selection Inference (RSI), a multiple-testing formulation for region assignment with false discovery rate (FDR) control. We studied two implementations of RSI: the empirical calibration-based method (RSI-EC), which relies on score thresholding and large-sample approximation, and the conformal selection-based



(a) RSI-EC



(b) RSI-CS

Figure 10: Cost-aware diagnostic plots for RSI-EC and RSI-CS at the target nominal FDR level $q = 0.3$. The two subfigures correspond to RSI-EC and RSI-CS, respectively, and rows correspond to the representative datasets HIVPROT and OX2. The three columns show the absolute operating characteristics, changes relative to the baseline $\eta = 0$, and changes in selected-set composition.

method (RSI-CS), which provides finite-sample FDR control under exchangeability.

Across 15 publicly available Kaggle QSAR benchmarks, RSI achieved reliable empirical error control and competitive power over a wide range of nominal levels in both limited-data and full-data settings. In smaller samples, RSI-CS generally yielded more stable FDR control, whereas RSI-EC was sometimes slightly more aggressive and attained marginally higher power. As the sample size increased, the two methods behaved more similarly, with both tracking the target error level closely while maintaining high power on most datasets. By contrast, the baseline two-stage construction was less reliable. Because it combines two independently generated rejection sets rather than formulating regional assignment as a single multiple-testing problem, it does not directly provide FDR control for the final regional assignment. We also observed a saturation effect at high nominal levels, where both FDP and power approached a plateau. This reflects an inherent limit of the screening problem: once all candidates separable under a given score and cutoff specification have effectively been identified, further relaxing the target level no longer materially changes the resulting selections.

We also explored a cost-aware extension of the RS framework for applications where different types of false discoveries incur asymmetric downstream costs. Rather than modifying the FDR target itself, this extension modifies the score construction to reflect asymmetric downstream costs while preserving the original regional selection objective. The resulting behavior was clearly dataset dependent. In the representative examples, RSI-EC reduced average cost while preserving or improving power, whereas RSI-CS produced stronger cost reductions but could lose power when the cost-aware downweighting became too aggressive. These findings suggest that the tuning parameter should be viewed as controlling a cost–power trade-off, rather than as a parameter to be made as large as possible. More broadly, the results show that the RS framework can accommodate not only uncertainty-aware region assignment, but also practical decision asymmetries that arise in resource-constrained screening pipelines.

Several directions remain for future work. First, RSI depends on the quality of the

underlying predictive model and, when applicable, the associated uncertainty estimates used in score construction. Improvements in calibration and predictive accuracy may directly translate into better screening performance. Second, the current multiple-testing guarantees rely on assumptions that may be violated in practice, for example through dependence induced by chemical similarity or shared model structure. Developing region-aware error control under dependence, without excessive conservativeness, remains an important open problem. Finally, our study focuses on static datasets with fixed cutoffs, whereas practical discovery pipelines are often iterative, with thresholds revised over time and experimental feedback arriving sequentially. Extending RSI to such adaptive settings is a natural direction for future work.

Overall, RS provides a statistically grounded framework for separating automatic decisions from defer-to-experiment cases in molecular screening, and RSI translates this idea into procedures with formal error control. By representing uncertainty through an explicit Indeterminate region, the framework offers a practical foundation for more resource-aware early-stage drug discovery.

Acknowledgments

This work was partially supported by CANSSI Collaborative Research Teams Grant, NSERC Discovery Grant (RGPIN-2024-06780) and FRQNT Team Research Project Grant (FRQNT 327788).

References

(1) Walters, W. P.; Murcko, M. A. Prediction of 'drug-likeness'. *Advanced Drug Delivery Reviews* **2002**, *54*, 255–271.

(2) Dietrich, J. A.; McKee, A. E.; Keasling, J. D. High-throughput metabolic engineering: advances in small-molecule screening and selection. *Annual Review of Biochemistry* **2010**, *79*, 563–590.

(3) Macarron, R.; Banks, M. N.; Bojanic, D.; Burns, D. J.; Cirovic, D. A.; Garyantes, T.; Green, D. V.; Hertzberg, R. P.; Janzen, W. P.; Paslay, J. W.; others Impact of high-throughput screening in biomedical research. *Nature Reviews Drug discovery* **2011**, *10*, 188–195.

(4) Carracedo-Reboredo, P.; Liñares-Blanco, J.; Rodríguez-Fernández, N.; Cedrón, F.; Novoa, F. J.; Carballal, A.; Maojo, V.; Pazos, A.; Fernandez-Lozano, C. A review on machine learning approaches and trends in drug discovery. *Computational and structural biotechnology journal* **2021**, *19*, 4538–4558.

(5) Naithani, U.; Guleria, V. Integrative computational approaches for discovery and evaluation of lead compound for drug design. *Frontiers in Drug Discovery* **2024**, *4*, 1362456.

(6) Svetnik, V.; Liaw, A.; Tong, C.; Culberson, J. C.; Sheridan, R. P.; Feuston, B. P. Random forest: a classification and regression tool for compound classification and QSAR modeling. *Journal of Chemical Information and Computer Sciences* **2003**, *43*, 1947–1958.

(7) Veerasamy, R.; Rajak, H.; Jain, A.; Sivadasan, S.; Varghese, C. P.; Agrawal, R. K. Validation of QSAR models — strategies and importance. *International Journal of Drug Design and Discovery* **2011**, *3*, 511–519.

(8) Ma, J.; Sheridan, R. P.; Liaw, A.; Dahl, G. E.; Svetnik, V. Deep neural nets as a method for quantitative structure–activity relationships. *Journal of Chemical Information and Modeling* **2015**, *55*, 263–274.

(9) Zakharov, A. V.; Varlamova, E. V.; Lagunin, A. A.; Dmitriev, A. V.; Muratov, E. N.; Fourches, D.; Kuz'min, V. E.; Poroikov, V. V.; Tropsha, A.; Nicklaus, M. C. QSAR modeling and prediction of drug–drug interactions. *Molecular Pharmaceutics* **2016**, *13*, 545–556.

(10) Svensson, F.; Aniceto, N.; Norinder, U.; Cortes-Ciriano, I.; Spjuth, O.; Carlsson, L.; Bender, A. Conformal regression for quantitative structure – activity relationship modeling – quantifying prediction uncertainty. *Journal of Chemical Information and Modeling* **2018**, *58*, 1132–1140.

(11) Mansouri, K.; Cariello, N. F.; Korotcov, A.; Tkachenko, V.; Grulke, C. M.; Sprankle, C. S.; Allen, D.; Casey, W. M.; Kleinstreuer, N. C.; Williams, A. J. Open-source QSAR models for pKa prediction using multiple machine learning approaches. *Journal of Cheminformatics* **2019**, *11*, 1–20.

(12) Dahl, G. E.; Jaitly, N.; Salakhutdinov, R. Multi-task neural networks for QSAR predictions. *arXiv preprint arXiv:1406.1231* **2014**,

(13) Truchon, J.-F.; Bayly, C. I. Evaluating virtual screening methods: good and bad metrics for the "early recognition" problem. *Journal of Chemical Information and Modeling* **2007**, *47*, 488–508.

(14) Benjamini, Y.; Hochberg, Y. Controlling the false discovery rate: a practical and powerful approach to multiple testing. *Journal of the Royal Statistical Society: Series B (Methodological)* **1995**, *57*, 289–300.

(15) Sheridan, R. P.; McMasters, D. R.; Voigt, J. H.; Wildey, M. J. eCounterscreening: using QSAR predictions to prioritize testing for off-target activities and setting the balance between benefit and risk. *Journal of Chemical Information and Modeling* **2015**, *55*, 231–238.

(16) Sheridan, R. P. Three useful dimensions for domain applicability in QSAR models using random forest. *Journal of Chemical Information and Modeling* **2012**, *52*, 814–823.

(17) Sheridan, R. P. Using random forest to model the domain applicability of another random forest model. *Journal of Chemical Information and Modeling* **2013**, *53*, 2837–2850.

(18) Jin, Y.; Candès, E. J. Selection by prediction with conformal p-values. *Journal of Machine Learning Research* **2023**, *24*, 1–41.

(19) Bates, S.; Candès, E.; Lei, L.; Romano, Y.; Sesia, M. Testing for outliers with conformal p-values. *The Annals of Statistics* **2023**, *51*, 149–178.

(20) Lei, J.; G'Sell, M.; Rinaldo, A.; Tibshirani, R. J.; Wasserman, L. Distribution-free predictive inference for regression. *Journal of the American Statistical Association* **2018**, *113*, 1094–1111.

(21) Shafer, G.; Vovk, V. A tutorial on conformal prediction. *Journal of Machine Learning Research* **2008**, *9*.

(22) Bai, T.; Tang, P.; Xu, Y.; Svetnik, V.; Yang, B.; Khalili, A.; Yu, X.; Yang, A. Y. Conformal selection for efficient and accurate compound screening in drug discovery. *Journal of Chemical Information and Modeling* **2025**, *65*, 13070–13085.

(23) Bai, T.; Zhao, Y.; Yu, X.; Yang, A. Y. Multivariate conformal selection. International Conference on Machine Learning. 2025; pp 2535–2559.

(24) Hao, Q.; Liao, W.; Jing, B.; Wei, H. Multi-condition conformal selection. International Conference on Learning Representation. 2026.

(25) Goeman, J. J.; Solari, A. Multiple testing for exploratory research. *Statistical Science* **2011**, 584–597.

(26) Benjamini, Y.; Hochberg, Y. Multiple hypotheses testing with weights. *Scandinavian Journal of Statistics* **1997**, *24*, 407–418.

(27) QSAR-data for Drug Discovery. Kaggle Dataset, `https://www.kaggle.com/datasets/vivek153/data-root`.

(28) Liu, R.; Wang, H.; Glover, K. P.; Feasel, M. G.; Wallqvist, A. Dissecting machine-learning prediction of molecular activity: is an applicability domain needed for quantitative structure–activity relationship

models based on deep neural networks? *Journal of Chemical Information and Modeling* **2018**, *59*, 117–126.

(29) Kato, Y.; Hamada, S.; Goto, H. Validation study of QSAR/DNN models using the competition datasets. *Molecular Informatics* **2020**, *39*, 1900154.

(30) Bishop, C. M.; Bishop, H. *Deep Learning: Foundations and Concepts*; Springer Nature, 2023.

Appendices

Appendix A: Proof of Regional Monotonicity

Proof. We verify the regional monotonicity property, namely,

$$V(x, y') \leq V(x, y), \qquad \forall x \in \mathcal{X},\ y' \in \mathcal{R}_0,\ y \in \mathcal{R}_1.$$

Recall that

$$V(x, y) = M \mathbb{1}\{y \in \mathcal{R}_1 \setminus \partial\mathcal{R}_1\} - \hat{\pi}(x), \qquad M > 0.$$

For any fixed $x \in \mathcal{X}$ and any $y' \in \mathcal{R}_0$, we have $y' \notin \mathcal{R}_1 \setminus \partial\mathcal{R}_1$, and hence $V(x, y') = -\hat{\pi}(x)$.

In Setting I, $\mathcal{R}_1 = (c_1, c_2)$ and $\partial\mathcal{R}_1 = \{c_1, c_2\}$, so $\mathcal{R}_1 \setminus \partial\mathcal{R}_1 = \mathcal{R}_1$. Therefore, for any $y \in \mathcal{R}_1$, we have $V(x, y) = M - \hat{\pi}(x)$, which implies

$$V(x, y') = -\hat{\pi}(x) \leq M - \hat{\pi}(x) = V(x, y).$$

In Setting II, $\mathcal{R}_1 = (-\infty, c_1] \cup [c_2, \infty)$ and $\mathcal{R}_1 \setminus \partial\mathcal{R}_1 = (-\infty, c_1) \cup (c_2, \infty)$. If $y \in (-\infty, c_1) \cup (c_2, \infty)$, then $V(x, y) = M - \hat{\pi}(x)$, so again

$$V(x, y') = -\hat{\pi}(x) \leq M - \hat{\pi}(x) = V(x, y).$$

If $y \in \partial\mathcal{R}_1 = \{c_1, c_2\}$, then $V(x, y) = -\hat{\pi}(x) = V(x, y')$. Hence, in all cases, $V(x, y') \leq V(x, y)$. Therefore, V satisfies the regional monotonicity property in both Settings I and II. □

Appendix B: Additional Numerical Studies

This appendix reports supplementary robustness and sensitivity analyses that extend the main experiments. We focus on three sets of variants:

- **B.1 Model choice.** We replace the default predictor with alternative baselines,

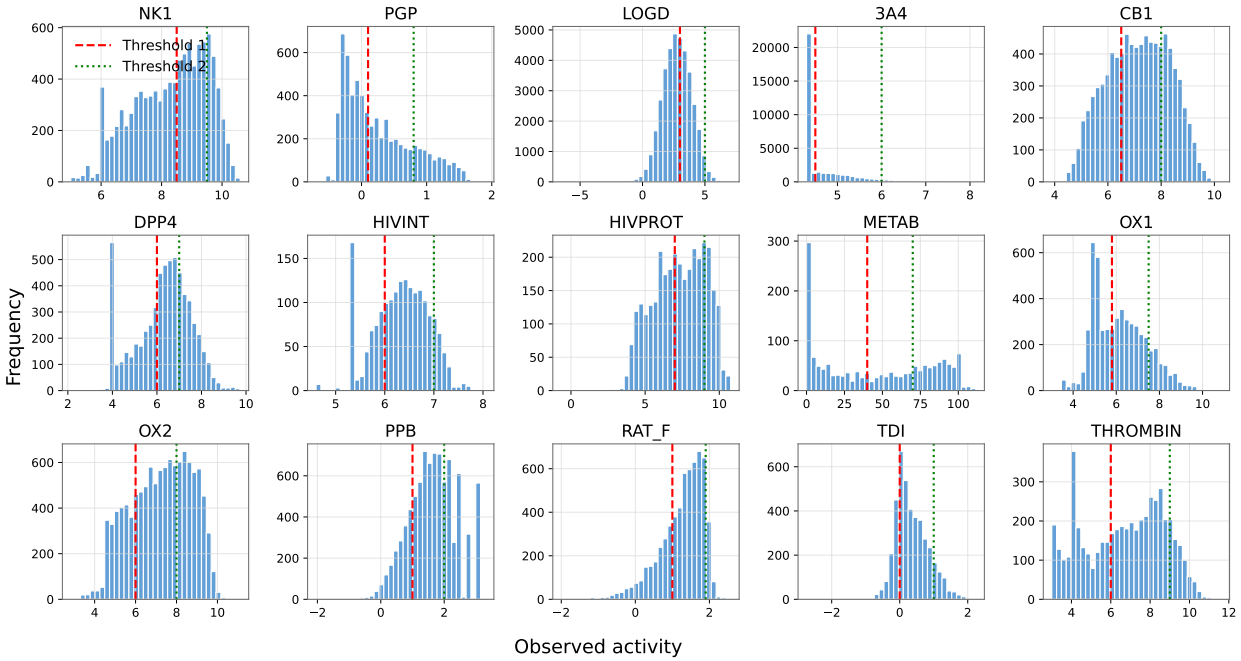


Figure A1: Empirical activity distributions for the 15 Kaggle QSAR datasets. Each panel shows a histogram of the observed activity values with the dataset-specific cutoffs (c_1, c_2) overlaid (dashed: c_1; dotted: c_2). These cutoffs define the True Fail ($y \leq c_1$), True Indeterminate ($c_1 < y < c_2$), and True Pass ($y \geq c_2$) regions used in the RSI experiments.

including Linear Regression and a Multilayer Perceptron (MLP).

- **B.2 Nonconformity scores.** We compare multiple nonconformity score constructions and evaluate their impact on screening efficiency and power.

- **B.3 Data preprocessing.** We examine the performance of Conformal Selection under different data transformations/preprocessing choices.

As background, Figure A1 summarizes the empirical activity distributions of the 15 datasets and the dataset-specific cutoffs (c_1, c_2) used to define the RS regions.

B.1 Screening Performance with Alternative Prediction Models

In this section, we apply different prediction models to investigate the robustness of FDR control for Conformal Selection and Empirical Calibration methods, and to compare their corresponding power. In addition to the random forest model employed in our main experi-

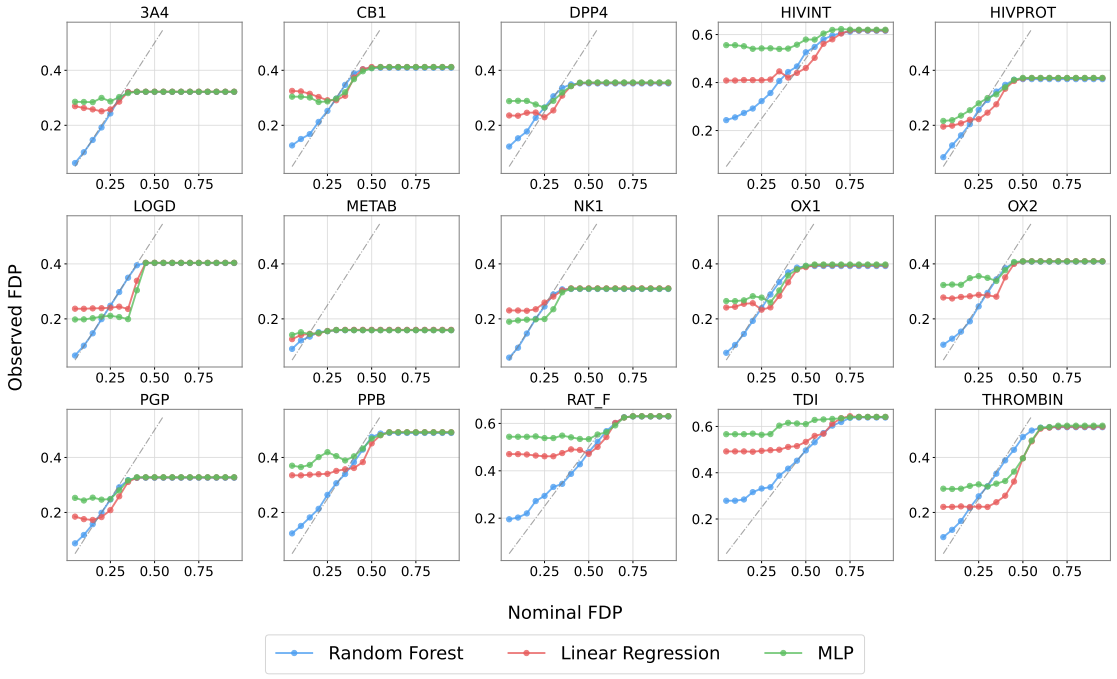
ments, we also consider two distinct alternatives: a linear regression model and a Multilayer Perceptron (MLP)[30]. The network architecture for the MLP consists of two fully-connected hidden layers, each containing 64 neurons.

Empirical Calibration-based method. We first evaluate the performance of the three prediction models using RSI-EC. Figure A2 presents the FDP control results, and Figure A3 shows the corresponding power.

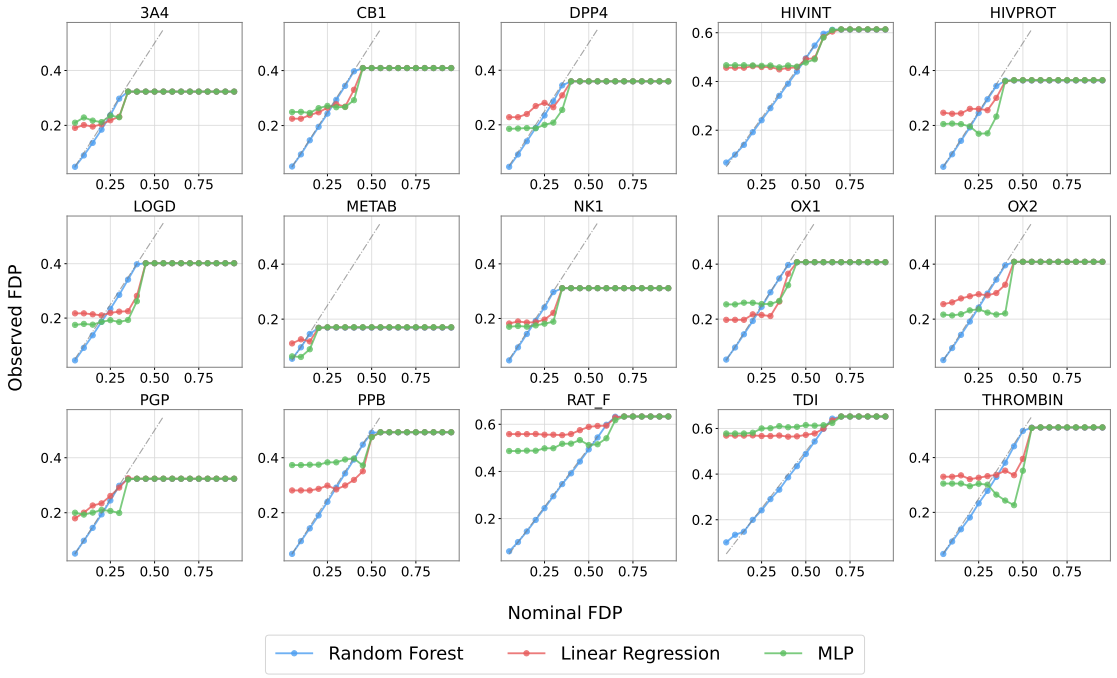
The results clearly indicate that the Random Forest model (blue curve) delivers the best overall performance. For FDP control (Figure A2), the Random Forest consistently keeps the observed FDP very close to the nominal FDP level (the gray dashed line) across the vast majority of datasets. Concurrently, it almost always maintains the highest power (Figure A3). In contrast, the other two models are not as effective. The MLP (green curve) performs similarly to the Random Forest but is slightly inferior in FDP control on several datasets (e.g., DPP4, NK1, TDI). The Linear Regression model (red curve) performs particularly poorly. FDP Control (Figure A2): Its FDP control is highly unstable. On some datasets (e.g., 3A4, LOGD, RAT_F), the observed FDP far exceeds the nominal level (making it too liberal and yielding excessive false discoveries), while on others (e.g., CB1, NK1), it is overly conservative (FDP is far below the nominal level). Power (Figure A3): Its power is significantly lower than the other two models on almost all datasets.

Conformal selection-based method. We turn to the Conformal Selection method (shown in Figures A4 and A5).

For FDP control (Figure A4), both the Random Forest and MLP models demonstrate robust performance. Their observed FDP curves (blue and green) closely track the gray dashed line (representing perfect control) on both the 10% and 100% datasets. In contrast, the Linear Regression model (red curve) provides poor FDP control, but in the opposite direction. This issue is particularly clear in the 10% subset setting (Figure A4a), where its observed FDP is overly conservative on the vast majority of datasets (e.g., 3A4, CB1, DPP4).

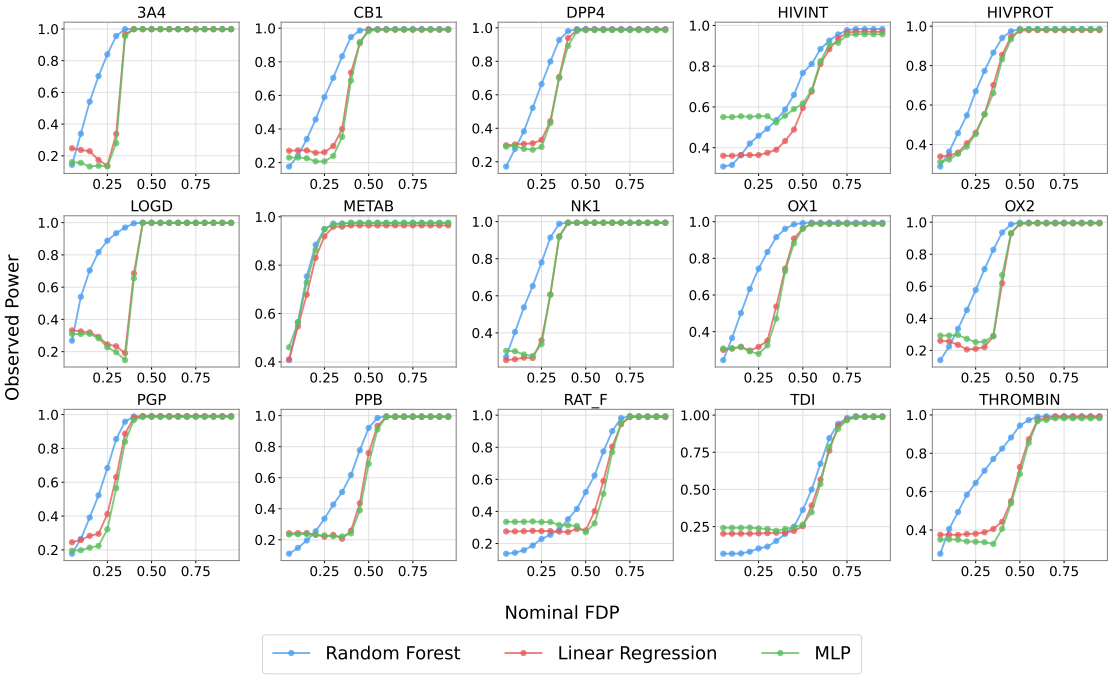


(a) FDP control on 10% subsets of the 15 Kaggle datasets.

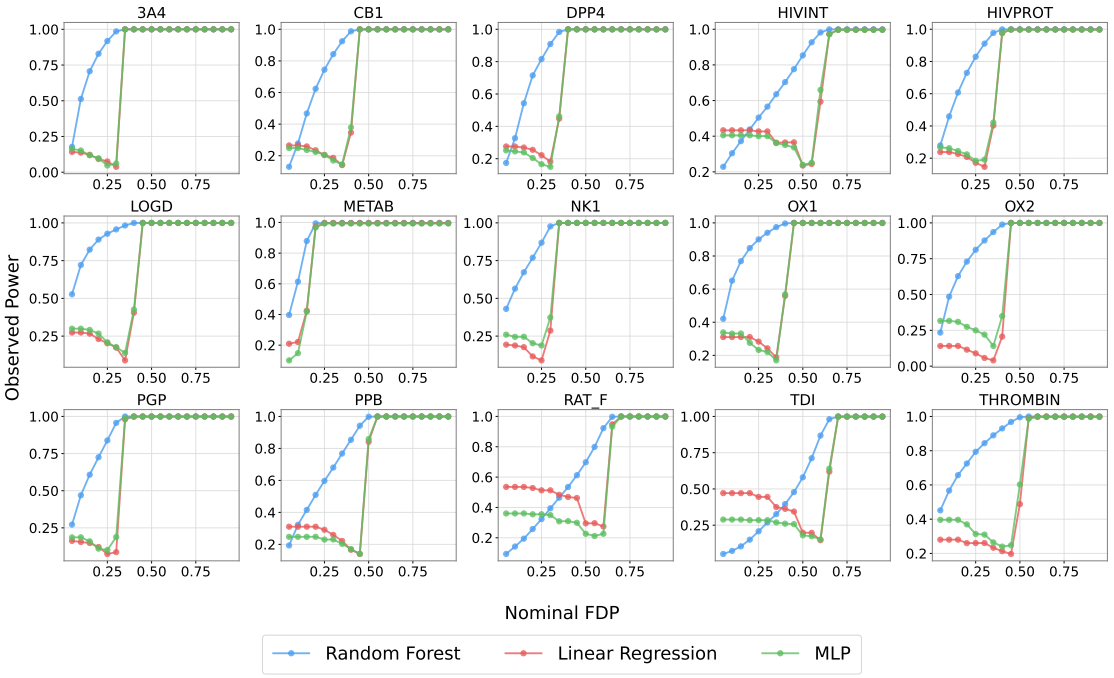


(b) FDP control on the entirety of the 15 Kaggle datasets.

Figure A2: FDP control of RSI-EC with Random Forest (blue), Linear Regression (red), and MLP (green) on (a) 10% subsets and (b) the full 15 datasets. Each subplot plots observed FDP against nominal FDP, with the gray dashed line indicating perfect control.



(a) Power on 10% subsets of the 15 Kaggle datasets.



(b) Power on the entirety of the 15 Kaggle datasets.

Figure A3: Power of RSI-EC with Random Forest (blue), Linear Regression (red), and MLP (green) on (a) 10% subsets and (b) the full 15 datasets. Each subplot plots observed Power against nominal FDP.

The red curve frequently falls far below the nominal FDP level, indicating that the method is failing to make discoveries. While this technically avoids exceeding the FDR, it confirms the model's ineffectiveness, which is also reflected in its extremely low power (shown in Figure A5).

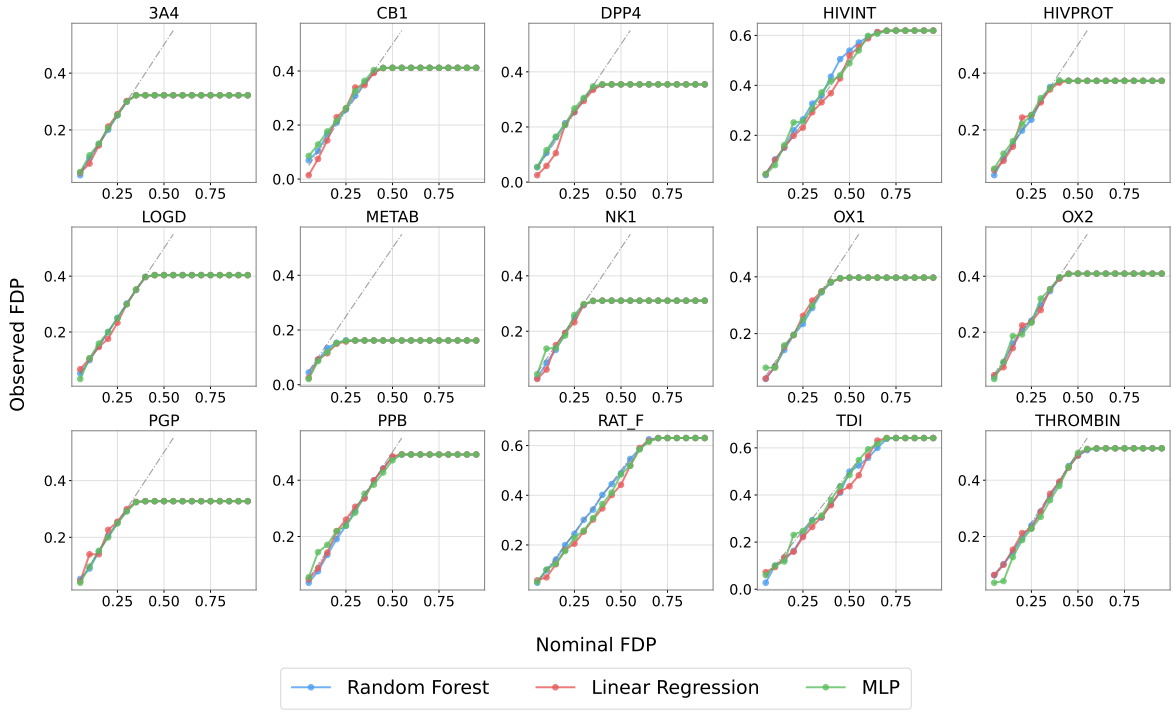
Regarding power (Figure A5), the Random Forest model achieves the best results, attaining the highest power on the vast majority of datasets. The MLP follows closely, performing nearly as well as the Random Forest. The Linear Regression model is the least ideal, exhibiting significantly lower power than the other two models across all datasets.

A key finding from this comparison is the superior robustness of Conformal Selection in FDP control. By comparing Figure A2 (RSI-EC) and Figure A4 (RSI-CS), we observe that FDP control of RSI-EC is susceptible to the quality of the prediction model. In Figure A2, only the Random Forest (blue) provides reliable control. The MLP (green) shows noticeable deviations on several datasets (e.g., DPP4, NK1), and the Linear Regression model (red) fails to maintain the nominal FDP level. In contrast, the RSI-CS (Figure A4) is far more stable. Both the Random Forest and the MLP models achieve near-perfect FDP control, with their respective curves (blue and green) closely tracking the gray dashed line. This suggests that RSI-CS is a more robust framework. It can provide reliable FDP guarantees even when paired with simpler or different model architectures (like MLP), as long as the model has sufficient predictive power. While both methods perform poorly with Linear Regression, this is likely due to the model's limited ability to capture the data complexity, as reflected in the low power shown in Figures A3 and A5.

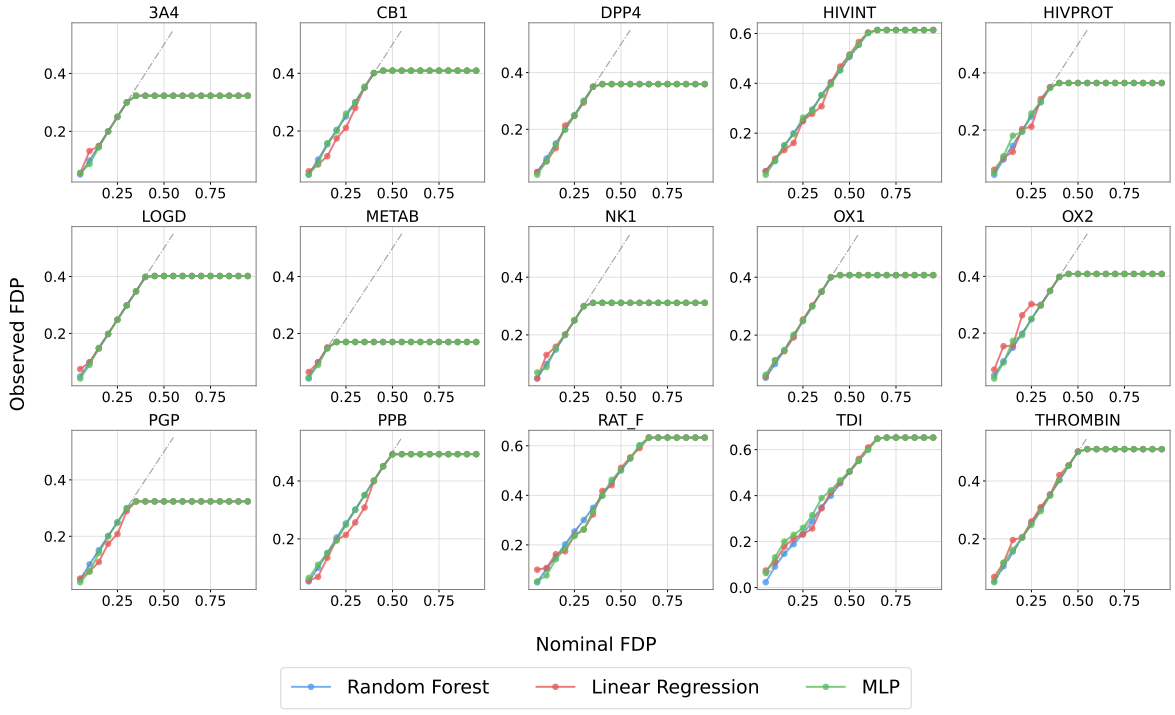
B.2 Comparing Different Nonconformity Scores

To assess the impact of score construction on empirical performance, we conducted a comparative study using several alternative score choices. In addition to the clipped score used in the main text, we also included residual-based heuristic baselines for comparison.

- Signed error score: for the calibration data, $V(X, Y) = Y - \hat{\mu}(X)$,

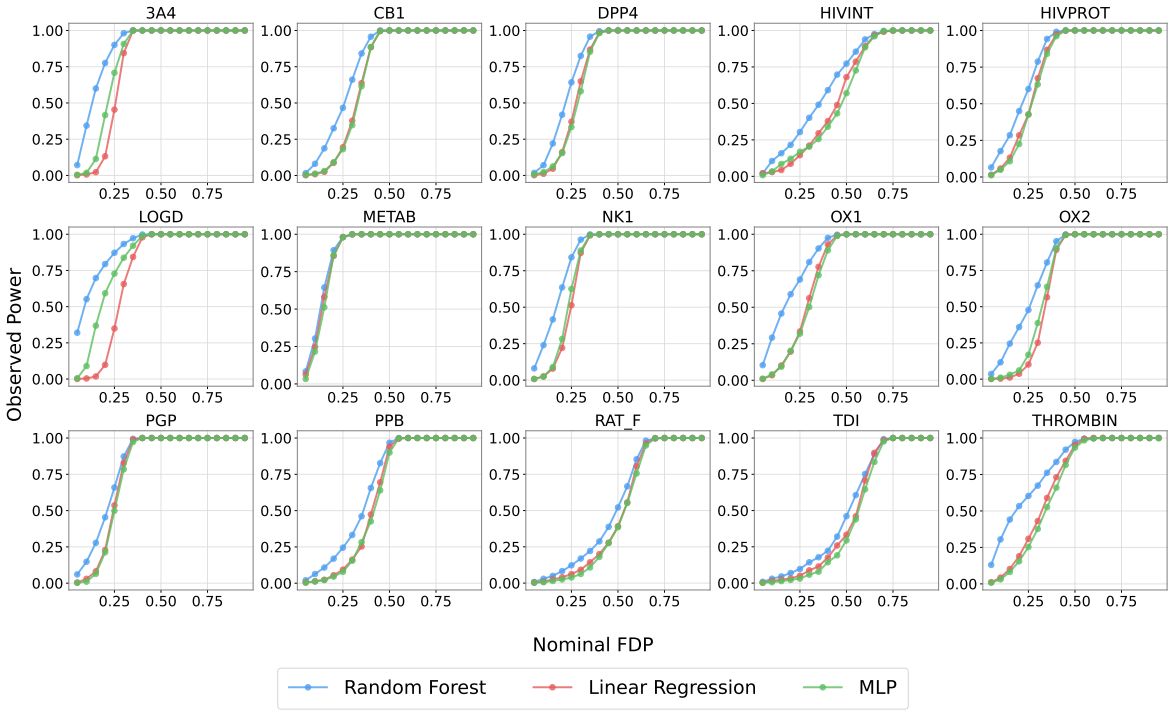


(a) FDP control on 10% subsets of the 15 Kaggle datasets.

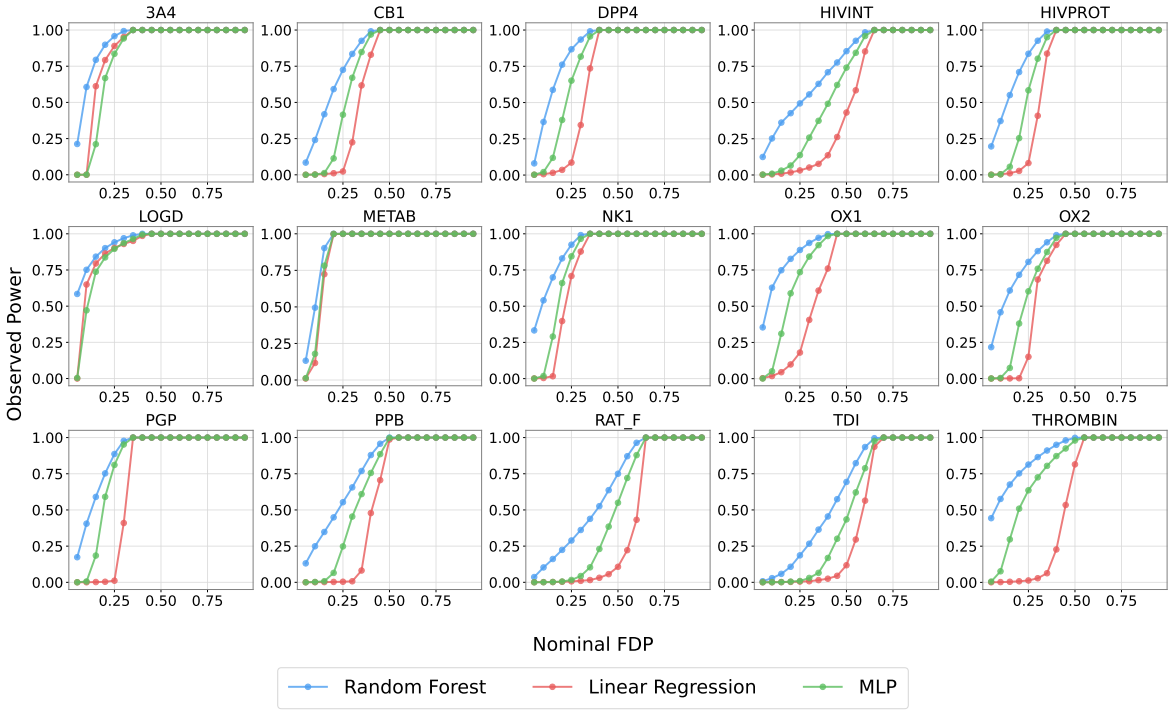


(b) FDP control on the entirety of the 15 Kaggle datasets.

Figure A4: FDP control for RSI-CS with Random Forest (blue), Linear Regression (red), and MLP (green) on (a) 10% subsets and (b) the full 15 datasets.



(a) Power on 10% subsets of the 15 Kaggle datasets.



(b) Power on the entirety of the 15 Kaggle datasets.

Figure A5: Power of RSI-CS with Random Forest (blue), Linear Regression (red), and MLP (green) on (a) 10% subsets and (b) the full 15 datasets.

for the test data $V(X, c) = \max\{c_1 - \hat{\mu}(X),\ \hat{\mu}(X) - c_2\}$.

- Uncertainty Signed error score: for the calibration data, $V(X, Y) = \frac{Y - \hat{\mu}(X)}{\hat{\sigma}_\mu(X)}$, for the test data $V(X, Y) = \frac{\max\{c_1 - \hat{\mu}(X),\, \hat{\mu}(X) - c_2\}}{\hat{\sigma}_\mu(X)}$.
- Clipped Score: $V(X, Y) = M \cdot \mathbf{1}\{Y \in \mathcal{R}_1 \setminus \partial\mathcal{R}_1\} - \hat{\pi}(X)$
- Uncertainty Clipped Score: $V(X, Y) = \frac{M \cdot \mathbf{1}\{Y \in \mathcal{R}_1 \setminus \partial\mathcal{R}_1\} - \hat{\pi}(X)}{\hat{\sigma}_\pi(X)}$

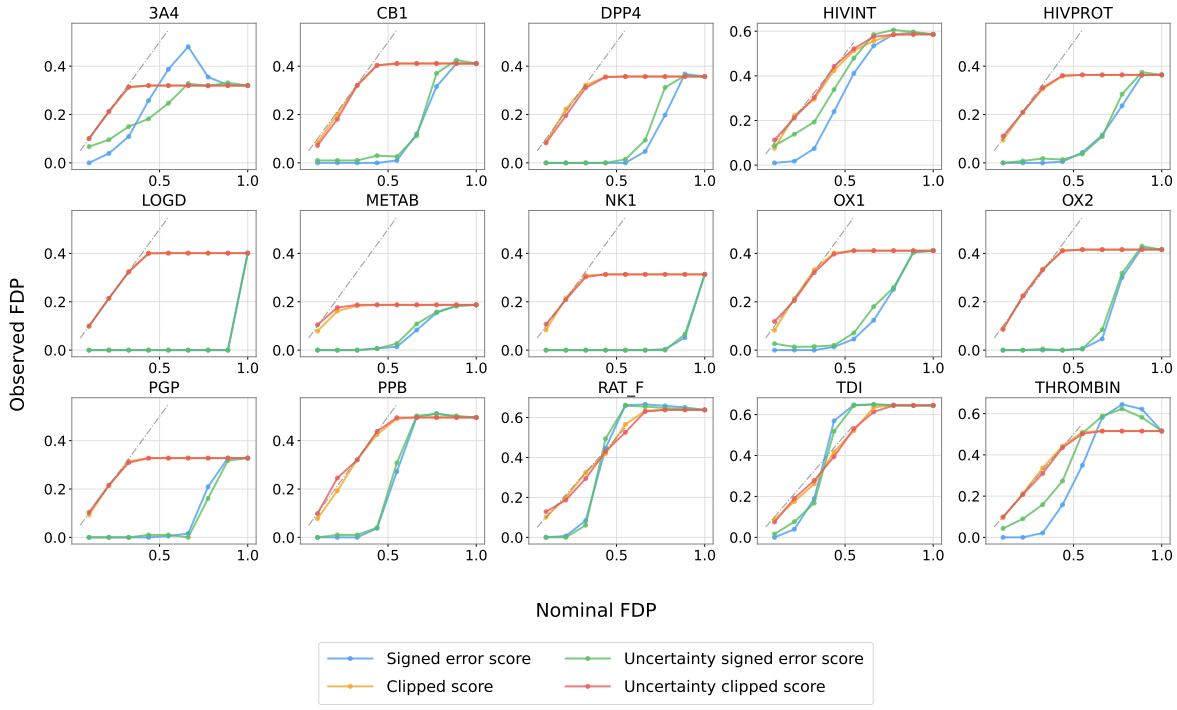
where $\hat{\mu}(X)$ is the prediction model for Y, $\hat{\sigma}_\mu(X)$ is its associated prediction uncertainty and $\mathcal{R}_1 = (-\infty, c_1] \cup [c_2, \infty)$. For the clipped scores, $\hat{\pi}(X)$ is a model estimating the probability $\mathbb{P}(Y \in \mathcal{R}_1 \mid X)$, and $\hat{\sigma}_\pi(X)$ is its corresponding uncertainty.

In Figure A6, all methods demonstrate valid control. Among them, the clipped score function can be considered the most efficient, as it tracks the nominal FDP boundary more closely without violating it, indicating a less conservative control strategy. In contrast, the signed score function performs poorly and is highly conservative, often yielding an observed FDP far below the nominal level. Furthermore, the inclusion of uncertainty does not affect the validity of FDP control. In Figure A7, the clipped score function is the top performer across almost all datasets, with the notable exception of RAT_F. For the score functions, adding uncertainty counterintuitively reduces their power.

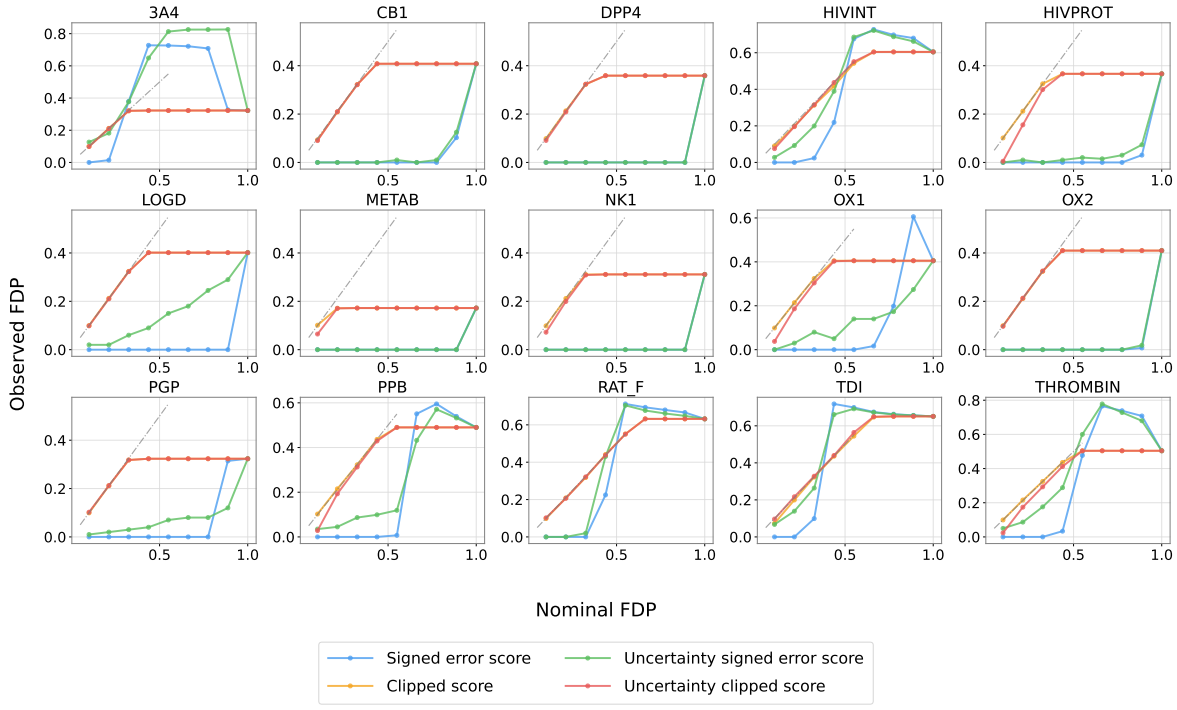
B.3 Conformal Selection with Different Data Transformations

We study how different choices of prediction target affect the performance of conformal selection. Specifically, we compare three modeling pipelines that differ in whether the model is trained to predict the continuous activity Y itself or the event $\{Y \in \mathcal{R}_1\}$, where $\mathcal{R}_1 = (-\infty, c_1] \cup [c_2, \infty)$. In each case, the fitted model output is used to rank compounds for conformal selection.

We consider the following three pipelines:

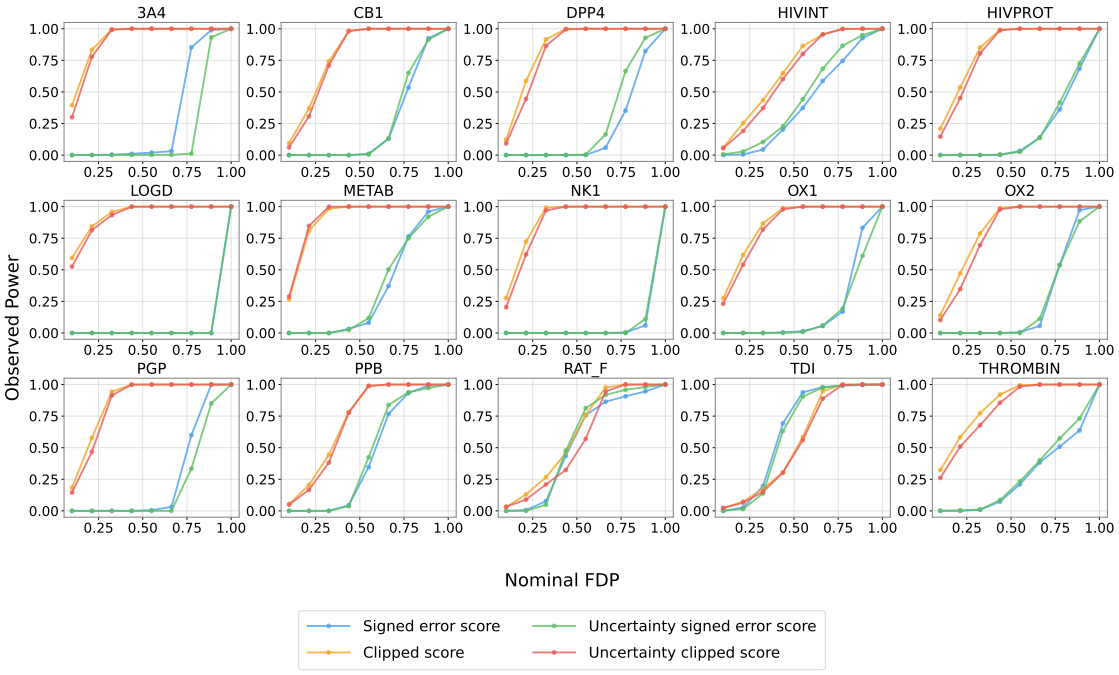


(a) FDP control on 10% subsets of the 15 Kaggle datasets.

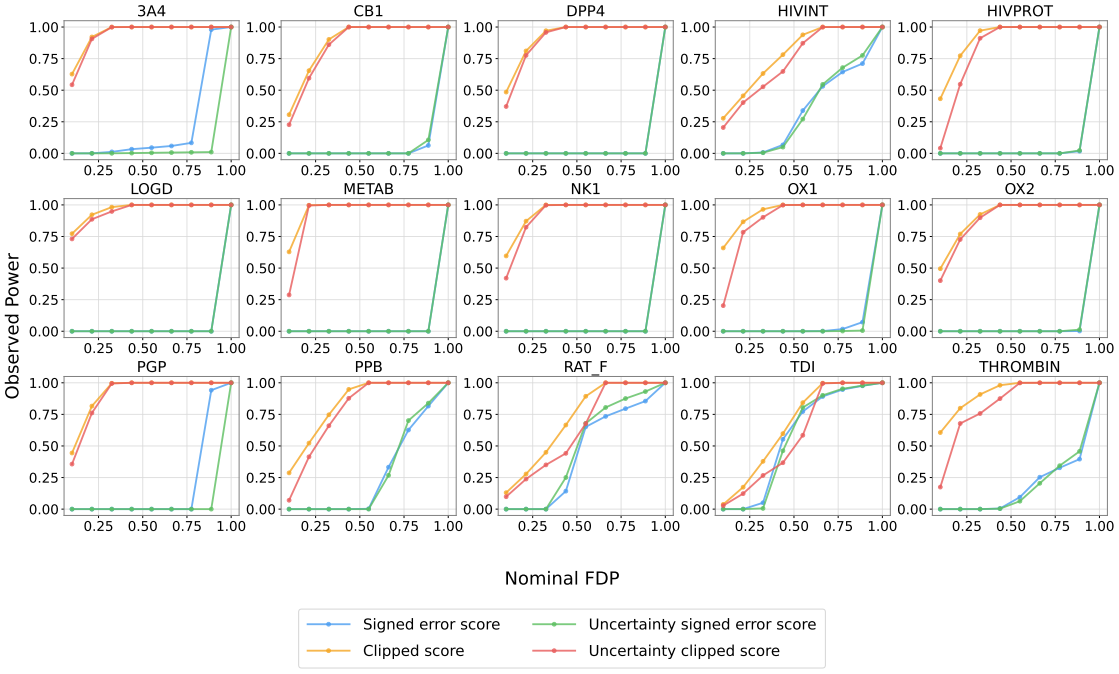


(b) FDP control on the entirety of the 15 Kaggle datasets.

Figure A6: FDP control of different scores under RSI-CS on (a) 10% subsets and (b) the full 15 Kaggle datasets, across nominal risk levels from 10% to 100%. Gray dashed lines indicate perfect risk control.



(a) Power on 10% subsets of the 15 Kaggle datasets.



(b) Power on the entirety of the 15 Kaggle datasets.

Figure A7: Power of different scores under RSI-CS. (a) 10% subsets of the 15 Kaggle datasets, and (b) the entirety of the datasets, with nominal risk levels varying from 10% to 100%.

- **Continuous regression.** We fit a regression model directly to the continuous response Y. The model is trained to estimate the conditional mean $\mu(X) = \mathbb{E}[Y \mid X]$, and the fitted value $\hat{\mu}(X)$ is used to rank compounds.

- **Binarization + classification.** We threshold the response according to the event $Y \in \mathcal{R}_1$ and train a probabilistic classifier to estimate $\pi(X) = \mathbb{P}(Y \in \mathcal{R}_1 \mid X)$. The estimated probability $\hat{\pi}_{\text{clf}}(X)$ is used to rank compounds.

- **Binarization + regression.** We threshold the activity according to the event $Y \in \mathcal{R}_1$ and fit a regression model to the resulting 0/1 outcome. The fitted value $\hat{\pi}_{\text{reg}}(X)$ is used to rank compounds.

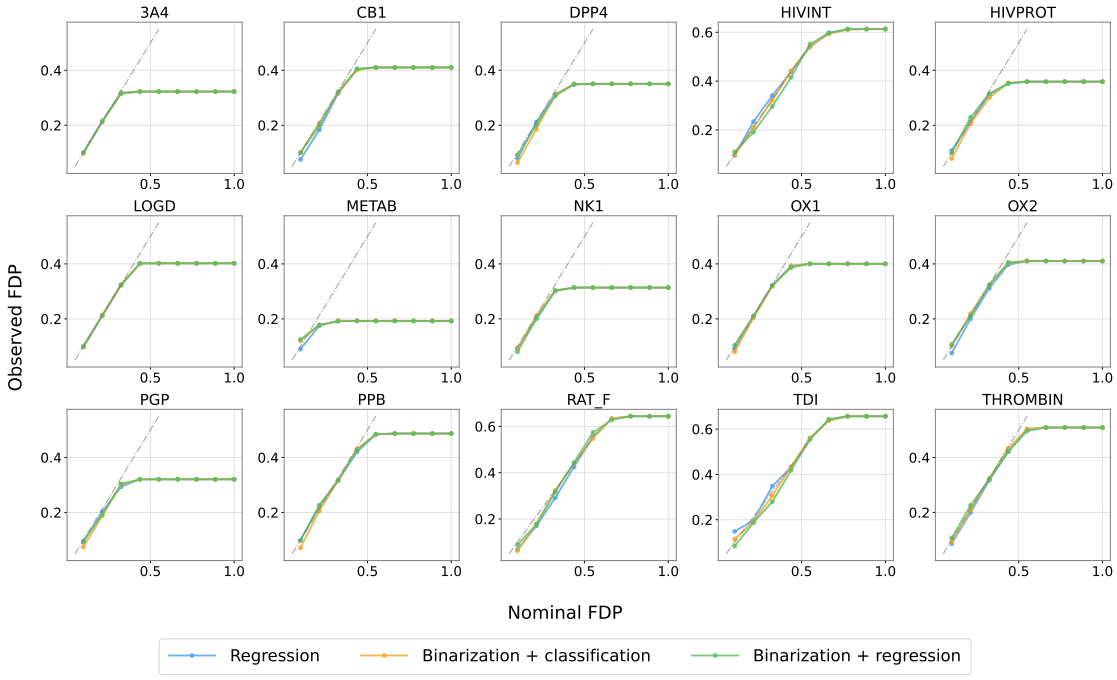
This comparison isolates the impact of (i) predicting the raw continuous activity versus the thresholded event of interest, and (ii) using a classification loss versus a regression loss after thresholding.

Figure A8 shows that, across all three pipelines, the observed FDP is at or below the nominal level over a wide range of target levels (from 10% to 100%), both on 10% subsets and on the full datasets. These results are empirically consistent with the intended FDP control of the conformal selection procedure under our experimental setup.

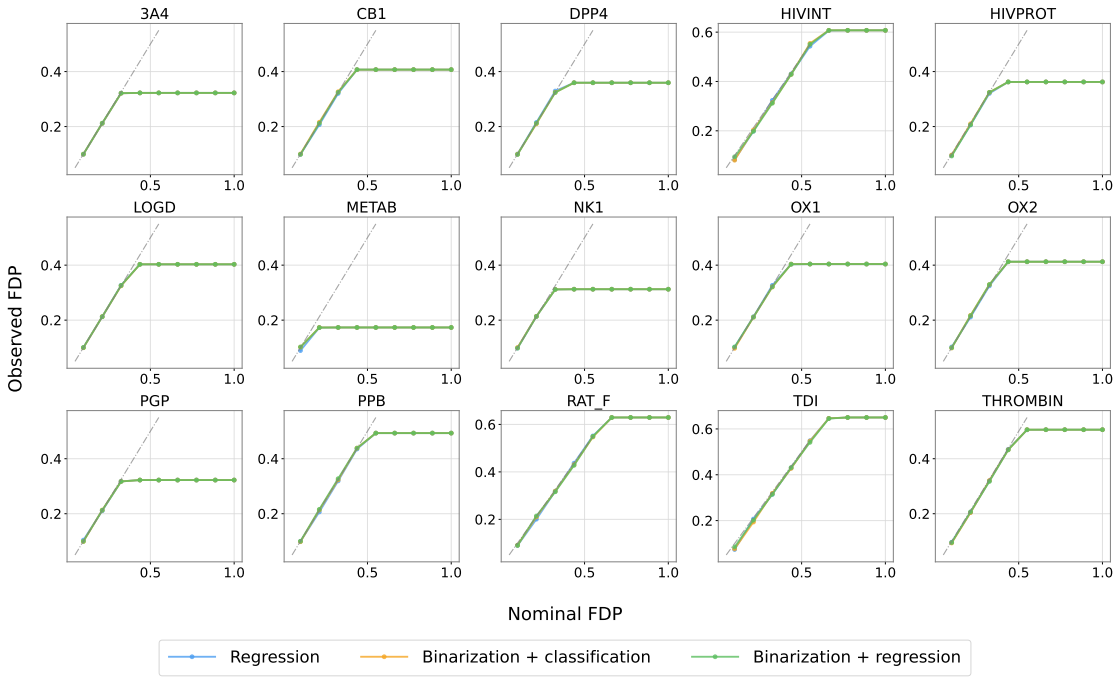
Figure A9 shows that the continuous-regression pipeline yields systematically lower power than the two binarization-based pipelines at comparable nominal FDP levels. A natural explanation is that the selection target is inherently threshold-based (i.e., detecting whether Y falls outside (c_1, c_2)), so modeling the binary event directly provides scores that are better aligned with the decision boundary, which improves ranking and leads to higher power.

Appendix C: Additional Numerical Results for Setting II

Under Setting II, Figure A10 displays observed FDP and power for RSI-CS, RSI-EC, and Baseline on 10% random subsets, while Figure A11 presents the corresponding results on the full datasets. In each subpanel, the x-axis shows the nominal target level and the y-axis

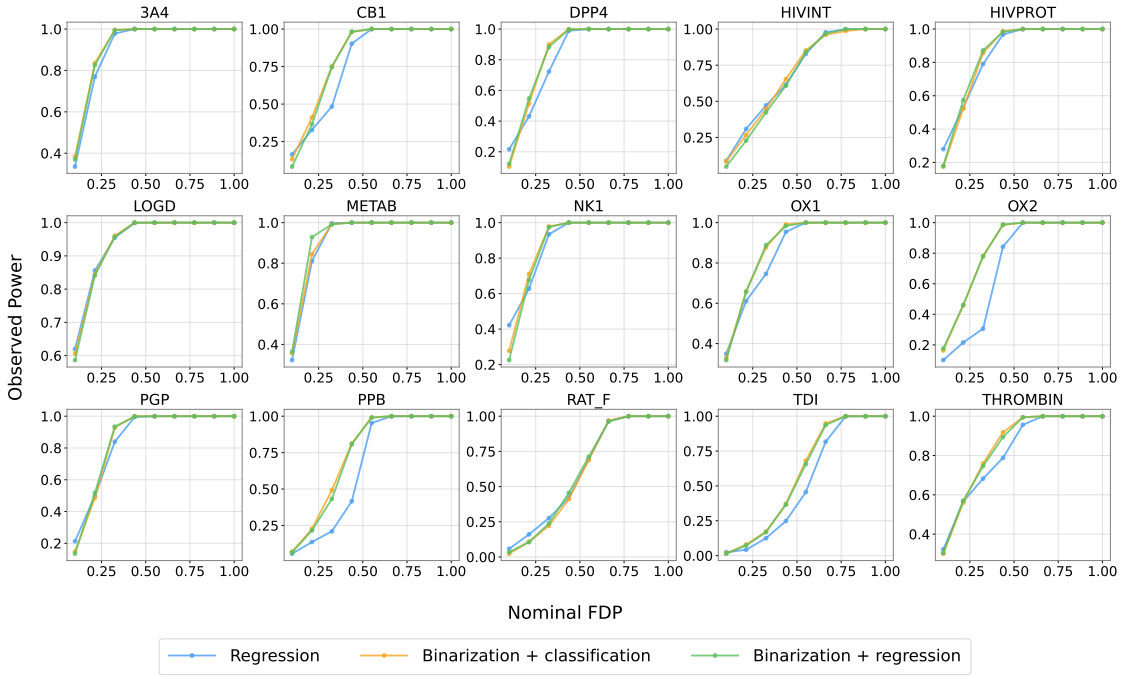


(a) FDP control on 10% subsets of the 15 Kaggle datasets.

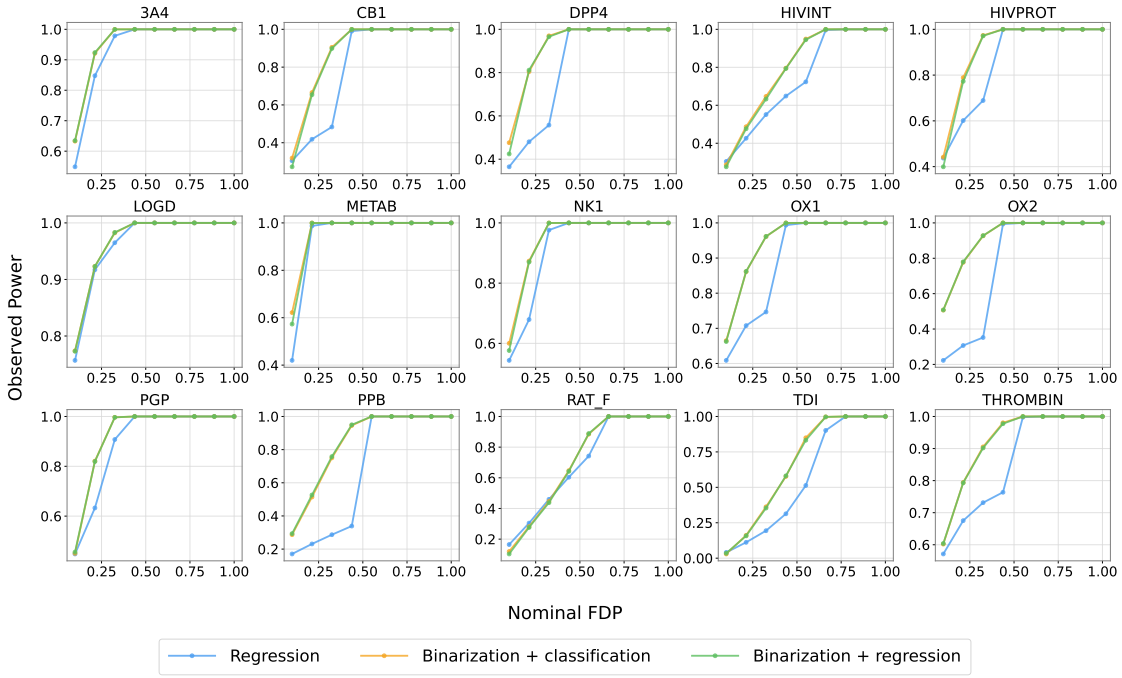


(b) FDP control on the entirety of the 15 Kaggle datasets.

Figure A8: FDP control of different data transformations under RSI-CS. (a) 10% subsets of the 15 Kaggle datasets, and (b) the entirety of the datasets, with nominal risk levels varying from 10% to 100%. The gray dashed lines represent perfect risk control, where the observed risk matches the specified risk level exactly.



(a) Power on 10% subsets of the 15 Kaggle datasets.



(b) Power on the entirety of the 15 Kaggle datasets.

Figure A9: Power of different data transformations under RSI-CS. (a) 10% subsets of the 15 Kaggle datasets, and (b) the entirety of the datasets, with nominal risk levels varying from 10% to 100%.

shows either observed FDP or power. The gray dashed line ($y = x$) indicates ideal FDP calibration.

Overall, both RSI-CS and RSI-EC continue to show favorable observed FDP behavior under Setting II while maintaining useful power across a broad range of nominal levels. The Baseline method is less reliable: its observed FDP is more erratic, and its power is generally lower. This is consistent with its construction, which combines two independently generated one-sided procedures rather than treating the clear-region assignment as a single coherent region-aware multiple-testing problem. As a result, it does not inherit the same structural justification as the proposed RSI procedures.

Appendix D: Complete Cost-Aware Delta Profiles

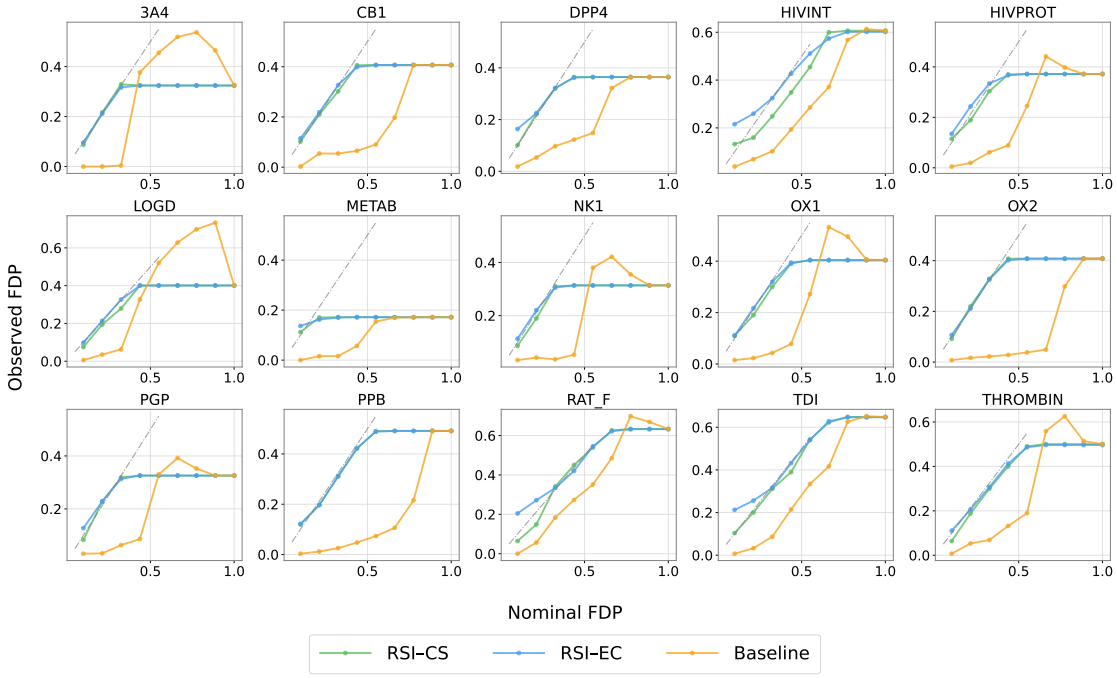
This appendix reports the complete cost-aware parameter-sweep results across all 15 datasets and all nominal FDR levels considered in the cost-aware experiment, $q \in \{0.1, 0.2, 0.3, 0.4, 0.5\}$. The main text focuses on a smaller set of representative datasets to make the selected-set mechanisms easier to interpret. Here, we provide the full dataset-level delta profiles for both RSI-EC and RSI-CS.

For each fixed dataset, method, and nominal level q, the horizontal axis is the cost-aware tuning parameter

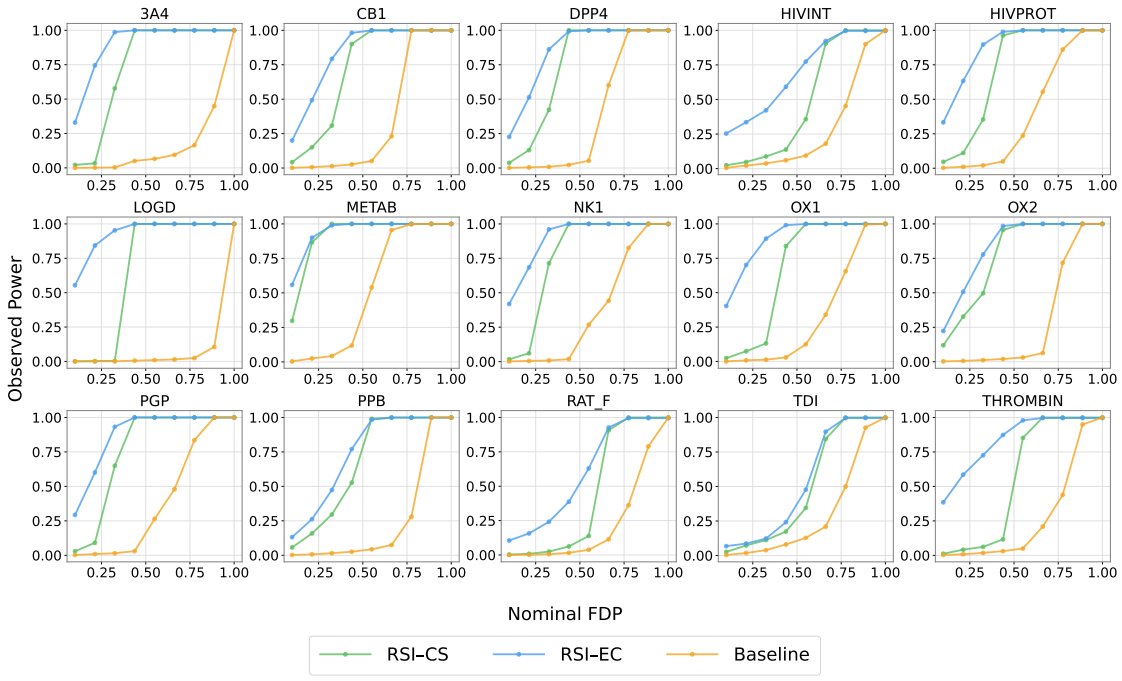
$$\eta \in \{0, 0.001, 0.002, 0.003, 0.005, 0.0075, 0.01, 0.015, 0.02, 0.03, 0.04, 0.05, 0.075, 0.1, 0.2\}$$

The baseline is the unadjusted score corresponding to $\eta = 0$. Let $A(\eta)$, $P(\eta)$, and $F(\eta)$ denote the average cost, empirical power, and observed FDP at tuning value η, respectively. The plotted delta metrics are

$$\Delta_{\text{cost}}(\eta) = 100\frac{A(0) - A(\eta)}{A(0)},$$

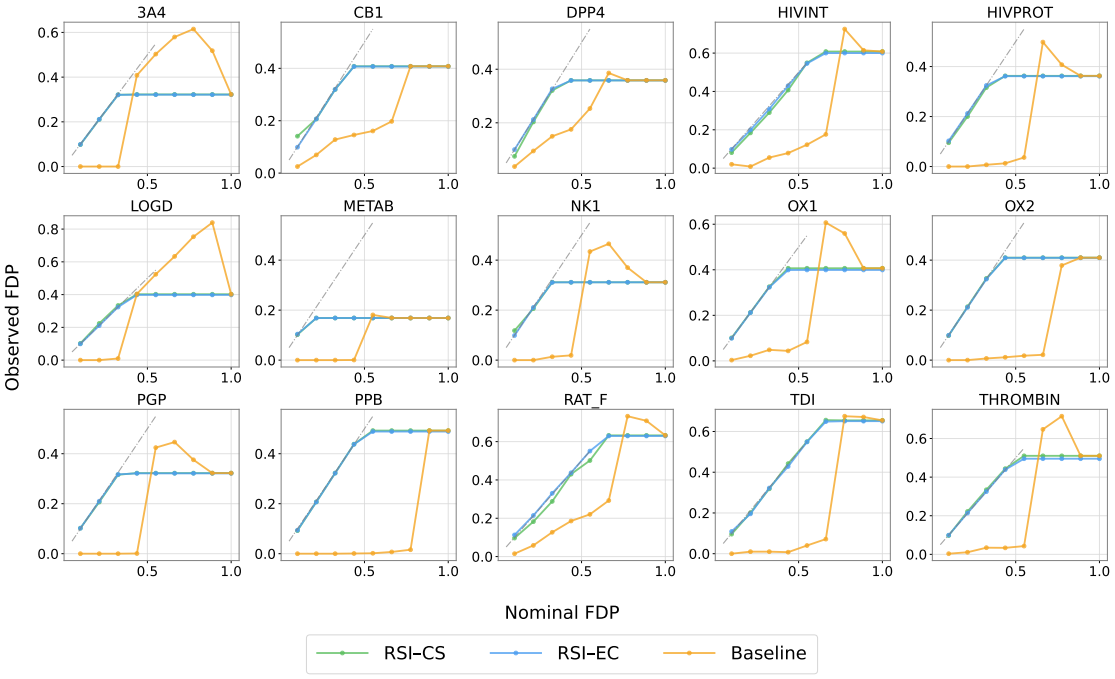


(a) FDP control on 10% subsets of the 15 Kaggle datasets.

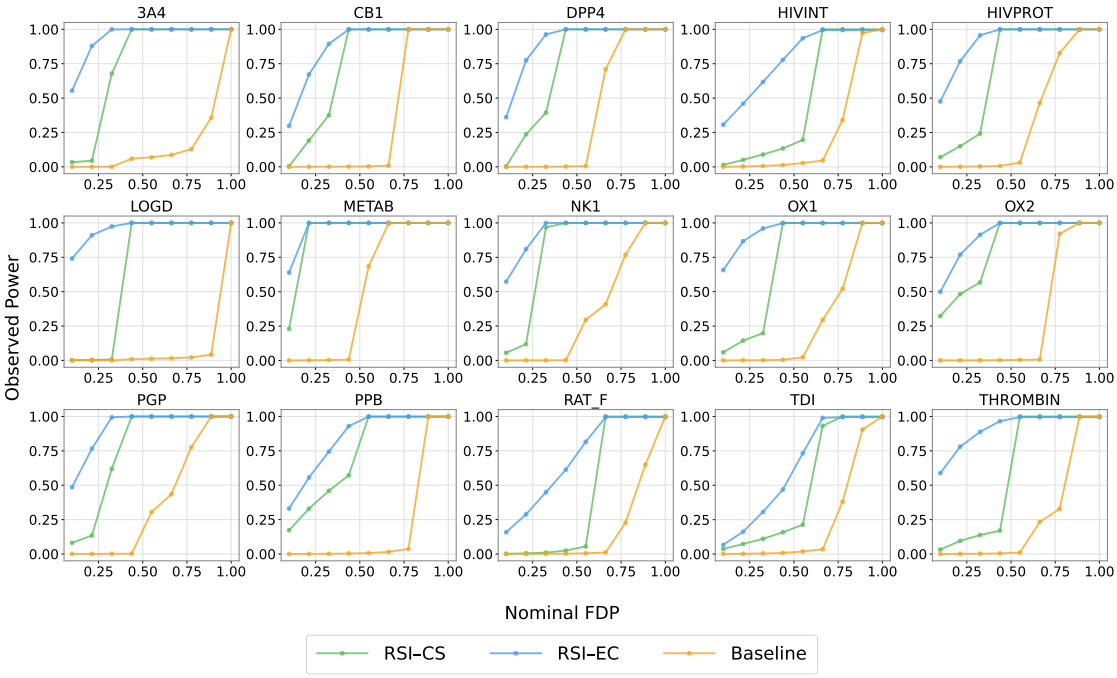


(b) Power on 10% subsets of the 15 Kaggle datasets.

Figure A10: Setting II results on 10% random subsets. The y-axis shows the observed FDP in panel (a) and power in panel (b), and the x-axis shows the nominal target level (10%–100%). The gray dashed diagonal ($y = x$) indicates perfect calibration, where the observed FDP equals the nominal level.



(a) FDP control on the entirety of the 15 Kaggle datasets.



(b) Power on the entirety of the 15 Kaggle datasets.

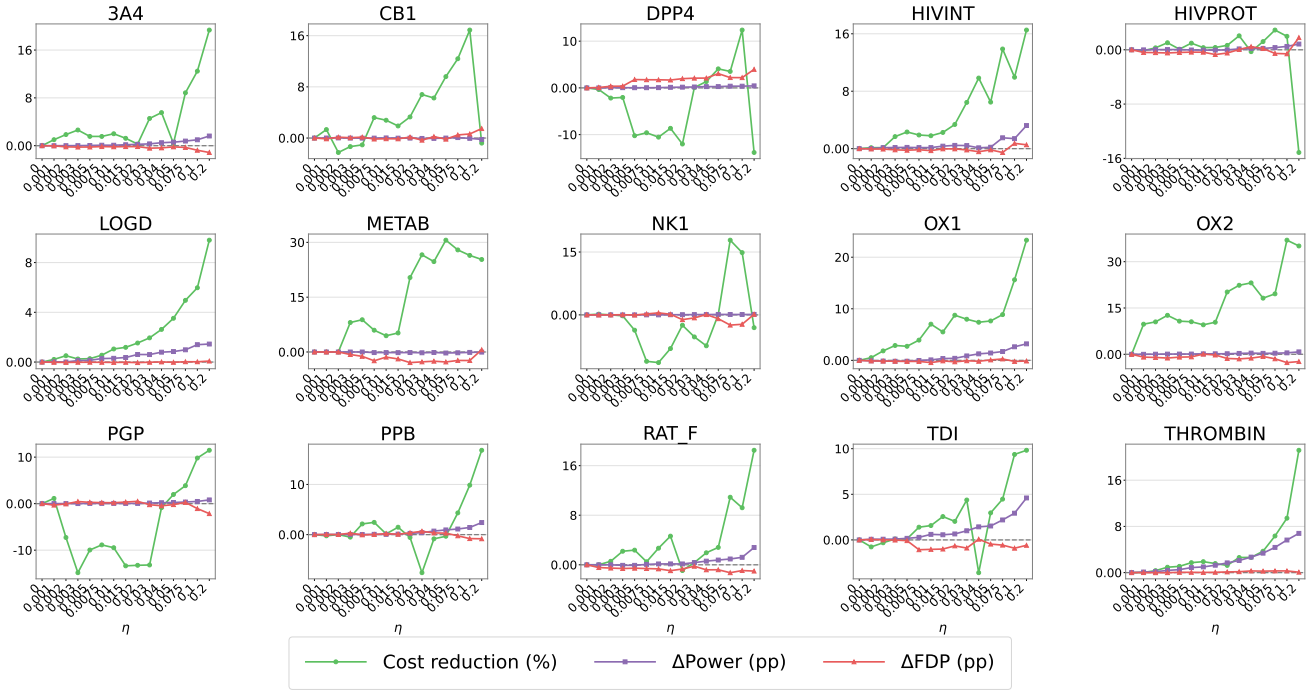
Figure A11: Setting II results on the full datasets. The y-axis shows the observed FDP in panel (a) and power in panel (b), and the x-axis shows the nominal target level (10%–100%). The gray dashed diagonal ($y = x$) indicates perfect calibration for FDP.

and

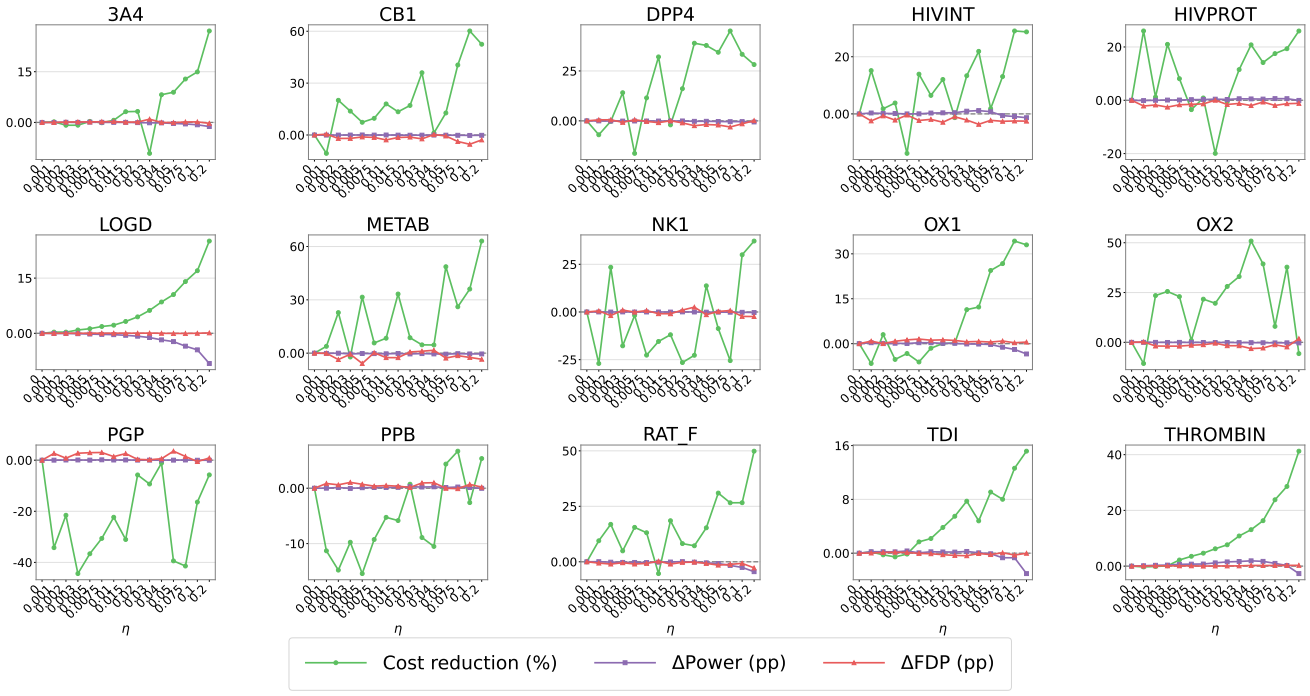
$$\Delta_{\text{power}}(\eta) = 100\{P(\eta) - P(0)\}, \qquad \Delta_{\text{FDP}}(\eta) = 100\{F(\eta) - F(0)\}.$$

Thus, cost is reported as a relative percentage reduction from the baseline, whereas power and FDP are reported as percentage-point changes. Positive values of Δ_{cost} indicate lower average cost than the baseline. Positive values of Δ_{power} indicate higher empirical power than the baseline, and positive values of Δ_{FDP} indicate a larger observed FDP than the baseline. The horizontal dashed line marks zero change.

The figures are organized by nominal level q. For each q, we place the RSI-EC and RSI-CS panels together to make the method comparison direct. Across many datasets, increasing η produces positive cost reduction while the observed FDP change remains close to zero. The power response is more dataset-dependent: in some datasets, power remains stable or increases mildly, whereas in others stronger cost adjustment is accompanied by power loss. Overall, these results support the main-text conclusion that cost-aware scoring can reduce average cost without visibly relaxing the observed FDP, but the magnitude of the cost reduction and the associated change in power depend on the dataset, method, and nominal FDR level.

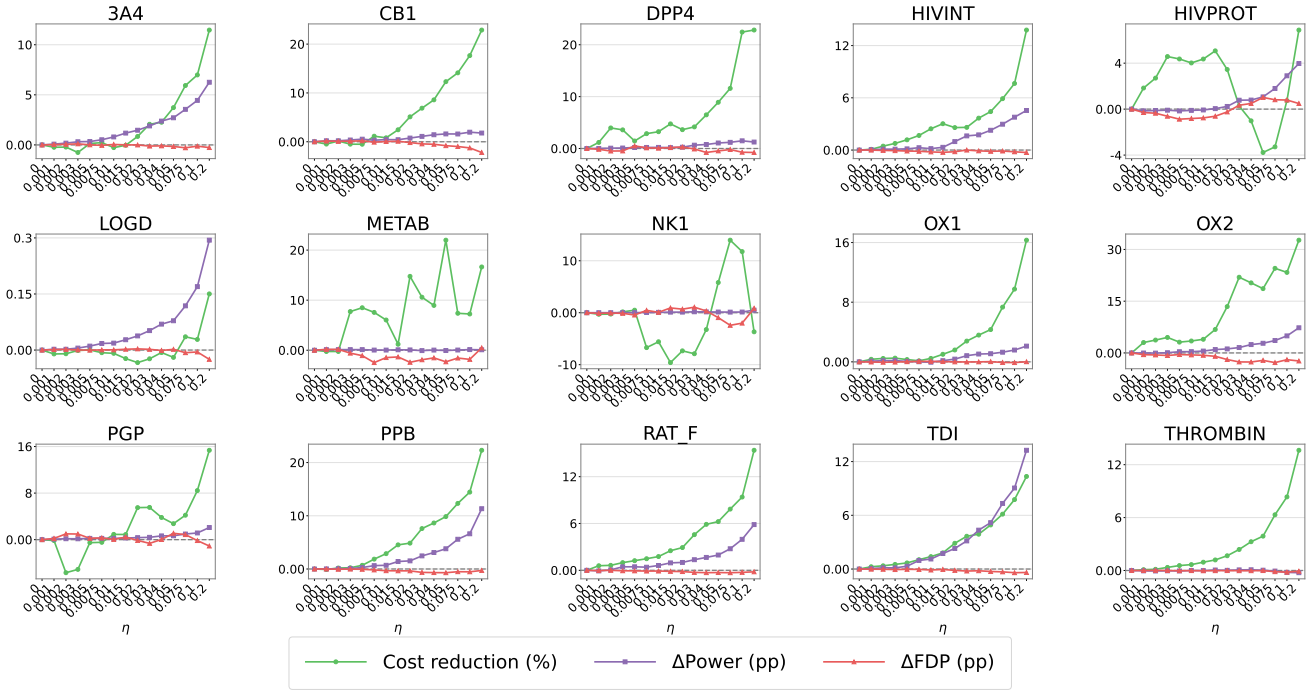


(a) RSI-EC.

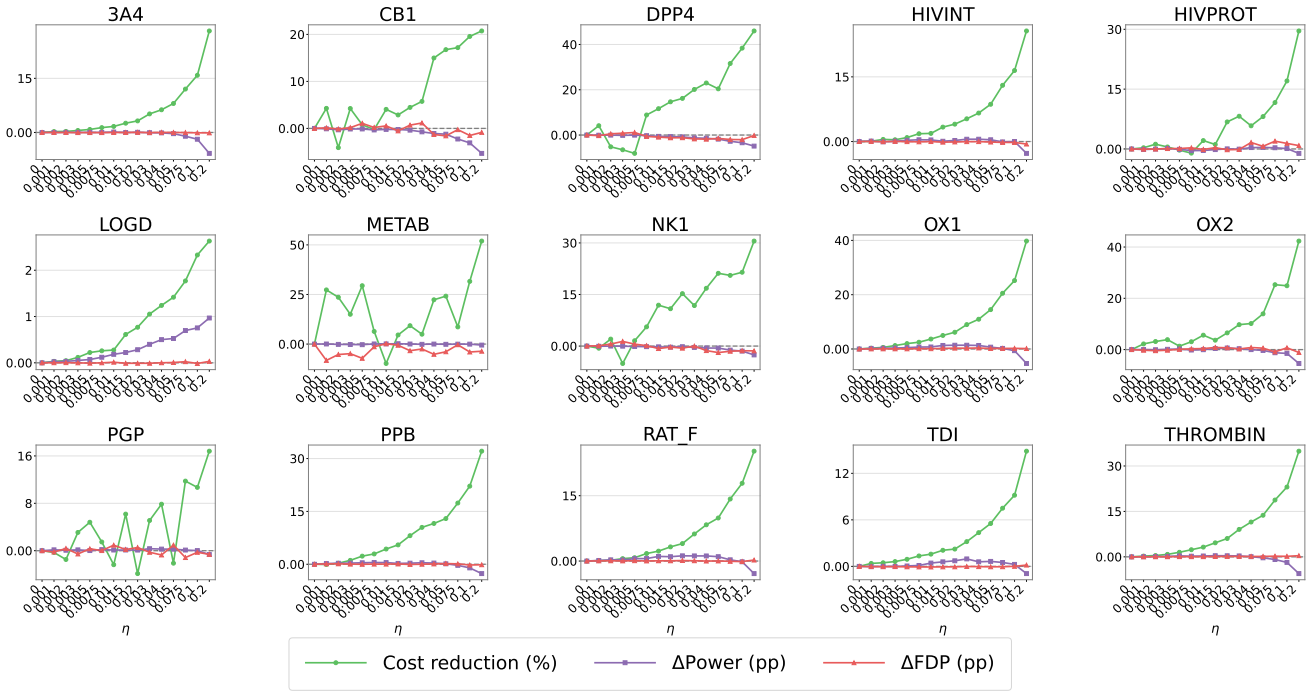


(b) RSI-CS.

Figure A12: Complete cost-aware delta profiles across all 15 datasets at the target nominal FDR level $q = 0.1$. The RSI-EC and RSI-CS results are shown together for direct comparison. Green curves show the relative percentage reduction in average cost, purple curves show the percentage-point change in empirical power, and red curves show the percentage-point change in observed FDP, all relative to $\eta = 0$.

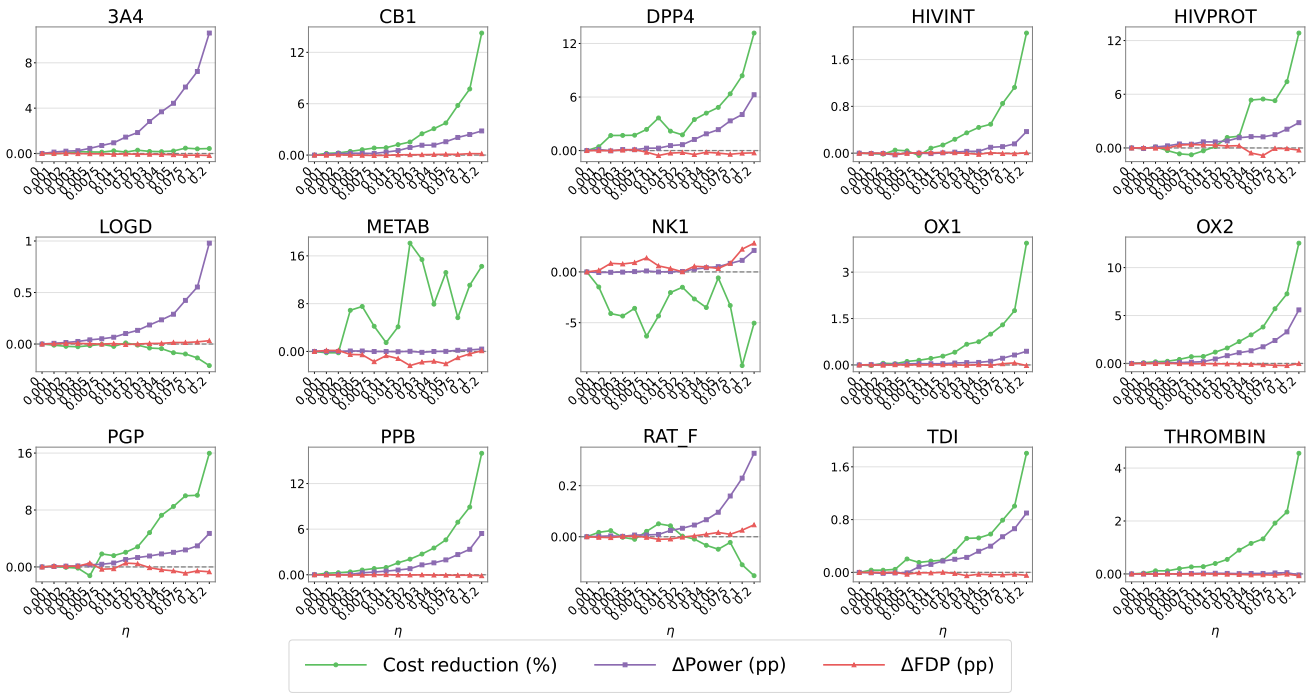


(a) RSI-EC.

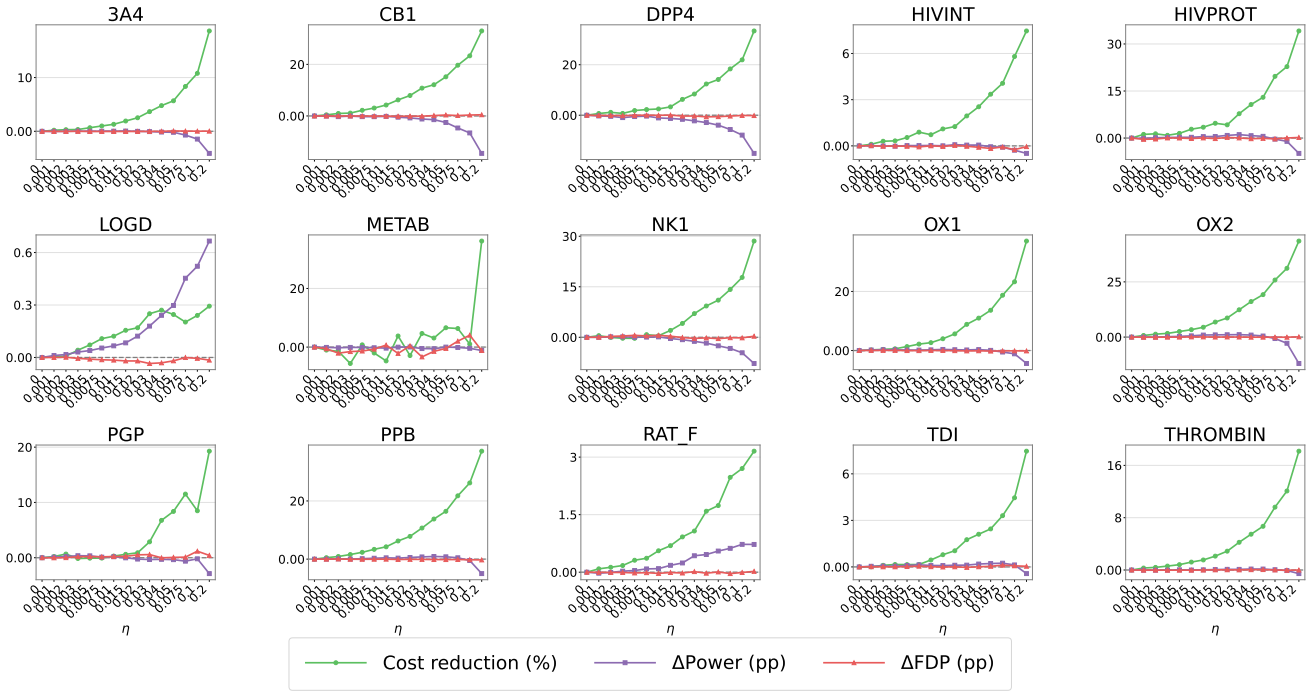


(b) RSI-CS.

Figure A13: Complete cost-aware delta profiles across all 15 datasets at the target nominal FDR level $q = 0.2$. The RSI-EC and RSI-CS results are shown together for direct comparison. Green curves show the relative percentage reduction in average cost, purple curves show the percentage-point change in empirical power, and red curves show the percentage-point change in observed FDP, all relative to $\eta = 0$.

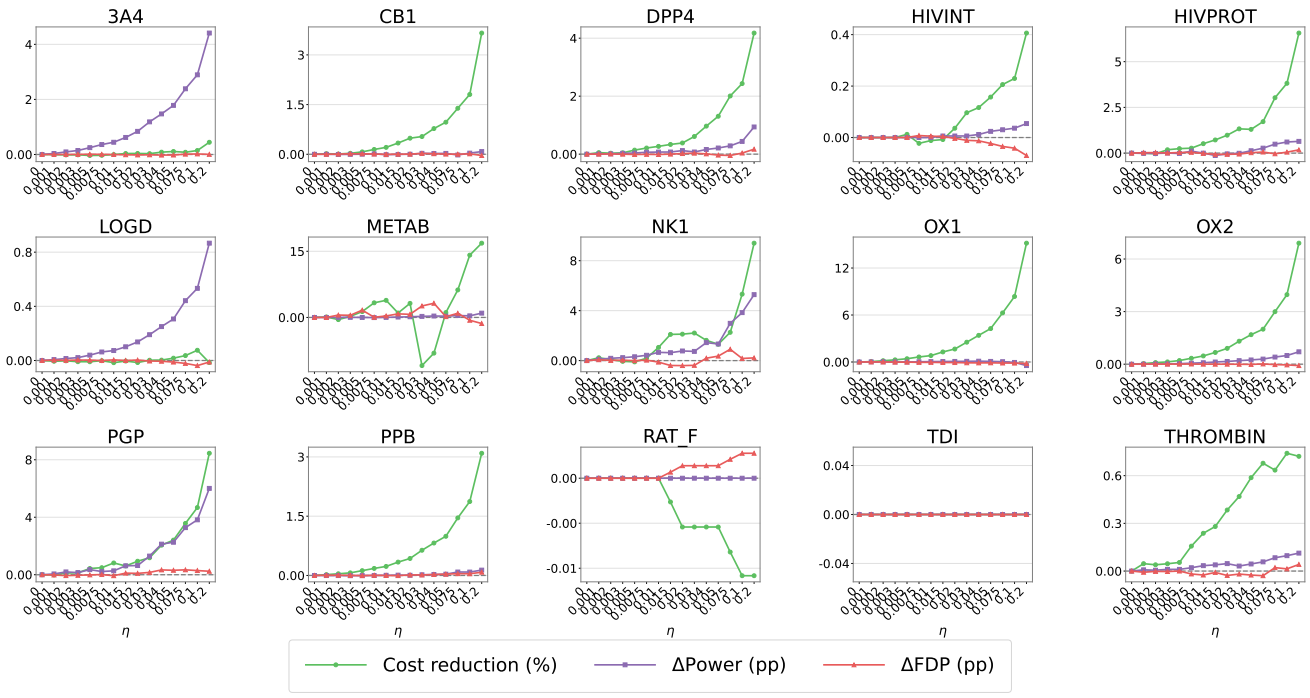


(a) RSI-EC.

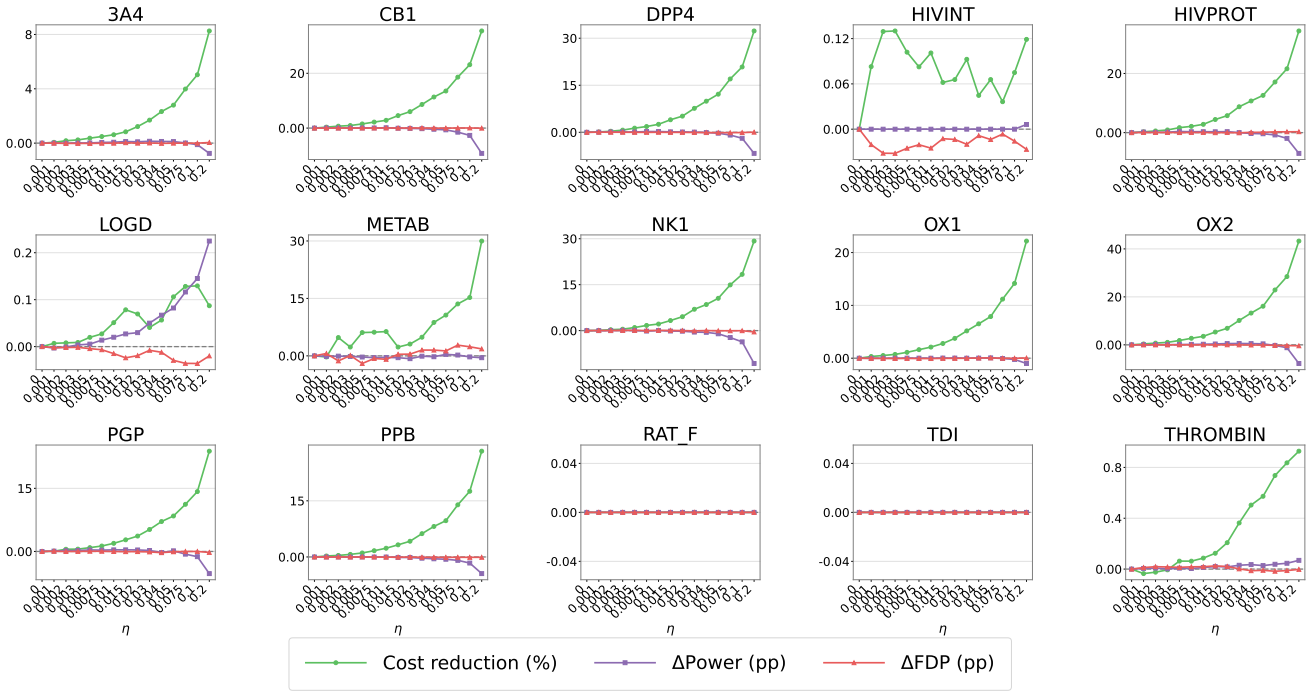


(b) RSI-CS.

Figure A14: Complete cost-aware delta profiles across all 15 datasets at the target nominal FDR level $q = 0.3$. The RSI-EC and RSI-CS results are shown together for direct comparison. Green curves show the relative percentage reduction in average cost, purple curves show the percentage-point change in empirical power, and red curves show the percentage-point change in observed FDP, all relative to $\eta = 0$.

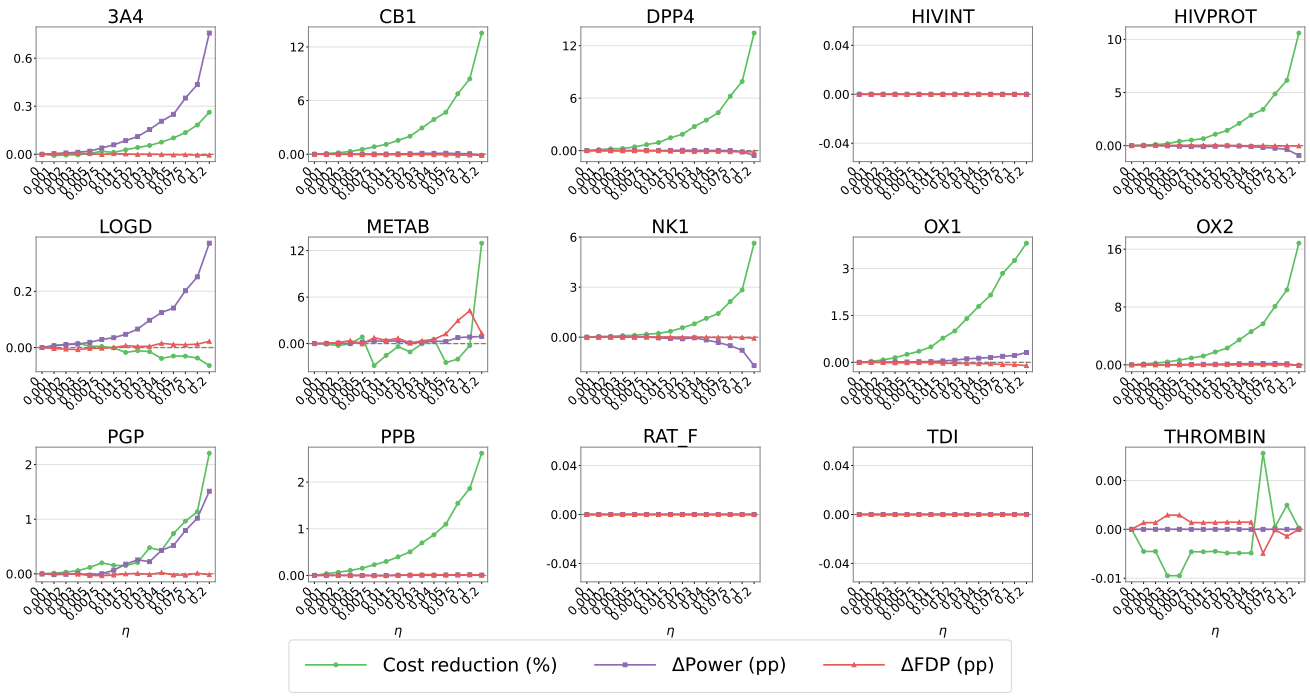


(a) RSI-EC.

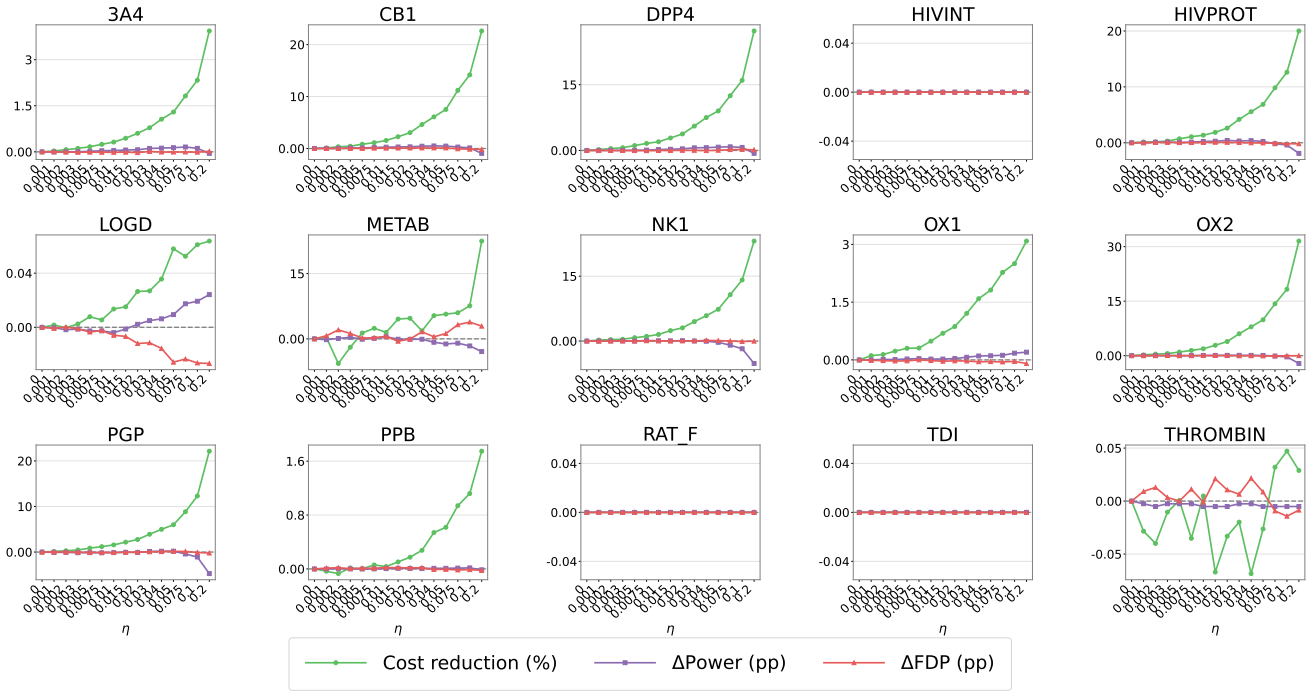


(b) RSI-CS.

Figure A15: Complete cost-aware delta profiles across all 15 datasets at the target nominal FDR level $q = 0.4$. The RSI-EC and RSI-CS results are shown together for direct comparison. Green curves show the relative percentage reduction in average cost, purple curves show the percentage-point change in empirical power, and red curves show the percentage-point change in observed FDP, all relative to $\eta = 0$.



(a) RSI-EC.



(b) RSI-CS.

Figure A16: Complete cost-aware delta profiles across all 15 datasets at the target nominal FDR level $q = 0.5$. The RSI-EC and RSI-CS results are shown together for direct comparison. Green curves show the relative percentage reduction in average cost, purple curves show the percentage-point change in empirical power, and red curves show the percentage-point change in observed FDP, all relative to $\eta = 0$.