APPENDIX

A BENJAMINI-HOCHBERG PROCEDURE

The Benjamini–Hochberg (BH) procedure (Benjamini & Hochberg, 1995; Benjamini & Yekutieli, 2001) is a statistical method designed to control the false discovery rate (FDR) when performing multiple hypothesis tests. Unlike conservative methods such as the Bonferroni correction, which aim to control the probability of making any false discovery (i.e., family-wise error rate), the BH procedure allows for a controlled proportion of false positives among all rejected hypotheses, offering greater power in settings with multiple simultaneous tests.

The BH procedure operates as follows:

1. Perform m hypothesis tests, resulting in m p-values. Sort these p-values in ascending order: $p_{(1)} \leq p_{(2)} \leq \cdots \leq p_{(m)}$.
2. Choose a desired FDR level q (e.g., $q = 0.10$).
3. For each sorted p-value $p_{(i)}$, calculate the threshold $\frac{i}{m} \cdot q$.
4. Identify the largest i such that $p_{(i)} \leq \frac{i}{m} \cdot q$.
5. Reject all null hypotheses corresponding to $p_{(1)}, p_{(2)}, \ldots, p_{(i)}$.

This adaptive approach balances discovery with type I error control. It is particularly useful in fields like genomics, neuroimaging, and other high-dimensional data analyses, where many tests are conducted simultaneously.

B EXPERIMENTAL SETUP FOR FIGURE 1

For data generation, we adopt the following simulation setting:

$$\mu(\boldsymbol{x}) = 4x_1 \cdot \mathbb{1}(x_2 > 0) \cdot \max(0.5, x_3) + 4x_1 \cdot \mathbb{1}(x_2 \leq 0) \cdot \min(-0.5, x_3).$$

The noise is modeled as $\epsilon_i \sim N(0, \sigma^2)$ with homogeneous variance, where $\sigma = 0.3$. The training and calibration set sizes are $|\mathcal{D}_{\text{train}}| = |\mathcal{D}_{\text{calib}}| = 1000$, and the test set size is $|\mathcal{D}_{\text{test}}| = 500$. We use gradient boosting to fit the regressor μ, implemented using the scikit-learn Python library. The experiment evaluates the total number of reject-to-accept samples across 500 sequentially generated data points, with results averaged over 30 independent runs. In essence, the experiment tracks how often previously selected data points are later deselected by the employed method.

C OMITTED PROOFS

C.1 PROOF OF THEOREM 5

Theorem C.1 (Restatement of Theorem 5)**.** *The OCS-ARC is an ARC procedure. In particular, the rejection sets $\{\mathcal{R}\}_{t=1}^{T}$ outputted by Algorithm 1 satisfy*

$$\mathcal{R}_1 \subseteq \mathcal{R}_2 \subseteq \cdots \subseteq \mathcal{R}_T$$

Proof. We first show that $k_t^* \leq k_{t+1}^*$ for all $t \in \mathbb{N}^+$. Recall that k_t^* is defined as

$$k_t^* = \max\left\{ k \in [t] : \sum_{j=1}^{t} \mathbb{1}\{p_j \leq kq\gamma_j\} \geq k \right\}.$$

It follows that

$$\sum_{j=1}^{t+1} \mathbb{1}\{p_j \leq k_t^* q\gamma_j\} = \sum_{j=1}^{t} \mathbb{1}\{p_j \leq k_t^* q\gamma_j\} + \mathbb{1}\{p_{t+1} \leq k_t^* q\gamma_{t+1}\} \geq k_t^* + \mathbb{1}\{p_{t+1} \leq k_t^* q\gamma_{t+1}\} \geq k_t^*$$

This implies that

$$k_t^* \in \left\{ k \in [t+1] : \sum_{j=1}^{t+1} \mathbb{1}\{p_j \leq kq\gamma_j\} \geq k \right\}$$

Recall that k_{t+1}^* is defined as

$$k_{t+1}^* = \max \left\{ k \in [t+1] : \sum_{j=1}^{t+1} \mathbb{1}\{p_j \leq kq\gamma_j\} \geq k \right\}.$$

Thus, we can find that $k_t^* \leq k_{t+1}^*$. Next, we show that once a conformal p-value is selected, it remains selected in all future rounds. Suppose that $j \in \mathcal{R}_t$, i.e., $p_j \leq k_t^* q\gamma_j$. Then, for any $t^{'} \geq t$, since $k_t^* \leq k_{t'}^*$ we have

$$p_j \leq k_t^* q\gamma_j \quad \Rightarrow \quad p_j \leq k_{t'}^* q\gamma_j,$$

which indicates that $j \in \mathcal{R}_{t'}$. Therefore, once a sample is selected, it remains in the selection set at all future times. Therefore, we conclude that OCS-ARC is an ARC procedure. □

C.2 Proof of Theorem 6

Theorem C.2 (Restatement of Theorem 6)**.** *Consider a monotone non-conformity score function V. Assume that the calibration data $\{(\boldsymbol{X}_i, Y_i)\}_{i=1}^n$ and the test data $\{(\boldsymbol{X}_{n+t}, Y_{n+t})\}_{t=1}^T$ are independently and identically distributed. Then, for any nominal level $q \in (0,1)$ and timestep t, the selection set $\mathcal{R}_t$ constructed by Algorithm 1 satisfies* $\mathrm{FDR}_t \leq q$.

Proof. Recall that in Algorithm 1, a data sample is selected, i.e., $j \in \mathcal{R}_t$ if and only if $p_j \leq k_t^* q\gamma_j$. Then, the FDR can be decomposed as

$$\begin{aligned}
\mathrm{FDR}_t = \mathbb{E}\left[\frac{|\mathcal{R}_t \cap \mathcal{H}_t^0|}{|\mathcal{R}_t|}\right] &= \mathbb{E}\left[\frac{1}{k_t^*}\sum_{j=1}^{t} \mathbb{1}\{j \in \mathcal{R}_t, Y_{n+j} \leq c_j\}\right] \\
&= \sum_{j=1}^{t} \mathbb{E}\left[\frac{1}{k_t^*}\mathbb{1}\{p_j \leq k_t^* q\gamma_j, Y_{n+j} \leq c_j\}\right]
\end{aligned}$$

Let $\mathcal{R}_{j\to *}$ be the rejection set obtained by setting p_j to p_j^* while keeping others fixed. By Lemma 5 of (Jin & Candès, 2023), we have $p_j \geq p_j^*$ on the event $\{Y_{n+t} \leq c_t\}$. Thus, $p_j \leq k_t^* q\gamma_j$ implies that $p_j^* \leq k_t^* q\gamma_j$. In other words, if $j \in \mathcal{R}_t$, then setting p_j to p_j^* does not change the rejection set. It follows that

$$\begin{aligned}
\frac{1}{k_t^*}\mathbb{1}\{p_j \leq k_j^* q\gamma_j, Y_{n+j} \leq c_j\} &= \frac{1}{|\mathcal{R}_{j\to *}|}\mathbb{1}\{p_j \leq k_j^* q\gamma_j, Y_{n+j} \leq c_j\} \\
&\leq \frac{1}{|\mathcal{R}_{j\to *}|}\mathbb{1}\{p_j^* \leq k_j^* q\gamma_j, Y_{n+j} \leq c_j\},
\end{aligned}$$

which implies that

$$\mathrm{FDR}_t \leq \sum_{j=1}^{t} \mathbb{E}\left[\frac{1}{|\mathcal{R}_{j\to *}|}\mathbb{1}\{p_j^* \leq k_t^* q\gamma_j, Y_{n+j} \leq c_j\}\right] \leq \sum_{j=1}^{t} \mathbb{E}\left[\frac{1}{|\mathcal{R}_{j\to *}|}\mathbb{1}\{p_j^* \leq k_t^* q\gamma_j\}\right]$$

By Lemma 5 of Jin & Candès (2023), we have $(p_1, \cdots, p_{j-1}, p_j^*, p_{j+1}, \cdots, p_t)$ is PRDS (Bates et al., 2023) on p_j^*. Then, Proposition 3.6 of Blanchard & Roquain (2008) gives that

$$\mathbb{E}\left[\frac{1}{|\mathcal{R}_{j\to *}|}\mathbb{1}\{p_j^* \leq k_t^* q\gamma_j\}\right] \leq \frac{k_t^* q\gamma_j}{|\mathcal{R}_{j\to *}|} = q\gamma_j$$

Therefore, we can conclude that $\mathrm{FDR}_t \leq \sum_{j=1}^{t} \gamma_j q \leq q$. □

C.3 PROOF OF THEOREM 7

Theorem C.3 (Restatement of Theorem 7)**.** *Consider a monotone non-conformity score function V. Assume that $\{V_1, \cdots, V_n, V_{n+t}\}$ are exchangeable conditional on $\{\hat{V}_{n+t'}\}_{t'=1,t'\neq t}^{T}$, and $\{V_{t'}\}_{t'=1}^{n+T}$ have no ties almost surely. Then, for any nominal level $q \in (0,1)$ and timestep t, the selection set $\mathcal{R}_t$ constructed by Algorithm 1 satisfies $\mathrm{FDR}_t \leq q$.*

Proof. Since we assume $\{V_i\}_{i=1}^{n+T}$ have no ties almost surely, the oracle and practical conformal p-value can be rewritten as

$$p_t^* = \frac{1}{n+1}\left[1 + \sum_{i=1}^{n} \mathbb{1}\{V_i \leq V_{n+t}\}\right], \quad p_t = \frac{1}{n+1}\left[1 + \sum_{i=1}^{n} \mathbb{1}\{V_i \leq \hat{V}_{n+t}\}\right]$$

For $t = 1, \cdots, T$, we define modified conformal p-values:

$$p_t^{(j)} = \frac{1}{n+1}\left[\sum_{i=1}^{n} \mathbb{1}\{V_i \leq \hat{V}_{n+t}\} + \mathbb{1}\{V_{n+j} \leq \hat{V}_{n+t}\}\right]$$

Also define $\mathcal{R}_t = \mathcal{R}(p_1, \cdots, p_t) \subseteq \{1, \cdots, t\}$ as the selection set, by taking p-values $p_1, \cdots, p_t$. In the sequel, we will compare $\mathcal{R}(p_1, \cdots, p_t)$ with $\mathcal{R}(p_1^{(j)}, \cdots, p_{j-1}^{(j)}, p_j^*, p_{j+1}^{(j)}, p_t)$, on the event $\{j \in \mathcal{R}_t, Y_{n+j} \leq c_j\}$. First, on this event, since V is monotone, we have $p_j^* \leq p_j$. For the remaining p-values $\{p_l^{(j)}\}_{l=1,l\neq j}^{t}$, since we assume the scores have no ties, we consider two cases:

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

$$p_j^{(j)} = \frac{1}{n+1}\left[\sum_{i=1}^{n} \mathbb{I}\{V_i \leq \hat{V}_{n+l}\} + \mathbb{I}\{V_{n+j} \leq \hat{V}_{n+l}\}\right] = \frac{1}{n+1}\left[\sum_{i=1}^{n} \mathbb{I}\{V_i \leq \hat{V}_{n+l}\} + 1\right] = p_l.$$

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

$$p_j^{(j)} \leq \frac{1}{n+l}\left[\sum_{i=1}^{n} \mathbb{I}\{V_i \leq \hat{V}_{n+l}\} + 1\right] \leq \frac{1}{n+l}\left[\sum_{i=1}^{n} \mathbb{I}\{V_i \leq \hat{V}_{n+j}\} + 1\right] = p_j.$$

Since $j \in \mathcal{R}_t$, by the construction of the selection set in Algorithm 1, $l \in \mathcal{R}_t$ as p_l has a smaller rank when ordering the p-values.

To summarize, if we replace $(p_1, \cdots, p_t)$ by $(p_1^{(j)}, \cdots, p_j^{(j)}, \cdots, p_t^{(j)})$ on the event $\{j \in \mathcal{R}_t, Y_{n+j} \leq c_j\}$, such a replacement does not modify any of those p-values p_i if they satisfied $p_i \geq p_j$. Also, for all p-values with $p_i \leq p_j$ including i itself ($i = j$), their replaced values $p_i^{(j)}$ 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 $j \in \mathcal{R}_t$ 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

$$\mathcal{R}_t^{(j)} := \mathcal{R}(p_1^{(j)}, \cdots, p_{j-1}^{(j)}, p_j^*, p_{j+1}^{(j)}, \cdots, p_t) = \mathcal{R}(p_1, \cdots, p_t) = \mathcal{R}_t$$

on the event $\{j \in \mathcal{R}_t, Y_{n+j} \leq c_j\}$. By decomposing the FDR, we have

$$\mathrm{FDR}_t = \mathbb{E}\left[\frac{1}{k_t^*}\sum_{l=1}^{t} \mathbb{1}\{Y_{n+l} \leq c_l, l \in \mathcal{R}_t\}\right] \leq \sum_{l=1}^{t} \mathbb{E}\left[\frac{1}{k_t^*}\mathbb{1}\{l \in \mathcal{R}_t\}\right] = \sum_{l=1}^{t} \mathbb{E}\left[\frac{1}{k_t^*}\mathbb{1}\{l \in \mathcal{R}_t^{(j)}\}\right]$$

Also, on the event $\{j \in \mathcal{R}_t^{(j)}\}$, sending p_j^* to zero does not change the rejection set. We have

$$\mathcal{R}_t^{j\to 0} := \mathcal{R}(p_1^{(j)}, \cdots, p_{j-1}^{(j)}, 0, p_{j+1}^{(j)}, \cdots, p_t) = \mathcal{R}(p_1^{(j)}, \cdots, p_{j-1}^{(j)}, p_j^*, p_{j+1}^{(j)}, \cdots, p_t) = \mathcal{R}_t^{(j)}$$

Thus,

$$\mathrm{FDR}_t \le \sum_{l=1}^{t} \mathbb{E}\left[\frac{1}{k_t^*}\mathbb{1}\{l \in \mathcal{R}_t^{(j)}\}\right] = \sum_{l=1}^{t} \mathbb{E}\left[\frac{1}{k_t^*}\mathbb{1}\{l \in \mathcal{R}_t^{j\to 0}\}\right] = \sum_{l=1}^{t} \mathbb{E}\left[\frac{\mathbb{1}\{p_l \le k_t^* q\gamma_l\}}{k_t^*}\right]$$
$$\le \sum_{j=1}^{t} \mathbb{E}\left[\frac{\mathbb{1}\{p_l^* \le k_t^* q\gamma_l\}}{k_t^*}\right]$$

By definition, $\{p_l^{(j)}\}_{l=1,l\neq j}^{t}$ 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 $\{\hat{V}_{n+t'}\}_{t'=1,t'\neq t}^{T}$, the distribution of p_j^j is independent from the ordering of $\{V_i\}_{i=1}^{n+1} \cup \{V_{n+j}\}$ conditioned on the (unordered) set $[V_1, \cdots, V_n, V_{n+j}] \cup \{\hat{V}_{n+t'}\}_{t'=1,t'\neq t}^{T}$. Also, conditioned on $\{\hat{V}_{n+t'}\}_{t'=1,t'\neq t}^{T}$, $\mathcal{R}_t^{j\to *}$ only depends on $[V_1, \cdots, V_n, V_{n+j}]$ and p_j^* only depends on the ordering of $\{V_1, \cdots, V_n, V_{n+j}\}$. This implies that $\mathcal{R}_t^{j\to *}$ is independent on p_j^* conditioned on $[\hat{V}_1, \cdots, \hat{V}_n, \hat{V}_{n+j}]$ and $\{\hat{V}_{n+t'}\}_{t'\neq j}$. Therefore, by the conservative property of conformal p-values and conditional independence,

$$\mathbb{P}\left\{p_j^{(j)} \le q k_t^* \gamma_j \middle| [V_1, \ldots, V_n, V_{n+j}] \cup \{\hat{V}_{n+l}\}_{l=1,l\neq j}^{T}\right\} \le q k_t^* \gamma_j$$

By the law of total expectation,

$$\mathbb{E}\left[\frac{\mathbb{1}\{p_l^* \le k_t^* q\gamma_l\}}{k_t^*}\right] = \mathbb{E}\left[\mathbb{E}\left[\frac{\mathbb{1}\{p_l^* \le k_t^* q\gamma_l\}}{k_t^*}\middle| [V_1, \ldots, V_n, V_{n+j}] \cup \{\hat{V}_{n+t'}\}_{t'=1,t'\neq t}^{T}\right]\right] \le q\gamma_l$$

Since when $k_t^* = 0$, we have $\mathbb{1}\{p_l^* \le q k_t^* \gamma_l\} = 0$. Then, we have $\mathrm{FDR}_t \le \sum_{l=1}^{t} q\gamma_l \le q$. □

C.4 Proof of Theorem 11

Theorem C.4 (Restatement of Theorem 11)**.** *Consider a regional monotone conformity score function V. Assume that the calibration data $\{(X_i, Y_i)\}_{i=1}^{n}$ and the test data $\{(X_{n+t}, Y_{n+t})\}_{t=1}^{T}$ are independently and identically distributed. Then, for any nominal level $q \in (0,1)$ and timestep t, the selection set $\mathcal{R}_t$ constructed by Algorithm 2 satisfies $\mathrm{FDR}_t \le q$.*

Proof. Recall that in Algorithm 1, a data sample is sampled, i.e., $j \in \mathcal{R}_t$ if and only if $p_j \le k_t^* q\gamma_j$. Then, the FDR can be decomposed as

$$\mathrm{FDR}_t = \mathbb{E}\left[\frac{|\mathcal{R}_t \cap \mathcal{H}_t^0|}{|\mathcal{R}_t|}\right] = \mathbb{E}\left[\frac{1}{k_t^*}\sum_{j=1}^{t}\mathbb{1}\{j \in \mathcal{R}_t, \boldsymbol{Y}_{n+j} \in R^c\}\right]$$
$$= \sum_{j=1}^{t}\mathbb{E}\left[\frac{1}{k_t^*}\mathbb{1}\{p_j \le k_t^* q\gamma_j, \boldsymbol{Y}_{n+j} \in R^c\}\right]$$

Let $\mathcal{R}_{j\to *}$ be the rejection set obtained by setting p_j to p_j^* while keeping others fixed. On the event $\{Y_{n+t} \le c_t\}$, the regional monotonicity follows that $V(\boldsymbol{X}_{n+t}, \boldsymbol{Y}_{n+t}) \le V(\boldsymbol{X}_{n+t}, \boldsymbol{r}_t)$, indicating that $p_j \ge p_j^*$. Thus, $p_j \le k_t^* q\gamma_j$ implies that $p_j^* \le k_t^* q\gamma_j$. In other words, if $j \in \mathcal{R}_t$, then setting p_j to p_j^* does not change the rejection set. It follows that

$$\frac{1}{k_t^*}\mathbb{1}\{p_j \le k_j^* q\gamma_j, \boldsymbol{Y}_{n+j} \in R^c\} = \frac{1}{|\mathcal{R}_{j\to *}|}\mathbb{1}\{p_j \le k_j^* q\gamma_j, \boldsymbol{Y}_{n+j} \in R^c\}$$
$$\le \frac{1}{|\mathcal{R}_{j\to *}|}\mathbb{1}\{p_j^* \le k_j^* q\gamma_j, \boldsymbol{Y}_{n+j} \in R^c\},$$

which implies that

$$\mathrm{FDR}_t \le \sum_{j=1}^{t}\mathbb{E}\left[\frac{1}{|\mathcal{R}_{j\to *}|}\mathbb{1}\{p_j^* \le k_t^* q\gamma_j, \boldsymbol{Y}_{n+j} \in R^c\}\right] \le \sum_{j=1}^{t}\mathbb{E}\left[\frac{1}{|\mathcal{R}_{j\to *}|}\mathbb{1}\{p_j^* \le k_t^* q\gamma_j\}\right]$$

By Lemma 5 of Jin & Candès (2023), we have $(p_1, \cdots, p_{j-1}, p_j^*, p_{j+1}, \cdots, p_t)$ is PRDS (Bates et al., 2023) on p_j^*. Then, Proposition 3.6 of Blanchard & Roquain (2008) gives that

$$\mathbb{E}\left[\frac{1}{|\mathcal{R}_{j\to *}|}\mathbb{1}\{p_j^* \le k_t^* q\gamma_j\}\right] \le \frac{k_t^* q\gamma_j}{|\mathcal{R}_{j\to *}|} = q\gamma_j$$

Therefore, we can conclude that $\mathrm{FDR}_t \le \sum_{j=1}^{t} \gamma_j q \le q$. □

C.5 Proof of Theorem E.2

Theorem C.5 (Restatement of Theorem E.2)**.** *The OB is an ARC procedure. In particular, the rejection sets $\{\mathcal{R}\}_{t=1}^T$ outputted by OB satisfy*

$$\mathcal{R}_1 \subseteq \mathcal{R}_2 \subseteq \cdots \subseteq \mathcal{R}_T$$

Proof. It suffices to show that the sample that is selected in previous timesteps will not be deselected later. Suppose that at timestep t, the corresponding conformal p-value p_t is selected, i.e., $p_t \le q\gamma_t$. Then, for any timestep $t^{'} \ge t$, we still have $p_t \le q\gamma_t$ and thus p_t is selected. Therefore, we conclude that OB is an ARC procedure. □

D Multivariate Online Conformal Selection with Accept-to-Reject Changes

D.1 Pseudo-algorithms

In this section, we present the pseudo-algorithms of Multivariate Online Conformal Selection with Accept-to-Reject Changes (dubbed mOCS-ARC).

Algorithm 1 Multivariate Online Conformal Selection with Accept-to-Reject Changes (mOCS-ARC)

Require: Calibration data $\mathcal{D}_{cal} = \{(\boldsymbol{X}_i, \boldsymbol{Y}_i)\}_{i=1}^n$, test data $\mathcal{D}_{test} = \{(\boldsymbol{X}_{n+t}, \boldsymbol{Y}_{n+t})\}_{t=1}^T$, user-defined target region R, FDR nominal level $q \in (0,1)$, monotone conformity score $V : \mathcal{X} \times \mathcal{Y} \to \mathbb{R}$, real numbers $\{\gamma_t\}_{t=1}^T$.

1: Compute $V_i = V(\boldsymbol{X}_i, \boldsymbol{Y}_i)$ for $i = 1, \cdots, n$.
2: **for** $t = 1, \cdots, T$ **do**
3: Compute $\hat{V}_{n+t} = V(\boldsymbol{X}_{n+t}, \boldsymbol{r}_t)$, for any $\boldsymbol{r}_t \in R$.
4: Construct conformal p-value p_t as in Eq. (3).
5: Compute online BH procedure threshold:

$$k_t^* = \max\left\{k \in [t] : \sum_{j=1}^{t} \mathbb{1}\{p_j \le kq\gamma_j\} \ge k\right\}$$

6: Construct selection set: $\mathcal{R}_t = \{j \in [t] : p_j \le k_t^* q\gamma_j\}$
7: **end for**

D.2 Numerical experiments

To rigorously assess the performance of our methods in the context of multivariate regression, we combine two distinct simulation settings to generate multiple response variables based on the input features. The relationships between the response variables and the predictors are defined as follows:

- **Setting 1:** $\mu_1(x) = 4x_1 \cdot \mathbb{1}(x_2 > 0) \cdot \max(0.5, x_3) + 4x_1 \cdot \mathbb{1}(x_2 \le 0) \cdot \min(-0.5, x_3)$
- **Setting 2:**$\mu_2(x) = 5x_1x_2 + e^{x_4 - 1}$.

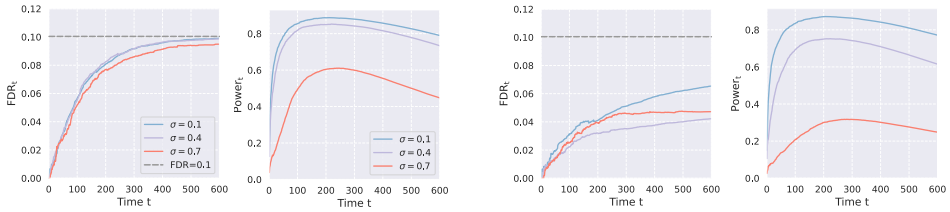
In essence, we integrate the two data-generating processes used in prior experiments to produce multiple responses. Gradient boosting is employed to fit the multivariate regression model across all

configurations. For the nonconformity score, we utilize two distinct multivariate score functions as proposed by Bai et al. (2025), which are formulated as follows:

1. (REGULAR) $D_1(\boldsymbol{z}, R^c) = D_2(\boldsymbol{z}, R^c) = \inf_{\boldsymbol{s}\in R^c} \|\boldsymbol{z} - \boldsymbol{s}\|_p$,
2. (CLIPPED) $D_1(\boldsymbol{z}, R^c) = M \cdot \mathbb{1}\{\boldsymbol{z} \notin R^c \cup \partial R\}, D_2(\boldsymbol{z}, R^c) = \inf_{\boldsymbol{s}\in R^c} \|\boldsymbol{z} - \boldsymbol{s}\|_p$,

where M is a large constant that serves as a relaxation of infinity. These scores generalize the *signed error* score and the *clipped* score (Jin & Candès, 2023), respectively. For brevity, we refer to mOCS-ARC with `CLIPPED` as CLIPPED, and mOCS-ARC with `REGULAR` as REGULAR.

All other settings remain consistent with those in the main synthetic experiment. The results are presented in Figure 1. We empirically evaluate the FDR at different timesteps by averaging the FDP over 300 runs. The FDR remains below 0.1 across all configurations, regardless of the score function and noise level. Similar to the findings in univariate regression, CLIPPED consistently shows the highest realized FDR across all settings while remaining closer to the nominal level, thereby achieving higher Power. These results affirm the validity of our approach for this setting.



(a) CLIPPED (b) REGULAR

Figure 1: **Synthetic experiments for multivariate regression with varying noise levels.** The target FDR is set to 0.10, and gradient boosting is used to fit the multivariate regression model across all configurations. Subplots (a) and (b) show the results for CLIPPED and REGULAR, respectively.

E Online Bonferroni correction with Accept-to-Reject Changes

To the best of our knowledge, this work is the first to extend conformal selection to the online ARC setting, for which *no prior baseline exists*. Bonferroni correction, defined as

- Bonferroni: Select all $p_t \leq q/m$,

is a common baseline method in conformal selection (Jin & Candès, 2023; Bai et al., 2025); however, it does not satisfy the ARC property. This can be illustrated by the following example:

Example E.1. Now consider nominal FDR level $q = 0.1$, and conformal p-values $p_1 = 0.1$, $p_2 = 0.2$. The Bonferroni correction yields $\mathcal{R}_1 = \{1\}$ and $\mathcal{R}_2 = \emptyset$. This shows that Bonferroni correction would change its earlier selection into deselection (item 1).

To establish a comparison, we introduce a baseline by adapting the Bonferroni correction to the ARC framework, referred to as *Online Bonferroni correction with ARC* (budded OB).

- OB: Select all $p_t \leq q\gamma_t$ with score function `CLIP`.

OB employs the same $\{\gamma_t\}_{t=1}^T$ as in OCS-ARC. In the following theorem, we demonstrate that OB satisfies an ARC procedure.

Theorem E.2. *The OB is an ARC procedure. In particular, the rejection sets $\{\mathcal{R}\}_{t=1}^T$ outputted by OB satisfy*

$$\mathcal{R}_1 \subseteq \mathcal{R}_2 \subseteq \cdots \subseteq \mathcal{R}_T$$

The proof is presented in Appendix C.5. To empirically validate this observation, we conduct synthetic experiments across various models, recording the number of data points that transition from selected to deselected (i.e., reject-to-accept). We adopt the experimental setup described in Appendix B. The results, presented in Figure 2, confirm that this framework satisfies the ARC property.

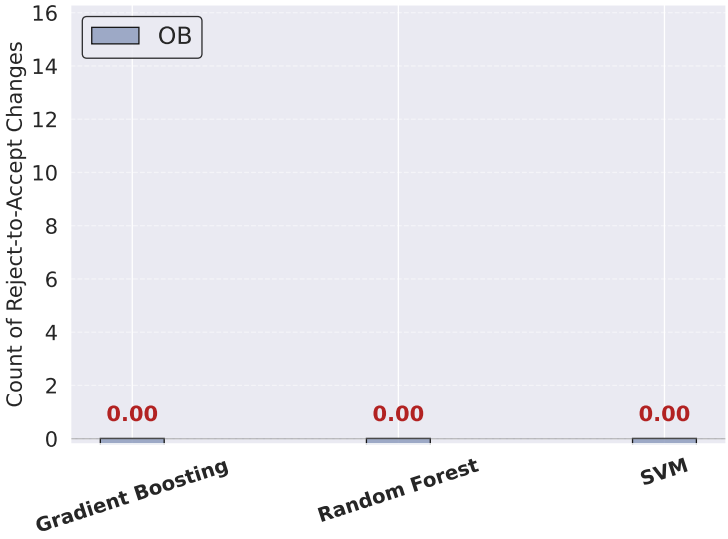


Figure 2: This experiment measures the total number of reject-to-accept samples for OB on 500 sequentially arriving data points from a synthetic dataset across three models. Results are averaged over 30 independent runs. The total number is 0 for all three employed base models.

F Additional Results

F.1 Results of Numerical experiments on support vector machine

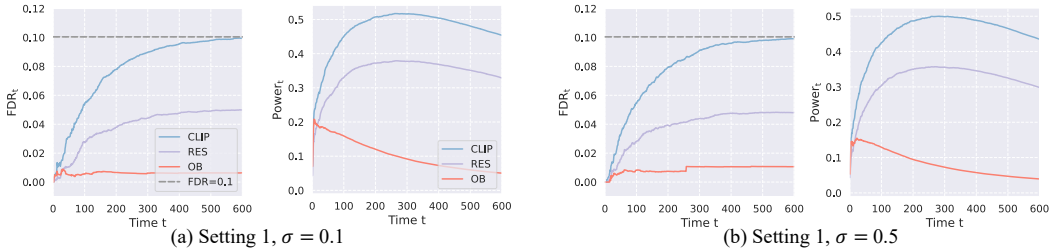
The results for gradient boosting are presented in the main paper, while the results for SVM are shown in Figure 3. All other setups remain consistent. We empirically evaluate the FDR at different timesteps by averaging the FDP over 300 runs. Among the two nonconformity scores, CLIP consistently demonstrates the highest realized FDR across all settings, remaining closer to the nominal level. Furthermore, the FDR stays below 0.1 in all configurations, regardless of the score function, noise level, or data-generating process. For Power, we assess it by calculating the average proportion of correct selections among all positive samples up to timestep t across all replicates. CLIP consistently achieves higher Power; however, the power decreases as noise strength increases across all settings. This decline occurs because increased noise impairs the model's ability to accurately fit the data, raising the fundamental difficulty of prediction. In summary, the results demonstrate the validity and robustness of our proposed approach across a diverse range of scenarios.

F.2 Results on Drug Property Prediction

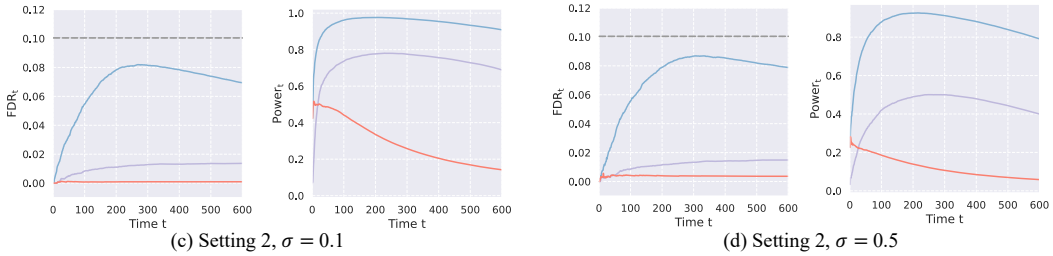
We consider the task of predicting drug properties for a specific protein target associated with HIV. The data may arrive incrementally over time, either from ongoing experiments or public repositories. Machine learning models are typically trained on a representative subset of the drug library screened via high-throughput screening and are then used to predict the activity of newly arriving compounds to identify promising candidates. Given that selected drugs may undergo further costly analyses, FDR control becomes crucial in this context.

Setup We utilize a real-world HIV screening dataset available on GitHub[1], consisting of 41,127 entries. The dataset is randomly split into training, calibration, and test sets with a ratio of 8:1:1. For the base predictive model, we train a small neural network for only 3 epochs; while using more complex or pre-trained deep networks could potentially improve power, it is not the primary focus of this study. We aim to select as many data points with a label of 1 as possible while controlling the FDR rate at a predefined level. Additionally, the target FDR for this experiment is set to 0.2, and CLIP is employed as the nonconformity score function.

[1] https://github.com/ying531/conformal-selection



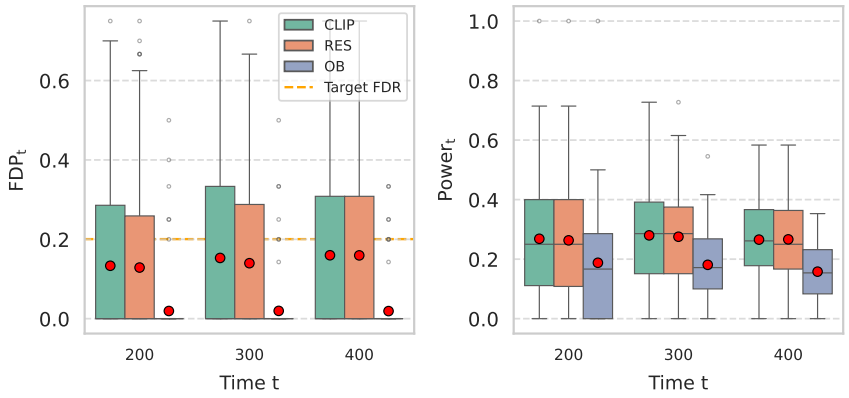
(a) Setting 1, $\sigma = 0.1$ (b) Setting 1, $\sigma = 0.5$



(c) Setting 2, $\sigma = 0.1$ (d) Setting 2, $\sigma = 0.5$

Figure 3: **Synthetic data experiments with varying noise levels across different data-generating processes.** The target FDR is set to 0.10, and SVM is used to fit the regression model across all configurations. Subplots (a) and (b) present results for simulation setting 1 (CLIP and RES, respectively), while subplots (c) and (d) correspond to simulation setting 2.

Results The results are presented in Figure 4. We evaluate the FDP$_t$ at $t = 200, 300, 400$ over 100 replications. Similar to the results observed in the recruitment task, the average FDP$_t$ remains below 0.2 at each evaluated timestep. We observe that our methods consistently outperform the baseline OB under all scores and simulation settings. However, due to the increased complexity of this task and the limited capacity of applied base model, the Power$_t$ is relatively lower compared to the previous experiment across different timesteps. In summary, the results of this experiment demonstrate the validity and effectiveness of our OCS-ARC in the context of drug property prediction.



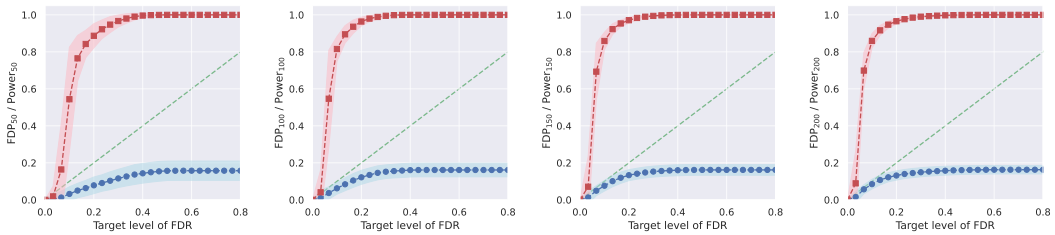
(a) FDP (b) Power

Figure 4: **Results for drug property prediction.** The target FDR is set to 0.20, and a three-layer neural network is trained for classification. FDP$_t$ and Power$_t$ are reported over 100 independent runs at timesteps 200, 300, 400.

F.3 Results on CoQA

Setup The setups remain consistent with the experiment on TriviaQA.

Results Results indicate that our method maintains the average FDP$_t$ below each target FDR at all evaluated timesteps. Additionally, it achieves high Power$_t$, effectively selecting the majority of



(a) t = 50 (b) t = 100 (c) t = 150 (d) t = 200

Figure 5: **Results for the application of OCS-ARC to question answering with LLMs** on the CoQA Dataset. Experiments are conducted using LLaMA-2-13B-chat, with the target FDR varied from 0 to 0.80 in fixed steps over 25 increments. FDP_t and Power_t are reported over 100 independent runs at timesteps 50, 100, 150, and 200.

high-quality answers. Overall, OCS-ARC offers a model-agnostic candidate selection framework that can be effortlessly integrated with any black-box model, including LLMs.

G Conformal Alignment

When applied to LLMs, our method is grounded in Conformal Alignment (Gui et al., 2024). In this section, we provide a concise overview of the framework and the entire procedure involved.

Overview Conformal Alignment is a versatile framework that establishes criteria for determining when to trust the outputs of foundation models, providing finite-sample, distribution-free guarantees. While it is rooted in the concept of conformal prediction, it differs from standard conformal prediction in that it does not focus on exploring the sampling space to construct a prediction set intended to encompass a desired output. Instead, this framework utilizes quantified confidence as a tool to identify specific units whose generated outputs can be deemed reliable. Given any model and any alignment criterion, Conformal Alignment guarantees that a predetermined proportion of the selected units' model outputs indeed satisfy the criterion. Here, "alignment" refers to a desired property of an output that may vary depending on the context. Specifically, let $m \in \mathbb{N}_+$ represent the number of new units awaiting outputs. With a specified error level $\alpha \in (0, 1)$, Conformal Alignment effectively controls the false discovery rate when selecting units with trustworthy outputs, as

$$\text{FDR} = \mathbb{E}\left[\frac{\sum_{i=1}^{m} \mathbb{1}\{i \text{ selected and not aligned}\}}{\sum_{i=1}^{m} \mathbb{1}\{i \text{ selected}\}}\right] \leq \alpha, \tag{1}$$

where the expectation is taken over the randomness of the data. FDR control offers an interpretable measure of the quality of selected deployable units. This method builds upon the Conformal Selection framework (Jin & Candès, 2023), leveraging a holdout set of "high-quality" reference data to guide the selection with calibrated statistical confidence for trusted outputs. Equation (1) holds in finite-sample as long as the holdout data are exchangeable with the new units. In addition, by selecting rather than modifying outputs, this approach preserves the nformativeness of the original outputs, and remains lightweight, e.g., avoiding the need to retrain large models.

Workflow Given a dataset $\mathcal{D}$ containing reference information, the described procedure initiates by dividing $\mathcal{D}$ into two subsets: the training set $\mathcal{D}_{\text{tr}}$ and the calibration set $\mathcal{D}_{\text{calib}}$. Notably, $\mathcal{D}_{\text{tr}}$ and $\mathcal{D}_{\text{calib}}$ may also represent the indices of the units within these sets, contingent on the context. Next, a model $g : \mathcal{X} \to \mathbb{R}$ is trained on $\mathcal{D}_{\text{tr}}$ to predict the alignment score based on the input X, potentially incorporating information from f. Following this, the procedure generates model outputs $f(X_i)$ and computes the predicted alignment scores $\hat{A}_i = g(X_i)$ for each $i \in [n+m]$. The pairs $(A_i, \hat{A}_i)_{i \in \mathcal{D}_{\text{calib}}}$ are then used to formulate the selection set. In line with the framework of *Conformal Selection* (Jin & Candès, 2023), statistical evidence is collected to assess trustworthy outputs through hypothesis testing. For each $j \in [m]$, the null hypothesis is defined as

$$H_j : A_{n+j} \leq c.$$

Rejecting H_j indicates evidence that the true alignment score of unit j exceeds the threshold c, suggesting that the generated output $f(X_{n+j})$ is aligned. Consequently, the task of selecting aligned units reduces to the simultaneous testing of the m hypotheses specified in the equation above. To facilitate this, the framework constructs the *conformal p-value*. For any $j \in [m]$, the conformal p-value is given by:

$$p_j = \frac{1 + \sum_{i \in \mathcal{D}_{\text{cal}}} \mathbb{1}\, A_i \leq c, \hat{A}_i \geq \hat{A}_{n+j}}{|\mathcal{D}_{\text{cal}}| + 1}. \tag{2}$$

It has been demonstrated in Jin & Candès (2023) that when the test unit is exchangeable with the calibration units, the above-defined p-value is valid in the sense that

$$\mathbb{P}(p_j \leq t, A_{n+j} \leq c) \leq t \quad \text{for any } t \in (0, 1).$$

Intuitively, when a generated output is likely to be aligned, one would expect $\hat{A}_{n+j}$ to exhibit a large magnitude, resulting in a smaller p_j. Thus, a small p_j leads to the rejection of the null hypothesis, signifying a sufficiently large alignment score, with the threshold for p-values established by the BH procedure (Benjamini & Hochberg, 1995). Specifically, let $p_{(1)} \leq \cdots \leq p_{(m)}$ denote the ordered statistics of the p-values; the rejection set of the BH applied to the conformal p-values is defined as

$$\mathcal{S} = \{j \in [m] : p_j \leq \frac{\alpha k^*}{m}\},$$

where

$$k^* = \max\left\{k \in [m] : p_{(k)} \leq \frac{\alpha k}{m}\right\},$$

with the convention that $\max \varnothing = 0$.

The complete procedure is outlined in Algorithm 2. For the sake of notational simplicity, the units are denoted as $Z_i = (X_i, E_i)$ for all $i \in [n + m]$, recognizing that E_i is not observable for $i > n$.

Algorithm 2 Conformal Alignment

Require: Pre-trained foundation model f; alignment score function $\mathcal{A}$; reference dataset $\mathcal{D} = (X_i, E_i)_{i=1}^n$; test dataset $\mathcal{D}_{\mathcal{D}_{\text{test}}} = (X_{n+j})_{j=1}^m$; algorithm for fitting alignment predictor $\mathcal{G}$; alignment level c; target FDR level α.

1: Compute the alignment score $A_i = \mathcal{A}(f(X_i), E_i), \forall i \in \mathcal{D}$.
2: Randomly split $\mathcal{D}$ into two disjoint sets: the training set $\mathcal{D}_{\text{tr}}$ and the calibration set $\mathcal{D}_{\text{cal}}$.
3: Fit the alignment score predictor with $\mathcal{D}_{\text{tr}}$: $g \leftarrow \mathcal{G}(\mathcal{D}_{\text{tr}})$.
4: Compute the predicted alignment score: $\hat{A}_i \leftarrow g(X_i), \forall i \in \mathcal{D}_{\text{cal}} \cup \mathcal{D}_{\mathcal{D}_{\text{test}}}$.
5: **for** $j \in [m]$ **do**
6: Compute the conformal p-values p_j according to Equation equation 2.
7: **end for**
8: Apply BH to the conformal p-values: $\mathcal{S} \leftarrow \text{BH}(p_1 \ldots, p_m)$.

Ensure: The selected units $\mathcal{S}$.

REFERENCES

Tian Bai, Yue Zhao, Xiang Yu, and Archer Y Yang. Multivariate conformal selection. In *International Conference on Machine Learning*. PMLR, 2025.

Stephen Bates, Emmanuel Candès, Lihua Lei, Yaniv Romano, and Matteo Sesia. Testing for outliers with conformal p-values. *The Annals of Statistics*, 2023.

Yoav Benjamini and Yosef Hochberg. Controlling the false discovery rate: a practical and powerful approach to multiple testing. *Journal of the Royal statistical society: series B (Methodological)*, 1995.

Yoav Benjamini and Daniel Yekutieli. The control of the false discovery rate in multiple testing under dependency. *Annals of statistics*, 2001.

Gilles Blanchard and Etienne Roquain. Two simple sufficient conditions for fdr control. 2008.

Yu Gui, Ying Jin, and Zhimei Ren. Conformal alignment: Knowing when to trust foundation models with guarantees. In *The Thirty-eighth Annual Conference on Neural Information Processing Systems*, 2024.

Ying Jin and Emmanuel J Candès. Selection by prediction with conformal p-values. *Journal of Machine Learning Research*, 2023.