Multivariate Conformal Selection

Tian Bai [1] **Yue Zhao** [2] **Xiang Yu** [3] **Archer Y. Yang** [1 4]

Abstract

Selecting high-quality candidates from large datasets is critical in applications such as drug discovery, precision medicine, and alignment of large language models (LLMs). While Conformal Selection (CS) provides rigorous uncertainty quantification, it is limited to univariate responses and scalar criteria. To address this issue, we propose Multivariate Conformal Selection (mCS), a generalization of CS designed for multivariate response settings. Our method introduces regional monotonicity and employs multivariate nonconformity scores to construct conformal p-values, enabling finite-sample False Discovery Rate (FDR) control. We present two variants: `mCS-dist`, using distance-based scores, and `mCS-learn`, which learns optimal scores via differentiable optimization. Experiments on simulated and real-world datasets demonstrate that mCS significantly improves selection power while maintaining FDR control, establishing it as a robust framework for multivariate selection tasks.

1. Introduction

Selecting a subset of promising candidates from a large pool is crucial across various scientific and real-world applications. In drug discovery, researchers search vast chemical spaces to identify compounds with strong effects, such as high binding affinity to a specific target (Szymański et al., 2011; Scannell et al., 2022; Sheridan et al., 2015; Zhang et al., 2025). Similarly, precision medicine aims to identify positive individual treatment effects (Lei & Candès, 2021), and post-hoc certification of large language model (LLM) outputs seeks to retain only trustworthy generations that meet user-defined criteria (Gui et al., 2024). In these settings, true test responses (e.g., binding affinity or alignment score) are often unavailable, requiring selection to rely on machine learning model predictions. Since selected targets drive downstream decisions, quantifying selection uncertainty is essential for maintaining efficiency. Conformal selection (Jin & Candès, 2023) provides a model-agnostic framework for selection with uncertainty quantification by extending conformal prediction (Vovk et al., 2005) to multiple hypothesis testing, using conformal p-values (Bates et al., 2023) and multiple testing corrections. It has shown promise in real-world drug discovery (Bai et al., 2024) and LLM alignment (Gui et al., 2024).

However, existing conformal selection methods are limited to univariate responses and selection targets of the form $y > c$, where c is a user-defined threshold. Many real-world applications require selection based on multiple interdependent criteria. For instance, LLM outputs must simultaneously satisfy alignment requirements such as fairness, safety, and correctness (Bai et al., 2022), which are better represented as multivariate alignment scores. This highlights the need for a principled selection method with uncertainty quantification in multivariate settings.

In this paper, we extend the conformal selection framework to multivariate response settings. To ensure finite-sample *false discovery rate* (FDR) control in our procedure, we generalize the concept of monotonicity (Jin & Candès, 2023) to *regional monotonicity* for the nonconformity function. We propose two types of nonconformity scores that satisfy this property: (i) distance-based nonconformity scores for regular-shaped and convex target regions, and (ii) a learning-based method for optimizing nonconformity scores. The latter approach leverages a loss function that penalizes either the smooth selection size or the conformal p-value to learn an optimal score. This method is particularly effective when the dimension of responses is large or when the target region is irregular or nonconvex. Through experiments on simulated and real-world datasets, both variants of mCS demonstrates enhanced selection power over baseline methods while ensuring finite-sample FDR control.

2. Background and Related Work

Problem Setup We let $\boldsymbol{x} \in \mathbb{R}^p$ represent the p-dimensional features, and let $\boldsymbol{y} \in \mathbb{R}^d$ denote the d-

[1] Department of Mathematics and Statistics, McGill University, Montreal, Canada [2] Department of Mathematics, University of York, York, UK [3] MRL, Merck & Co., Inc., Rahway, NJ, USA [4] Mila - Quebec AI Institute, Montreal, Quebec, Canada. Correspondence to: Archer Y. Yang <archer.yang@mcgill.ca>.

Proceedings of the 42^{nd} International Conference on Machine Learning, Vancouver, Canada. PMLR 267, 2025. Copyright 2025 by the author(s).

dimensional multivariate response variables. We consider a training dataset $\mathcal{D}_{\text{train}} = \{\boldsymbol{x}_i, \boldsymbol{y}_i\}_{i=1}^n$ and a test dataset $\mathcal{D}_{\text{test}} = \{\boldsymbol{x}_{n+j}\}_{j=1}^m$, where the corresponding test responses $\{\boldsymbol{y}_{n+j}\}_{j=1}^m$ are unobserved. We further assume that the combined set of training and test samples $\{\boldsymbol{x}_i, \boldsymbol{y}_i\}_{i=1}^{n+m}$ are drawn i.i.d.[1] from an unknown, arbitrary distribution $\mathcal{D}_{X\times Y}$.

We formulate the selection problem as follows: Given a predefined d-dimensional closed region $R \subseteq \mathbb{R}^d$, our goal is to identify a subset of indices $\mathcal{S} \subseteq \{1, \dots, m\}$ from $\mathcal{D}_{\text{test}}$ such that as many test observations $j \in \mathcal{S}$ satisfy $\boldsymbol{y}_{n+j} \in R$ as possible, while controlling the FDR (Benjamini & Hochberg, 1995) below a user-specified level q. The FDR is defined as the expected proportion of false discoveries ($j \in \mathcal{S}$ but $\boldsymbol{y}_{n+j} \notin R$) among all selected observations

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

where $\mathcal{H}_0 = \{j : \boldsymbol{y}_{n+j} \notin R\}$, with the convention that $0/0 = 0$ in the fraction above. This criterion measures the overall Type-I error rate of the selection procedure.

The overall Type-II error of selection can be quantified by the *power*, defined as the expected proportion of desirable observations ($\boldsymbol{y}_{n+j} \in R$) that are correctly selected,

$$\text{Power} = \mathbb{E}\left[\frac{|\mathcal{S} \cap \mathcal{H}_1|}{|\mathcal{H}_1|}\right] \tag{2}$$

where $\mathcal{H}_1 = \{j : \boldsymbol{y}_{n+j} \in R\}$. The Type-II error of selection is therefore $(1 - \text{Power})$. An ideal selection procedure should aim to maximize the power while keeping the FDR below the specified nominal level.

Conformal Prediction Conformal prediction (CP) (Vovk et al., 2005) is a popular framework for uncertainty quantification that constructs prediction intervals on a per-sample basis. Assuming exchangeable calibration and test data, CP provides prediction sets $\widehat{C}_{1-\alpha}(\boldsymbol{x})$ with finite-sample coverage guarantees for $\alpha \in (0, 1)$:

$$\mathbb{P}(y \in \widehat{C}_{1-\alpha}(\boldsymbol{x})) \geq 1 - \alpha.$$

Although CP was originally designed for univariate responses, numerous studies have proposed multivariate generalizations (Kuleshov et al., 2018; Bates et al., 2021; Messoudi et al., 2022; Johnstone & Cox, 2021; Feldman et al., 2023; Park et al., 2024; Klein et al., 2025). However, these multivariate CP methods are not directly applicable to our selection problem. The primary objective of CP – constructing confidence sets for predictions – does not naturally align with selection tasks. Specifically, the potentially complex shapes of the multivariate CP sets $\widehat{C}_{1-\alpha}$ may be incompatible with the pre-defined target region R.

Even in the simpler cases, such as when the response is binary, using CP for selection introduces a multiplicity issue (Jin & Candès, 2023). In this context, CP only controls the per-comparison error rate (PCER), which differs from the FDR. PCER also measures the Type-I error, and is defined as (1) with the denominator replaced by m, the size of the test dataset. By definition, the PCER is always smaller than the FDR, making it a less stringent error control criterion. As a result, procedures controlling PCER may fail to meet the stricter requirements of FDR control.

Conformal Selection Conformal selection (CS) (Jin & Candès, 2023) is a model-agnostic selection framework that guarantees finite-sample FDR control. However, CS only considers the univariate response case ($d = 1$) and assumes that the selection region takes the form $[c, +\infty)$, where c is a predefined threshold. In this setting, CS formulates one hypothesis test per candidate:

$$H_{0j} : y_{n+j} \leq c \quad \text{vs.} \quad H_{1j} : y_{n+j} > c.$$

Rejecting the null hypothesis H_{0j} indicates that the j-th test sample is selected, as its response is deemed to exceed the threshold c.

CS uses nonconformity scores to guide its selection process. A nonconformity measure quantifies how atypical (or nonconforming) an observation is, based on the relationship between inputs and responses. For calibration samples, the nonconformity scores are $V_i = V(\boldsymbol{x}_i, y_i)$, and for test samples, they are $\widehat{V}_{n+j} = V(\boldsymbol{x}_{n+j}, c)$, where c replaces the unobserved y_{n+j}. These scores are then used to compute conformal p-values through a rank-based comparison of $\widehat{V}_{n+j}$ against the calibration scores $V_1, \dots, V_n$. A lower rank of $\widehat{V}_{n+j}$ relative to the calibration scores provides stronger evidence for rejecting H_{0j}.

To determine the final selected subset $\mathcal{S}$, CS applies the Benjamini-Hochberg (BH) procedure (Benjamini & Hochberg, 1995; 1997), a widely used method for controlling FDR in multiple testing setting, to the set of conformal p-values. The use of the BH procedure ensures that the overall Type-I error rate is kept below the specified level.

3. Multivariate Conformal Selection

In this section, we introduce the key concepts, procedures, and theoretical foundations of multivariate Conformal Selection (mCS).

First, for each d-dimensional multivariate response $\boldsymbol{y}_{n+j}$,

[1] Later, we will relax the i.i.d. assumption to exchangeability conditions.

mCS performs the following hypothesis test:

$$H_{0j} : \boldsymbol{y}_{n+j} \in R^c \quad \text{vs.} \quad H_{1j} : \boldsymbol{y}_{n+j} \in R,$$

where R represents an arbitrary pre-defined closed target region in $\mathbb{R}^d$. Multivariate responses $\boldsymbol{y}_{n+j}$ can represent either regression or classification outcomes. In this paper, we focus on the more challenging regression setting. For a discussion of the classification case, please see Appendix B.1. The mCS process consists of three main steps:

1. **Training:** Construct a multivariate predictive model $\hat{\mu}$ for $\boldsymbol{y}$. This model can be obtained using any suitable machine learning algorithm.

2. **Calibration:** Build a regionally monotone multivariate nonconformity function based on $\hat{\mu}$, and evaluate this function on the calibration dataset and test dataset. Subsequently, we compute the conformal p-values for each test sample.

3. **Thresholding:** Apply the Benjamini-Hochberg (BH) procedure as in the original CS procedure to the set of conformal p-values, yielding the final selection set $\mathcal{S}$.

Both the training step and the calibration step rely on the labeled dataset $\mathcal{D}_{\text{train}}$. In the case where a model $\hat{\mu}$ is already available, mCS can be directly applied using all labeled data for calibration. Otherwise, the training data is divided into two subsets: one for model training (the proper training dataset) and the other for calibration. For simplicity, in the following discussion, we assume that $\hat{\mu}$ is already available and all training data $\{\boldsymbol{x}_i, \boldsymbol{y}_i\}_{i=1}^n$ are used for calibration so that $\mathcal{D}_{\text{cal}} = \mathcal{D}_{\text{train}}$.

The conformal p-values (Bates et al., 2023) are used to perform the hypothesis tests. If the true responses $\{\boldsymbol{y}_{n+j}\}_{j=1}^m$ were observed, the *oracle* conformal p-value would be defined as

$$p_j^* = \frac{\sum_{i=1}^n \mathbb{1}\{V_i < V_{n+j}\} + U_j(1 + \sum_{i=1}^n \mathbb{1}\{V_i = V_{n+j}\})}{n+1} \tag{3}$$

where $V_i = V(\boldsymbol{x}_i, \boldsymbol{y}_i)$ for $i = 1, \ldots, n+m$ and V is a multivariate nonconformity function based on $\hat{\mu}$, and $U_j \sim \text{Unif}(0,1)$ is an independent random variable for the tie-breaking of the nonconformity scores. We defer the specific choice of V to later sections.

The evaluation of the oracle p-value p_j^* is infeasible, because that in the above definition (3), the computation of $V_{n+j} = V(\boldsymbol{x}_{n+j}, \boldsymbol{y}_{n+j})$ requires knowledge of the unobserved response $\boldsymbol{y}_{n+j}$. To address this issue, we replace V_{n+j} with $\widehat{V}_{n+j} = V(\boldsymbol{x}_{n+j}, \boldsymbol{r}_{n+j})$, where $\boldsymbol{r}_{n+j}$ is an arbitrarily chosen point in the target region R, yielding the (practical) conformal p-values:

$$p_j = \frac{\sum_{i=1}^n \mathbb{1}\{V_i < \widehat{V}_{n+j}\} + U_j(1 + \sum_{i=1}^n \mathbb{1}\{V_i = \widehat{V}_{n+j}\})}{n+1}. \tag{4}$$

Standard results from conformal inference ensure that the oracle conformal p-values p_j^* follow the $\text{Unif}(0,1)$ distribution, in particular implying that they are conservative in the sense that p_j^* as a random variable has a super-uniform distribution on $[0,1]$, satisfying $\mathbb{P}(p_j^* \leq \alpha) \leq \alpha$ (Bates et al., 2023). To guarantee that the practical conformal p-values p_j also maintain this conservativeness, the nonconformity function V must satisfy a property that we introduce as *regional monotonicity*. With this property the p_j's in the resulting mCS procedure will ensure the control of the FDR. We formally define the regional monotonicity as follows:

Definition 3.1 (Regional Monotonicity)**.** A nonconformity score $V : \mathcal{X} \times \mathcal{Y} \to \mathbb{R}$ satisfy the regional monotone property if $V(\boldsymbol{x}, \boldsymbol{y}') \leq V(\boldsymbol{x}, \boldsymbol{y})$ for any $\boldsymbol{x} \in \mathcal{X}$, $\boldsymbol{y}' \in R^c$ and $\boldsymbol{y} \in R$.

Definition 3.1 leads to the conservativeness of p_j. The following proposition formalizes this result, with proof available in Appendix A.1.

Proposition 3.2. *Given that the calibration data $\{(\boldsymbol{x}_i, \boldsymbol{y}_i)\}_{i=1}^n$ together with the j-th data test data point $(\boldsymbol{x}_{n+j}, \boldsymbol{y}_{n+j})$ are exchangeable for $j \in \{1, \ldots, m\}$, regionally monotone nonconformity scores V ensures that the conformal p-value p_j defined in* (4) *is conservative in the following sense,*

$$\mathbb{P}(p_j \leq \alpha \text{ and } j \in \mathcal{H}_0) \leq \alpha, \quad \text{for all } \alpha \in (0,1). \tag{5}$$

Remark 3.3 (Clarification on conservativeness)*.* The conservativeness described in (5) differs from the conventional notion of statistical conservativeness, which is not conditional on the event $j \in \mathcal{H}_0$. Due to the inherent randomness in $\boldsymbol{y}_{n+j}$ within our hypothesis tests, an unknown dependency exists between the p-values and the event $j \in \mathcal{H}_0$. Consequently, the standard form of conservativeness does not hold in this context.

Remark 3.4 (Univariate monotonicity as a special case)*.* Regional monotonicity generalizes the univariate monotonicity concept introduced in CS (Jin & Candès, 2023), which was originally defined for nonconformity scores increasing in their second argument. The original definition is restricted to the univariate case, where the target region R is specified as $(c, +\infty)$. For a univariate nonconformity score $V : \mathcal{X} \times \mathcal{Y} \to \mathbb{R}$, monotonicity implies regional monotonicity. Specifically, for any $y \in R^c = (-\infty, c]$ and $y' \in R$, it holds that $V(\boldsymbol{x}, y) \leq V(\boldsymbol{x}, y')$ for all $\boldsymbol{x} \in \mathcal{X}$. Moreover, even in the univariate case, the original definition of monotonicity across the entire domain $\mathcal{X}$ is unnecessary and monotonicity across the regions R^c and R suffices.

Once a regionally monotone multivariate nonconformity score is defined, conformal p-values can be computed using (4). Leveraging these conformal p-values, mCS again applies the Benjamini-Hochberg procedure (Benjamini & Hochberg, 1995) to construct a selection set $\mathcal{S}$. The complete approach is outlined in Algorithm 1.

```
Algorithm 1 mCS: Multivariate Conformal Selection
Input: Calibration data $\mathcal{D}_{\text{cal}} = \{(\boldsymbol{x}_i, \boldsymbol{y}_i)\}_{i=1}^n$, test data
    $\mathcal{D}_{\text{test}} = \{\boldsymbol{x}_{n+j}\}_{j=1}^m$, target target region $R$, FDR nominal level $q \in (0,1)$,
    regionally monotone nonconformity score $V : \mathcal{X} \times \mathcal{Y} \to \mathbb{R}$.
Output: Selection set $\mathcal{S}$.
1: Compute $V_i = V(\boldsymbol{x}_i, \boldsymbol{y}_i)$ for $i = 1, \ldots, n$, and $\widehat{V}_{n+j} = V(\boldsymbol{x}_{n+j}, \boldsymbol{r}_{n+j})$ for $j = 1, \ldots, m$ with $\boldsymbol{r}_{n+j} \in R$.
2: Construct conformal $p$-values $\{p_i\}_{i=1}^m$ as in (4).
3: (BH procedure) Compute the BH selection threshold $k^* = \max\{k \in \mathbb{Z}_{\geq 0} : \sum_{j=1}^m \mathbb{1}\{p_j \leq qk/m\} \geq k\}$,
   and return the selection set as $\mathcal{S} = \{j : p_j \leq qk^*/m\}$.
```

The following theorem shows that Algorithm 1 controls FDR in finite sample. The proof can be found in Appendix A.2.

Theorem 3.5. *Suppose V is a regionally monotone nonconformity score, and for any $j \in \{1, \ldots, m\}$, the random variables $V_1, \ldots, V_n, V_{n+j}$ are exchangeable conditioned on $\{\widehat{V}_{n+\ell}\}_{\ell \neq j}$. Then, for any $q \in (0,1)$, the output $\mathcal{S}$ of mCS satisfies FDR $\leq q$.*

4. Choices of Nonconformity Score

While the previous sections demonstrate that FDR-controlled selection can be performed using any regionally monotone nonconformity score, the selection power heavily depends on the quality of the chosen score. Although related studies have explored this issue in the context of CP (Romano et al., 2019; Kivaranovic et al., 2020; Sesia & Candès, 2020), there has been limited focus on the choice of scores specific for CS.

In this section, we introduce two types of nonconformity score that satisfies the regional monotonicity. As a result of Theorem 3.5, applying Algorithm 1 with the purposed scores would guarantee FDR control.

4.1. `mCS-dist`: Distance-based Scores

For multivariate selection, we propose distance-based nonconformity scores of the following form:

$$V(\boldsymbol{x}, \boldsymbol{y}) = D_1(\boldsymbol{y}, R^c) - D_2(\hat{\mu}(\boldsymbol{x}), R^c) \tag{6}$$

where $\hat{\mu}$ is a trained predictive model of $\boldsymbol{y}$, and D_1 and D_2 are distance functions. Two examples of distance-based nonconformity score are as follows:

$$1.\text{(regular)} \quad D_1(\boldsymbol{z}, R^c) = D_2(\boldsymbol{z}, R^c) = \inf_{\boldsymbol{s} \in R^c} \|\boldsymbol{z} - \boldsymbol{s}\|_p. \tag{7}$$

$$2.\text{(clipped)} \quad \begin{aligned} D_1(\boldsymbol{z}, R^c) &= M \cdot \mathbb{1}\{\boldsymbol{z} \notin R^c \cup \partial R\}, \\ D_2(\boldsymbol{z}, R^c) &= \inf_{\boldsymbol{s} \in R^c} \|\boldsymbol{z} - \boldsymbol{s}\|_p, \end{aligned} \tag{8}$$

where M is a large constant that serves as a relaxation of infinity. We discuss the role of M in the following paragraphs. These scores generalize the *signed error* score and the *clipped* score (Jin & Candès, 2023), respectively.

This formulation enables the decomposition of $V(\boldsymbol{x}, \boldsymbol{y})$ into two terms. The first term D_1 inherently ensures the regional monotonicity of the score V (Definition 3.1), as D_1 is a distance function satisfying $0 = D_1(\boldsymbol{y}', R^c) \leq D_1(\boldsymbol{y}, R^c)$ for any $\boldsymbol{y}' \in R^c$ and $\boldsymbol{y} \in R$. Meanwhile, the second term D_2 measures the distance between the predicted responses of the points and R^c; this term is designed to increase as $\hat{\mu}(\boldsymbol{x})$ moves away from R^c, ensuring that the test points with the predicted responses having large distance from R^c will yield smaller test scores $\widehat{V}_{n+j}$. According to (4), these points will have smaller conformal p-values and are more likely to be rejected (selected) by the BH procedure. Note that when the predictive model $\hat{\mu}$ outputs an estimated conditional distribution $\widehat{P}(\boldsymbol{y}|\boldsymbol{x})$ – as in classification, conditional density estimation, or Bayesian models – the second term D_2 can be replaced by the predicted probability of being in the target region: $\boldsymbol{y} \in R$, i.e. $\int \mathbb{1}\{\boldsymbol{y} \in R\} \mathrm{d}\widehat{P}(\boldsymbol{y}|\boldsymbol{x})$. This serves the same purpose: points with high predicted probability of satisfying the selection criterion will receive lower scores and are more likely to be selected.

We note that computing $\widehat{V}_{n+j} = V(\boldsymbol{x}_{n+j}, \boldsymbol{r}_{n+j})$ in (4) requires choosing a point $\boldsymbol{r}_{n+j} \in R$ to ensure FDR control. Although any point in R works, selecting $\boldsymbol{r}_{n+j}$ on the boundary ∂R is optimal for maximizing selection power for both regular and clipped scores. In this case, for any fixed $\boldsymbol{x}_{n+j}$, the test score $\widehat{V}_{n+j}$ achieves its minimum, ensuring it to be smaller than a larger proportion of calibration scores $V_i = V(\boldsymbol{x}_i, \boldsymbol{y}_i)$. For example, with the clipped score, if $\boldsymbol{r}_{n+j} \in \partial R$, then the first term D_1 becomes 0 in $\widehat{V}_{n+j}$, while for any calibration samples with $\boldsymbol{y}_i \in R$, we have $D_1 = M$ (except when $\boldsymbol{y}_i$ lies exactly on ∂R, which occurs with zero probability.) This yields smaller p-values p_j and enables more samples to be selected. We require M to be large for the same reason.

The following asymptotic analysis provides a theoretical basis for preferring the second, clipped score (8) over the first score (7) when choosing a distance-based nonconformity score. This result extends the original CS framework (Jin & Candès, 2023) to the multivariate setting, as formalized in the following theorem:

Theorem 4.1. *Let V be any fixed regionally monotone* *nonconformity score, and suppose $\{(\boldsymbol{x}_i, \boldsymbol{y}_i)\}_{i=1}^{n+m}$ are exchangeable from distribution $\mathcal{D}_{X\times Y}$. Let $(\boldsymbol{x}, \boldsymbol{y})$ denote a random pair also drawn from $\mathcal{D}_{X\times Y}$, and define $F(v, u) = \mathbb{P}(V(\boldsymbol{x}, \boldsymbol{y}) < v) + u \cdot \mathbb{P}(V(\boldsymbol{x}, \boldsymbol{y}) = v)$ for any $v \in \mathbb{R}$ and $u \in [0, 1]$. Assuming the choice of $\boldsymbol{r}_{n+j} \equiv \boldsymbol{r} \in R$ is fixed, define*

$$t^* = \sup\left\{t \in [0,1] : \frac{t}{\mathbb{P}(F(V(\boldsymbol{x}, \boldsymbol{r}), U) \le t)} \le q\right\}. \quad (9)$$

Suppose that, for any sufficiently small $\epsilon > 0$, there exists some $t \in (t^* - \epsilon, t^*)$ such that $\frac{t}{\mathbb{P}(F(V(\boldsymbol{x},\boldsymbol{r}),U)\le t)} \le q$. Then the output $\mathcal{S}$ of Algorithm 1 from input $\{(\boldsymbol{x}_i, \boldsymbol{y}_i)\}_{i=1}^{n} \cup \{\boldsymbol{x}_{n+j}\}_{j=1}^{m}$ satisfies

$$\lim_{n,m\to\infty} \text{FDR} = \frac{\mathbb{P}(F(V(\boldsymbol{x}, \boldsymbol{r}), U) \le t^*, \boldsymbol{y} \in R^c)}{\mathbb{P}(F(V(\boldsymbol{x}, \boldsymbol{r}), U) \le t^*)}, \quad \text{and}$$
$$\lim_{n,m\to\infty} \text{Power} = \frac{\mathbb{P}(F(V(\boldsymbol{x}, \boldsymbol{r}), U) \le t^*, \boldsymbol{y} \in R)}{\mathbb{P}(\boldsymbol{y} \in R)}. \quad (10)$$

We omit the proof of Theorem 4.1, as it closely mirrors Proposition 7 in Jin & Candès (2023). An intuitive explanation can be found in Appendix B.2. Leveraging a characterization of the BH procedure (Storey et al., 2004), the theorem establishes that the asymptotics of FDR and power can be precisely achieved by replacing each term in (1) and (2) with its population counterpart. Notably, in this context, t^* represents the population version of the BH rejection threshold for the p-values.

Theorem 4.1 indicates that the second, clipped score (8) is preferable to the first score (7) for achieving higher power. According to (10), since the value $V(\boldsymbol{x}, \boldsymbol{r}) = -\inf_{\boldsymbol{s}\in R^c} \|\boldsymbol{r} - \boldsymbol{s}\|_p$ is identical for both scores assuming $\boldsymbol{r} \in \partial R$, it suffices to compare their asymptotic BH thresholds t^*. The score with the larger t^* achieves higher asymptotic power and is therefore more effective. In the definition of t^* in (9), the fraction can be rewritten as

$$G_V(t) \equiv \frac{t}{\mathbb{P}(F(V(\boldsymbol{x}, \boldsymbol{r}), U) \le t)} = \frac{\mathbb{P}(F(V(\boldsymbol{x}, \boldsymbol{y}), U) \le t)}{\mathbb{P}(F(V(\boldsymbol{x}, \boldsymbol{r}), U) \le t)} = \frac{\mathbb{P}(V(\boldsymbol{x}, \boldsymbol{y}) \le s)}{\mathbb{P}(V(\boldsymbol{x}, \boldsymbol{r}) \le s)}$$

when s is the inverse of $F(\cdot, U)$ at t, assuming it exists. The first equality follows from the fact that $F(V(\boldsymbol{x}, \boldsymbol{y}), U) \sim \text{Unif}(0, 1)$, while the second arises from the monotonicity of F with respect to its first argument v. To maximize t^* in (9) (equivalently, the inverse s^*), an effective score should yield a larger $V(\boldsymbol{x}, \boldsymbol{y})$ relative to $V(\boldsymbol{x}, \boldsymbol{r})$, thereby reducing $G_V(t)$ for a fixed t. This, in turn, results in a larger t^* when computing $\sup\{t : G_V(t) \le q\}$ in (9).

This criterion is precisely satisfied by the clipped score. An alternative justification for favoring the clipped score, based on maximizing the realized FDR, is provided in Appendix B.2, along with further discussions on Theorem 4.1. For an empirical comparison of the performance of the two scores (7) and (8), refer to Appendix C.2.1.

4.2. `mCS-learn`: Learning-based Nonconformity Scores

The two distance-based nonconformity scores introduced in the previous section offer straightforward and practical solutions for many scenarios. However, their effectiveness, particularly in the design of the second distance term $D_2(\cdot)$, is limited in some cases. For example, our numerical simulations indicate that `mCS-dist` would only achieve suboptimal power when R is a nonconvex set; see Appendix C.2.3 for further details. Furthermore, when the target region R is irregular, constructing a closed-form distance function can be challenging, leading to higher computational costs and potential inaccuracies.

To address these challenges, we propose an alternative method `mCS-learn`, which leverages a loss function that penalizes either the smooth selection size or conformal p-value to learn an optimal nonconformity score within the following family:

$$V^\theta(\boldsymbol{x}, \boldsymbol{y}) = M \cdot \mathbb{1}\{\boldsymbol{y} \notin R^c \cup \partial R\} - f_\theta(\boldsymbol{x}, \boldsymbol{y}; R) \quad (11)$$

where M is a large constant and $f_\theta : \mathcal{X} \times \mathcal{Y} \to \mathbb{R}$ is a flexible function parametrized by θ, that can be chosen from a specific machine learning model class, such as kernel machines, gradient boosting models or neural networks, etc. The first term, an indicator function identical to $D_1(\cdot)$ in (8), ensures regional monotonicity in Definition 3.1 and boosts selection power, as suggested in Section 4.1. The second term generalizes the distance term $D_2(\cdot)$ from (6), offering a more expressive framework for constructing optimal nonconformity scores.

The following result demonstrates the expressiveness of the family (11) by showing that it can include the optimal nonconformity score for any selection task. The proof is provided in Appendix A.3.

Proposition 4.2. *Let $\{(\boldsymbol{x}_i, \boldsymbol{y}_i)\}_{i=1}^{n+m}$ be sampled i.i.d. from a distribution, and assume a fixed choice of $\boldsymbol{r}_{n+j} \equiv \boldsymbol{r}$. Under Algorithm 1, for any nominal FDR level q and target region R, there exists a function f^* such that the score constructed using f^* in (11) maximizes the number of selected samples (and thus the power) among all scores with FDR control.*

Remark 4.3 (Subfamilies of the score class). A notable subfamily of (11) is

$$M \cdot \mathbb{1}\{\boldsymbol{y} \notin R^c \cup \partial R\} - f_\theta(\hat{\mu}(\boldsymbol{x}); R), \quad (12)$$

where f_θ depends solely on the prediction $\hat{\mu}(\boldsymbol{x})$. This subfamily includes the clipped distance-based score from Section 4.1 as a special case. Here, regional monotonicity is guaranteed for any constant M, but the score can also be generalized further by allowing f to depend on $\boldsymbol{x}$ as well:

$$M \cdot \mathbb{1}\{\boldsymbol{y} \notin R^c \cup \partial R\} - f_\theta(\boldsymbol{x}, \hat{\mu}(\boldsymbol{x}); R).$$

In contrast, the broader family defined in (11) offers greater flexibility by incorporating $\boldsymbol{y}$ in the second term. However, this added expressiveness requires a sufficiently large M to preserve regional monotonicity. Specifically, M must satisfy $M > 2|f_\theta(\boldsymbol{x}, \boldsymbol{y}; R)|$. In practice, M is chosen to be sufficiently large to ensure that this inequality holds across the entire dataset.

Remark 4.4 (Incorporating pretrained models). This could be achieved in several ways, such as using the predictions of $\hat{\mu}$ as inputs to f_θ, or train f_θ as a prediction on top of $\hat{\mu}$ when both models are implemented as neural networks. While such practice does not increase the expressiveness of the score family, it often facilitates the training of f_θ, as $\hat{\mu}(\boldsymbol{x})$ estimates $\boldsymbol{y}$ and is very informative for selection. Since f_θ can directly learn the data and the selection task, `mCS-learn` can still perform well when $\hat{\mu}$ is poorly fitted; see Appendix C.2.4.

To identify an optimal function within the family (11), we introduce a differentiable loss function that mimics the inherently non-differentiable mCS procedure. The "hard" sorting and ranking operations in the mCS workflow are replaced with their smooth, differentiable counterparts (Blondel et al., 2020; Cuturi et al., 2019). We adopt the implementation introduced in Blondel et al. (2020), with ℓ_2 regularization and regularization strength set to 0.1. The resulting loss function is then used for a chosen machine learning method to train f_θ. Specifically, we partition the calibration data into three batches $\mathcal{D}_{\text{cal}} = \mathcal{D}_{f\text{-train}} \cup \mathcal{D}_{f\text{-val}} \cup \mathcal{D}'_{\text{cal}}$, where $\mathcal{D}_{f\text{-train}}$ and $\mathcal{D}_{f\text{-val}}$ are used for training and validating f_θ, respectively. Upon completion of training and validation, Algorithm 1 can be applied with $\mathcal{D}'_{\text{cal}}$ as the calibration dataset for V^θ to generate the final selection set $\mathcal{S}$.

Training Step. The training loss function is defined based on the smoothed selection size. We denote the softened rank of an element $a \in A$ within the set A by soft-rank$(a; A)$. We randomly partition $\mathcal{D}_{f\text{-train}}$ into two subsets $\mathcal{D}_{f\text{-train1}}$ and $\mathcal{D}_{f\text{-train2}}$, and we assume $\mathcal{D}_{f\text{-train1}} = \{(\boldsymbol{x}_i, \boldsymbol{y}_i)\}_{i=1}^{n'}$ and $\mathcal{D}_{f\text{-train2}} = \{(\boldsymbol{x}_{n'+j}, \boldsymbol{y}_{n'+j})\}_{j=1}^{m'}$ for notational simplicity. We define the smooth conformal p-value $\bar{p}_j^\theta$ for $j = 1, \ldots, m'$ as

$$\bar{p}_j^\theta = \frac{\text{soft-rank}\big(\widehat{V}_{n'+j}^\theta; \{V_i^\theta\}_{i=1}^{n'} \cup \{\widehat{V}_{n'+j}^\theta\}\big)}{n'+1}. \tag{13}$$

where $V_i^\theta := V^\theta(\boldsymbol{x}_i, \boldsymbol{y}_i)$ are computed on $\mathcal{D}_{f\text{-train1}}$ and $\widehat{V}_{n'+j}^\theta := V^\theta(\boldsymbol{x}_{n'+j}, \boldsymbol{r}_{n'+j})$ are computed on $\mathcal{D}_{f\text{-train2}}$, with V^θ in (11). Here, $\mathcal{D}_{f\text{-train1}}$ and $\mathcal{D}_{f\text{-train2}}$ serve as the calibration dataset and test dataset respectively, to obtain the smoothened conformal p-value for training.

Next, to smooth the BH procedure, we first apply the soft-sorting operation to the smooth p-values $\bar{p}_j^\theta$ to obtain their corresponding ranks a_j^θ,

$$a_j^\theta = \text{soft-rank}\big(\bar{p}_j^\theta; \{\bar{p}_k^\theta\}_{k=1}^{m'}\big), \tag{14}$$

and then compute the softened selection size $\overline{S}(\theta)$ as

$$\overline{S}(\theta) \overset{(i)}{=} \log \sum_{j=1}^{m'} e^{a_j^\theta s_j^\theta}, \text{ where } s_j^\theta \overset{(ii)}{=} \sigma\Big(\frac{q \cdot a_j^\theta / m' - \bar{p}_j^\theta}{\tau}\Big).$$

In the above equation, the sigmoid function σ with temperature coefficient τ in (ii) serves as a smooth approximation of the indicator function $\mathbb{1}\{\bar{p}_j^\theta < q a_j^\theta / m'\}$, while the log-sum-exp function in (i) approximates the element-wise max function $\max(a_1^\theta s_1^\theta, \ldots, a_{m'}^\theta s_{m'}^\theta)$. This formulation ensures that $\overline{S}(\theta)$ closely approximates the BH selection size, as the selection size is defined by the largest rank with a p-value below the threshold.

To maximize the selection size, the loss function for learning the `mCS-learn` score can be defined as

$$L_1(\theta) = -\overline{S}(\theta). \tag{15}$$

While above formulation of the loss L_1 is intuitive in its attempt to approximate the final selection size using differentiable functions, the inclusion of two soft sorting steps may reduce numerical stability and impede the training process.

In the mCS procedure, the BH procedure is solely intended for multiplicity correction for FDR control, and thus is not necessarily required in the formulation of the loss function, whose primary objective is to learn the function f_θ. A simpler yet effective alternative loss function directly penalizes the smooth p-values $\bar{p}_j^\theta$ through the following loss function:

$$L_2(\theta) = \sum_{j=1}^{m'} \bar{p}_j^\theta \big[\mathbb{1}(\boldsymbol{y}_{n+j} \in R) - \gamma \cdot \mathbb{1}(\boldsymbol{y}_{n+j} \in R^c)\big]. \tag{16}$$

where $\bar{p}_j^\theta$ is minimized when the j-th sample is deemed desirable, as indicated by the first term. The second term, scaled by a balancing coefficient γ, ensures that the p-value is not uniformly small but becomes relatively larger for less favorable samples. This approach eliminates the need to estimate the BH threshold via a secondary soft-ordering (14), leading to improved numerical stability and enhanced overall performance. A comparison of the two loss functions can be found in Appendix C.2.5. After computing the loss in each epoch, we can follow the standard backpropagation procedure to train θ.

Validation Step. To avoid overfitting f_θ, we perform an additional model selection procedure using a hold-out dataset $\mathcal{D}_{f\text{-val}}$. Specifically, for each epoch $t = 1, \ldots, T$ of the backpropagation procedure, we apply K random partitions on $\mathcal{D}_{f\text{-val}}$ to obtain $\mathcal{D}^{(k)}_{f\text{-val1}}$ and $\mathcal{D}^{(k)}_{f\text{-val2}}$ for $k = 1, \ldots, K$. For each k, we then apply Algorithm 1 with the setting $\mathcal{D}_{\text{cal}} := \mathcal{D}^{(k)}_{f\text{-val1}}$ and $\mathcal{D}_{\text{test}} := \mathcal{D}^{(k)}_{f\text{-val2}}$, and record the validation power $\rho_k(t)$. We then compute the average validation power $\overline{\text{Power}}(t) = \sum_{k=1}^{K} \rho_k(t)/K$.

In the end, we select $\bar{t}$ from $\{1, \ldots, T\}$ to be the epoch with the highest average validation selection power $\overline{\text{Power}}(t)$, and deploy the associated model $f_{\theta_{\bar{t}}}$ for the final selection.

Finally, Algorithm 2 details the complete learning procedure for `mCS-learn` scores. The code for reproduction can be found at `https://github.com/Tian-Bai/mcs`.

Algorithm 2 `mCS-learn` Learning Procedure

```
Input: Training data $\mathcal{D}_{f\text{-train}}$, validation data $\mathcal{D}_{f\text{-val}}$, target region $R$, FDR level $q \in (0,1)$, other hyperparameters.
Output: Trained nonconformity function $f_\theta$.
Initialize parameters $\theta = \theta_0$.
for epoch $t = 1, \ldots, T$ do
    Training Step
    Randomly partition $\mathcal{D}_{f\text{-train}}$ into two disjoint subsets $\mathcal{D}_{f\text{-train1}}$ and $\mathcal{D}_{f\text{-train2}}$.
    Use the current $f_\theta$ to obtain $V_i^\theta$ from $\mathcal{D}_{f\text{-train1}}$ and $\widehat{V}^\theta_{n'+j}$ from $\mathcal{D}_{f\text{-train2}}$.
    Compute the smooth conformal $p$-values $\bar{p}_j^\theta$ (13).
    Compute the loss function using (15) or (16).
    Back-propagate to update the parameters $\theta = \theta_t$.
    Validation Step
    Apply $K$ random partitions on $\mathcal{D}_{f\text{-val}}$ to obtain $\mathcal{D}^{(k)}_{f\text{-val1}}$ and $\mathcal{D}^{(k)}_{f\text{-val2}}$ for each $k = 1, \ldots, K$.
    For each $k$, apply Algorithm 1 for score function $V^\theta$ with the setting $\mathcal{D}_{\text{cal}} := \mathcal{D}^{(k)}_{f\text{-val1}}$, $\mathcal{D}_{\text{test}} := \mathcal{D}^{(k)}_{f\text{-val2}}$ and compute validation power $\rho_k(t)$.
    Compute $\overline{\text{Power}}(t) = \sum_{k=1}^{K} \rho_k(t)/K$.
end for
Determine $\bar{t} = \arg\max_t \overline{\text{Power}}(t)$ and return $f_{\theta_{\bar{t}}}$.
```

5. Simulation Studies

5.1. Baseline Methods

While the standard CS approach is originally designed for univariate settings and cannot be directly applied to multivariate selections, appropriate adaptations can be devised. In this section, we introduce several naïve methods directly adapted from CS to address the multivariate case. Later, we employ these adapted methods as baselines and compare their performance against our proposed method.

In the scenario when the target region $R \subseteq \mathbb{R}^d$ is rectangular, the overall selection criterion can be decomposed, allowing each dimension to be evaluated independently. By applying CS separately to each dimension, we obtain d selection sets $\mathcal{S}_1, \ldots, \mathcal{S}_d$, where each $\mathcal{S}_k$ contains observations satisfying the k-th corresponding marginal criterion. The final selection set S is then given by the intersection of these individual sets, $\mathcal{S} = \cap_{k=1}^{d} \mathcal{S}_k$. We refer to this approach as `CS_int`.

It can be shown that `CS_int`, when each CS subroutine is conducted at a nominal level q, fails to control the FDR at or below q. This issue is analogous to the intersection hypothesis testing (IHT) problem in statistics. A common approach to address this is the Bonferroni correction (Dunn, 1961), which adjusts the nominal level of each subroutine to a lower threshold q/d. Then obtain the intersection of individual sets. However, this method is widely recognized as being overly conservative (Perneger, 1998; Westfall & Young, 1993). We refer to the Bonferroni-adjusted `CS_int` as `CS_ib`. An alternative to the Bonferroni correction is to adaptively account for intersection hypothesis testing. Rather than predefining the nominal levels for subroutines, we determine suitable values by validating on a hold-out dataset. This approach necessitates additional data splits to construct the hold-out set, and we refer to it as `CS_is`.

Beyond considering each dimension separately, another natural adaptation of CS involves transforming the response vector $\boldsymbol{y}$ before applying CS. Specifically, each response $\boldsymbol{y}_i$ is converted to a binary indicator reflecting whether it meets the selection criterion, defined as $\tilde{y}_i = \mathbb{1}\{\boldsymbol{y}_i \in R\}$. Under this transformation, the new selection threshold can be set to $c = 0$, as $\tilde{y}_i > 0$ is equivalent to $\boldsymbol{y}_i \in R$. We refer to this approach as `bi`.

5.2. Numerical Results

We compare the performance of `mCS-dist`, `mCS-learn` (abbreviated as `mCS-d` and `mCS-l` respectively) against the baseline methods outlined in Section 5.1.

In our data generation processes, covariates $\boldsymbol{x}$ are sampled uniformly from $\text{Unif}(-1,1)^p$ where p is the covariate dimension, and the responses $\boldsymbol{y}$ are generated as $\boldsymbol{y} = \mu(\boldsymbol{x}) + \boldsymbol{\epsilon}$, where μ denotes the regression function and $\boldsymbol{\epsilon}$ represents noise drawn from either a multivariate Gaussian or multivariate t-distribution. By varying the regression function, the size of response dimensions, and the choice of Gaussian or heavy-tailed noise, we create a range of selection problems with differing levels of difficulty.

We consider two selection tasks where the target region R

is defined as:

Task 1. The (shifted) first orthant, $R = \{\boldsymbol{y} : y_k \geq c_k \ \ \forall k\}$,

Task 2. A sphere centered at $\boldsymbol{c}$, $R = \{\boldsymbol{y} : \|\boldsymbol{y} - \boldsymbol{c}\|_2 \leq r\}$.

These two specific tasks are particularly relevant in applications, as they simulate scenarios where (1) d criteria must be simultaneously satisfied or (2) an instance must be sufficiently close to a specific point $\boldsymbol{c}$ to be deemed acceptable.

In our simulation, the coefficients $\boldsymbol{c}$ and r for each selection problem are chosen to ensure that approximately 15% to 35% of the data points lie within R across all six data-generating processes. Detailed descriptions of the data generation process, model specifications, and the specific values of the coefficients are provided in Appendix C.1. We first train a support vector regression model $\hat{\mu}$ using 1000 data points, and use an additional labeled dataset of 1000 samples to construct selection sets for different methods in comparison. We evaluate the selection power and FDR using a test dataset of size 100. We adopt the clipped score (8) for `mCS-dist`, and adopt the loss function in (16) with balancing coefficient $\gamma = 0.5$ for `mCS-learn`. For `mCS-learn`, the calibration data is split to $\mathcal{D}_{f\text{-train}}, \mathcal{D}_{f\text{-val}}$ and $\mathcal{D}'_{\text{cal}}$ with ratio 8:1:1, and the model f_θ is formulated as a two-layer MLP with batch normalization. The response dimension is set to be $d = 30$, and nominal FDR level is set at $q = 0.3$. Number of iterations for validation is set to $K = 100$. The selection process is repeated across 100 iterations, with a new dataset generated independently for each iteration.

Table 1 and Table 2 summarize the experimental results for the first selection task. As shown in Table 1, `CS_int` substantially violates FDR control. `CS_is` provides only approximate FDR control, and in scenarios such as Setting 6, the FDR control may be compromised. In each setting, the red numbers in Table 2 indicate the highest achieved power among all methods that properly control the FDR. Among the four remaining methods that always maintain valid FDR control, our two proposed methods consistently achieve the best and second-best power, outperforming the baseline methods under all settings.

Table 3 summarizes similar results for the second selection task. Baseline methods `CS_int` and `CS_ib` are not included as they are not applicable to non-rectangular target regions. Figure 1 shows the realized FDR and power curves across varying nominal FDR levels, ranging from 0.05 to 0.5 in increments of 0.05. Results are shown exclusively for Setting 3 due to space constraints. Among the methods compared, `mCS-dist`, `mCS-learn`, and `bi` demonstrate consistent FDR control, as their respective curves remain below the black dashed line indicating the nominal FDR threshold. Notably, `mCS-learn` also achieves consistently higher power across all nominal levels.

Additional simulation results and ablation studies exploring various factors of the selection problem and the mCS algorithm are provided in Appendix C.2.

Table 1: Observed FDR for Task 1 (shifted first orthant R). The nominal FDR level is $q = 0.3$.

Setting	`CS_int`	`CS_ib`	`CS_is`	`bi`	`mCS-d`	`mCS-l`
1	0.773	0.000	0.280	0.240	0.277	0.251
2	0.801	0.000	0.133	0.300	0.315	0.266
3	0.724	0.000	0.204	0.295	0.264	0.278
4	0.811	0.000	0.255	0.309	0.277	0.315
5	0.810	0.000	0.300	0.266	0.308	0.239
6	0.778	0.000	0.374	0.245	0.287	0.258

Table 2: Observed power for Task 1 (shifted first orthant R).

Setting	`CS_int`	`CS_ib`	`CS_is`	`bi`	`mCS-d`	`mCS-l`
1	1.000	0.000	0.406	0.324	0.555	0.325
2	1.000	0.000	0.012	0.069	0.104	0.108
3	1.000	0.000	0.039	0.059	0.068	0.102
4	1.000	0.000	0.124	0.194	0.324	0.198
5	1.000	0.000	0.019	0.035	0.060	0.042
6	1.000	0.000	0.101	0.027	0.046	0.034

Table 3: Observed FDR and power for Task 2 (spherical R). The nominal FDR level is $q = 0.3$.

	FDR			Power		
Setting	`bi`	`mCS-d`	`mCS-l`	`bi`	`mCS-d`	`mCS-l`
1	0.255	0.265	0.279	0.636	0.760	0.534
2	0.260	0.263	0.273	0.343	0.405	0.421
3	0.216	0.223	0.263	0.115	0.134	0.179
4	0.316	0.286	0.254	0.192	0.333	0.180
5	0.307	0.291	0.273	0.137	0.170	0.189
6	0.292	0.283	0.207	0.055	0.063	0.061

6. Real Data Application

We apply the mCS framework to drug discovery, selecting drug candidates with desirable biological properties. Each candidate corresponds to a chemical compound, where the feature vector $\boldsymbol{x}$ encodes structural or chemical characteristics, and the multivariate response $\boldsymbol{y}$ represents biological properties. The multidimensional nature of $\boldsymbol{y}$ reflects the need to evaluate multiple biological criteria before advancing a compound. Ensuring FDR control improves downstream processes, such as wet-lab validation.

We employ an imputed public ADMET dataset compiled from multiple sources (Wenzel et al., 2019; Iwata et al., 2022; Kim et al., 2023; Watanabe et al., 2018; Falcón-Cano et al., 2022; Esposito et al., 2020; Braga et al., 2015; Aliagas et al., 2022; Perryman et al., 2020; Meng et al., 2022; Vermeire et al., 2022), comprising $n = 22805$ compounds with $d = 15$ biological assay responses. We focus on three different selection tasks with the following target regions:

Task 1. The (shifted) first orthant, $R = \{\boldsymbol{y} : y_k \geq c_k \ \ \forall k\}$,

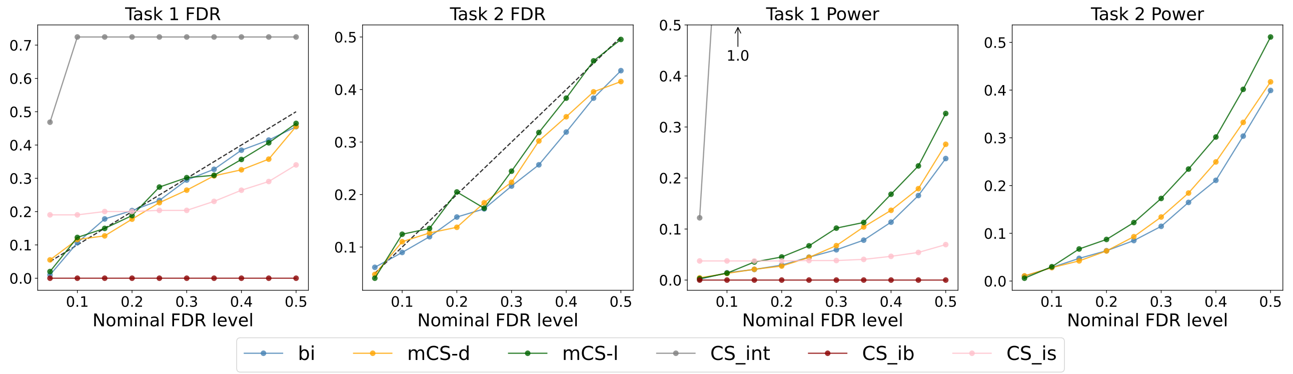

Figure 1: Observed FDR and power across varying nominal levels for Task 1 (shifted first orthant R) and 2 (spherical R).

Task 2. A sphere centered at $\boldsymbol{c}$, $R = \{\boldsymbol{y} : \|\boldsymbol{y} - \boldsymbol{c}\|_2 \leq r\}$,

Task 3. The complement of a sphere centered at $\boldsymbol{c}$, $R = \{\boldsymbol{y} : \|\boldsymbol{y} - \boldsymbol{c}\|_2 \geq r'\}$.

We included the two tasks introduced in the numerical simulations (Section 5), and we also designed the third task to evaluate the performance of various methods under a nonconvex target region with real data. Each of the selection task is designed so that approximately 15%–30% of the compounds qualify as acceptable. Further details on the dataset and problem setup, including response descriptions and cutoffs, can be found in Appendix D.

In this experiment, the underlying predictor $\hat{\mu}$ is specified as a drug property prediction model from the `DeepPurpose` Python package (Huang et al., 2020) with `Morgan` drug encoding. We train the model using $n_{\text{train}} = 12000$ samples, provide $n_{\text{cal}} = 8000$ samples for calibration and reserves the remaining data of size $n_{\text{test}} = 2805$ as test data. We keep the configuration and hyperparameters of the methods unchanged as in Section 5. Two nominal levels $q = 0.3$ and $q = 0.5$ are considered, and the selection processes for each method are repeated across 500 iterations.

Table 4: Observed FDR of different methods with real data.

Task	q	CS_int	CS_ib	CS_is	bi	mCS-d	mCS-l
1	0.3	0.760	0.000	0.303	0.038	0.304	0.275
2	0.3	–	–	–	0.000	0.300	0.293
3	0.3	–	–	–	0.084	0.301	0.296
1	0.5	0.782	0.393	0.496	0.040	0.499	0.488
2	0.5	–	–	–	0.000	0.499	0.498
3	0.5	–	–	–	0.084	0.501	0.497

Table 5: Observed power of methods with real data.

Task	q	CS_int	CS_ib	CS_is	bi	mCS-d	mCS-l
1	0.3	0.993	0.000	0.019	0.000	0.006	0.010
2	0.3	–	–	–	0.000	0.278	0.086
3	0.3	–	–	–	0.000	0.410	0.431
1	0.5	1.000	0.003	0.225	0.000	0.433	0.193
2	0.5	–	–	–	0.000	0.759	0.515
3	0.5	–	–	–	0.000	0.449	0.589

Tables 4 and Table 5 summarize the FDR and power of the competing methods respectively. Consistent with our numerical simulation, the methods `bi`, `mCS-dist`, and `mCS-learn` all demonstrate valid FDR control. Among the methods guaranteed to control FDR, `mCS-dist` and `mCS-learn` consistently achieve the best and the second-best power across all tasks and nominal levels. Notably, `mCS-learn` exhibits superior performance under nonconvex target regions, corroborating the results presented in Appendix C.2.3. We note that although the baseline method `bi` performed well in the simulation settings, it barely selected any compound in the current task. This outcome may be attributed to the suboptimal performance of the underlying binary classification model, which achieved an F1 score of only 0.31.

7. Conclusion

We propose multivariate conformal selection, an extension of conformal selection to multivariate response settings. Our experiments demonstrate that mCS significantly improves selection power while maintaining rigorous FDR control, outperforming existing baselines across simulated and real-world datasets. The flexibility of mCS makes it a valuable tool for selective tasks involved in diverse fields including drug discovery. Looking forward, we anticipate that the mCS framework can be further extended to handle additional practical challenges, including settings with hierarchical or structured responses. By addressing these challenges, mCS has the potential to further enhance its applicability and impact across diverse scientific and industrial domains.

Impact Statement

This work does not present any foreseeable societal consequence.

Acknowledgments

This work was supported by Natural Sciences and Engineering Research Council (NSERC) Discovery Grant (RGPIN-2024-06780) and FRQNT Team Research Project Grant (FRQ-NT 327788).

References

Aliagas, I., Gobbi, A., Lee, M.-L., and Sellers, B. D. Comparison of logp and logd correction models trained with public and proprietary data sets. *Journal of Computer-Aided Molecular Design*, 36(3):253–262, 2022.

Bai, T., Tang, P., Xu, Y., Svetnik, V., Khalili, A., Yu, X., and Yang, A. Conformal selection for efficient and accurate compound screening in drug discovery. *ChemRxiv*, 2024. doi: 10.26434/chemrxiv-2024-pf3ph-v2. URL https://chemrxiv.org/engage/chemrxiv/article-details/67cfbb1c81d2151a02f4efc6.

Bai, Y., Jones, A., Ndousse, K., Askell, A., Chen, A., DasSarma, N., Drain, D., Fort, S., Ganguli, D., Henighan, T., et al. Training a helpful and harmless assistant with reinforcement learning from human feedback. *arXiv preprint arXiv:2204.05862*, 2022.

Bates, S., Angelopoulos, A., Lei, L., Malik, J., and Jordan, M. Distribution-free, risk-controlling prediction sets. *Journal of the ACM (JACM)*, 68(6):1–34, 2021.

Bates, S., Candès, E., Lei, L., Romano, Y., and Sesia, M. Testing for outliers with conformal p-values. *The Annals of Statistics*, 51(1), February 2023. ISSN 0090-5364. doi: 10.1214/22-aos2244. URL http://dx.doi.org/10.1214/22-AOS2244.

Benjamini, Y. and Hochberg, Y. Controlling the false discovery rate: A practical and powerful approach to multiple testing. *Journal of the Royal Statistical Society. Series B (Methodological)*, 57(1):289–300, 1995. ISSN 00359246. URL http://www.jstor.org/stable/2346101.

Benjamini, Y. and Hochberg, Y. Multiple hypotheses testing with weights. *Scandinavian Journal of Statistics*, 24(3): 407–418, 1997. doi: https://doi.org/10.1111/1467-9469.00072. URL https://onlinelibrary.wiley.com/doi/abs/10.1111/1467-9469.00072.

Blondel, M., Teboul, O., Berthet, Q., and Djolonga, J. Fast differentiable sorting and ranking. In *International Conference on Machine Learning*, pp. 950–959. PMLR, 2020.

Braga, R. C., Alves, V. M., Silva, M. F., Muratov, E., Fourches, D., Lião, L. M., Tropsha, A., and Andrade, C. H. Pred-herg: A novel web-accessible computational tool for predicting cardiac toxicity. *Molecular Informatics*, 34(10):698–701, 2015.

Cuturi, M., Teboul, O., and Vert, J.-P. Differentiable ranking and sorting using optimal transport. *Advances in Neural Information Processing Systems*, 32, 2019.

Dunn, O. J. Multiple comparisons among means. *Journal of the American Statistical Association*, 56(293):52–64, 1961.

Esposito, C., Wang, S., Lange, U. E., Oellien, F., and Riniker, S. Combining machine learning and molecular dynamics to predict p-glycoprotein substrates. *Journal of Chemical Information and Modeling*, 60(10):4730–4749, 2020.

Etemadi, N. and Kaminski, M. Strong law of large numbers for 2-exchangeable random variables. *Statistics & probability letters*, 28(3):245–250, 1996.

Falcón-Cano, G., Molina, C., and Cabrera-Pérez, M. Á. Reliable prediction of caco-2 permeability by supervised recursive machine learning approaches. *Pharmaceutics*, 14(10):1998, 2022.

Feldman, S., Bates, S., and Romano, Y. Calibrated multiple-output quantile regression with representation learning. *Journal of Machine Learning Research*, 24(24):1–48, 2023.

Gui, Y., Jin, Y., and Ren, Z. Conformal alignment: Knowing when to trust foundation models with guarantees. *Advances in Neural Information Processing Systems*, 2024.

Heid, E., Greenman, K. P., Chung, Y., Li, S.-C., Graff, D. E., Vermeire, F. H., Wu, H., Green, W. H., and McGill, C. J. Chemprop: a machine learning package for chemical property prediction. *Journal of Chemical Information and Modeling*, 64(1):9–17, 2023.

Huang, K., Fu, T., Glass, L. M., Zitnik, M., Xiao, C., and Sun, J. Deeppurpose: a deep learning library for drug–target interaction prediction. *Bioinformatics*, 36(22-23): 5545–5547, 2020.

Iwata, H., Matsuo, T., Mamada, H., Motomura, T., Matsushita, M., Fujiwara, T., Maeda, K., and Handa, K. Predicting total drug clearance and volumes of distribution using the machine learning-mediated multimodal method through the imputation of various nonclinical data. *Journal of Chemical Information and Modeling*, 62 (17):4057–4065, 2022.

Jin, Y. and Candès, E. J. Selection by prediction with conformal p-values. *Journal of Machine Learning Research*, 24(244):1–41, 2023.

Johnstone, C. and Cox, B. Conformal uncertainty sets for robust optimization. In *Conformal and Probabilistic Prediction and Applications*, pp. 72–90. PMLR, 2021.

Kim, S., Chen, J., Cheng, T., Gindulyte, A., He, J., He, S., Li, Q., Shoemaker, B. A., Thiessen, P. A., Yu, B., et al. Pubchem 2023 update. *Nucleic Acids Research*, 51(D1): D1373–D1380, 2023.

Kivaranovic, D., Johnson, K. D., and Leeb, H. Adaptive, distribution-free prediction intervals for deep networks. In *International Conference on Artificial Intelligence and Statistics*, pp. 4346–4356. PMLR, 2020.

Klein, M., Bethune, L., Ndiaye, E., and Cuturi, M. Multivariate conformal prediction using optimal transport. *arXiv preprint arXiv:2502.03609*, 2025.

Kuleshov, A., Bernstein, A., and Burnaev, E. Conformal prediction in manifold learning. In *Conformal and Probabilistic Prediction and Applications*, pp. 234–253. PMLR, 2018.

Lei, L. and Candès, E. J. Conformal inference of counterfactuals and individual treatment effects. *Journal of the Royal Statistical Society Series B: Statistical Methodology*, 83(5):911–938, 2021.

Meng, J., Chen, P., Wahib, M., Yang, M., Zheng, L., Wei, Y., Feng, S., and Liu, W. Boosting the predictive performance with aqueous solubility dataset curation. *Scientific Data*, 9(1):71, 2022.

Messoudi, S., Destercke, S., and Rousseau, S. Ellipsoidal conformal inference for multi-target regression. In *Conformal and Probabilistic Prediction with Applications*, pp. 294–306. PMLR, 2022.

Park, J. W., Tibshirani, R., and Cho, K. Semiparametric conformal prediction. *arXiv preprint arXiv:2411.02114*, 2024.

Perneger, T. V. What's wrong with bonferroni adjustments. *BMJ*, 316(7139):1236–1238, 1998.

Perryman, A. L., Inoyama, D., Patel, J. S., Ekins, S., and Freundlich, J. S. Pruned machine learning models to predict aqueous solubility. *ACS Omega*, 5(27):16562–16567, 2020.

Romano, Y., Patterson, E., and Candes, E. Conformalized quantile regression. *Advances in Neural Information Processing Systems*, 32, 2019.

Scannell, J. W., Bosley, J., Hickman, J. A., Dawson, G. R., Truebel, H., Ferreira, G. S., Richards, D., and Treherne, J. M. Predictive validity in drug discovery: what it is, why it matters and how to improve it. *Nature Reviews Drug Discovery*, 21(12):915–931, 2022.

Sesia, M. and Candès, E. J. A comparison of some conformal quantile regression methods. *Stat*, 9(1):e261, 2020.

Sheridan, R. P., McMasters, D. R., Voigt, J. H., and 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*, 55(2):231–238, 2015.

Storey, J. D., Taylor, J. E., and Siegmund, D. Strong control, conservative point estimation and simultaneous conservative consistency of false discovery rates: a unified approach. *Journal of the Royal Statistical Society Series B: Statistical Methodology*, 66(1):187–205, 2004.

Szymański, P., Markowicz, M., and Mikiciuk-Olasik, E. Adaptation of high-throughput screening in drug discovery—toxicological screening tests. *International Journal of Molecular Sciences*, 13(1):427–452, 2011.

Vermeire, F. H., Chung, Y., and Green, W. H. Predicting solubility limits of organic solutes for a wide range of solvents and temperatures. *Journal of the American Chemical Society*, 144(24):10785–10797, 2022.

Vovk, V., Gammerman, A., and Shafer, G. *Algorithmic Learning in a Random World*. Springer, 2005. URL https://api.semanticscholar.org/CorpusID:118783209.

Watanabe, R., Esaki, T., Kawashima, H., Natsume-Kitatani, Y., Nagao, C., Ohashi, R., and Mizuguchi, K. Predicting fraction unbound in human plasma from chemical structure: improved accuracy in the low value ranges. *Molecular Pharmaceutics*, 15(11):5302–5311, 2018.

Wenzel, J., Matter, H., and Schmidt, F. Predictive multitask deep neural network models for adme-tox properties: learning from large data sets. *Journal of Chemical Information and Modeling*, 59(3):1253–1268, 2019.

Westfall, P. H. and Young, S. S. *Resampling-Based Multiple Testing: Examples and Methods for p-value Adjustment*, volume 279. John Wiley & Sons, 1993.

Yang, K., Swanson, K., Jin, W., Coley, C., Eiden, P., Gao, H., Guzman-Perez, A., Hopper, T., Kelley, B., Mathea, M., et al. Analyzing learned molecular representations for property prediction. *Journal of Chemical Information and Modeling*, 59(8):3370–3388, 2019.

Zhang, K., Yang, X., Wang, Y., Yu, Y., Huang, N., Li, G., Li, X., Wu, J. C., and Yang, S. Artificial intelligence in drug development. *Nature Medicine*, pp. 1–15, 2025.

A. Technical Proofs

A.1. Proof of Proposition 3.2

Proof of Proposition 3.2. Since V is fixed, the nonconformity scores $V_1, \dots, V_n$ and V_{n+j} are exchangeable. By a standard result from conformal inference Vovk et al. (2005, Proposition 2.4), the oracle p-value p_j^* defined as in (3) is uniform distributed with value ranging in $(0,1)$, and $\mathbb{P}(p_j^* \le \alpha) \le \alpha$. This gives

$$\mathbb{P}(p_j^* \le \alpha \text{ and } j \in \mathcal{H}_0) \le \alpha.$$

When the null hypothesis H_{0j} is true, $\boldsymbol{y}_{n+j} \in R^c$. Since $\boldsymbol{r}_{n+j} \in R$, by the regional monotone property, we have $V_{n+j} = V(\boldsymbol{x}_{n+j}, \boldsymbol{y}_{n+j}) \le V(\boldsymbol{x}_{n+j}, \boldsymbol{r}_{n+j}) = \widehat{V}_{n+j}$. We then have $p_j^* \le p_j$ by definition, and

$$\mathbb{P}(p_j \le \alpha \text{ and } j \in \mathcal{H}_0) \le \mathbb{P}(p_j^* \le \alpha \text{ and } j \in \mathcal{H}_0) \le \alpha.$$

□

A.2. Proof of Theorem 3.5

Proof of Theorem 3.5. We adapt the proof of Theorem 6 in (Jin & Candès, 2023). In the proof that follows, we fix index $j \in \{1, \dots, m\}$. For notational simplicity, only in this proof we deal with the *deterministic* conformal p-values,

$$p_j = \frac{1}{n+1}\big[1 + \sum_{i=1}^{n} \mathbb{1}\{V_i \le \widehat{V}_{n+j}\}\big]$$

We note that the deterministic conformal p-values are only valid when the scores $\{V_i\}_{i=1}^n \cup \{V_{n+j}\}_{j=1}^m$ have no ties almost surely, and therefore we also make this assumption. We highlight that the validity of the *random* conformal p-value does not rely on this statement. We also define the corresponding deterministic oracle p-values,

$$p_j^* = \frac{1}{n+1}\big[1 + \sum_{i=1}^{n} \mathbb{1}\{V_i \le V_{n+j}\}\big].$$

This version of p-values are also conservative by standard results in conformal inference (Bates et al., 2023; Jin & Candès, 2023). For $l = 1, \dots, m$, we define a set of slightly modified p-values,

$$p_l^{(j)} = \frac{1}{n+1}\big[\sum_{i=1}^{n} \mathbb{1}\{V_i \le \widehat{V}_{n+l}\} + \mathbb{1}\{V_{n+j} \le \widehat{V}_{n+l}\}\big].$$

Also define $\mathcal{S}(a_1, \dots, a_m) \subseteq \{1, \dots, m\}$ as the rejection index set obtained by the Benjamini-Hochberg procedure, from p-values taking values $a_1, \dots, a_m$. Then, the output of mCS is $\mathcal{S}(p_1, \dots, p_m)$.

In the sequel, we will compare $\mathcal{S}(p_1, \dots, p_m)$ to

$$\mathcal{S}(p_1^{(j)}, \dots, p_{j-1}^{(j)}, p_j^*, p_{j+1}^{(j)}, \dots, p_m^{(j)})$$

for the test sample j that is falsely rejected, i.e.

$$j \in \mathcal{S}, Y_{n+j} \in R^c.$$

First, on this event, we have $V_{n+j} = V(\boldsymbol{x}_{n+j}, \boldsymbol{y}_{n+j}) \le V(\boldsymbol{x}_{n+j}, \boldsymbol{r}_{n+j}) = \widehat{V}_{n+j}$ and $p_j^* \le p_j$. For the remaining p-values, $\{p_l^{(j)}\}_{l \ne j}$, we consider two cases:

(i) If $\widehat{V}_{n+l} \ge \widehat{V}_{n+j}$, then $p_l \ge p_j$. In this case, we also have $\widehat{V}_{n+l} \ge V_{n+j}$ as $\widehat{V}_{n+j} \ge V_{n+j}$. This implies

$$p_l^{(j)} = \frac{1}{n+1}\big[\sum_{i=1}^{n} \mathbb{1}\{V_i \le \widehat{V}_{n+l}\} + \mathbb{1}\{V_{n+j} \le \widehat{V}_{n+l}\}\big] = \frac{1}{n+1}\big[\sum_{i=1}^{n} \mathbb{1}\{V_i \le \widehat{V}_{n+l}\} + 1\big] = p_l.$$

(ii) If $\widehat{V}_{n+l} < \widehat{V}_{n+j}$, then $p_l \leq p_j$. We also have

$$p_l^{(j)} \leq \frac{1}{n+1}\big[\sum_{i=1}^{n} \mathbb{1}\{V_i \leq \widehat{V}_{n+l}\} + 1\big] \leq \frac{1}{n+1}\big[\sum_{i=1}^{n} \mathbb{1}\{V_i \leq \widehat{V}_{n+j}\} + 1\big] = p_j.$$

Since $j \in \mathcal{S}$, by the definition of Benjamini-Hochberg procedure $l \in \mathcal{S}$ as p_l has smaller rank when ordering the p-values.

To summarize, if we replace $(p_1, \ldots, p_m)$ by $(p_1^{(j)}, \ldots, p_{j-1}^{(j)}, p_j^*, p_{j+1}^{(j)}, \ldots, p_m^{(j)})$ on the event $\{j \in \mathcal{S}, Y_{n+j} \in R^c\}$, such a replacement does not modify any of those p-values p_l if they satisfied $p_l \geq p_j$. Also, for all p-values p_l with $p_l \leq p_j$ including j itself ($l = j$), their replaced values $p_l^{(j)}$ are still no greater than p_j. Since all p-values are no larger than their original values after the replacements, the size of rejection set must not decrease. On the other hand, since $j \in \mathcal{S}$ and no p-values larger than p_j are modified, no new hypotheses can be rejected by the new set of p-values. We conclude that

$$\mathcal{S}_j^* := \mathcal{S}(p_1^{(j)}, \ldots, p_{j-1}^{(j)}, p_j^*, p_{j+1}^{(j)}, \ldots, p_m^{(j)}) = \mathcal{S}(p_1, \ldots, p_m) = \mathcal{S}$$

on the event $\{Y_{n+j} \in R^c, j \in \mathcal{S}\}$. By decomposing the FDR we have

$$\begin{aligned}
\text{FDR} &= \mathbb{E}\Big[\frac{\sum_{j=1}^{m} \mathbb{1}\{Y_{n+j} \in R^c\}\mathbb{1}\{j \in \mathcal{S}\}}{\max(1, |\mathcal{S}|)}\Big] \\
&= \sum_{j=1}^{m}\sum_{k=1}^{m} \frac{1}{k}\mathbb{E}\big[\mathbb{1}\{|\mathcal{S}| = k\}\mathbb{1}\{Y_{n+j} \in R^c\}\mathbb{1}\{p_j \leq \frac{qk}{m}, j \in \mathcal{S}\}\big] \\
&\leq \sum_{j=1}^{m}\sum_{k=1}^{m} \frac{1}{k}\mathbb{E}\big[\mathbb{1}\{|\mathcal{S}_j^*| = k\}\mathbb{1}\{Y_{n+j} \in R^c\}\mathbb{1}\{p_j^* \leq \frac{qk}{m}\}\big] \\
&\leq \sum_{j=1}^{m}\sum_{k=1}^{m} \frac{1}{k}\mathbb{E}\big[\mathbb{1}\{|\mathcal{S}_j^*| = k\}\mathbb{1}\{p_j^* \leq \frac{qk}{m}\}\big] \\
&= \sum_{j=1}^{m}\sum_{k=1}^{m} \frac{1}{k}\mathbb{E}\big[\mathbb{1}\{|\mathcal{S}_j^*| = k\}\mathbb{1}\{j \in \mathcal{S}_j^*\}\big].
\end{aligned}$$

The last equality is again by the property of the Benjamini-Hochberg procedure. Also, by its step-up nature, sending p_j^* to zero does not change the rejection set if the corresponding hypothesis of p_j^* is rejected, i.e. on the event $\{j \in \mathcal{S}_j^*\}$. We have

$$\mathcal{S}_j^* = \mathcal{S}(p_1^{(j)}, \ldots, p_{j-1}^{(j)}, p_j^*, p_{j+1}^{(j)}, \ldots, p_m^{(j)}) = \mathcal{S}(p_1^{(j)}, \ldots, p_{j-1}^{(j)}, 0, p_{j+1}^{(j)}, \ldots, p_m^{(j)}) =: \mathcal{S}_{j\to 0}^*$$

Thus,

$$\text{FDR} \leq \sum_{j=1}^{m}\sum_{k=1}^{m} \frac{1}{k}\mathbb{E}\big[\mathbb{1}\{|\mathcal{S}_{j\to 0}^*| = k\}\mathbb{1}\{j \in \mathcal{S}_j^*\}\big] = \sum_{j=1}^{m} \mathbb{E}\Big[\frac{\mathbb{1}\{p_j^* \leq q|\mathcal{S}_j^*|/m\}}{\max(1, |\mathcal{S}_{j\to 0}^*|)}\Big] \leq \sum_{j=1}^{m} \mathbb{E}\Big[\frac{\mathbb{1}\{p_j^* \leq q|\mathcal{S}_{j\to 0}^*|/m\}}{\max(1, |\mathcal{S}_{j\to 0}^*|)}\Big]$$

By definition, $\{p_l^{(j)}\}_{l\neq j}$ is invariant after permuting $\{V_i\}_{i=1}^n \cup \{V_{n+j}\}$. Since $\{V_i\}_{i=1}^n \cup \{V_{n+j}\}$ are exchangeable conditioned on $\{\widehat{V}_{n+l}\}_{l\neq j}$, the distribution of $\{p_l^{(j)}\}_{l\neq j}$ is independent from the ordering of $\{V_i\}_{i=1}^n \cup \{V_{n+j}\}$ conditioned on the (unordered) set $[V_1, \ldots, V_n, V_{n+j}] \cup \{\widehat{V}_{n+l}\}_{l\neq j}$. Also, conditioned on $\{\widehat{V}_{n+l}\}_{l\neq j}$, $\mathcal{S}_{j\to 0}^*$ only depends on $\{p_l^{(j)}\}_{l\neq j}$ which is in turn only dependent on the unordered set $[V_1, \ldots, V_n, V_{n+j}]$, and p_j^* only depends on the ordering of $\{V_i\}_{i=1}^n \cup \{V_{n+j}\}$. This implies that $\mathcal{S}_{j\to 0}^*$ is independent on p_j^* conditioned on the (unordered) set $[V_1, \ldots, V_n, V_{n+j}]$ and $\{\widehat{V}_{n+l}\}_{l\neq j}$. Therefore, by the conservative property of conformal p-values and conditional independence,

$$\mathbb{P}\Big(p_j^* \leq \frac{q|\mathcal{S}_{j\to 0}^*|}{m} \,\Big|\, [V_1, \ldots, V_n, V_{n+j}] \cup \{\widehat{V}_{n+l}\}_{l\neq j}\Big) \leq \frac{q|\mathcal{S}_{j\to 0}^*|}{m}.$$

By the law of total expectation,

$$\begin{aligned}\mathbb{E}\Big[\frac{\mathbb{1}\{p_j^* \le q|\mathcal{S}_{j\to 0}^*|/m\}}{\max(1,|\mathcal{S}_{j\to 0}^*|)}\Big] &= \mathbb{E}\Big[\mathbb{E}\Big[\frac{\mathbb{1}\{p_j^* \le q|\mathcal{S}_{j\to 0}^*|/m\}}{\max(1,|\mathcal{S}_{j\to 0}^*|)}\ \Big|\ [V_1,\ldots,V_n,V_{n+j}]\cup\{\widehat{V}_{n+l}\}_{l\neq j}\Big]\Big] \\ &\le \mathbb{E}\Big[\mathbb{1}\{|\mathcal{S}_{j\to 0}^*|\neq 0\}\frac{q|\mathcal{S}_{j\to 0}^*|}{m|\mathcal{S}_{j\to 0}^*|}\ \Big|\ [V_1,\ldots,V_n,V_{n+j}]\cup\{\widehat{V}_{n+l}\}_{l\neq j}\Big] \\ &\le \frac{q}{m}.\end{aligned}$$

Since when $|\mathcal{S}_{j\to 0}^*| = 0$, $\mathbb{1}\{p_j^* \le q|\mathcal{S}_{j\to 0}^*|/m\} = 0$. Now, the proof is concluded by summing over every $j = 1, \ldots, m$. □

A.3. Proof of Proposition 4.2

Proof of Proposition 4.2. For any specific dataset and selection problem, let V^{opt} denote the optimal nonconformity score that controls the FDR and maximize the number of selected samples. Define $V_i^{\text{opt}} = V^{\text{opt}}(\boldsymbol{x}_i, \boldsymbol{y}_i)$ for calibration samples and $\widehat{V}_{n+j}^{\text{opt}} = V^{\text{opt}}(\boldsymbol{x}_{n+j}, \boldsymbol{r}_{n+j})$ for test samples. Consider the following score W within the family (11):

$$W(\boldsymbol{x}, \boldsymbol{y}) = M \cdot \mathbb{1}\{\boldsymbol{y} \notin R^c \cup \partial R\} + V^{\text{opt}}(\boldsymbol{x}, \boldsymbol{y}).$$

Picking $\boldsymbol{r}_{n+j} \equiv \boldsymbol{r} \in \partial R$, the test nonconformity scores $\widehat{W}_{n+j}$ satisfy:

$$\widehat{W}_{n+j} = W(\boldsymbol{x}_{n+j}, \boldsymbol{r}_{n+j}) = V^{\text{opt}}(\boldsymbol{x}_{n+j}, \boldsymbol{r}_{n+j}) = \widehat{V}_{n+j}^{\text{opt}},$$

and for each calibration score,

$$W_i = W(\boldsymbol{x}_i, \boldsymbol{y}_i) \ge V^{\text{opt}}(\boldsymbol{x}_i, \boldsymbol{y}_i) = V_i^{\text{opt}}.$$

Therefore, by replacing V^{opt} with W, the calibration nonconformity scores $W_1, \ldots, W_n$ may increase, while the test nonconformity scores $\widehat{W}_{n+j}$ remain unchanged. According to (4), this leads to a decrease in each conformal p-value p_j, which in turn increase the value k^* in Algorithm 1. This means that we would select more samples, by the definition of $\mathcal{S}$. Consequently, we conclude that there must exist a score within the family (11) that achieves the optimal selection size. A similar argument applies to the maximization selection power, thereby concluding the proof. □

B. Deferred Discussions

B.1. Discussions on Classification Responses

In our paper, we choose to focus primarily on the regression response case for multivariate conformal selection. In fact, the selection problem for classification responses (univariate or multivariate) can be directly reduced to the univariate conformal selection framework introduced by Jin & Candès (2023):

For the univariate classification setting, suppose the response space is composed of classes $\mathcal{Y} = \cup_{k=1}^K C_k$ with target region $R = \cup_{k=1}^s C_k$ (with $s < K$). Then, by defining a binary response: $\tilde{y}_i = \mathbb{1}\{y_i \in R\}$, the original selection problem directly translates into a univariate conformal selection problem, where we select samples with $\tilde{y}_i = 1$.

For multivariate classification case, e.g. suppose that bivariate responses $\boldsymbol{y}_i$ are drawn from a joint class space $\mathcal{Y} = (\mathcal{Y}^{(1)}, \mathcal{Y}^{(2)}) = \cup_{k,\ell}(C_k^{(1)}, C_\ell^{(2)})$, and the target region $R = \cup_{(k,\ell)\in\mathcal{I}}(C_k^{(1)}, C_\ell^{(2)})$. Again, we define a binary indicator $\tilde{y}_i = \mathbb{1}\{\boldsymbol{y}_i \in R\}$, converting the original multivariate selection problem into a standard univariate selection task:

$$H_{0j} : \tilde{y}_{n+j} < 0.5 \quad \text{versus} \quad H_{1j} : \tilde{y}_{n+j} \ge 0.5.$$

Since $P(\tilde{y}_i = 1) = P(\boldsymbol{y}_i \in R)$, there is a direct correspondence between the multivariate and univariate nonconformity scores:

$$V(\boldsymbol{x}, \boldsymbol{y}) = M \cdot \mathbb{1}\{\boldsymbol{y} \in R\} - \widehat{P}(\boldsymbol{y} \in R|\boldsymbol{x})$$

and

$$V(\boldsymbol{x}, \tilde{y}) = M \cdot \mathbb{1}\{\tilde{y} \ge 0.5\} - \tilde{\mu}(\boldsymbol{x})$$

where $\tilde{\mu}(\boldsymbol{x}) \equiv \widehat{P}(\tilde{y} = 1|\boldsymbol{x})$. Moreover, regional monotonicity is simply the usual monotone condition of univariate conformal selection: $V(\boldsymbol{x}, \tilde{y} = 0) \leq V(\boldsymbol{x}, \tilde{y} = 1)$.

Thus, every classification-based selection task can be naturally and effectively solved using existing univariate conformal selection methods.

B.2. Additional Discussions on Theorem 4.1 and Advantages of The Clipped Nonconformity Score

In this section, we first provide interpretations about Theorem 4.1 and offer an alternative perspective on the advantages of the clipped nonconformity score.

We first note that the (practical) conformal p-values defined in (4) can be rewritten as

$$p_j = \frac{1}{n+1}\sum_{i=1}^{n} \mathbb{1}\{V_i < \widehat{V}_{n+j}\} + \frac{1}{n+1}U_j \sum_{i=1}^{n} \mathbb{1}\{V_i = \widehat{V}_{n+j}\} + \frac{1}{n+1}U_j.$$

By the Glivenko-Cantelli theorem under exchangeability (Etemadi & Kaminski, 1996), as $n \to \infty$,

$$\sup_{t\in\mathbb{R}} \left| \frac{1}{n+1}\sum_{i=1}^{n} \mathbb{1}\{V_i < t\} + \frac{1}{n+1}U_j \sum_{i=1}^{n} \mathbb{1}\{V_i = t\} + \frac{1}{n+1}U_j - \mathbb{P}(V(\boldsymbol{x},\boldsymbol{y}) < t) - U_j \cdot \mathbb{P}(V(\boldsymbol{x},\boldsymbol{y}) = t) \right| \overset{a.s.}{\to} 0.$$

Replacing t in the above by $\widehat{V}_{n+j} = V(\boldsymbol{x}_{n+j}, \boldsymbol{r})$ yields

$$\begin{aligned} p_j &\overset{a.s.}{\to} \mathbb{P}(V(\boldsymbol{x},\boldsymbol{y}) < V(\boldsymbol{x}_{n+j},\boldsymbol{r})|V(\boldsymbol{x}_{n+j},\boldsymbol{r})) + U_j \cdot \mathbb{P}(V(\boldsymbol{x},\boldsymbol{y}) = V(\boldsymbol{x}_{n+j},\boldsymbol{r})|V(\boldsymbol{x}_{n+j},\boldsymbol{r})) = F(V(\boldsymbol{x}_{n+j},\boldsymbol{r}), U_j) \\ &\overset{d}{\sim} F(V(\boldsymbol{x},\boldsymbol{r}), U). \end{aligned}$$

Therefore, $F(V(\boldsymbol{x},\boldsymbol{r}), U)$ is also a conservative random variable.

The quantity t^* in Theorem 4.1 can also be viewed as the asymptotic counterpart of a corresponding finite-sample quantity. A characterization of the BH procedure (Storey et al., 2004) states that the rejection set $\mathcal{S} = \{j : p_j \leq \tau^*\}$, where the BH rejection threshold τ^* is defined as

$$\tau^* = \sup\left\{ t \in [0,1] : \frac{t}{\frac{1}{m}\sum_{j=1}^{m} \mathbb{1}\{p_j \leq t\}} \leq q \right\}.$$

By the fact that $\frac{1}{m}\sum_{j=1}^{m} \mathbb{1}\{p_j \leq t\} - \mathbb{P}(p_j \leq t) \overset{p}{\to} 0$ as $m \to \infty$ (which follows due to the the weak law of large numbers for triangular arrays), and our earlier characterization of $F(V(\boldsymbol{x},\boldsymbol{r}), U_j)$ which implies $\mathbb{P}(p_j \leq t) \to \mathbb{P}(F(V(\boldsymbol{x},\boldsymbol{r}), U) \leq t)$ as $n \to \infty$, the fraction in the definition of τ^* satisfies (in the limit $n, m \to \infty$)

$$\frac{t}{\frac{1}{m}\sum_{j=1}^{m} \mathbb{1}\{p_j \leq t\}} \overset{p}{\to} \frac{t}{\mathbb{P}(F(V(\boldsymbol{x},\boldsymbol{r}), U) \leq t)}.$$

This establishes t^* as the asymptotic limit of the BH rejection threshold τ^*, which further implies

$$\frac{1}{m}\sum_{j=1}^{m} \mathbb{1}\{p_j \leq \tau^*\} \overset{p}{\to} \mathbb{P}(F(V(\boldsymbol{x},\boldsymbol{r}), U) \leq t^*) \text{ and } \frac{1}{m}\sum_{j=1}^{m} \mathbb{1}\{p_j \leq \tau^* \text{ and } \boldsymbol{y}_{n+j} \in R\} \overset{p}{\to} \mathbb{P}(F(V(\boldsymbol{x},\boldsymbol{r}), U) \leq t^*, \boldsymbol{y} \in R).$$

Theorem 4.1 then gives the asymptotic version of FDR and power:

$$\begin{aligned} \text{FDR} &= \mathbb{E}\left[\frac{|\mathcal{S} \cap \{j : \boldsymbol{y}_{n+j} \in R^c\}|}{|\mathcal{S}|}\right] = \mathbb{E}\left[\frac{|\{j : p_j \leq \tau^*\} \cap \{j : \boldsymbol{y}_{n+j} \in R^c\}|}{|\{j : p_j \leq \tau^*\}|}\right] \\ &= \mathbb{E}\left[\frac{\frac{1}{m}\sum_{j=1}^{m} \mathbb{1}\{p_j \leq \tau^* \text{ and } \boldsymbol{y}_{n+j} \in R^c\}}{\frac{1}{m}\sum_{j=1}^{m} \mathbb{1}\{p_j \leq \tau^*\}}\right] \\ &\to \frac{\mathbb{P}(F(V(\boldsymbol{x},\boldsymbol{r}), U) \leq t^*, \boldsymbol{y} \in R^c)}{\mathbb{P}(F(V(\boldsymbol{x},\boldsymbol{r}), U) \leq t^*)}, \end{aligned}$$

$$\begin{aligned}
\text{Power} &= \mathbb{E}\left[\frac{|\mathcal{S}\cap\{j : \boldsymbol{y}_{n+j}\in R\}|}{|\{j : \boldsymbol{y}_{n+j}\in R\}|}\right] = \mathbb{E}\left[\frac{|\{j : p_j \le \tau^*\}\cap\{j : \boldsymbol{y}_{n+j}\in R\}|}{|\{j : \boldsymbol{y}_{n+j}\in R\}|}\right] \\
&= \mathbb{E}\left[\frac{\frac{1}{m}\sum_{j=1}^m \mathbb{1}\{p_j \le \tau^* \text{ and } \boldsymbol{y}_{n+j}\in R\}}{\frac{1}{m}\sum_{j=1}^m \mathbb{1}\{\boldsymbol{y}_{n+j}\in R\}}\right] \\
&\to \frac{\mathbb{P}(F(V(\boldsymbol{x},\boldsymbol{r}),U)\le t^*, \boldsymbol{y}\in R)}{\mathbb{P}(\boldsymbol{y}\in R)}.
\end{aligned}$$

Theorem 4.1 suggests that the clipped score should be preferred from the perspective of maximizing the realized (asymptotic) FDR (which is still controlled below the nominal level). The asymptotic FDR is bounded through the following chain of inequalities:

$$\begin{aligned}
\frac{\mathbb{P}(F(V(\boldsymbol{x},\boldsymbol{r}),U)\le t^*, \boldsymbol{y}\in R^c)}{\mathbb{P}(F(V(\boldsymbol{x},\boldsymbol{r}),U)\le t^*)} &\le \frac{\mathbb{P}(F(V(\boldsymbol{x},\boldsymbol{y}),U)\le t^*, \boldsymbol{y}\in R^c)}{\mathbb{P}(F(V(\boldsymbol{x},\boldsymbol{r}),U)\le t^*)} \\
&\overset{(*)}{\le} \frac{\mathbb{P}(F(V(\boldsymbol{x},\boldsymbol{y}),U)\le t^*)}{\mathbb{P}(F(V(\boldsymbol{x},\boldsymbol{r}),U)\le t^*)} = \frac{t^*}{\mathbb{P}(F(V(\boldsymbol{x},\boldsymbol{r}),U)\le t^*)}. \qquad (17)
\end{aligned}$$

By the definition of t^*, the final term in (17) is at most the nominal level q. If the bounds in this chain of inequalities could be further tightened, the realized FDR would more closely align with the nominal level q, effectively allowing for greater selection power as more of the FDR budget could be utilized.

Using the clipped score tightens the inequality $(*)$ in (17). Assuming the score function V is clipped, we observe that

$$\frac{\mathbb{P}(F(V(\boldsymbol{x},\boldsymbol{y}),U)\le t^*)}{\mathbb{P}(F(V(\boldsymbol{x},\boldsymbol{r}),U)\le t^*)} - \frac{\mathbb{P}(F(V(\boldsymbol{x},\boldsymbol{y}),U)\le t^*, \boldsymbol{y}\in R^c)}{\mathbb{P}(F(V(\boldsymbol{x},\boldsymbol{r}),U)\le t^*)} = \frac{\mathbb{P}(F(V(\boldsymbol{x},\boldsymbol{y}),U)\le t^*, \boldsymbol{y}\in R)}{\mathbb{P}(F(V(\boldsymbol{x},\boldsymbol{r}),U)\le t^*)}$$

is approximately zero, since for $\boldsymbol{y}\in R$, the clipped score satisfies $V(\boldsymbol{x},\boldsymbol{y})\approx M$, implying that $\mathbb{P}(F(V(\boldsymbol{x},\boldsymbol{y}),U)\le t^*, \boldsymbol{y}\in R)\approx 0$, because that $F(v,u)$ is monotone with respect to its first argument v. Notably, the large constant M is originally introduced in conformal selection as a relaxation of infinity, and these approximations indeed hold. This phenomenon is verified through our extra simulations in Appendix C.2.1.

C. Additional Details for Numerical Simulations (Section 5)

C.1. Data Generating Processes and Configuration of the Selection Tasks

The configuration of the true regression function μ and noise term $\boldsymbol{\epsilon}\in\mathbb{R}^d$ can be found in Table 6.

For the noise term, the degree of freedom for the t-distribution scenario is set to $\nu = 3$, and the scale matrix Σ is specified as a matrix with diagonal entries of 0.5 and off-diagonal entries of 0.05. The second column denotes the k-th output entry of the true regression function μ, which relates to the k-th coordinate y_k of the response $\boldsymbol{y}$ and the k-th coordinate ϵ_k of the noise $\boldsymbol{\epsilon}$ as $y_k = [\mu(\boldsymbol{x})]_k + \epsilon_k$. In all of our simulations, we take $p = 10$ (recall that $\boldsymbol{x}\sim \text{Unif}(-1,1)^p$ but $\boldsymbol{y}\in\mathbb{R}^d$). If, when generating y_k the value of x_ℓ (the ℓ-th coordinate of $\boldsymbol{x}$) for $\ell > p$ is needed, we take $x_\ell = x_{((\ell-1) \bmod p)+1}$.

Table 7 summarizes the values of coefficient vectors ($\boldsymbol{c}$ and $\boldsymbol{r}$) that define the selection tasks for each response dimension. Within each vector, all entries share the same value. For instance, if c_k is listed as 1, it indicates that $\boldsymbol{c} = (1,1,\ldots,1)$.

The six settings we consider differ in two key aspects that influence the difficulty of selection. First, the regression function μ exhibits varying degrees of nonlinearity across settings. Specifically, Settings 1 and 4 are linear, while Settings 2 and 5 have weak nonlinearity, and Settings 3 and 6 exhibit strong nonlinearity. The degree of nonlinearity affects the predictive accuracy of the estimated regression function $\hat{\mu}$, which in turn impacts the selection difficulty. Second, the distribution of the noise term $\boldsymbol{\epsilon}$ differs across settings, leading to variations in the conditional variance. In Settings 1, 2, and 3, $\boldsymbol{\epsilon}$ follows a multivariate Gaussian distribution, whereas in Settings 4, 5, and 6, it follows a multivariate t-distribution. The conditional variance $\text{cov}(\boldsymbol{\epsilon}) = \text{Var}(\boldsymbol{y}\mid\boldsymbol{x})$ in the latter three settings is $\frac{\nu}{\nu-2}\Sigma = 3\Sigma$, which is three times larger than that of the former three settings, thereby increasing the difficulty of selection.

To better illustrate the data distributions, Figure 2 presents example scatter plots of the response vector $\boldsymbol{y}$ for the six settings, with the response dimension set to 2 for visualization purposes. Higher-dimensional cases are not displayed as they are

Table 6: True regression functions and noise distributions

Setting	$[\mu(\cdot)]_k$	$\boldsymbol{\epsilon}$
1	$x_k - \frac{1}{2}x_{k+1} + x_{k+2} + \frac{3}{2}$	$\mathcal{N}(0, \Sigma)$
2	$x_k + x_{k+2}^2 + \frac{1}{2}$	$\mathcal{N}(0, \Sigma)$
3	$\mathbb{1}\{x_k x_{k+1} > 0\} \cdot \mathbb{1}\{x_{k+2} > 0.5\} \cdot (0.25 + x_{k+2})$ $\mathbb{1}\{x_k x_{k+1} \leq 0\} \cdot \mathbb{1}\{x_{k+2} \leq 0.5\} \cdot (x_{k+2} - 0.25) + 0.75$	$\mathcal{N}(0, \Sigma)$
4	Same as Setting 1	$t_\nu(0, \Sigma)$
5	Same as Setting 2	$t_\nu(0, \Sigma)$
6	Same as Setting 3	$t_\nu(0, \Sigma)$

Table 7: Coefficients for Selection Task 1 and 2

Response Dimension	c_k (Task 1)	c_k (Task 2)	r_k (Task 2)
2	1	2	1.5
5	0.2	2	2.6
10	-0.2	2	4.1
30	-0.6	2	7.5

more challenging to interpret. Settings within the same column share the same conditional expectation $\mathbb{E}(\boldsymbol{y} \mid \boldsymbol{x})$, while those in the lower row exhibit greater noise and more dispersed scatter patterns due to the multivariate t-distributed noise $\boldsymbol{\epsilon}$. The target regions for the two tasks, defined based on the coefficients in Table 7, are highlighted as red and yellow shaded areas in the plots.

C.2. Extra Simulated Experiments

C.2.1. Comparing Two Distance-Based Scores

In Tables 8 and 9, we compare the performance of the two distance-based scores (7) and (8) in `mCS-dist`. All other configurations are kept the same as in Section 5.

Table 8: Observed FDR of `mCS-dist` with score (7) and (8) for Task 1 and 2.

	Task 1		Task 2	
Setting	Score (7)	Clipped Score (8)	Score (7)	Clipped Score (8)
1	0.154	0.277	0.102	0.265
2	0.269	0.314	0.127	0.263
3	0.241	0.265	0.150	0.223
4	0.202	0.278	0.226	0.286
5	0.282	0.308	0.240	0.292
6	0.265	0.287	0.220	0.283

Across a wide range of tasks and experimental settings, the clipped score (8) consistently demonstrates superior performance.

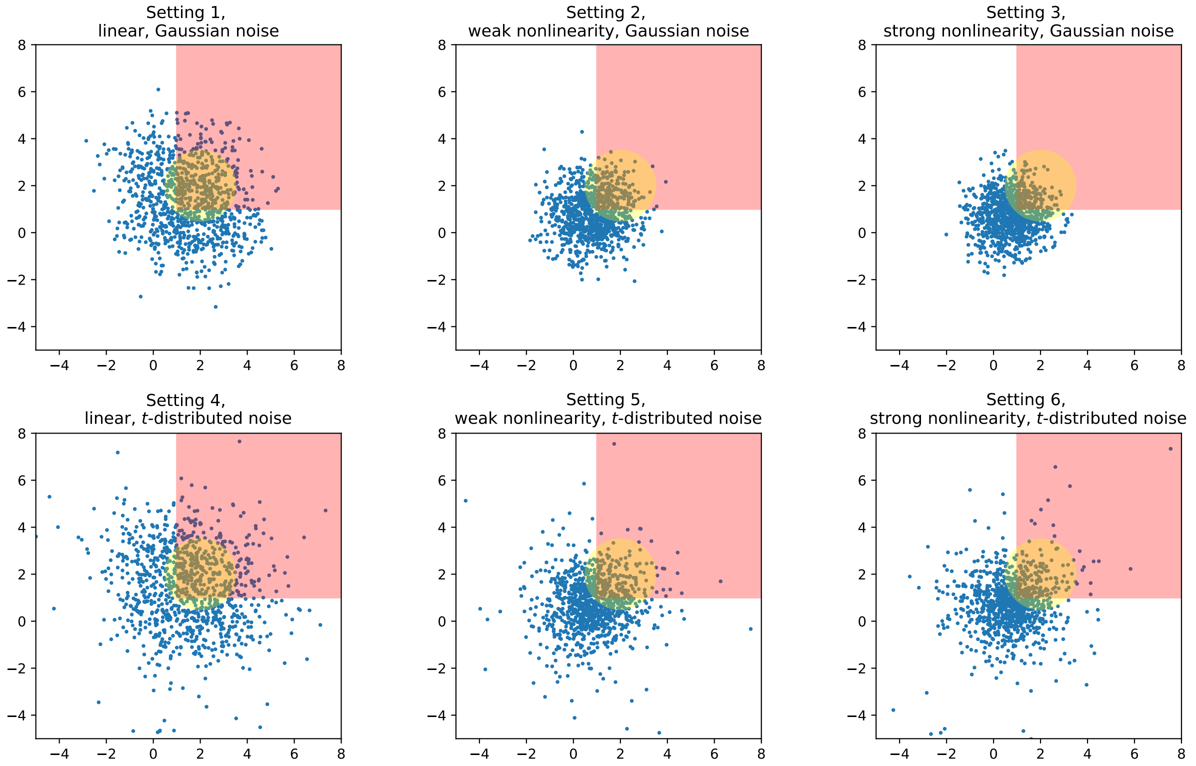

Figure 2: Scatter plots of 1,000 i.i.d. samples from our data generating processes. Each point represents a sample, with the x- and y-axes corresponding to its responses entries $\boldsymbol{y}_1$ and $\boldsymbol{y}_2$, respectively. The red and the yellow shaded areas represents the two target regions for the two selection tasks.

Table 9: Observed power of mCS-dist with score (7) and (8) for Task 1 and 2.

	Task 1		Task 2	
Setting	Score (7)	Clipped Score (8)	Score (7)	Clipped Score (8)
1	0.305	0.555	0.440	0.760
2	0.069	0.104	0.208	0.405
3	0.046	0.068	0.075	0.134
4	0.191	0.324	0.222	0.333
5	0.052	0.060	0.093	0.170
6	0.038	0.046	0.048	0.063

Although both scoring methods guarantee valid FDR control, the clipped score tends to exhibit a higher observed FDR. This finding is consistent with the asymptotic analysis in Appendix B.2, which suggests that the clipped score can leverage a larger portion of the available FDR budget.

C.2.2. Varying Response Dimension

The main article included results with response dimension $d = 30$. In this section, we access the performance of `mCS-dist` and `mCS-learn` in lower dimensional response settings $d \in \{2, 5, 10\}$, examining whether `mCS-learn` continues to outperform competing methods for smaller values of d. To simplify the analysis, we focus on Setting 3 for these experiments. All other configurations are kept unchanged.

Tables 10 and 11 show that even for smaller d, `mCS-learn` and `mCS-dist` continue to outperform the other baseline methods, achieving the best and the second-best power respectively (while controlling FDR). Notably, the performance gap of `mCS-learn` over competing methods becomes more pronounced as the response dimension increases, suggesting that

Table 10: Observed FDR of different methods for lower response dimensions.

	Task 1						Task 2		
d	CS_int	CS_ib	CS_is	bi	mCS-d	mCS-l	bi	mCS-d	mCS-l
2	0.244	0.000	0.142	0.303	0.268	0.281	0.270	0.299	0.302
5	0.706	0.000	0.255	0.290	0.324	0.294	0.246	0.290	0.302
10	0.735	0.000	0.342	0.311	0.276	0.310	0.257	0.265	0.311
30	0.724	0.000	0.204	0.295	0.264	0.278	0.216	0.223	0.263

Table 11: Observed power of different methods for lower response dimensions.

	Task 1						Task 2		
d	CS_int	CS_ib	CS_is	bi	mCS-d	mCS-l	bi	mCS-d	mCS-l
2	0.029	0.000	0.022	0.032	0.047	0.063	0.050	0.068	0.069
5	1.000	0.000	0.067	0.041	0.049	0.083	0.050	0.057	0.072
10	1.000	0.000	0.055	0.044	0.058	0.069	0.062	0.069	0.106
30	1.000	0.000	0.039	0.059	0.068	0.102	0.115	0.134	0.179

mCS-learn is particularly advantageous for high-dimensional settings.

C.2.3. Nonconvex Target Region

To examine the suitability of our methods for tasks involving more irregular target regions, we conduct experiments on two additional tasks in which the target region R is nonconvex:

3. The complement of the (shifted) orthant, $R = \{\boldsymbol{y} : y_k \geq c_k \ \ \forall k\}^c = \{\boldsymbol{y} : y_k < c_k \ \text{ for some } k\}$,
4. The complement of a sphere centered at $\boldsymbol{c}$, $R = \{\boldsymbol{y} : \|\boldsymbol{y} - \boldsymbol{c}\|_2 > r\}$.

To ensure that a reasonable proportion (15%-35%) of responses $\boldsymbol{y}$ fall within the selection region, the coefficients c_k and r are defined differently from their counterparts in Task 1 and 2. Table 12 presents the specific values, where each scalar, following the convention in Table 7, represents a vector whose entries are all equal to that scalar. All other setups, configurations, and model hyperparameters are the same as in Section 5.2.

Table 12: Coefficients for Selection Task 3 and 4

Response Dimension	c_k (Task 3)	c_k (Task 3)	r_k (Task 4)
2	-0.5	2	3
5	-0.8	2	4
10	1.1	2	5.5
30	1.6	2	9.5

Table 13 and Table 14 summarize the observed power and FDR for the two additional tasks respectively, where the dimension is $d = 30$ and nominal FDR level is $q = 0.3$. For the nonconvex tasks, mCS-dist demonstrates inferior performance compared to the baseline bi. In contrast, mCS-learn performs comparably to bi in relatively easy settings and surpasses bi in more challenging scenarios, such as Settings 3 and 6. Consequently, mCS-learn emerges as the preferred choice when R is nonconvex or otherwise irregular.

C.2.4. Other Pretrained Models $\hat{\mu}$

We also test the performance of various methods when the underlying model $\hat{\mu}$ is less predictive. To simulate this effect, we train $\hat{\mu}$ as a linear model instead of support vector machines as in Section 5. All other setups remain unchanged.

Table 13: Observed FDR of different methods for Task 3 and 4.

	Task 3			Task 4		
Setting	bi	mCS-d	mCS-l	bi	mCS-d	mCS-l
1	0.321	0.267	0.280	0.327	0.298	0.304
2	0.312	0.228	0.332	0.381	0.271	0.286
3	0.334	0.257	0.276	0.385	0.276	0.275
4	0.314	0.328	0.277	0.298	0.298	0.284
5	0.273	0.266	0.279	0.305	0.288	0.313
6	0.237	0.255	0.303	0.335	0.263	0.265

Table 14: Observed power of different methods for Task 3 and 4.

	Task 3			Task 4		
Setting	bi	mCS-d	mCS-l	bi	mCS-d	mCS-l
1	0.305	0.052	0.255	0.763	0.311	0.555
2	0.008	0.011	0.017	0.098	0.015	0.089
3	0.018	0.005	0.017	0.043	0.012	0.044
4	0.307	0.050	0.241	0.589	0.233	0.446
5	0.024	0.001	0.034	0.114	0.017	0.117
6	0.023	0.001	0.040	0.036	0.016	0.062

Table 15 and Table 16 summarize the FDR and power of different methods when $\hat{\mu}$ is a fitted linear model. Although overall performance declines for most procedures, mCS-dist and mCS-learn remain comparatively stronger than the baselines. In particular, mCS-learn appears less sensitive to the imperfect linear fit, thanks to its data-adaptive nonconformity score V^{θ}, which is learned based on the model $\hat{\mu}$ and the observed data.

Table 15: Observed FDR of different methods for lower response dimensions.

	Task 1						Task 2		
Setting	CS_int	CS_ib	CS_is	bi	mCS-d	mCS-l	bi	mCS-d	mCS-l
1	0.776	0.101	0.266	0.274	0.264	0.267	0.257	0.273	0.276
2	0.798	0.000	0.230	0.269	0.299	0.288	0.245	0.222	0.275
3	0.742	0.000	0.210	0.312	0.235	0.206	0.232	0.223	0.287
4	0.803	0.000	0.267	0.318	0.290	0.282	0.307	0.286	0.260
5	0.813	0.000	0.174	0.306	0.279	0.272	0.273	0.298	0.285
6	0.779	0.000	0.268	0.259	0.306	0.264	0.292	0.267	0.279

Table 16: Observed power of different methods for lower response dimensions.

	Task 1						Task 2		
Setting	CS_int	CS_ib	CS_is	bi	mCS-d	mCS-l	bi	mCS-d	mCS-l
1	1.000	0.204	0.467	0.195	0.550	0.379	0.063	0.842	0.583
2	1.000	0.000	0.024	0.077	0.071	0.089	0.386	0.337	0.419
3	1.000	0.000	0.026	0.066	0.074	0.079	0.129	0.138	0.148
4	1.000	0.000	0.203	0.120	0.309	0.186	0.040	0.408	0.188
5	1.000	0.000	0.016	0.036	0.048	0.051	0.176	0.135	0.162
6	1.000	0.000	0.028	0.042	0.039	0.048	0.056	0.056	0.073

C.2.5. Comparison of Loss Functions in mCS-learn

Here we compare the performance of mCS-learn with loss L_1 (15) and loss L_2 (16). The response dimension is $d = 30$, and the nominal FDR level is $q = 0.3$.

Table 17: Observed FDR of mCS-learn with loss L_1 and L_2.

	Task 1		Task 2	
Setting	L_1	L_2	L_1	L_2
1	0.255	0.251	0.276	0.279
2	0.279	0.267	0.278	0.273
3	0.299	0.278	0.331	0.263
4	0.274	0.315	0.280	0.254
5	0.210	0.239	0.278	0.273
6	0.296	0.258	0.361	0.207

Table 18: Observed power of mCS-learn with loss L_1 and L_2.

	Task 1		Task 2	
Setting	L_1	L_2	L_1	L_2
1	0.040	0.325	0.051	0.534
2	0.024	0.109	0.111	0.421
3	0.022	0.102	0.041	0.179
4	0.034	0.199	0.019	0.180
5	0.009	0.042	0.041	0.189
6	0.012	0.034	0.032	0.061

As noted in the main article, employing the L_2 loss obviates the need for an additional round of smooth ranking, thereby enhancing both numerical stability and training efficacy. This explains the superior power observed in Table 18 when mCS-learn utilizes the L_2 loss.

C.2.6. Comparison of Score Families in mCS-learn

When learning f_θ in mCS-learn, various forms of data may be utilized as inputs, giving rise to different score families. The simplest approach draws exclusively on the covariates $\boldsymbol{x}$ as features, making it applicable even in scenarios where the model $\hat{\mu}$ is not available. Alternatively, incorporating the response $\boldsymbol{y}$ expands the family to (11). When $\hat{\mu}$ is available, its predictions $\hat{\mu}(\boldsymbol{x})$ can also be included as input. Although this does not increase the overall expressiveness of the family, it can accelerate the training process. In this section, we evaluate the performance of the following score families with varying complexity:

1. Covariate only: $M \cdot \mathbb{1}\{\boldsymbol{y} \notin R^c \cup \partial R\} - f_\theta(\boldsymbol{x}; R)$.
2. Prediction only: $M \cdot \mathbb{1}\{\boldsymbol{y} \notin R^c \cup \partial R\} - f_\theta(\hat{\mu}(\boldsymbol{x}); R)$.
3. Covariate and Prediction: $M \cdot \mathbb{1}\{\boldsymbol{y} \notin R^c \cup \partial R\} - f_\theta(\boldsymbol{x}, \hat{\mu}(\boldsymbol{x}); R)$.
4. Full Family (12) : $M \cdot \mathbb{1}\{\boldsymbol{y} \notin R^c \cup \partial R\} - f_\theta(\boldsymbol{x}, \boldsymbol{y}; R)$.
5. All available information: $M \cdot \mathbb{1}\{\boldsymbol{y} \notin R^c \cup \partial R\} - f_\theta(\boldsymbol{x}, \hat{\mu}(\boldsymbol{x}), \boldsymbol{y}; R)$.

As discussed in the main article, families 4 and 5 incorporate $\boldsymbol{y}$ as an input, which compromises the regional monotonicity of the score unless M is sufficiently large. Table 19 summarizes the FDR and power of mCS-learn with different families, with Setting 3. For both Task 1 and Task 2, families 4 and 5 violated FDR control when $M = 10^3$. Among the three subfamilies that maintained valid FDR control, family 3 exhibited the best performance.

Table 19: Observed FDR and power of mCS-learn with different score families.

	FDR		Power	
Family	Task 1	Task 2	Task 1	Task 2
1	0.298	0.237	0.108	0.165
2	0.260	0.291	0.096	0.175
3	0.278	0.263	0.102	0.179
4	0.711	0.594	0.972	0.807
5	0.667	0.494	0.782	0.536

D. Additional Details for Real Data Application (Section 6)

D.1. Overview of Drug Discovery Data and Configuration of the Selection Tasks

The drug discovery dataset we used in Section 6 is compiled from various public sources (Wenzel et al., 2019; Iwata et al., 2022; Kim et al., 2023; Watanabe et al., 2018; Falcón-Cano et al., 2022; Esposito et al., 2020; Braga et al., 2015; Aliagas et al., 2022; Perryman et al., 2020; Meng et al., 2022; Vermeire et al., 2022). Because the integrated data contained missing values, we employed Chemprop (Yang et al., 2019; Heid et al., 2023) to impute these entries. The resulting imputed dataset was then used in all subsequent experiments. The processed dataset contains $n = 22805$ data points, and can be found at https://github.com/Tian-Bai/mcs.

We list the names, units and cutoffs of the responses (only relevant to the first selection task) in the imputed dataset in Table 20, and provide detailed descriptions of their biological significance and drug discovery relevance in Figure 4. Recall that in the first task we consider target regions of the shape $R = \{\boldsymbol{y} : y_k \geq c_k \ \ \forall k\}$, where the cutoffs are the values c_k defining the selection problem. Figure 3 shows the distribution of these 15 responses, with vertical red lines indicating their corresponding cutoffs. Approximately 21% of the test dataset compounds exceed all 15 thresholds, thereby qualifying for selection.

For the second task, the target region is defined as a sphere $\{\boldsymbol{y} : \|\boldsymbol{y} - \boldsymbol{c}\|_2 \leq r\}$. For convenience, we take the center of the sphere the same as the cutoffs c_k in task 1, and let $r = 2.4$. Under this definition, approximately 24% of the compounds qualify for selection. Similarly, for the third task where the target region is the complement of a sphere $\{\boldsymbol{y} : \|\boldsymbol{y} - \boldsymbol{c}\|_2 \leq r'\}$, we adopt the same center $\boldsymbol{c}$ and set $r' = 3.4$. This choice ensures that 18% of the compounds qualify for selection.

Table 20: List of responses in the drug discovery dataset.

Name	Unit	Cutoff c_k
CL_microsome_human	log 10 (mL/min/kg)	4
CL_microsome_mouse	log 10 (mL/min/kg)	4
CL_microsome_rat	log 10 (mL/min/kg)	4
CL_total_dog	log 10 (mL/min/kg)	0.5
CL_total_human	log 10 (mL/min/kg)	0
CL_total_monkey	log 10 (mL/min/kg)	0.5
CL_total_rat	log 10 (mL/min/kg)	1
CYP2C8_inhibition	log 10 (nMolar IC50)	3.5
CYP2C9_inhibition	log 10 (nMolar IC50)	3.5
CYP2D6_inhibition	log 10 (nMolar IC50)	3.5
CYP3A4_inhibition	log 10 (nMolar IC50)	3.5
Papp_Caco2	log 10 (10^{-6}cm/s)	0.8
Pgp_human	log 10 (efflux ratio)	-0.2
hERG_binding	log 10 (nMolar IC50)	3.5
LogD_pH_7.4	log 10 (M/M)	2

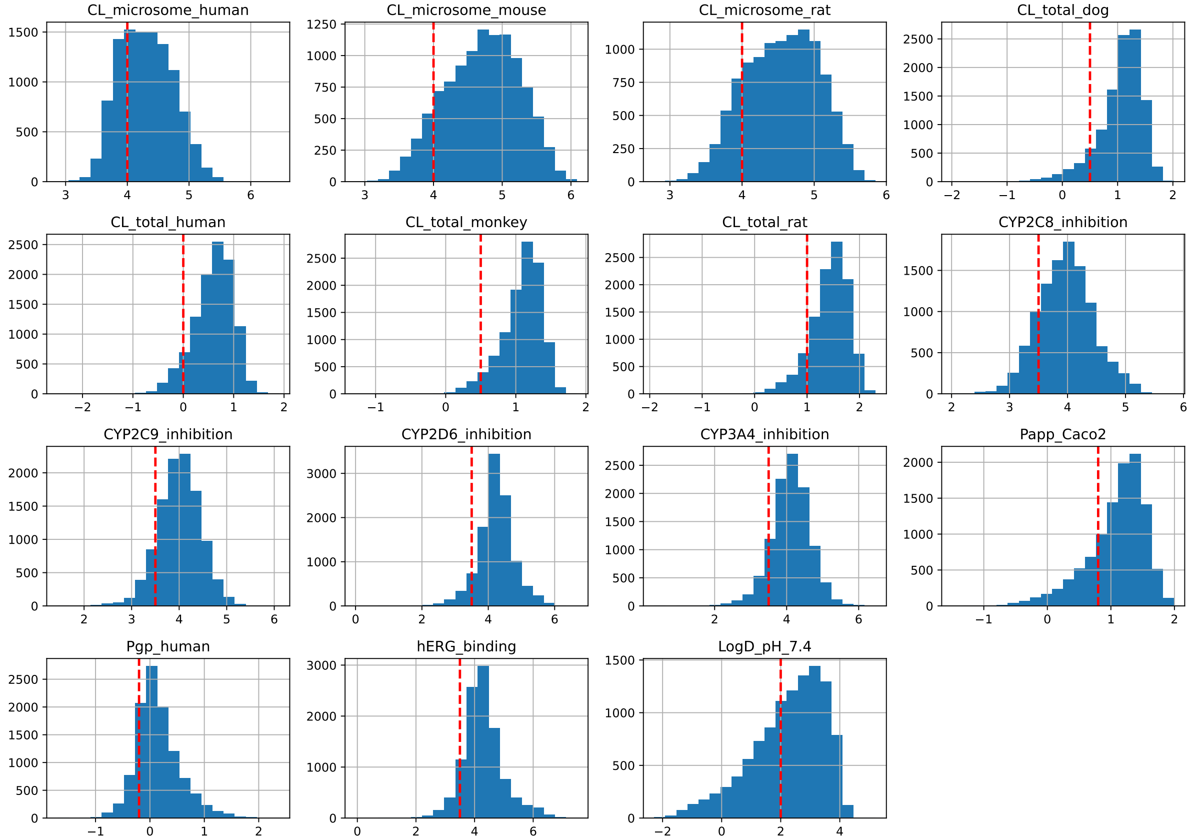

Figure 3: Histograms of 15 responses in the drug dataset. The vertical red lines denote the corresponding cutoffs (for the first task) for each response.

Figure 4: Description and drug discovery relevance of the responses.

CL_microsome_human:

Intrinsic metabolic clearance in human liver microsomes. (Help predict rate of liver metabolism in human through the in vitro in vivo correlation.)

CL_microsome_mouse:

Intrinsic metabolic clearance in mouse liver microsomes. (Help predict rate of liver metabolism in human through the in vitro in vivo correlation.)

CL_microsome_rat:

Intrinsic metabolic clearance in rat liver microsomes. (Help predict rate of liver metabolism in human through the in vitro in vivo correlation.)

CL_total_dog:

Total body clearance measured in vivo in dogs. (Drug exposure in this species, crucial for determine human dose regimens through translational modeling.)

CL_total_human:

Total body clearance measured in vivo in humans (Drug exposure in this species, crucial for determine human dose regimens through translational modeling.)

CL_total_monkey:

Total body clearance measured in vivo in monkeys. (Drug exposure in this species, crucial for determine human dose regimens through translational modeling.)

CL_total_rat:

Total body clearance measured in vivo in rats. (Drug exposure in this species, crucial for determine human dose regimens through translational modeling.)

CYP2C8_inhibition:

Inhibition potential against human CYP2C8 enzyme. (Assessment of risk of drug-drug interactions (DDIs) involving drugs metabolized by CYP2C8. Important for safety assessment.)

CYP2C9_inhibition:

Inhibition potential against human CYP2C9 enzyme. (Assessment of drug-drug interactions (DDIs) involving drugs metabolized by CYP2C9 (e.g., warfarin). Important for safety assessment.)

CYP2D6_inhibition:

Inhibition potential against human CYP2D6 enzyme. (Assessment of drug-drug interactions (DDIs) involving drugs metabolized by CYP2D6 (e.g., some antidepressants, beta-blockers). Important for safety assessment.)

CYP3A4_inhibition:

Inhibition potential against human CYP3A4 enzyme. (Assessment of drug-drug interactions (DDIs) involving drugs metabolized by CYP3A4 (a very common pathway). Important for safety assessment.)

Papp_Caco2:

Apparent permeability across Caco-2 cell monolayer. (Potential for oral absorption by crossing the intestinal wall. High Papp suggests good absorption is likely.)

Pgp_human:

Interaction with human P-glycoprotein (Pgp) efflux transporter. (Assess if the drug is pumped out of cells by Pgp, potentially limiting oral absorption and brain penetration.)

hERG_binding:

Binding/inhibition of the hERG potassium channel. (Assessment of risk of cardiac toxicity (QT prolongation, arrhythmias). A critical safety screen; high binding is a major safety concern.)

LogD_pH_7.4:

Logarithm of the distribution coefficient at pH 7.4. (Predicts drug's lipophilicity (fat vs. water solubility) under physiological conditions. Influences absorption, distribution, membrane crossing, and CNS penetration.)

D.2. Extra Experiments on Real Data

D.2.1. Numerical Stability of Different Methods

In this section, we test the numerical stability of `mCS-dist` and `mCS-learn` on real data. To do this, unlike previous experiments where we sample new data or randomly partition data for each iteration, we fix the pretrained model $\hat{\mu}$, the calibration data $\mathcal{D}_{\text{cal}}$ and the test data $\mathcal{D}_{\text{test}}$. This is to avoid the variability from data sampling or pretraining, and only consider the inherent variance of the methods. Due to the high computation cost of retraining f_θ, we also keep f_θ the same across different iterations. We keep all other setups and configurations unchanged as in Section 6.

Table 21: Observed standard error of FDRs for different methods with real data.

Task	q	CS_int	CS_ib	CS_is	bi	mCS-d	mCS-l
1	0.3	0.000	0.000	0.248	0.000	0.000	0.000
2	0.3	–	–	–	0.000	0.000	0.002
3	0.3	–	–	–	0.000	0.021	0.004
1	0.5	0.000	0.000	0.032	0.000	0.001	0.004
2	0.5	–	–	–	0.000	0.000	0.002
3	0.5	–	–	–	0.000	0.028	0.005

Table 22: Observed standard error of powers for different methods with real data.

Task	q	CS_int	CS_ib	CS_is	bi	mCS-d	mCS-l
1	0.3	0.000	0.000	0.022	0.000	0.000	0.000
2	0.3	–	–	–	0.000	0.000	0.001
3	0.3	–	–	–	0.000	0.005	0.002
1	0.5	0.000	0.000	0.056	0.000	0.000	0.004
2	0.5	–	–	–	0.000	0.000	0.004
3	0.5	–	–	–	0.000	0.011	0.005

For all methods, the standard errors for both FDR and power remain low, with the exception of `CS_is` at a nominal level of $q = 0.3$. This phenomenon likely arises from the unstable nature of the subroutine nominal level search at lower user-specified nominal levels. The failure of `CS_is` to control the FDR at this low nominal level (Figure 1) further corroborates this interpretation.