PARTIAL CONJUNCTION CONFORMAL SELECTION

Anonymous authors
Paper under double-blind review

ABSTRACT

Conformal selection (CS) provides finite-sample guarantees for False Discovery Rate (FDR) control when selecting a subset of high-quality candidates from a large pool. Existing multivariate conformal selection (MCS) methods are restricted to the "all-or-nothing" rule (i.e., samples are selected only if all criteria are satisfied) and do not address the practical need for r-out-of-d conjunction rules. In this study, we show that directly applying unadjusted partial conjunction combinations to componentwise conformal p-values can fail to preserve FDR control in the CS setting. To address this, we propose Partial Conjunction Conformal Selection (*PCCS*), a unified framework that accommodates both the global r-out-of-d and block-level r_0-out-of-L conjunction rules. In particular, we introduce two types of novel nonconformity scores to construct conformal p-values and then run the Benjamini–Hochberg (BH) procedure, thereby providing theoretical guarantees for finite-sample FDR control. Simulated and real-world experiments demonstrate that *PCCS* maintains valid FDR control while substantially improving selection power, establishing a versatile framework for selection tasks such as drug property prediction, LLM response filtering, and monocular depth estimation.

1 INTRODUCTION

Selecting a subset of candidates from a large pool is essential in various real-world applications, such as drug discovery (Sheridan et al., 2015), LLM alignment (Gui et al., 2024), and monocular depth estimation (Hao et al., 2026). In practical scenarios, true labels (e.g., chemical structure, alignment score, or depth value) for test candidates are often unavailable, so selection must rely on machine learning model predictions instead. This gives rise to the significance of conformal selection (Jin & Candès, 2023) (CS), a systematic framework for candidate selection with uncertainty quantification by extending conformal prediction (Vovk et al., 2005). CS employs conformal p-values (Bates et al., 2023) and multiple testing corrections to control the false discovery rate (FDR).

However, existing multivariate CS (MCS) approaches mainly consider the case $y_k \in R_k$ for all $k = 1, \ldots, d$, where $\mathbf{y} \in \mathbb{R}^d$ and R_k is a user-defined target region. This all-or-nothing rule can be overly restrictive in practical applications. In drug discovery, compounds satisfying at least r out of d desired properties (for a user-specified threshold $1 \leq r \leq d$) are often preferred to retain more candidates for downstream wet-lab validation, which exemplifies the need for the global r-out-of-d rule. Notably, MCS corresponds to the special case with $r = d$. Moreover, the criteria may naturally be organized into L groups representing distinct practical requirements. For example, in drug discovery, the criteria may be grouped into safety, efficacy, and pharmacokinetic properties. A compound may be considered promising if it satisfies the within-group requirement for at least two of these three groups, where group ℓ is satisfied when at least r_ℓ of its d_ℓ properties meet their target criteria. This two-level requirement defines the block-level r_0-out-of-L conjunction rule.

In this paper, we extend the CS framework to accommodate both the global r-out-of-d and block-level r_0-out-of-L conjunction rules. We show that directly applying standard unadjusted partial conjunction combinations (Benjamini & Heller, 2008; Futschik et al., 2019) to componentwise conformal p-values can fail to preserve FDR control in the CS setting. To address this, we propose Partial Conjunction Conformal Selection (*PCCS*), which introduces two types of conjunction-aware nonconformity scores for the global r-out-of-d rule and the block-level r_0-out-of-L rule, both equipped with rigorous theoretical guarantees. These scores are used to compute conformal p-values, to which the Benjamini–Hochberg (BH) procedure (Benjamini & Hochberg, 1995; Benjamini & Yekutieli, 2001) is then applied for selection.

Reviewers: please read the Reviewer Guidelines (iclr.cc/Conferences/2027/ReviewerGuidelines) and the AI Policy for Reviewers (iclr.cc/Conferences/2027/AIPolicyForReviewers). If you used AI to expand, edit, or polish your review, please provide the input text to the LLM. Better still, consider skipping the LLM and submitting your original text: we, and the authors, are much more interested in your unedited thoughts than in what an LLM has to say. AI-assisted or not, you are putting your name and reputation behind your review: LLM-generated falsehoods, hallucinations or misrepresentations are subject to disciplinary action, which may include desk-rejecting all papers you have authored.

To verify the effectiveness of our method, we conduct extensive experiments on both simulated and real-world datasets. For real-world scenarios, we apply *PCCS* to drug property prediction, LLM response filtering, and monocular depth estimation. The results demonstrate that our proposed method maintains valid FDR control while achieving higher selection power than FDR-controlling baselines in most examined scenarios. For instance, in the drug property prediction experiment on the 15-dimensional ADMET dataset of Bai et al. (2025), *PCCS* achieves a power of 0.977 at $r = 11$, compared with 0.830 for the best-performing baseline. Notably, even though additionally requiring all 6 safety-critical properties to hold (block-level rule with $r_1 = 6$ and $r_2 = 9$) reduces *PCCS*'s power from 0.977 at $r = 11$ to 0.210, this remains substantially higher than the power of merely 0.012 at $r = 15$, thereby retaining more candidates for downstream wet-lab validation.

Our contributions are summarized as follows:

- We are the first to incorporate partial conjunction rules into the conformal selection setting to develop a general and flexible selection framework. We show that directly applying unadjusted partial conjunction combinations to componentwise conformal p-values can fail to control the FDR in the CS setting.
- We propose Partial Conjunction Conformal Selection (*PCCS*), a simple and effective method with finite-sample FDR control. *PCCS* designs two conjunction-aware nonconformity scores to construct conformal p-values and applies the BH procedure.
- Simulations show that *PCCS* maintains valid FDR control while improving selection power over the baseline. The effectiveness is further validated on real-world tasks, including drug property prediction, LLM response filtering, and monocular depth estimation.

2 PRELIMINARY

Problem Setup. We let $\mathcal{X} \subset \mathbb{R}^p$ denote the p-dimensional feature space and $\mathcal{Y} \subset \mathbb{R}^d$ denote the d-dimensional response space. We consider a training dataset $\mathcal{D}_{\text{train}}$, used only to fit the predictors $\{\widehat{f}_k\}_{k=1}^d$, a calibration dataset with n samples, $\mathcal{D}_{\text{cal}} = \{(\mathbf{x}_i, \mathbf{y}_i)\}_{i=1}^n$, and a test dataset with m inputs, $\mathcal{D}_{\text{test}} = \{\mathbf{x}_{n+j}\}_{j=1}^m$, where the corresponding test responses $\{\mathbf{y}_{n+j}\}_{j=1}^m$ are unobserved. We assume that the calibration and test pairs $\{(\mathbf{x}_i, \mathbf{y}_i)\}_{i=1}^{n+m}$ are drawn *i.i.d.* from an unknown underlying distribution $\mathcal{P}$ on $\mathcal{X} \times \mathcal{Y}$, independently of $\mathcal{D}_{\text{train}}$.

The selection problem can be formulated as follows. Given a predefined d-dimensional target region $R \subseteq \mathbb{R}^d$, our goal is to identify a subset $S \subseteq \{1, \ldots, m\}$ from the test set $\mathcal{D}_{\text{test}}$ such that as many selected test observations as possible satisfy $\mathbf{y}_{n+j} \in R$, while controlling the FDR (Benjamini & Hochberg, 1995) below a user-specified level q. The FDR is the expected false discovery proportion (FDP), i.e., the proportion of selected observations with $\mathbf{y}_{n+j} \notin R$:

$$\text{FDR} = \mathbb{E}[\text{FDP}], \quad \text{FDP} = \frac{\sum_{j=1}^m \mathbf{1}\{j \in S, \mathbf{y}_{n+j} \notin R\}}{1 \vee |S|}. \tag{1}$$

where $1 \vee |S| = \max\{1, |S|\}$. It is equally important to assess whether a method retains a sufficient number of selections that genuinely meet the target criteria. We define the Power as the expected proportion of desirable observations (i.e. $\mathbf{y}_{n+j} \in R$) that are correctly selected:

$$\text{Power} = \mathbb{E}\left[\frac{\sum_{j=1}^m \mathbf{1}\{j \in S, \mathbf{y}_{n+j} \in R\}}{1 \vee \sum_{j=1}^m \mathbf{1}\{\mathbf{y}_{n+j} \in R\}}\right]. \tag{2}$$

An ideal selection procedure should simultaneously keep the FDR below the user-specified level q while maximizing the power.

Conformal Selection. Conformal Selection (CS) is a model-agnostic selection framework with finite-sample FDR control. In the univariate setting (Jin & Candès, 2023), a test sample j is selected when the null $H_{0j} : y_{n+j} \notin R$ is rejected against $H_{1j} : y_{n+j} \in R$, where the target region $R \subseteq \mathbb{R}$ may be a half-line $(c, +\infty)$, an interval (c_1, c_2), a union $(-\infty, c_1) \cup (c_2, +\infty)$, or a more complex set. Multivariate conformal selection (MCS) (Bai et al., 2025; Hao et al., 2026) extends this to a vector response $\mathbf{y}_{n+j} = (y_{n+j,1}, \ldots, y_{n+j,d})$ with the hypothesis $H_{0j} : \mathbf{y}_{n+j} \notin R$ vs.

$H_{1j} : \mathbf{y}_{n+j} \in R$, where $R = R_1 \times \cdots \times R_d$ is a predefined region and each per-dimension region $R_k \subseteq \mathbb{R}$ may itself be a half-line, an interval, a union of intervals, or a more complex set.

To construct the selection set S, CS computes a nonconformity score $V(\mathbf{x}, \mathbf{y})$ for each calibration sample, evaluates $\widehat{V}_{n+j} = V(\mathbf{x}_{n+j}, \boldsymbol{c}_{n+j})$ for each test sample by substituting the unobserved response with predefined threshold $\boldsymbol{c}_{n+j}$, and derives conformal p-values $\{p_j\}$ via rank-based comparison against the calibration scores. The Benjamini–Hochberg (BH) procedure (Benjamini & Hochberg, 1995; Benjamini & Yekutieli, 2001) is then applied to these p-values to yield a selection set S with $\mathrm{FDR} \leq q$.

3 Partial Conjunction Conformal Selection

Selection Target. Existing MCS methods select samples whose d-dimensional response vector satisfies all d criteria. Partial Conjunction Conformal Selection (*PCCS*) relaxes the standard "all-or-nothing" rule by requiring at least r of the d criteria to be satisfied, rather than requiring all criteria to hold simultaneously. PCCS encompasses two fundamental structures: the global r-out-of-d conjunction rule and its block-level r_0-out-of-L generalization. More complex selection structures can be constructed by combining these basic rules. Specifically, given a user-specified threshold $r \in \{1, \ldots, d\}$, the global r-out-of-d conjunction rule is defined as

$$H_{0j}^{r/d} : \sum_{k=1}^{d} \mathbf{1}\left\{y_{n+j,k} \in R_k\right\} < r \quad \text{vs.} \quad H_{1j}^{r/d} : \sum_{k=1}^{d} \mathbf{1}\left\{y_{n+j,k} \in R_k\right\} \geq r. \tag{3}$$

The threshold r interpolates between two extremes: $r = 1$ gives the disjunction rule, where the response needs only satisfy a single criterion; $r = d$ gives the conjunction rule, where all d criteria must hold simultaneously, recovering the standard MCS setting. The partial conjunction rule $1 < r < d$ requires at least r of the d criteria to be satisfied, providing a flexible trade-off between the permissive $r = 1$ case and the stringent $r = d$ case.

The global r-out-of-d conjunction rule in equation 3 can itself be applied across groups of criteria, giving rise to a block-level conjunction structure. Specifically, let the index set $\{1, \ldots, d\}$ be partitioned into L disjoint blocks $G_1, \ldots, G_L$, with $\sum_{\ell=1}^{L} |G_\ell| = d$, inducing a corresponding partition of the response vector y_{n+j}. Let $r_\ell \in \{1, \ldots, |G_\ell|\}$ be the user-specified threshold for block G_ℓ, so that block ℓ is regarded as satisfied whenever at least r_ℓ of its $|G_\ell|$ criteria hold. Given an additional block-level threshold $r_0 \in \{1, \ldots, L\}$, a candidate is considered desirable if at least r_0 out of the L blocks are satisfied. The block-level r_0-out-of-L conjunction hypothesis is

$$\begin{aligned} H_{0j}^{r_0/L} &: \sum_{\ell=1}^{L} \mathbf{1}\left\{\sum_{k \in G_\ell} \mathbf{1}\left\{y_{n+j,k} \in R_k\right\} \geq r_\ell\right\} < r_0 \\ H_{1j}^{r_0/L} &: \sum_{\ell=1}^{L} \mathbf{1}\left\{\sum_{k \in G_\ell} \mathbf{1}\left\{y_{n+j,k} \in R_k\right\} \geq r_\ell\right\} \geq r_0. \end{aligned} \tag{4}$$

Limitations of Standard Partial Conjunction Testing in CS. A seemingly straightforward approach to the global r-out-of-d conjunction rule is to apply classical partial conjunction procedures (Benjamini & Heller, 2008; Futschik et al., 2019) to componentwise conformal p-values. Specifically, for each test observation j, we compute a nonconformity score $V_{j,k}$ and the corresponding conformal p-value $p_{j,k}$ for each component $k \in [d]$, and then combine these componentwise p-values into a partial conjunction p-value for selection.

However, the classical validity guarantees of such procedures do not automatically carry over to CS. For component k of test candidate j, conformal selection provides the joint validity property $\mathbb{P}\left(p_{j,k} \leq \alpha, H_{0j,k}\right) \leq \alpha$ for every $\alpha \in [0, 1]$, where $H_{0j,k}$ denotes the event that the k-th component null holds. In contrast, standard partial conjunction procedures are typically justified when the base hypotheses are fixed, so that each null p-value is super-uniform given the configuration of true nulls. When $\mathbb{P}(H_{0j,k}) > 0$, joint validity only implies $\mathbb{P}\left(p_{j,k} \leq \alpha \mid H_{0j,k}\right) \leq \frac{\alpha}{\mathbb{P}(H_{0j,k})}$, which may exceed α when $\mathbb{P}(H_{0j,k}) < 1$. Thus, componentwise joint validity alone is insufficient to extend the FDR guarantees of partial conjunction procedures to the CS setting. The following proposition illustrates this limitation for the direct unadjusted order-statistic combination $p_{(r)}$.

Proposition 1. *For every $d \geq 2$ and $r \in [d]$, there exists a joint distribution of $(H_{0,1}, \ldots, H_{0,d}, p_1, \ldots, p_d)$ such that $\mathbb{P}(p_k \leq \alpha, H_{0,k}) \leq \alpha$ for every $\alpha \in [0,1]$ and $k \in [d]$ but the unadjusted combination $p_{(r)}$ satisfies*

$$\mathbb{P}\Big(p_{(r)} \leq \alpha, H_0^{r/d}\Big) = \begin{cases} 1-(1-\alpha)^d, & r = 1, \\ \min\left\{\dfrac{d}{d-r+1}\alpha, 1\right\}, & 2 \leq r \leq d. \end{cases}$$

Consequently, $\mathbb{P}\Big(p_{(r)} \leq \alpha, H_0^{r/d}\Big) > \alpha$ for every $\alpha \in (0,1)$ when $r = 1$, and for every $\alpha \in (0, (d-r+1)/d)$ when $2 \leq r \leq d$. Hence, with a single test candidate $m = 1$, applying BH at any such nominal level q yields $\mathrm{FDR} > q$.

The two cases illustrate distinct issues. When $r = 1$, all component nulls hold almost surely, and the failure of $p_{(1)}$ results from the absence of a multiplicity adjustment. When $r = d$, the Bonferroni partial conjunction p-value reduces to $p_{(d)}$, which is valid when each component p-value is conditionally super-uniform; under componentwise joint validity alone, however, $\mathbb{P}(p_{(d)} \leq \alpha, H_0^{d/d})$ can be as large as $\min\{d\alpha, 1\}$.

4 Proposed Method

4.1 Addressing global r-out-of-d conjunction rule

To address the global r-out-of-d conjunction rule, define $r^*(\mathbf{y}) = \sum_{k=1}^d \mathbb{1}\{y_k \in R_k\}$ and $\mathcal{R}^{r/d} = \{\mathbf{y} \in \mathcal{Y} : r^*(\mathbf{y}) \geq r\}$. For each coordinate $k \in [d]$, let $D_k\big(\widehat{f}_k(\mathbf{x}), R_k\big)$ denote the signed margin of $\widehat{f}_k(\mathbf{x})$ relative to R_k, which is positive if and only if $\widehat{f}_k(\mathbf{x}) \in R_k$, zero on the boundary of R_k, and negative otherwise. Let $D^{\downarrow}_{[1]}(\mathbf{x}) \geq \cdots \geq D^{\downarrow}_{[d]}(\mathbf{x})$ denote the decreasing order statistics of $\Big\{D_k\big(\widehat{f}_k(\mathbf{x}), R_k\big) : k \in [d]\Big\}$. We define the conjunction-aware nonconformity score as

$$V^{r/d}(\mathbf{x}, \mathbf{y}) = M\mathbb{1}\{r^*(\mathbf{y}) \geq r\} - D^{\downarrow}_{[r]}(\mathbf{x}), \tag{5}$$

where $M > 0$ is a fixed constant. Finite-sample validity holds for any M, since the null-branch substitution is exact regardless of M; a positive M ensures regional monotonicity, and all experiments use $M = 1000$. The signed margins $D_k\big(\widehat{f}_k(\mathbf{x}), R_k\big)$, which equal zero on the boundary of R_k, are summarized in Table 4: we adopt existing signed margins for the half-line $(c, +\infty)$ and the interval (c_1, c_2) from Jin & Candès (2023) and Hao et al. (2026), respectively, and give new signed margins for the union $(-\infty, c_1) \cup (c_2, +\infty)$ and a more complex piecewise region.

Remark 4.1. *The r-th largest signed margin $D^{\downarrow}_{[r]}(\mathbf{x})$ provides a scalar summary of the predictive evidence for the r-out-of-d criterion, since $D^{\downarrow}_{[r]}(\mathbf{x}) > 0$ is equivalent to $\sum_{k=1}^d \mathbb{1}\left\{D_k\big(\widehat{f}_k(\mathbf{x}), R_k\big) > 0\right\} \geq r$. Thus, $D^{\downarrow}_{[r]}(\mathbf{x})$ is positive precisely when at least r fitted responses lie in their respective target regions, and a larger value indicates stronger predictive evidence that the global r-out-of-d criterion is satisfied.*

When the true response $\mathbf{y}_i$ is observed, the score $V^{r/d}(\mathbf{x}_i, \mathbf{y}_i)$ can be computed directly. For brevity, write $V_i^{r/d} = V^{r/d}(\mathbf{x}_i, \mathbf{y}_i)$. The corresponding oracle conformal p-value is

$$p_j^* = \frac{1 + \sum_{i=1}^n \mathbb{1}\{V_i^{r/d} \leq V_{n+j}^{r/d}\}}{n+1}. \tag{6}$$

In practice, the test response $\mathbf{y}_{n+j}$ is unobserved, so $V_{n+j}^{r/d}$ cannot be computed directly. We evaluate the score under the null branch $r^*(\mathbf{y}_{n+j}) < r$, for which the indicator in Equation 5 equals zero.

Definition 1 (Null-branch substitution)**.** *For every $\mathbf{x} \in \mathcal{X}$, define the practical conjunction-aware score as*

$$\widehat{V}^{r/d}(\mathbf{x}) = -D^{\downarrow}_{[r]}(\mathbf{x}), \tag{7}$$

and write $\widehat{V}_{n+j}^{r/d} = \widehat{V}^{r/d}(\mathbf{x}_{n+j})$ for each test observation $j \in [m]$.

The following proposition establishes the regional monotonicity of the proposed score.

Proposition 2. *For every $M > 0$, the conjunction-aware nonconformity score $V^{r/d}$ in Equation 5 is regionally monotone with respect to $\mathcal{R}^{r/d}$. Specifically, for every $\mathbf{x} \in \mathcal{X}$, $\mathbf{y} \in \mathcal{R}^{r/d}$, and $\mathbf{y}' \notin \mathcal{R}^{r/d}$, $V^{r/d}(\mathbf{x}, \mathbf{y}') < V^{r/d}(\mathbf{x}, \mathbf{y})$.*

Indeed, $V^{r/d}(\mathbf{x}, \mathbf{y}) = M - D^{\downarrow}_{[r]}(\mathbf{x})$ for $\mathbf{y} \in \mathcal{R}^{r/d}$, whereas $V^{r/d}(\mathbf{x}, \mathbf{y}') = -D^{\downarrow}_{[r]}(\mathbf{x})$ for $\mathbf{y}' \notin \mathcal{R}^{r/d}$. The conclusion follows immediately from $M > 0$. More importantly, the null-branch substitution is exact under the null hypothesis.

Lemma 1. *For every $(\mathbf{x}, \mathbf{y})$ satisfying $\mathbf{y} \notin \mathcal{R}^{r/d}$, $V^{r/d}(\mathbf{x}, \mathbf{y}) = \widehat{V}^{r/d}(\mathbf{x})$.*

The practical conformal p-value is

$$p_j = \frac{1 + \sum_{i=1}^{n} \mathbb{1}\left\{V_i^{r/d} \le \widehat{V}_{n+j}^{r/d}\right\}}{n+1}. \tag{8}$$

A smaller p_j provides stronger evidence that at least r of the d criteria are satisfied. Because $\widehat{V}_{n+j}^{r/d} = -D^{\downarrow}_{[r]}(\mathbf{x}_{n+j})$, stronger predictive evidence produces a smaller practical score and a conformal p-value that is no larger. By Lemma 1, on the null event $H_{0j}^{r/d}$, we have $\widehat{V}_{n+j}^{r/d} = V_{n+j}^{r/d}$ and $p_j = p_j^*$.

Proposition 3. *Suppose that the fitted predictors $\{\widehat{f}_k\}_{k=1}^{d}$ are fixed or trained using an independent training sample, and that the calibration observations $\{(\mathbf{x}_i, \mathbf{y}_i)\}_{i=1}^{n}$, together with the j-th test observation $(\mathbf{x}_{n+j}, \mathbf{y}_{n+j})$, are exchangeable. Then, for every $\alpha \in (0,1)$, $\mathbb{P}\left(p_j \le \alpha,\ H_{0j}^{r/d}\right) \le \alpha$.*

Given the practical conformal p-values $\{p_j\}_{j=1}^{m}$, we apply the Benjamini–Hochberg procedure (Benjamini & Hochberg, 1995; Benjamini & Yekutieli, 2001) at level q to obtain the selection set $\mathcal{S}$.

Theorem 1. *Let $M > 0$, and suppose that the practical scores $\{\widehat{V}_{n+j}^{r/d}\}_{j=1}^{m}$ are constructed using the null-branch substitution in Definition 1. For every $j \in [m]$, assume that $V_1^{r/d}, \ldots, V_n^{r/d}, V_{n+j}^{r/d}$ are exchangeable conditionally on $\{\widehat{V}_{n+\ell}^{r/d}\}_{\ell \neq j}$. Then, for every $q \in (0,1)$, the selection set $\mathcal{S}$ returned by Algorithm 3 satisfies* $\mathrm{FDR} \le q$.

Theorem 1 provides finite-sample FDR control under the stated conditional exchangeability assumption. This assumption holds when the calibration and test pairs are i.i.d. and the predictors are fitted on an independent training sample. The marginal joint validity in Proposition 3 alone does not establish FDR control for BH. While Theorem 1 establishes finite-sample validity, it does not explicitly characterize the selection power or the degree of conservativeness of the procedure. We therefore provide the following asymptotic characterization of the FDR and power as $n, m \to \infty$.

Theorem 2. *Let $V^{r/d}$ be the conjunction-aware nonconformity score defined in equation 5. Suppose that $\{(\mathbf{x}_i, \mathbf{y}_i)\}_{i=1}^{n+m}$ are i.i.d. samples from $\mathcal{P}$, and that the fitted predictors $\{\widehat{f}_k\}_{k=1}^{d}$ are fixed or trained using an independent training sample. Let $(\mathbf{x}, \mathbf{y}) \sim \mathcal{P}$ and $F_V(v) = \mathbb{P}\big(V^{r/d}(\mathbf{x}, \mathbf{y}) \le v\big)$. Let $\widehat{V}^{r/d}(\mathbf{x}) = -D^{\downarrow}_{[r]}(\mathbf{x})$ denote the null-branch practical score defined in Definition 1, and define the limiting practical p-value by $p^{\infty} = F_V\Big(\widehat{V}^{r/d}(\mathbf{x})\Big)$. Define $r^*(\mathbf{y}) = \sum_{k=1}^{d} \mathbb{1}\{y_k \in R_k\}$ and $\mathcal{R}^{r/d} = \{\mathbf{y} : r^*(\mathbf{y}) \ge r\}$, and let $G(t) = \mathbb{P}(p^{\infty} \le t)$. The population BH threshold is $t_q = \sup\{t \in [0,1] : t \le qG(t)\}$. Assume that G is continuous on $[0,1]$, $t_q > 0$, $G(t_q) > 0$, and $\mathbb{P}(r^*(\mathbf{y}) \ge r) > 0$. Assume further that t_q is a stable population BH crossing: there exists $\varepsilon > 0$ such that $t < qG(t)$ for all $t \in (t_q - \varepsilon, t_q)$ and $t > qG(t)$ for all $t \in (t_q, t_q + \varepsilon)$. Let $\mathcal{S}_q$ be the selection set obtained by applying BH at level q to the practical conformal p-values in equation 8. Then, as $n, m \to \infty$,*

$$\mathrm{FDR}(\mathcal{S}_q) \longrightarrow \frac{\mathbb{P}(p^{\infty} \le t_q,\ r^*(\mathbf{y}) < r)}{\mathbb{P}(p^{\infty} \le t_q)}, \qquad \mathrm{Power}(\mathcal{S}_q) \longrightarrow \frac{\mathbb{P}(p^{\infty} \le t_q,\ r^*(\mathbf{y}) \ge r)}{\mathbb{P}(r^*(\mathbf{y}) \ge r)}.$$

4.2 Addressing the Block-Level r_0-out-of-L Conjunction Rule

Building on the global r-out-of-d construction, we extend *PCCS* to the block-level r_0-out-of-L conjunction rule introduced in equation 4. Let $G_1, \ldots, G_L$ be a partition of $[d]$, let $r_\ell \in$

$\{1, \ldots, |G_\ell|\}$ denote the local conjunction threshold for block G_ℓ, and let $r_0 \in [L]$ denote the required number of satisfied blocks. For each block $\ell \in [L]$, define its observed satisfaction indicator as $B_\ell(\mathbf{y}) = \mathbb{1}\left\{\sum_{k \in G_\ell} \mathbb{1}\{y_k \in R_k\} \geq r_\ell\right\}$. The block-level target region is $\mathcal{R}^{r_0/L} = \left\{\mathbf{y} \in \mathcal{Y} : \sum_{\ell=1}^{L} B_\ell(\mathbf{y}) \geq r_0\right\}$.

For each block ℓ, let $D^{\downarrow}_{\ell,[1]}(\mathbf{x}) \geq \cdots \geq D^{\downarrow}_{\ell,[|G_\ell|]}(\mathbf{x})$ denote the decreasing order statistics of $\left\{D_k\big(\widehat{f}_k(\mathbf{x}), R_k\big) : k \in G_\ell\right\}$. The block-specific signed margin is $D^{\downarrow}_{\ell,[r_\ell]}(\mathbf{x})$. In particular, $D^{\downarrow}_{\ell,[r_\ell]}(\mathbf{x}) > 0$ is equivalent to $\sum_{k \in G_\ell} \mathbb{1}\left\{D_k\big(\widehat{f}_k(\mathbf{x}), R_k\big) > 0\right\} \geq r_\ell$. Let $D^{\downarrow}_{\text{block},[1]}(\mathbf{x}) \geq \cdots \geq D^{\downarrow}_{\text{block},[L]}(\mathbf{x})$ denote the decreasing order statistics of $\left\{D^{\downarrow}_{\ell,[r_\ell]}(\mathbf{x}) : \ell \in [L]\right\}$. It follows that $D^{\downarrow}_{\text{block},[r_0]}(\mathbf{x}) > 0$ is equivalent to $\sum_{\ell=1}^{L} \mathbb{1}\left\{D^{\downarrow}_{\ell,[r_\ell]}(\mathbf{x}) > 0\right\} \geq r_0$. Thus, $D^{\downarrow}_{\text{block},[r_0]}(\mathbf{x})$ provides a scalar signed margin for the block-level r_0-out-of-L criterion. Given $M > 0$, we therefore define the block-level conjunction-aware nonconformity score as

$$V^{r_0/L}(\mathbf{x}, \mathbf{y}) = M\mathbb{1}\left\{\sum_{\ell=1}^{L} B_\ell(\mathbf{y}) \geq r_0\right\} - D^{\downarrow}_{\text{block},[r_0]}(\mathbf{x}). \tag{9}$$

Proposition 4. *For every $M > 0$, the block-level score $V^{r_0/L}$ in equation 9 is regionally monotone with respect to $\mathcal{R}^{r_0/L}$. Specifically, for every $\mathbf{x} \in \mathcal{X}$, $\mathbf{y} \in \mathcal{R}^{r_0/L}$, and $\mathbf{y}' \notin \mathcal{R}^{r_0/L}$, $V^{r_0/L}(\mathbf{x}, \mathbf{y}') < V^{r_0/L}(\mathbf{x}, \mathbf{y})$.*

As in the global r-out-of-d construction, we evaluate the block-level score under the null branch and define $\widehat{V}^{r_0/L}_{n+j} = -D^{\downarrow}_{\text{block},[r_0]}(\mathbf{x}_{n+j})$. Under the block-level null hypothesis $H^{r_0/L}_{0j} : \sum_{\ell=1}^{L} B_\ell(\mathbf{y}_{n+j}) < r_0$, the substitution is exact: $\widehat{V}^{r_0/L}_{n+j} = V^{r_0/L}_{n+j}$. Substituting $\widehat{V}^{r_0/L}_{n+j}$ and the calibration scores $V^{r_0/L}_i = V^{r_0/L}(\mathbf{x}_i, \mathbf{y}_i)$ into equation 8 yields the practical block-level conformal p-values, to which we apply the same BH procedure as in Algorithm 3 at level q to obtain the final selection set.

Theorem 3. *Under the assumptions of Theorem 1, with $V^{r/d}$ and $\widehat{V}^{r/d}_{n+j}$ replaced by $V^{r_0/L}$ and $\widehat{V}^{r_0/L}_{n+j}$, the selection set obtained by applying BH at level q to the block-level practical conformal p-values satisfies $\mathrm{FDR} \leq q$ for every $q \in (0, 1)$.*

5 Simulation Studies

Setup. We evaluate our method on simulated regression datasets. The p-dimensional feature vector $\mathbf{x}$ is sampled from $\mathrm{Unif}(-1, 1)^p$, and the d-dimensional response vector is generated as $\mathbf{y} = f(\mathbf{x}) + \epsilon$, where f is the regression function and ϵ denotes the random noise. Different experimental settings are constructed by varying the regression function and noise distribution. Each dataset is split into training, calibration, and test sets of 1000 samples each. We use multivariate linear regression (MLR) as the default predictive model. In each replication, we generate independent training, calibration, validation, and test sets of 1000 samples each; the validation set is used only to fix the target thresholds (Appendix A.3)..

Baselines. For the global r-out-of-d conjunction rule, we compare our method with several standard partial conjunction testing baselines, including the *Bonferroni* method (Benjamini & Heller, 2008) and the *Omnibus* method (Futschik et al., 2019). For block-level r_0-out-of-L conjunction rule, we consider four baselines. The *IT* and *IB* baselines both apply the global r-out-of-d conjunction CS procedure separately to each block and intersect the individual sets, differing only in how the error budget is allocated: *IT* uses a uniform nominal level q, while *IB* sets each subroutine's level to q/L. The *IS* baseline determines per-block nominal levels by calibrating on a held-out validation set. Following Bai et al. (2025), the *BinCS* baseline reduces the multivariate response to a binary indicator, $\tilde{y}_{n+j} = \mathbf{1}\{\sum_{k=1}^{d} \mathbf{1}\{y_{n+j,k} \in R_k\} \geq r\}$ for the global r-out-of-d conjunction rule or $\tilde{y}_{n+j} = \mathbf{1}\{\sum_{\ell=1}^{L} \mathbf{1}\{\sum_{k \in G_\ell} \mathbf{1}\{y_{n+j,k} \in R_k\} \geq r_\ell\} \geq r_0\}$ for the block-level r_0-out-of-L conjunction rule, and applies a standard univariate CS procedure (Jin & Candès, 2023).

Table 1: Empirical FDR and power for the global r-out-of-d conjunction rule across varying values of the conjunction parameter r, with the dimension fixed at $d = 5$. The data-generating process follows simulation setting 1. The nominal FDR level is $q = 0.30$, MLR is used, and the target region is $R_k = (c_k, \infty)$ for all $k = 1, \ldots, 5$. Results are reported as mean $\pm$ standard deviation. Boldface indicates the best-performing method among those that control the FDR.

Metric	r	*Bonferroni*	*Omnibus*	*BinCS*	*PCCS* (Ours)
FDR	1	0.201 ± 0.044	0.239 ± 0.042	0.295 ± 0.136	0.297 ± 0.041
	2	0.038 ± 0.048	0.120 ± 0.051	0.306 ± 0.059	0.296 ± 0.050
	3	0.008 ± 0.024	0.125 ± 0.044	0.306 ± 0.047	0.301 ± 0.040
	4	0.074 ± 0.051	0.262 ± 0.066	0.304 ± 0.048	0.300 ± 0.047
	5	0.701 ± 0.019	0.701 ± 0.019	0.299 ± 0.059	0.295 ± 0.037
Power	1	0.345 ± 0.087	0.453 ± 0.071	0.131 ± 0.095	$\mathbf{0.570 \pm 0.090}$
	2	0.087 ± 0.087	0.272 ± 0.103	0.442 ± 0.095	$\mathbf{0.637 \pm 0.087}$
	3	0.026 ± 0.055	0.341 ± 0.092	0.572 ± 0.075	$\mathbf{0.681 \pm 0.069}$
	4	0.242 ± 0.139	0.648 ± 0.118	0.599 ± 0.079	$\mathbf{0.706 \pm 0.066}$
	5	1.000 ± 0.000	1.000 ± 0.000	0.391 ± 0.094	$\mathbf{0.679 \pm 0.054}$

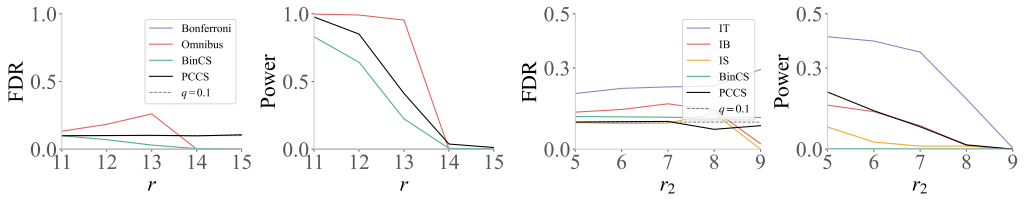
Table 2: Empirical FDR and power for the block-level r_0-out-of-L conjunction rule under the configurations described in Table 5. The two blocks satisfy $|G_1| = |G_2| = 5$, giving a total dimension of $d = 10$. The data-generating process follows simulation setting 1. The nominal FDR level is $q = 0.30$, MLR is used, and the coordinate-wise target region is $R_k = (c_k, \infty)$ for all $k = 1, \ldots, d$. Results are reported as mean $\pm$ standard deviation over 500 simulation replications. Boldface indicates the best-performing method among those that control the FDR.

Metric	Task	*IT*	*IB*	*IS*	*BinCS*	*PCCS* (Ours)
FDR	Task 1	0.485 ± 0.032	0.257 ± 0.044	0.294 ± 0.045	0.297 ± 0.088	0.303 ± 0.053
	Task 2	0.489 ± 0.028	0.262 ± 0.044	0.290 ± 0.055	0.313 ± 0.106	0.303 ± 0.055
	Task 3	0.475 ± 0.028	0.247 ± 0.040	0.293 ± 0.041	0.299 ± 0.064	0.301 ± 0.044
	Task 4	0.484 ± 0.031	0.254 ± 0.051	0.289 ± 0.050	0.284 ± 0.122	0.295 ± 0.038
Power	Task 1	0.819 ± 0.041	0.415 ± 0.069	0.442 ± 0.071	0.273 ± 0.115	$\mathbf{0.510 \pm 0.090}$
	Task 2	0.784 ± 0.040	0.376 ± 0.059	0.413 ± 0.097	0.169 ± 0.094	$\mathbf{0.479 \pm 0.096}$
	Task 3	0.837 ± 0.034	0.473 ± 0.064	0.529 ± 0.071	0.452 ± 0.098	$\mathbf{0.596 \pm 0.077}$
	Task 4	0.807 ± 0.048	0.372 ± 0.070	0.442 ± 0.076	0.150 ± 0.097	$\mathbf{0.513 \pm 0.099}$

Main Results. Table 1 reports the empirical FDR and power of *PCCS* and the baseline methods under the global r-out-of-d conjunction rule, with the dimension fixed at $d = 5$ and the conjunction level varying from $r = 1$ to $r = 5$. Across all values of r, *PCCS* maintains an observed FDR close to the nominal level of 0.30, ranging from 0.296 to 0.301. *BinCS* also maintains an FDR close to the nominal level. Although Bonferroni and Omnibus are conservative for $r = 1, \ldots, 4$, both fail to control the FDR under the full conjunction rule ($r = d = 5$), attaining an FDR of 0.701. Among the methods that maintain FDR control, *PCCS* consistently achieves the highest power, improving over *BinCS* from 0.131 to 0.570 for the disjunction rule ($r = 1$) and from 0.391 to 0.679 for the full conjunction rule ($r = 5$). These results demonstrate that *PCCS* provides consistently strong selection power while maintaining FDR near the nominal level and suggest that standard conjunction-testing methods cannot be directly applied within the conformal selection framework.

Table 2 summarizes the results for the four block-level r_0-out-of-L conjunction rules. Across all four tasks, *PCCS* maintains empirical FDR close to the nominal level of 0.30, while achieving the highest power among the methods that maintain FDR near or below the nominal level. The *IT* baseline substantially violates FDR control in every task, despite achieving high power. In contrast, *IB*, *IS*, and *BinCS* generally maintain FDR near or below the nominal level, but their power is consistently lower than that of *PCCS*. For example, in Task 1, *PCCS* achieves a mean power of 0.510, outperforming *IB* (0.415), *IS* (0.442), and *BinCS* (0.273). This represents a power improvement of 0.068 over the second-best FDR-controlling baseline, *IS*.

Ablation Studies. We conduct a series of ablation studies to evaluate *PCCS* under diverse experimental conditions. First, Table 6 shows that *PCCS* maintains strong performance when different predictive models are employed. Second, Tables 7–9 demonstrate that *PCCS* performs consistently



(a) Global r-out-of-d Conjunction Rule (b) Block-Level r_0-out-of-L Conjunction Rule

Figure 1: Observed FDR and power on the ADMET dataset across varying conjunction thresholds (r and r_2) for the global r-out-of-d and block-level r_0-out-of-L conjunction rules, with nominal FDR level $q = 0.10$. Here, r_1 and r_2 denote the block-specific thresholds for the two blocks G_1 (of size d_1) and G_2 (of size d_2) under the block-level r_0-out-of-L conjunction rule, with $r_1 = d_1 = 6$ fixed and r_2 varying.

across different nominal FDR levels, $q \in \{0.05, 0.10, 0.20\}$. Third, Tables 10–12 show that its performance remains robust across different data-generating processes. Fourth, Tables 13–14 demonstrate that *PCCS* maintains strong selection performance across different response dimensions d. Fifth, Table 1 and Tables 15–17 show that *PCCS* performs consistently across diverse target regions R_k, including a half-line $(c, +\infty)$, an interval (c_1, c_2), a union of two intervals $(-\infty, c_1) \cup (c_2, +\infty)$, and a more complex region $(-\infty, c_1) \cup (c_2, c_3) \cup (c_4, +\infty)$. Overall, *PCCS* achieves the highest power among the FDR-controlling methods while maintaining FDR close to the nominal level.

6 Real Data Applications

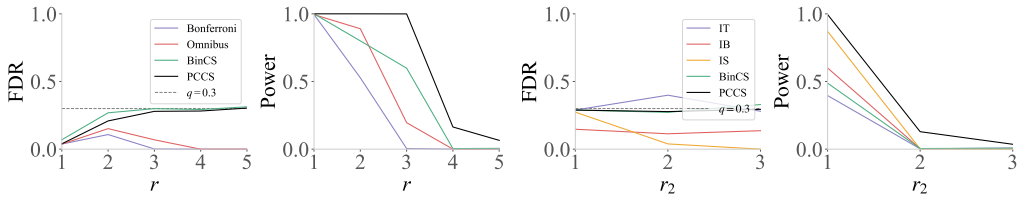
Multi-property Molecular Selection. We apply the *PCCS* method to identify drug candidates with desirable multi-property profiles, where FDR control improves downstream wet-lab validation efficiency. We use an ADMET dataset from Bai et al. (2025) with response dimension $d = 15$, and DeepPurpose (Huang et al., 2020) with Morgan drug encoding as the predictive model. In the global r-out-of-d conjunction rule, a candidate is selected if at least r out of d criteria are satisfied. In the block-level r_0-out-of-L conjunction rule, responses are partitioned into a hard-constraint group ($d_1 = 6$, safety-critical endpoints) and a soft-constraint group ($d_2 = 9$, pharmacokinetic endpoints); selection requires all hard constraints and at least r_2 soft constraints to be met. We use $n_{\text{train}} = 12000$, $n_{\text{cal}} = 8000$, $n_{\text{test}} = 2805$, target FDR $q = 0.1$, and 500 independent iterations.

Figure 1 summarizes the FDR and power of all methods. Both *BinCS* and *PCCS* maintain valid FDR control, with *PCCS* consistently achieving the highest power across all values of r under the global r-out-of-d rule and all values of r_2 (with $r_1 = 6$ fixed) under the block-level r_0-out-of-L rule, where it outperforms *BinCS* in every setting. Requiring all $d = 15$ criteria to be jointly satisfied is often overly stringent, yielding a power of only 1.2% and thus retaining very few candidates for wet-lab validation. By relaxing the requirement to $r = 11$, *PCCS* raises the power to 97.7%, illustrating the practical value of the r-out-of-d flexibility that MCS (Bai et al., 2025) does not offer.

Response Filtering. We apply *PCCS* to select high-quality LLM responses on HelpSteer2 (Wang et al., 2024a;b). For each response, HelpSteer2 provides human-annotated scores ranging from 0 to 4 for five attributes: *Helpfulness*, *Correctness*, *Coherence*, *Complexity*, and *Verbosity*. An attribute is regarded as satisfied if its human-annotated score exceeds 2 (i.e., $R_k = (2, 4]$), and selection performance is evaluated against these human annotations. We set $n_{\text{cal}} = 5000$ and $n_{\text{test}} = 2593$, and $q = 0.3$. Results are averaged over 500 repeated random splits.

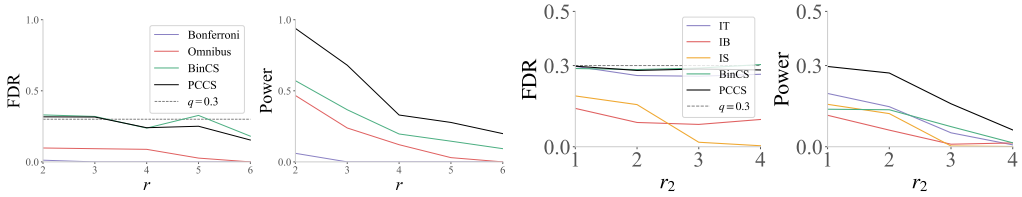
Figure 2 reports the empirical FDR and power based on the predictive scores generated by DeepSeek-V4.1-Flash (Xu et al., 2026), with the human annotations used as ground truth. Under both conjunction rules, *PCCS* maintains FDR near the nominal level $q = 0.30$ and achieves the highest power among the empirically FDR-controlling methods. Its power decreases as r or r_2 increases, but remains nonzero under the most stringent settings, whereas the competing methods either lose nearly all power or exceed the nominal FDR level.

Monocular Depth Estimation. We extend the *PCCS* framework to monocular depth estimation, with the goal of selecting images for which the ground-truth depths of at least r prede-



(a) Global r-out-of-d Conjunction Rule (b) Block-Level r_0-out-of-L Conjunction Rule

Figure 2: Observed FDR and power on HelpSteer2 under the global r-out-of-d and block-level r_0-out-of-L conjunction rules, with nominal FDR level $q = 0.30$. For the block-level rule, $G_1 = \{\textit{Helpfulness}, \textit{Correctness}\}$ is the mandatory block with $r_1 = d_1 = 2$, while $G_2 = \{\textit{Coherence}, \textit{Complexity}, \textit{Verbosity}\}$ requires at least $r_2 \in \{1, 2, 3\}$ of its $d_2 = 3$ criteria to be satisfied. Both blocks are required, corresponding to $r_0 = L = 2$.



(a) Global r-out-of-d Conjunction Rule (b) Block-Level r_0-out-of-L Conjunction Rule

Figure 3: Observed FDR and power on the NYU Depth V2 dataset for the global r-out-of-6 and block-level r_0-out-of-2 conjunction rules, with nominal FDR level $q = 0.30$. Here, r denotes the required number of regions for the global rule, and r_2 denotes the threshold for the four left/right regions under the block-level rule, with the two center regions fixed as mandatory ($r_1 = d_1 = 2$).

fined spatial regions satisfy a target criterion. We define the six-dimensional response $\mathbf{y} = (y_{\mathrm{LT}}, y_{\mathrm{LB}}, y_{\mathrm{CT}}, y_{\mathrm{CB}}, y_{\mathrm{RT}}, y_{\mathrm{RB}})$, where each component is the average ground-truth depth in the left-top (LT), left-bottom (LB), center-top (CT), center-bottom (CB), right-top (RT), and right-bottom (RB) region, respectively. For region k, we set the target region as $R_k = (3, +\infty)$ meters; under the global r-out-of-6 rule, an image is desirable if at least r of the six regions exceed 3 meters. We further consider a block-level r_0-out-of-2 rule, where the two center regions form a mandatory block, while the remaining four peripheral regions form a second block desirable if at least r_2 of them satisfy the criterion. We evaluate on NYU Depth V2 (Silberman et al., 2012) with ResNet-50 (He et al., 2016) features and a Ridge regressor, setting $n_{\mathrm{train}} = 1073$, $n_{\mathrm{cal}} = 188$, $n_{\mathrm{test}} = 188$, $q = 0.3$, and averaging over 500 trials.

Figure 3 reports results for *PCCS* under both the global r-out-of-6 rule and the block-level r_0-out-of-2 rule. Under the global rule, *PCCS* maintains empirical FDR near $q = 0.30$ while achieving higher power than *BinCS* across all r. Under the block-level rule, *PCCS* again controls FDR near the nominal level, whereas *IT* exceeds it due to uncorrected intersection bias; *IB* and *IS* remain conservative but at much lower power, and *BinCS* controls FDR at near-zero power. *PCCS* achieves the highest power among FDR-controlling methods in both settings, though power declines for all methods as the conjunction requirement (larger r or r_2) leaves fewer eligible images.

7 Conclusion

We propose Partial Conjunction Conformal Selection (*PCCS*), a general and flexible framework for selective tasks that extends multivariate conformal selection to conjunction hypotheses. In particular, we introduce two types of nonconformity scores to construct conformal p-values, followed by the BH procedure, thereby enabling finite-sample FDR control with theoretical guarantees. Extensive simulated and real-data experiments show that *PCCS* maintains empirical FDR near the nominal level while achieving higher selection power. *PCCS* can be applied to selection tasks across diverse domains, including drug discovery, LLM response filtering, and monocular depth estimation.

AI USE STATEMENT

We used a large language model solely for improving the fluency and readability of the manuscript. The model was not involved in research ideation, experimental design, data analysis, or result interpretation. All scientific contributions and substantive content were solely produced by the authors.

REPRODUCIBILITY STATEMENT

We provide comprehensive technical details in the paper, including the methodology (Section 4), theoretical proofs (Appendix A.2), and experimental setup (Appendix A.3). The code is also provided in the supplementary material.

REFERENCES

Ignacio Aliagas, Alberto Gobbi, Man-Ling Lee, and Benjamin D Sellers. 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.

Anastasios Angelopoulos, Stephen Bates, Jitendra Malik, and Michael I Jordan. Uncertainty sets for image classifiers using conformal prediction. *arXiv preprint arXiv:2009.14193*, 2020.

Tian Bai and Ying Jin. Optimized conformal selection: Powerful selective inference after conformity score optimization. *arXiv preprint arXiv:2411.17983*, 2024.

Tian Bai, Yue Zhao, Xiang Yu, and Archer Y. Yang. Multivariate conformal selection. In *Forty-second International Conference on Machine Learning*, 2025. URL `https://openreview.net/forum?id=g2tr7nA4pS`.

Stephen Bates, Emmanuel Candès, Lihua Lei, Yaniv Romano, and Matteo Sesia. Testing for outliers with conformal p-values. *The Annals of Statistics*, 51(1):149–178, 2023.

Yoav Benjamini and Ruth Heller. Screening for partial conjunction hypotheses. *Biometrics*, 64(4): 1215–1222, 2008.

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)*, 57(1):289–300, 1995.

Yoav Benjamini and Daniel Yekutieli. The control of the false discovery rate in multiple testing under dependency. *Annals of statistics*, pp. 1165–1188, 2001.

Rodolpho C Braga, Vinicius M Alves, Meryck FB Silva, Eugene Muratov, Denis Fourches, Luciano M Lião, Alexander Tropsha, and Carolina H Andrade. Pred-herg: A novel web-accessible computational tool for predicting cardiac toxicity. *Molecular informatics*, 34(10):698–701, 2015.

Sacha Braun, Eugène Berta, Michael I Jordan, and Francis Bach. Multivariate conformal prediction via conformalized gaussian scoring. *arXiv preprint arXiv:2507.20941*, 2025.

John J Cherian, Isaac Gibbs, and Emmanuel J Candès. Large language model validity via enhanced conformal prediction methods. *Advances in Neural Information Processing Systems*, 37:114812–114842, 2024.

Lahav Dabah and Tom Tirer. On temperature scaling and conformal prediction of deep classifiers. *arXiv preprint arXiv:2402.05806*, 2024.

Monitirtha Dey, Anna Vesely, and Thorsten Dickhaus. A new approach to partial conjunction analysis in neuroimaging. *arXiv preprint arXiv:2508.03675*, 2025.

Thorsten Dickhaus, Ruth Heller, Anh-Tuan Hoang, and Yosef Rinott. A procedure for multiple testing of partial conjunction hypotheses based on a hazard rate inequality. *Bernoulli*, 32(1): 274–298, 2026.

Carmen Esposito, Shuzhe Wang, Udo EW Lange, Frank Oellien, and Sereina Riniker. Combining machine learning and molecular dynamics to predict p-glycoprotein substrates. *Journal of chemical information and modeling*, 60(10):4730–4749, 2020.

Gabriela Falcón-Cano, Christophe Molina, and Miguel Ángel Cabrera-Pérez. Reliable prediction of caco-2 permeability by supervised recursive machine learning approaches. *Pharmaceutics*, 14 (10):1998, 2022.

Andreas Futschik, Thomas Taus, and Sonja Zehetmayer. An omnibus test for the global null hypothesis. *Statistical methods in medical research*, 28(8):2292–2304, 2019.

Ulysse Gazin, Ruth Heller, Etienne Roquain, and Aldo Solari. Powerful batch conformal prediction for classification. *arXiv preprint arXiv:2411.02239*, 2024.

Yu Gui, Ying Jin, and Zhimei Ren. Conformal alignment: Knowing when to trust foundation models with guarantees. *Advances in Neural Information Processing Systems*, 37:73884–73919, 2024.

Yu Gui, Ying Jin, Yash Nair, and Zhimei Ren. Acs: An interactive framework for conformal selection. *arXiv preprint arXiv:2507.15825*, 2025.

Qingyang Hao, Wenbo Liao, Bingyi Jing, and Hongxin Wei. Multi-condition conformal selection. In *The Fourteenth International Conference on Learning Representations*, 2026. URL https://openreview.net/forum?id=giL8Q1V26J.

Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pp. 770–778, 2016.

Esther Heid, Kevin P Greenman, Yunsie Chung, Shih-Cheng Li, David E Graff, Florence H Vermeire, Haoyang Wu, William H Green, and Charles J McGill. Chemprop: a machine learning package for chemical property prediction. *Journal of chemical information and modeling*, 64(1): 9–17, 2024.

Kexin Huang, Tianfan Fu, Lucas M Glass, Marinka Zitnik, Cao Xiao, and Jimeng Sun. Deeppurpose: a deep learning library for drug–target interaction prediction. *Bioinformatics*, 36(22-23): 5545–5547, 2020.

Hiroaki Iwata, Tatsuru Matsuo, Hideaki Mamada, Takahisa Motomura, Mayumi Matsushita, Takeshi Fujiwara, Kazuya Maeda, and Koichi Handa. 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.

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

Bikram Karmakar, Dylan S Small, and Paul R Rosenbaum. Reinforced designs: Multiple instruments plus control groups as evidence factors in an observational study of the effectiveness of catholic schools. *Journal of the American Statistical Association*, 116(533):82–92, 2021.

Yuko Kato, David MJ Tax, and Marco Loog. A review of nonconformity measures for conformal prediction in regression. *Conformal and probabilistic prediction with applications*, pp. 369–383, 2023.

Navdeep Kaur, Lachlan McPheat, Alessandra Russo, Anthony G Cohn, and Pranava Madhyastha. An empirical study of conformal prediction in llm with asp scaffolds for robust reasoning. *arXiv preprint arXiv:2503.05439*, 2025.

Sunghwan Kim, Jie Chen, Tiejun Cheng, Asta Gindulyte, Jia He, Siqian He, Qingliang Li, Benjamin A Shoemaker, Paul A Thiessen, Bo Yu, et al. Pubchem 2023 update. *Nucleic acids research*, 51(D1):D1373–D1380, 2023.

Michal Klein, Louis Bethune, Eugene Ndiaye, and Marco Cuturi. Multivariate conformal prediction using optimal transport. *arXiv preprint arXiv:2502.03609*, 2025.

Yonghoon Lee and Zhimei Ren. Selection from hierarchical data with conformal e-values. *arXiv preprint arXiv:2501.02514*, 2025.

Jing Lei and Larry Wasserman. Distribution-free prediction bands for non-parametric regression. *Journal of the Royal Statistical Society Series B: Statistical Methodology*, 76(1):71–96, 2014.

B Liang, L Zhang, and L Janson. Powerful partial conjunction hypothesis testing via conditioning. *Biometrika*, 112(4):asaf036, 10 2025. ISSN 1464-3510. doi: 10.1093/biomet/asaf036. URL https://doi.org/10.1093/biomet/asaf036.

Kangdao Liu, Huajun Xi, Chi Man Vong, and Hongxin Wei. Online conformal selection with accept-to-reject changes. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 40, pp. 23765–23773, 2026.

Zhonghua Liu, Jincheng Shen, Richard Barfield, Joel Schwartz, Andrea A Baccarelli, and Xihong Lin. Large-scale hypothesis testing for causal mediation effects with applications in genome-wide epigenetic studies. *Journal of the American Statistical Association*, 117(537):67–81, 2022.

Jintao Meng, Peng Chen, Mohamed Wahib, Mingjun Yang, Liangzhen Zheng, Yanjie Wei, Shengzhong Feng, and Wei Liu. Boosting the predictive performance with aqueous solubility dataset curation. *Scientific Data*, 9(1):71, 2022.

Louis Meyer and Wenkai Xu. A kernel nonconformity score for multivariate conformal prediction. *arXiv preprint arXiv:2604.21595*, 2026.

Yash Nair, Ying Jin, James Yang, and Emmanuel Candes. Diversifying conformal selections. *arXiv preprint arXiv:2506.16229*, 2025.

Alexander L Perryman, Daigo Inoyama, Jimmy S Patel, Sean Ekins, and Joel S Freundlich. Pruned machine learning models to predict aqueous solubility. *ACS omega*, 5(27):16562–16567, 2020.

Robert P Sheridan, Daniel R McMasters, Johannes H Voigt, and Mary Jo Wildey. 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.

Nathan Silberman, Derek Hoiem, Pushmeet Kohli, and Rob Fergus. Indoor segmentation and support inference from rgbd images. In *European conference on computer vision*, pp. 746–760. Springer, 2012.

Yam Tawachi and Bracha Laufer-Goldshtein. Multi-dimensional conformal prediction. In *International Conference on Learning Representations*, volume 2025, pp. 35491–35520, 2025.

Florence H Vermeire, Yunsie Chung, and William H Green. 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.

Vladimir Vovk, Alexander Gammerman, and Glenn Shafer. *Algorithmic learning in a random world*. Springer, 2005.

Zhilin Wang, Alexander Bukharin, Olivier Delalleau, Daniel Egert, Gerald Shen, Jiaqi Zeng, Oleksii Kuchaiev, and Yi Dong. Helpsteer2-preference: Complementing ratings with preferences. *arXiv preprint arXiv:2410.01257*, 2024a.

Zhilin Wang, Yi Dong, Olivier Delalleau, Jiaqi Zeng, Gerald Shen, Daniel Egert, Jimmy J Zhang, Makesh N Sreedhar, and Oleksii Kuchaiev. Helpsteer 2: Open-source dataset for training top-performing reward models. *Advances in Neural Information Processing Systems*, 37:1474–1501, 2024b.

Jan Wenzel, Hans Matter, and Friedemann Schmidt. 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.

Anyi Xu, B Li, Bangcai Lin, Bing Xue, BingCheng Xian, Bingzheng Xu, Bochao Wu, Bowei Zhang, Boyi Deng, CC Yu, et al. Deepseek-v4. 1-flash: Pushing the limits of kv cache compression. *arXiv preprint arXiv:2609.19969*, 2026.

Chen Xu, Hanyang Jiang, and Yao Xie. Conformal prediction for multi-dimensional time series by ellipsoidal sets. *arXiv preprint arXiv:2403.03850*, 2024.

Kevin Yang, Kyle Swanson, Wengong Jin, Connor Coley, Philipp Eiden, Hua Gao, Angel Guzman-Perez, Timothy Hopper, Brian Kelley, Miriam Mathea, et al. Analyzing learned molecular representations for property prediction. *Journal of chemical information and modeling*, 59(8):3370, 2019.

A APPENDIX

A.1 RELATED WORK

Conformal Selection Conformal prediction (CP) (Vovk et al., 2005) is a popular framework for uncertainty quantification that constructs prediction sets with finite-sample coverage guarantees under the assumption of exchangeability. CP has been generalized to the multivariate setting by a number of studies (Gazin et al., 2024; Xu et al., 2024; Braun et al., 2025; Klein et al., 2025; Tawachi & Laufer-Goldshtein, 2025; Meyer & Xu, 2026), and applied across domains, including regression (Lei & Wasserman, 2014; Kato et al., 2023), classification (Angelopoulos et al., 2020; Dabah & Tirer, 2024), and text generation (Cherian et al., 2024; Kaur et al., 2025).

cfBH (Jin & Candès, 2023) extends the CP framework to selection tasks, which require controlling the FDR over a selected set rather than constructing prediction sets with coverage guarantees. Subsequent work has extended this foundation to address diverse practical challenges. For instance, conformal e-values (Lee & Ren, 2025) address the selection problem for hierarchical data, DACS (Nair et al., 2025) constructs diverse selection sets, OptCS (Bai & Jin, 2024) enables data-driven model optimization without compromising validity, and ACS (Gui et al., 2025) dynamically updates the prediction model during selection to boost power. Some studies generalize CS from the univariate to the multivariate setting (Bai et al., 2025; Hao et al., 2026; Liu et al., 2026). However, these multivariate CS methods target the full conjunction $\mathbf{y} \in R_1 \times \cdots \times R_d$ and are not directly applicable to partial conjunction rules. Our work extends this line of research by addressing conjunction hypotheses while maintaining theoretical guarantees, a setting that arises in various real-world applications.

Partial Conjunction Hypothesis Testing A partial conjunction hypothesis test combines information across a set of base hypotheses to determine whether some subset is nonnull. Partial conjunction hypothesis tests arise in diverse fields, including genomics (Liu et al., 2022; Liang et al., 2025), neuroimaging (Dey et al., 2025), and social policy evaluation (Karmakar et al., 2021). The combined p-value approach of Benjamini & Heller (2008) is one of the first methods proposed for this task, and it provides a procedure that controls the false discovery rate (FDR). Because the combined p-value approach can be highly conservative, several refinements have been proposed: Futschik et al. (2019) use cumulative sums of the transformed p-values, Liang et al. (2025) condition on certain order statistics of the base p-values, and Dickhaus et al. (2026) adopt a two-step scheme that performs conditional testing after filtering. However, these advances assume access to valid p-values, and do not naturally extend to the conformal selection setting, where only conformal p-values are available.

A.2 PROOF

A.2.1 PROOF OF PROPOSITION 1

Proof. **Case 1**: $r = 1$. Let all component nulls hold almost surely: $\mathbb{P}(H_{0,k}) = 1$ for every $k \in [d]$ and let $p_1, \ldots, p_d \overset{\text{i.i.d.}}{\sim} \text{Unif}(0,1)$. The 1-out-of-$d$ null $H_0^{1/d} = \bigcap_{k=1}^{d} H_{0,k}$ therefore holds almost surely. Moreover, for every $k \in [d]$ and $\alpha \in [0,1]$,

$$\mathbb{P}(p_k \le \alpha, H_{0,k}) = \mathbb{P}(p_k \le \alpha) = \alpha,$$

so componentwise joint validity holds. Since $p_{(1)} = \min_{k\in[d]} p_k$, independence gives

$$\begin{aligned}\mathbb{P}\Big(p_{(1)} \le \alpha, H_0^{1/d}\Big) &= \mathbb{P}\left(\min_{k\in[d]} p_k \le \alpha\right), \\ &= 1 - \mathbb{P}(p_k > \alpha,\ \forall k \in [d]), \\ &= 1 - (1-\alpha)^d.\end{aligned}$$

For $d \ge 2$ and $\alpha \in (0,1)$,

$$1 - (1-\alpha)^d > \alpha.$$

Case 2: $2 \le r \le d$. Let $\mathcal{N}$ be selected uniformly at random from all subsets of $[d]$ with cardinality $|\mathcal{N}| = d - r + 1$. Define the component-null events by $H_{0,k} = \{k \in \mathcal{N}\}$ and $H_{1,k} = H_{0,k}^c$.

Thus, exactly $d-r+1$ component nulls and $r-1$ component alternatives hold in every realization. Consequently,

$$\sum_{k=1}^{d} \mathbb{1}\{H_{1,k}\} = r-1 < r,$$

so $H_0^{r/d}$ holds almost surely.

By symmetry,

$$\mathbb{P}(H_{0,k}) = \frac{d-r+1}{d}, \quad k \in [d].$$

Let $U \sim \mathrm{Unif}(0,1)$ be independent of $\mathcal{N}$ and define $p_1 = \cdots = p_d = \frac{d-r+1}{d}U$. For every $k \in [d]$ and $\alpha \in [0,1]$,

$$\begin{aligned}
\mathbb{P}(p_k \le \alpha, H_{0,k}) &= \mathbb{P}(H_{0,k})\mathbb{P}\left(\frac{d-r+1}{d}U \le \alpha\right) \\
&= \frac{d-r+1}{d}\min\left\{\frac{d}{d-r+1}\alpha, 1\right\} \\
&= \min\left\{\alpha, \frac{d-r+1}{d}\right\} \\
&\le \alpha.
\end{aligned}$$

Hence, every component satisfies joint validity.

Because all component p-values are identical, $p_{(r)} = \frac{d-r+1}{d}U$. Since $H_0^{r/d}$ holds almost surely,

$$\begin{aligned}
\mathbb{P}\left(p_{(r)} \le \alpha, H_0^{r/d}\right) &= \mathbb{P}\left(\frac{d-r+1}{d}U \le \alpha\right) \\
&= \min\left\{\frac{d}{d-r+1}\alpha, 1\right\}.
\end{aligned}$$

For $\alpha \in \left(0, \frac{d-r+1}{d}\right)$, the minimum is attained by its first argument. Since $r \ge 2$ implies $\frac{d}{d-r+1} > 1$, we obtain

$$\mathbb{P}\left(p_{(r)} \le \alpha, H_0^{r/d}\right) = \frac{d}{d-r+1}\alpha > \alpha.$$

Finally, consider a single test candidate, $m = 1$. In this case, the BH procedure at level q rejects if and only if $p_{(r)} \le q$. In both constructions, $H_0^{r/d}$ holds almost surely, so every rejection is a false discovery. Therefore,

$$\mathrm{FDR} = \mathbb{P}\left(p_{(r)} \le q, H_0^{r/d}\right) > q$$

for every $q \in (0,1)$ when $r = 1$, and for every $q \in (0, (d-r+1)/d)$ when $2 \le r \le d$. □

A.2.2 Proof of Proposition 2

Proof. Fix $\mathbf{x} \in \mathcal{X}$, $\mathbf{y} \in \mathcal{R}^{r/d}$, and $\mathbf{y}' \notin \mathcal{R}^{r/d}$. Since $\mathbf{y} \in \mathcal{R}^{r/d}$,

$$V^{r/d}(\mathbf{x}, \mathbf{y}) = M - D^{\downarrow}_{[r]}(\mathbf{x}).$$

Since $\mathbf{y}' \notin \mathcal{R}^{r/d}$,

$$V^{r/d}(\mathbf{x}, \mathbf{y}') = -D^{\downarrow}_{[r]}(\mathbf{x}).$$

Therefore,

$$V^{r/d}(\mathbf{x}, \mathbf{y}) - V^{r/d}(\mathbf{x}, \mathbf{y}') = M > 0,$$

which proves $V^{r/d}(\mathbf{x}, \mathbf{y}') < V^{r/d}(\mathbf{x}, \mathbf{y})$. □

A.2.3 PROOF OF LEMMA 1

Proof. If $\mathbf{y} \notin \mathcal{R}^{r/d}$, then $r^*(\mathbf{y}) < r$, and hence

$$\mathbb{1}\{r^*(\mathbf{y}) \geq r\} = 0.$$

Therefore,

$$V^{r/d}(\mathbf{x}, \mathbf{y}) = -D^{\downarrow}_{[r]}(\mathbf{x}) = \widehat{V}^{r/d}(\mathbf{x}).$$

□

A.2.4 PROOF OF PROPOSITION 3

Proof. Fix $j \in [m]$. By Lemma 1, on the null event $H^{r/d}_{0j}$, the practical score equals the oracle score: $\widehat{V}^{r/d}_{n+j} = V^{r/d}_{n+j}$. Consequently, $p_j = p^*_j$ on $H^{r/d}_{0j}$, where $p^*_j = \frac{1+\sum_{i=1}^n \mathbb{1}\{V_i^{r/d} \leq V^{r/d}_{n+j}\}}{n+1}$ is the oracle conformal p-value. Therefore, for every $\alpha \in (0, 1)$,

$$\begin{aligned}\mathbb{P}\left(p_j \leq \alpha,\ H^{r/d}_{0j}\right) &= \mathbb{P}\left(p^*_j \leq \alpha,\ H^{r/d}_{0j}\right) \\ &\leq \mathbb{P}\left(p^*_j \leq \alpha\right).\end{aligned}$$

By exchangeability of $V^{r/d}_1, \ldots, V^{r/d}_n, V^{r/d}_{n+j}$, the deterministic rank-based conformal p-value p^*_j is super-uniform, including in the presence of ties. Hence,

$$\mathbb{P}\left(p^*_j \leq \alpha\right) \leq \alpha.$$

Combining the preceding inequalities gives

$$\mathbb{P}\left(p_j \leq \alpha,\ H^{r/d}_{0j}\right) \leq \alpha,$$

which proves the claim. □

A.2.5 PROOF OF THEOREM 1

Proof. For each $j \in [m]$, define the practical and oracle conformal p-values as

$$p_j = \frac{1 + \sum_{i=1}^n \mathbb{1}\{V_i^{r/d} \leq \widehat{V}^{r/d}_{n+j}\}}{n+1}, \qquad p^*_j = \frac{1 + \sum_{i=1}^n \mathbb{1}\{V_i^{r/d} \leq V^{r/d}_{n+j}\}}{n+1}.$$

Let $\mathcal{E}_j = \{j \in \mathcal{S},\ H^{r/d}_{0j}\} = \{j \in \mathcal{S},\ \mathbf{y}_{n+j} \notin \mathcal{R}^{r/d}\}$. By Lemma 1, on $H^{r/d}_{0j}$, we have $V^{r/d}_{n+j} = \widehat{V}^{r/d}_{n+j}$ and $p^*_j = p_j$. Fix $j \in [m]$. For every $\ell \in [m]$, define

$$p^{(j)}_\ell = \frac{\sum_{i=1}^n \mathbb{1}\{V_i^{r/d} \leq \widehat{V}^{r/d}_{n+\ell}\} + \mathbb{1}\{V^{r/d}_{n+j} \leq \widehat{V}^{r/d}_{n+\ell}\}}{n+1}.$$

Let $\mathcal{S}(a_1, \ldots, a_m)$ denote the BH rejection set constructed from $(a_1, \ldots, a_m)$, and define $\mathcal{S}^*_j = \mathcal{S}\left(p^{(j)}_1, \ldots, p^{(j)}_{j-1}, p^*_j, p^{(j)}_{j+1}, \ldots, p^{(j)}_m\right)$. We claim that $\mathcal{S}^*_j = \mathcal{S}$ on $\mathcal{E}_j$. To see this, consider any $\ell \neq j$. If $\widehat{V}^{r/d}_{n+\ell} \geq \widehat{V}^{r/d}_{n+j}$, then, on $\mathcal{E}_j$,

$$V^{r/d}_{n+j} = \widehat{V}^{r/d}_{n+j} \leq \widehat{V}^{r/d}_{n+\ell},$$

and hence $p^{(j)}_\ell = p_\ell$. If instead $\widehat{V}^{r/d}_{n+\ell} < \widehat{V}^{r/d}_{n+j}$, then $p_\ell \leq p_j$ and $p^{(j)}_\ell \leq p_\ell$. Since $j \in \mathcal{S}$, every such ℓ is already rejected by BH. Thus, only already rejected p-values are decreased, while all nonrejected p-values remain unchanged. The largest index satisfying the BH inequality therefore remains unchanged, proving $\mathcal{S}^*_j = \mathcal{S}$ on $\mathcal{E}_j$.

Next, define $\mathcal{S}^*_{j\to 0} = \mathcal{S}\left(p^{(j)}_1, \ldots, p^{(j)}_{j-1}, 0, p^{(j)}_{j+1}, \ldots, p^{(j)}_m\right)$. Whenever $j \in \mathcal{S}^*_j$, lowering the already rejected p^*_j to zero changes only its position within the rejected portion of the ordered p-values. All

nonrejected ordered p-values remain unchanged, so the largest index satisfying the BH inequality is unchanged. Hence, $\mathcal{S}^*_{j\to 0} = \mathcal{S}^*_j$ on $\{j \in \mathcal{S}^*_j\}$.

It follows that

$$\begin{aligned}\mathrm{FDR} &= \sum_{j=1}^{m} \mathbb{E}\left[\frac{\mathbb{1}\{H_{0j}^{r/d}, j \in \mathcal{S}\}}{1 \vee |\mathcal{S}|}\right] \\ &\le \sum_{j=1}^{m} \mathbb{E}\left[\frac{\mathbb{1}\{H_{0j}^{r/d}, j \in \mathcal{S}^*_j\}}{1 \vee |\mathcal{S}^*_j|}\right] \\ &\le \sum_{j=1}^{m} \mathbb{E}\left[\frac{\mathbb{1}\left\{p^*_j \le q|\mathcal{S}^*_{j\to 0}|/m\right\}}{1 \vee |\mathcal{S}^*_{j\to 0}|}\right]. \end{aligned} \tag{10}$$

The first inequality holds termwise: on $\mathcal{E}_j = \{H_{0j}^{r/d}, j \in \mathcal{S}\}$ we have $\mathcal{S}^*_j = \mathcal{S}$, while outside $\mathcal{E}_j$ the summand on the left is zero. The second inequality holds because $j \in \mathcal{S}^*_j$ implies $p^*_j \le q|\mathcal{S}^*_j|/m$ by the BH rule, and $\mathcal{S}^*_{j\to 0} = \mathcal{S}^*_j$ on $\{j \in \mathcal{S}^*_j\}$.

Let $\mathcal{G}_j$ be the sigma-field generated by the unordered multiset $[V_1^{r/d}, \dots, V_n^{r/d}, V_{n+j}^{r/d}]$ and the remaining practical test scores $\{\widehat{V}_{n+\ell}^{r/d}\}_{\ell \ne j}$. The set $\mathcal{S}^*_{j\to 0}$ is $\mathcal{G}_j$-measurable, because each $p_\ell^{(j)}$, $\ell \ne j$, is a symmetric function of the oracle scores. By the assumed conditional exchangeability, the deterministic oracle conformal p-value p^*_j is conditionally super-uniform given $\mathcal{G}_j$. Therefore, for every $\mathcal{G}_j$-measurable random variable $T_j \in [0,1]$, $\mathbb{P}\big(p^*_j \le T_j \mid \mathcal{G}_j\big) \le T_j$. Taking $T_j = \frac{q|\mathcal{S}^*_{j\to 0}|}{m}$ gives

$$\mathbb{P}\left(p^*_j \le \frac{q|\mathcal{S}^*_{j\to 0}|}{m} \,\middle|\, \mathcal{G}_j\right) \le \frac{q|\mathcal{S}^*_{j\to 0}|}{m}.$$

Applying the tower property gives

$$\begin{aligned}&\mathbb{E}\left[\frac{\mathbb{1}\left\{p^*_j \le q|\mathcal{S}^*_{j\to 0}|/m\right\}}{1 \vee |\mathcal{S}^*_{j\to 0}|}\right] \\ &= \mathbb{E}\left[\frac{\mathbb{P}\left(p^*_j \le q|\mathcal{S}^*_{j\to 0}|/m \mid \mathcal{G}_j\right)}{1 \vee |\mathcal{S}^*_{j\to 0}|}\right] \\ &\le \mathbb{E}\left[\frac{q|\mathcal{S}^*_{j\to 0}|/m}{1 \vee |\mathcal{S}^*_{j\to 0}|}\right] \le \frac{q}{m}.\end{aligned}$$

Summing over $j = 1, \dots, m$ yields $\mathrm{FDR} \le \sum_{j=1}^m \frac{q}{m} = q$.

□

A.2.6 Proof of Theorem 2

Proof. Define the empirical distribution function of the calibration scores by $\widehat{F}_n(v) = \frac{1}{n}\sum_{i=1}^n \mathbb{1}\{V_i^{r/d} \le v\}$. Because $V^{r/d}$ is fixed or constructed using an independent training sample, the calibration scores $\{V_i^{r/d}\}_{i=1}^n$ are i.i.d. with distribution function $F_V(v) = \mathbb{P}\big(V^{r/d}(\mathbf{x}, \mathbf{y}) \le v\big)$. Therefore, by the Glivenko–Cantelli theorem,

$$\sup_{v\in\mathbb{R}} \left|\widehat{F}_n(v) - F_V(v)\right| \longrightarrow 0 \qquad \text{almost surely.}$$

For each test observation j, the practical conformal p-value is

$$p_j = \frac{1 + \sum_{i=1}^n \mathbb{1}\{V_i^{r/d} \le \widehat{V}_{n+j}^{r/d}\}}{n+1} = \frac{1 + n\widehat{F}_n(\widehat{V}_{n+j}^{r/d})}{n+1}.$$

Define $p_j^\infty = F_V(\widehat{V}_{n+j}^{r/d})$. Then

$$\sup_{1\le j\le m} |p_j - p_j^\infty| \le \sup_{v\in\mathbb{R}} |\widehat{F}_n(v) - F_V(v)| + \frac{1}{n+1}$$
$$\longrightarrow 0 \qquad \text{almost surely.} \tag{11}$$

Since the test observations are i.i.d., $\{p_j^\infty\}_{j=1}^m$ are i.i.d. copies of $p^\infty = F_V\left(\widehat{V}^{r/d}(\mathbf{x})\right)$.

Let $\widehat{G}_m(t) = \frac{1}{m}\sum_{j=1}^m \mathbb{1}\{p_j \le t\}$ denote the empirical distribution function of the finite-sample practical p-values. By equation 11, the Glivenko–Cantelli theorem applied to $\{p_j^\infty\}_{j=1}^m$, and the continuity of G on $[0,1]$,

$$\sup_{t\in[0,1]} \left|\widehat{G}_m(t) - G(t)\right| \longrightarrow 0 \qquad \text{almost surely.}$$

Define the empirical BH threshold by

$$\widehat{t}_q = \max\left\{\frac{qk}{m} : p_{(k)} \le \frac{qk}{m}, \quad k = 1,\dots,m\right\},$$

with $\widehat{t}_q = 0$ if the set is empty. By the uniform convergence of $\widehat{G}_m$ to G and the assumed stability of the population BH crossing,

$$\widehat{t}_q \longrightarrow t_q \qquad \text{almost surely.}$$

Moreover, since G is continuous at t_q, $\mathbb{P}(p^\infty = t_q) = 0$. Let

$$H^0(t) = \mathbb{P}(p^\infty \le t,\, r^*(\mathbf{y}) < r)\,.$$

Since $|H^0(t) - H^0(s)| \le |G(t) - G(s)|$ for $s \le t$, H^0 is continuous at t_q. By the Glivenko–Cantelli theorem applied to the i.i.d. pairs $\{(p_j^\infty, \mathbb{1}\{r^*(\mathbf{y}_{n+j}) < r\})\}_{j=1}^m$, their empirical subdistribution converges uniformly to H^0. Moreover, equation 11 gives $\sup_{1\le j\le m} |p_j - p_j^\infty| \to 0$ almost surely. Combining these results with $\widehat{t}_q \to t_q$ and the continuity of H^0 and G at t_q yields

$$\frac{1}{m}\sum_{j=1}^m \mathbb{1}\{p_j \le \widehat{t}_q,\, r^*(\mathbf{y}_{n+j}) < r\} \longrightarrow \mathbb{P}(p^\infty \le t_q,\, r^*(\mathbf{y}) < r)\,, \tag{12}$$

$$\frac{|\mathcal{S}_q|}{m} \longrightarrow \mathbb{P}(p^\infty \le t_q). \tag{13}$$

Since $\mathbb{P}(p^\infty \le t_q) > 0$, equations 12–13 yield

$$\mathrm{FDP}(\mathcal{S}_q) \longrightarrow \frac{\mathbb{P}(p^\infty \le t_q,\, r^*(\mathbf{y}) < r)}{\mathbb{P}(p^\infty \le t_q)} \qquad \text{almost surely.}$$

Because $0 \le \mathrm{FDP}(\mathcal{S}_q) \le 1$, the bounded convergence theorem gives

$$\mathrm{FDR}(\mathcal{S}_q) \longrightarrow \frac{\mathbb{P}(p^\infty \le t_q,\, r^*(\mathbf{y}) < r)}{\mathbb{P}(p^\infty \le t_q)}.$$

Similarly, applying the same empirical-subdistribution argument with $r^*(\mathbf{y}_{n+j}) \ge r$ in place of $r^*(\mathbf{y}_{n+j}) < r$ gives

$$\frac{1}{m}\sum_{j=1}^m \mathbb{1}\{p_j \le \widehat{t}_q,\, r^*(\mathbf{y}_{n+j}) \ge r\} \longrightarrow \mathbb{P}(p^\infty \le t_q,\, r^*(\mathbf{y}) \ge r)$$

and

$$\frac{1}{m}\sum_{j=1}^m \mathbb{1}\{r^*(\mathbf{y}_{n+j}) \ge r\} \longrightarrow \mathbb{P}(r^*(\mathbf{y}) \ge r) > 0$$

almost surely. Hence, the realized true-positive proportion converges almost surely to

$$\frac{\mathbb{P}(p^\infty \le t_q,\, r^*(\mathbf{y}) \ge r)}{\mathbb{P}(r^*(\mathbf{y}) \ge r)}.$$

Since this proportion is bounded between zero and one, bounded convergence yields

$$\mathrm{Power}(\mathcal{S}_q) \longrightarrow \frac{\mathbb{P}(p^\infty \le t_q,\, r^*(\mathbf{y}) \ge r)}{\mathbb{P}(r^*(\mathbf{y}) \ge r)}.$$

This completes the proof. □

A.2.7 Proof of Proposition 4

Proof. Fix $\mathbf{x} \in \mathcal{X}$, $\mathbf{y} \in \mathcal{R}^{r_0/L}$, and $\mathbf{y}' \notin \mathcal{R}^{r_0/L}$. By Equation 9,

$$V^{r_0/L}(\mathbf{x}, \mathbf{y}) = M - D^{\downarrow}_{\text{block},[r_0]}(\mathbf{x}),$$

whereas $V^{r_0/L}(\mathbf{x}, \mathbf{y}') = -D^{\downarrow}_{\text{block},[r_0]}(\mathbf{x})$. Therefore,

$$V^{r_0/L}(\mathbf{x}, \mathbf{y}) - V^{r_0/L}(\mathbf{x}, \mathbf{y}') = M > 0,$$

which proves the result. □

A.2.8 Proof of Theorem 3

Proof. For each test observation j, define the oracle block-level conformal p-value as

$$p_j^* = \frac{1 + \sum_{i=1}^{n} \mathbb{1}\left\{V_i^{r_0/L} \le V_{n+j}^{r_0/L}\right\}}{n+1}.$$

On the block-level null event $H_{0j}^{r_0/L}$, $\sum_{\ell=1}^{L} B_\ell(\mathbf{y}_{n+j}) < r_0$. Hence, the indicator in Equation 9 equals zero, and therefore

$$V_{n+j}^{r_0/L} = -D^{\downarrow}_{\text{block},[r_0]}(\mathbf{x}_{n+j}) = \widehat{V}_{n+j}^{r_0/L}.$$

Consequently, $p_j = p_j^*$ on $H_{0j}^{r_0/L}$. Under the conditional exchangeability assumption in Theorem 1, the oracle block-level conformal p-value is conditionally super-uniform. The leave-one-out BH argument in the proof of Theorem 1 therefore applies with $V^{r/d}$ and $\widehat{V}^{r/d}$ replaced by $V^{r_0/L}$ and $\widehat{V}^{r_0/L}$, respectively. Thus,

$$\text{FDR} \le q.$$

□

A.3 Details of Simulation Studies

Data Generating Processes The synthetic datasets are systematically constructed under controlled conditions to evaluate methodological performance across varying degrees of complexity, ranging from linear to highly nonlinear relationships, combined with different noise characteristics. Following the simulation settings of Bai et al. (2025), we consider the regression functions f and noise distributions $\epsilon \in \mathbb{R}^d$ summarized in Table 3.

Features are drawn as $\mathbf{x} \sim \text{Unif}(-1, 1)^p$ with $p = 10$, and indices of $\mathbf{x}$ are taken modulo p; hence, for $d > p$ (e.g., $d = 30$), coordinates k and $k + p$ share the same mean function. The noise scale matrix Σ has diagonal entries 0.5 and off-diagonal entries 0.05, and the t settings use $\nu = 3$ degrees of freedom. In each replication, independent training, calibration, validation, and test sets of size 1000 are generated. Each response coordinate is standardized by its training mean and standard deviation, and a threshold shared across coordinates is chosen on the validation set so that approximately a fraction π of validation samples satisfies the selection rule: $\pi = 0.3$ for the half-line and interval regions and $\pi = 0.5$ for the union and complex regions. In the block-level experiments, responses are not standardized, and each block receives its own threshold with $\pi = 0.5$ for the within-block requirement. Since the validation set is independent of the calibration and test sets, the target region is fixed before these data are used.

All conjunction-aware scores use $M = 1000$. The ablation with support vector regression uses an RBF kernel with $\gamma = 0.1$. ***BinCS*** fits the same model family to the binary rule indicator (logistic regression in the block-level experiments). For the union and complex regions, the componentwise p-values used by the partial conjunction baselines combine the p-values of the individual pieces of R_k by a Bonferroni correction. The ***Omnibus*** null distribution is approximated with 10^4 Monte Carlo draws, and ***IS*** selects a common per-block level from 40 equally spaced values in $[q/20, q]$ on the validation set. Results are averaged over 100 replications.

Table 5: Configurations of the block-structured conjunction hypotheses. The two blocks satisfy $|G_1| = |G_2| = 5$. A partial conjunction block uses threshold $r_\ell = 3$.

Task	r_1	r_2
Disjunctive–Partial Conjunction	1	3
Disjunctive–Full Conjunction	1	5
Partial Conjunction–Partial Conjunction	3	3
Disjunctive–Disjunctive	1	1

Table 3: True regression functions and noise distributions

Setting	$[f(\cdot)]_k$	ϵ
1	$2x_k - \frac{1}{2}x_{k+1} + x_{k+2} + \frac{3}{2}$	$\mathcal{N}(0, \Sigma)$
2	$x_k + x_{k+2}^2 + \frac{1}{2}$	$\mathcal{N}(0, \Sigma)$
3	Same as Setting 1	$t_3(0, \Sigma)$
4	Same as Setting 2	$t_3(0, \Sigma)$

Target Region Table 4 summarizes the four target regions and their signed margins for each dimension. We constructed four distinct target regions, covering scenarios involving a half-line, an interval, a union of two unbounded intervals, and a more complex piecewise region. For the signed margin $D_k\left(\widehat{f}_k(\mathbf{x}), R_k\right)$, we adopt the existing signed margins for the half-line and interval regions from Jin & Candès (2023) and Hao et al. (2026), respectively, and introduce new signed margins for the union and complex regions. For the union region $R_k = (-\infty, c_{k,1}) \cup (c_{k,2}, \infty)$, we define the signed margin as $D_k(z, R_k) = \max\{c_{k,1} - z,\ z - c_{k,2}\}$. Under our sign convention, $D_k(z, R_k)$ is positive when $z \in R_k$, zero when $z \in \partial R_k$, and negative when $z \in (c_{k,1}, c_{k,2})$. Its absolute value equals the distance to the nearest boundary point. For the complex region $R_k = (-\infty, c_{k,1}) \cup (c_{k,2}, c_{k,3}) \cup (c_{k,4}, \infty)$ and $c_{k,1} < c_{k,2} < c_{k,3} < c_{k,4}$, we define $D_k(z, R_k) = \max\left\{c_{k,1} - z,\ \min\{z - c_{k,2},\ c_{k,3} - z\},\ z - c_{k,4}\right\}$. This signed margin is positive when $z \in R_k$, zero when $z \in \partial R_k$, and negative when z lies in either excluded gap $(c_{k,1}, c_{k,2})$ or $(c_{k,3}, c_{k,4})$.

Table 4: Configurations of the target region and the corresponding signed margin.

Region Type	Target Region R_k	Signed margin $D_k(z, R_k)$
Half-line Region	(c_k, ∞)	$z - c_k$
Interval Region	$(c_{k,1}, c_{k,2})$	$\min(z - c_{k,1},\ c_{k,2} - z)$
Union Region	$(-\infty, c_{k,1}) \cup (c_{k,2}, \infty)$	$\max(c_{k,1} - z,\ z - c_{k,2})$
Complex Region	$(-\infty, c_{k,1}) \cup (c_{k,2}, c_{k,3}) \cup (c_{k,4}, \infty)$	$\max(c_{k,1} - z,\ \min(z - c_{k,2},\ c_{k,3} - z),\ z - c_{k,4})$

Table 5 summarizes the four block-structured conjunction tasks. We consider two disjoint blocks G_1 and G_2 with $|G_1| = |G_2| = 5$ and $d = |G_1| + |G_2| = 10$. For a partial conjunction block, we set the required number of satisfied criteria to $r_\ell = 3$.

Table 6: Empirical FDR and power for the global r-out-of-d conjunction rule across varying values of r, with dimension fixed at $d = 5$. The data-generating process follows simulation setting 1. The nominal FDR level is $q = 0.30$, support vector regression is used, and the target region is $R_k = (c_k, \infty)$ for all $k = 1, \ldots, 5$. Results are reported as mean $\pm$ standard deviation. Boldface indicates the best-performing method among those that approximately control the FDR.

Metric	r	*Bonferroni*	*Omnibus*	*BinCS*	*PCCS*
FDR	1	0.222 ± 0.120	0.245 ± 0.059	0.292 ± 0.061	0.299 ± 0.049
	2	0.015 ± 0.040	0.087 ± 0.068	0.303 ± 0.049	0.295 ± 0.047
	3	0.000 ± 0.005	0.116 ± 0.046	0.301 ± 0.046	0.302 ± 0.043
	4	0.055 ± 0.051	0.237 ± 0.061	0.301 ± 0.046	0.298 ± 0.045
	5	0.701 ± 0.019	0.701 ± 0.019	0.306 ± 0.046	0.287 ± 0.040
Power	1	0.244 ± 0.112	0.375 ± 0.103	0.379 ± 0.125	$\mathbf{0.471 \pm 0.114}$
	2	0.018 ± 0.043	0.149 ± 0.122	0.570 ± 0.088	$\mathbf{0.596 \pm 0.101}$
	3	0.003 ± 0.013	0.284 ± 0.096	0.600 ± 0.081	$\mathbf{0.660 \pm 0.076}$
	4	0.167 ± 0.138	0.598 ± 0.118	0.629 ± 0.079	$\mathbf{0.690 \pm 0.072}$
	5	1.000 ± 0.000	1.000 ± 0.000	0.565 ± 0.084	$\mathbf{0.654 \pm 0.059}$

Table 7: Empirical FDR and power for the global r-out-of-d conjunction rule across varying values of the conjunction parameter r, with the dimension fixed at $d = 5$. The data-generating process follows simulation setting 1. The nominal FDR level is $q = 0.05$, multivariate linear regression is used, and the target region is $R_k = (c_k, \infty)$ for all $k = 1, \ldots, 5$. Results are reported as mean $\pm$ standard deviation. Boldface indicates the best-performing method among those that control the FDR.

Metric	r	*Bonferroni*	*Omnibus*	*BinCS*	*PCCS*
FDR	1	0.047 ± 0.183	0.013 ± 0.035	0.047 ± 0.162	0.048 ± 0.071
	2	0.000 ± 0.000	0.003 ± 0.016	0.044 ± 0.088	0.040 ± 0.051
	3	0.000 ± 0.000	0.001 ± 0.009	0.042 ± 0.054	0.041 ± 0.043
	4	0.000 ± 0.000	0.004 ± 0.011	0.041 ± 0.046	0.045 ± 0.041
	5	0.003 ± 0.010	0.004 ± 0.015	0.058 ± 0.131	0.051 ± 0.038
Power	1	0.002 ± 0.004	0.015 ± 0.031	0.003 ± 0.013	$\mathbf{0.037 \pm 0.050}$
	2	0.000 ± 0.000	0.004 ± 0.016	0.024 ± 0.039	$\mathbf{0.087 \pm 0.083}$
	3	0.000 ± 0.000	0.008 ± 0.019	0.095 ± 0.080	$\mathbf{0.115 \pm 0.091}$
	4	0.000 ± 0.000	0.032 ± 0.042	0.093 ± 0.073	$\mathbf{0.160 \pm 0.095}$
	5	0.016 ± 0.040	0.012 ± 0.036	0.031 ± 0.044	$\mathbf{0.189 \pm 0.093}$

Table 8: Empirical FDR and power for the global r-out-of-d conjunction rule across varying values of the conjunction parameter r, with the dimension fixed at $d = 5$. The data-generating process follows simulation setting 1. The nominal FDR level is $q = 0.10$, multivariate linear regression is used, and the target region is $R_k = (c_k, \infty)$ for all $k = 1, \ldots, 5$. Results are reported as mean $\pm$ standard deviation. Boldface indicates the best-performing method among those that control the FDR.

Metric	r	*Bonferroni*	*Omnibus*	*BinCS*	*PCCS*
FDR	1	0.079 ± 0.159	0.085 ± 0.072	0.077 ± 0.152	0.099 ± 0.120
	2	0.001 ± 0.007	0.012 ± 0.032	0.098 ± 0.086	0.095 ± 0.055
	3	0.000 ± 0.000	0.019 ± 0.040	0.098 ± 0.048	0.098 ± 0.046
	4	0.000 ± 0.000	0.024 ± 0.029	0.093 ± 0.053	0.102 ± 0.044
	5	0.049 ± 0.041	0.048 ± 0.044	0.112 ± 0.123	0.100 ± 0.039
Power	1	0.031 ± 0.057	0.089 ± 0.070	0.012 ± 0.021	$\mathbf{0.107 \pm 0.086}$
	2	0.001 ± 0.013	0.024 ± 0.041	0.084 ± 0.066	$\mathbf{0.204 \pm 0.098}$
	3	0.000 ± 0.000	0.045 ± 0.038	0.207 ± 0.085	$\mathbf{0.251 \pm 0.097}$
	4	0.001 ± 0.009	0.115 ± 0.070	0.203 ± 0.086	$\mathbf{0.320 \pm 0.088}$
	5	0.168 ± 0.096	0.158 ± 0.105	0.093 ± 0.064	$\mathbf{0.322 \pm 0.070}$

Table 9: Empirical FDR and power for the global r-out-of-d conjunction rule across varying values of the conjunction parameter r, with the dimension fixed at $d = 5$. The data-generating process follows simulation setting 1. The nominal FDR level is $q = 0.20$, multivariate linear regression is used, and the target region is $R_k = (c_k, \infty)$ for all $k = 1, \ldots, 5$. Results are reported as mean $\pm$ standard deviation. Boldface indicates the best-performing method among those that control the FDR.

Metric	r	*Bonferroni*	*Omnibus*	*BinCS*	*PCCS*
FDR	1	0.120 ± 0.070	0.165 ± 0.046	0.198 ± 0.197	0.191 ± 0.054
	2	0.009 ± 0.026	0.064 ± 0.053	0.205 ± 0.074	0.196 ± 0.049
	3	0.001 ± 0.009	0.061 ± 0.042	0.203 ± 0.049	0.192 ± 0.048
	4	0.012 ± 0.026	0.114 ± 0.045	0.203 ± 0.043	0.201 ± 0.038
	5	0.359 ± 0.110	0.357 ± 0.117	0.197 ± 0.070	0.196 ± 0.030
Power	1	0.165 ± 0.100	0.268 ± 0.091	0.048 ± 0.053	$\mathbf{0.328 \pm 0.099}$
	2	0.014 ± 0.036	0.131 ± 0.086	0.244 ± 0.101	$\mathbf{0.441 \pm 0.098}$
	3	0.004 ± 0.019	0.170 ± 0.080	0.421 ± 0.081	$\mathbf{0.507 \pm 0.094}$
	4	0.049 ± 0.076	0.371 ± 0.108	0.402 ± 0.082	$\mathbf{0.547 \pm 0.073}$
	5	0.737 ± 0.152	0.729 ± 0.164	0.232 ± 0.091	$\mathbf{0.529 \pm 0.064}$

Table 10: Empirical FDR and power for the global r-out-of-d conjunction rule across varying values of the conjunction parameter r, with the dimension fixed at $d = 5$. The data-generating process follows simulation setting 2. The nominal FDR level is $q = 0.30$, multivariate linear regression is used, and the target region is $R_k = (c_k, \infty)$ for all $k = 1, \ldots, 5$. Results are reported as mean $\pm$ standard deviation. Boldface indicates the best-performing method among those that control the FDR.

Metric	r	*Bonferroni*	*Omnibus*	*BinCS*	*PCCS*
FDR	1	0.228 ± 0.242	0.201 ± 0.155	0.276 ± 0.120	0.274 ± 0.107
	2	0.000 ± 0.000	0.000 ± 0.000	0.285 ± 0.133	0.295 ± 0.104
	3	0.000 ± 0.000	0.000 ± 0.000	0.294 ± 0.109	0.300 ± 0.115
	4	0.000 ± 0.000	0.000 ± 0.000	0.303 ± 0.089	0.298 ± 0.097
	5	0.702 ± 0.017	0.702 ± 0.017	0.300 ± 0.115	0.288 ± 0.125
Power	1	0.004 ± 0.011	0.014 ± 0.024	0.024 ± 0.048	$\mathbf{0.058 \pm 0.032}$
	2	0.000 ± 0.000	0.000 ± 0.001	0.084 ± 0.078	$\mathbf{0.144 \pm 0.083}$
	3	0.000 ± 0.000	0.000 ± 0.000	0.136 ± 0.090	$\mathbf{0.178 \pm 0.104}$
	4	0.000 ± 0.000	0.000 ± 0.000	0.121 ± 0.096	$\mathbf{0.175 \pm 0.078}$
	5	1.000 ± 0.000	1.000 ± 0.000	0.123 ± 0.086	$\mathbf{0.151 \pm 0.075}$

Table 11: Empirical FDR and power for the global r-out-of-d conjunction rule across varying values of the conjunction parameter r, with the dimension fixed at $d = 5$. The data-generating process follows simulation setting 3. The nominal FDR level is $q = 0.30$, multivariate linear regression is used, and the target region is $R_k = (c_k, \infty)$ for all $k = 1, \ldots, 5$. Results are reported as mean $\pm$ standard deviation. Boldface indicates the best-performing method among those that control the FDR.

Metric	r	*Bonferroni*	*Omnibus*	*BinCS*	*PCCS*
FDR	1	0.192 ± 0.100	0.244 ± 0.072	0.309 ± 0.187	0.297 ± 0.068
	2	0.000 ± 0.000	0.045 ± 0.074	0.297 ± 0.080	0.298 ± 0.048
	3	0.000 ± 0.000	0.026 ± 0.050	0.301 ± 0.054	0.298 ± 0.048
	4	0.000 ± 0.000	0.066 ± 0.085	0.300 ± 0.052	0.295 ± 0.049
	5	0.697 ± 0.020	0.697 ± 0.020	0.294 ± 0.077	0.296 ± 0.048
Power	1	0.136 ± 0.099	0.245 ± 0.105	0.065 ± 0.063	$\mathbf{0.340 \pm 0.117}$
	2	0.000 ± 0.003	0.037 ± 0.056	0.300 ± 0.113	$\mathbf{0.472 \pm 0.104}$
	3	0.000 ± 0.000	0.032 ± 0.055	0.507 ± 0.107	$\mathbf{0.548 \pm 0.094}$
	4	0.000 ± 0.000	0.098 ± 0.136	0.498 ± 0.088	$\mathbf{0.596 \pm 0.087}$
	5	1.000 ± 0.000	1.000 ± 0.000	0.294 ± 0.121	$\mathbf{0.545 \pm 0.087}$

Table 12: Empirical FDR and power for the global r-out-of-d conjunction rule across varying values of the conjunction parameter r, with the dimension fixed at $d = 5$. The data-generating process follows simulation setting 4. The nominal FDR level is $q = 0.30$, multivariate linear regression is used, and the target region is $R_k = (c_k, \infty)$ for all $k = 1, \ldots, 5$. Results are reported as mean $\pm$ standard deviation. Boldface indicates the best-performing method among those that control the FDR.

Metric	r	*Bonferroni*	*Omnibus*	*BinCS*	*PCCS*
FDR	1	0.218 ± 0.364	0.201 ± 0.283	0.287 ± 0.301	0.221 ± 0.311
	2	0.000 ± 0.000	0.000 ± 0.000	0.293 ± 0.195	0.300 ± 0.250
	3	0.000 ± 0.000	0.000 ± 0.000	0.313 ± 0.128	0.287 ± 0.146
	4	0.000 ± 0.000	0.000 ± 0.000	0.298 ± 0.120	0.293 ± 0.174
	5	0.698 ± 0.018	0.698 ± 0.018	0.270 ± 0.163	0.283 ± 0.164
Power	1	0.003 ± 0.007	0.010 ± 0.020	$\mathbf{0.017 \pm 0.025}$	0.010 ± 0.021
	2	0.000 ± 0.000	0.000 ± 0.001	0.036 ± 0.054	$\mathbf{0.058 \pm 0.052}$
	3	0.000 ± 0.000	0.000 ± 0.000	0.063 ± 0.080	$\mathbf{0.133 \pm 0.078}$
	4	0.000 ± 0.000	0.000 ± 0.000	0.121 ± 0.103	$\mathbf{0.147 \pm 0.087}$
	5	1.000 ± 0.000	1.000 ± 0.000	0.079 ± 0.071	$\mathbf{0.091 \pm 0.060}$

Table 13: Empirical FDR and power for the global r-out-of-d conjunction rule across varying values of the conjunction parameter r, with the dimension fixed at $d = 10$. The data-generating process follows simulation setting 1. The nominal FDR level is $q = 0.30$, multivariate linear regression is used, and the target region is $R_k = (c_k, \infty)$ for all $k = 1, \ldots, 10$. Results are reported as mean $\pm$ standard deviation. Boldface indicates the best-performing method among those that approximately control the FDR.

Metric	r	*Bonferroni*	*Omnibus*	*BinCS*	*PCCS*
FDR	1	0.183 ± 0.099	0.264 ± 0.048	0.300 ± 0.198	0.303 ± 0.057
	3	0.000 ± 0.000	0.079 ± 0.055	0.303 ± 0.058	0.301 ± 0.049
	5	0.000 ± 0.000	0.108 ± 0.047	0.303 ± 0.046	0.301 ± 0.046
	7	0.000 ± 0.000	0.249 ± 0.065	0.288 ± 0.047	0.288 ± 0.045
	10	0.697 ± 0.021	0.697 ± 0.021	0.302 ± 0.083	0.297 ± 0.049
Power	1	0.119 ± 0.091	0.306 ± 0.077	0.058 ± 0.060	$\mathbf{0.393 \pm 0.114}$
	3	0.000 ± 0.000	0.114 ± 0.059	0.396 ± 0.103	$\mathbf{0.562 \pm 0.090}$
	5	0.000 ± 0.000	0.220 ± 0.083	0.588 ± 0.089	$\mathbf{0.613 \pm 0.095}$
	7	0.000 ± 0.000	0.576 ± 0.121	0.594 ± 0.085	$\mathbf{0.641 \pm 0.073}$
	10	1.000 ± 0.000	1.000 ± 0.000	0.192 ± 0.084	$\mathbf{0.546 \pm 0.084}$

Table 14: Empirical FDR and power for the global r-out-of-d conjunction rule across varying values of the conjunction parameter r, with the dimension fixed at $d = 30$. The data-generating process follows simulation setting 1. The nominal FDR level is $q = 0.30$, multivariate linear regression is used, and the target region is $R_k = (c_k, \infty)$ for all $k = 1, \ldots, 30$. Results are reported as mean $\pm$ standard deviation. Boldface indicates the best-performing method among those that approximately control the FDR.

Metric	r	*Bonferroni*	*Omnibus*	*BinCS*	*PCCS*
FDR	1	0.067 ± 0.149	0.338 ± 0.034	0.276 ± 0.195	0.287 ± 0.043
	8	0.000 ± 0.000	0.212 ± 0.031	0.296 ± 0.046	0.299 ± 0.035
	15	0.000 ± 0.000	0.328 ± 0.027	0.295 ± 0.045	0.302 ± 0.043
	22	0.000 ± 0.000	0.600 ± 0.019	0.288 ± 0.041	0.299 ± 0.039
	30	0.699 ± 0.020	0.699 ± 0.020	0.287 ± 0.075	0.303 ± 0.043
Power	1	0.013 ± 0.018	0.579 ± 0.042	0.066 ± 0.066	$\mathbf{0.571 \pm 0.087}$
	8	0.000 ± 0.000	0.524 ± 0.045	0.553 ± 0.085	$\mathbf{0.709 \pm 0.065}$
	15	0.000 ± 0.000	0.766 ± 0.036	0.749 ± 0.063	$\mathbf{0.753 \pm 0.061}$
	22	0.000 ± 0.000	0.987 ± 0.009	0.698 ± 0.062	$\mathbf{0.760 \pm 0.055}$
	30	1.000 ± 0.000	1.000 ± 0.000	0.240 ± 0.093	$\mathbf{0.674 \pm 0.061}$

Table 15: Empirical FDR and power for the global r-out-of-d conjunction rule across varying values of the conjunction parameter r, with the dimension fixed at $d = 5$. The data-generating process follows simulation setting 1. The nominal FDR level is $q = 0.30$, multivariate linear regression is used, and the coordinate-wise target region is $R_k = (c_{\text{low}}, c_{\text{high}})$ for all $k = 1, \ldots, 5$. Results are reported as mean $\pm$ standard deviation. Boldface indicates the best-performing method among those that approximately control the FDR.

Metric	r	*Bonferroni*	*Omnibus*	*BinCS*	*PCCS*
FDR	1	0.174 ± 0.345	0.112 ± 0.270	0.402 ± 0.435	0.255 ± 0.345
	2	0.000 ± 0.000	0.000 ± 0.000	0.318 ± 0.410	0.240 ± 0.303
	3	0.000 ± 0.000	0.000 ± 0.000	0.275 ± 0.400	0.300 ± 0.150
	4	0.000 ± 0.000	0.000 ± 0.000	0.285 ± 0.408	0.288 ± 0.084
	5	0.703 ± 0.020	0.703 ± 0.020	0.260 ± 0.406	0.298 ± 0.067
Power	1	0.002 ± 0.005	0.002 ± 0.005	0.003 ± 0.007	$\mathbf{0.005 \pm 0.011}$
	2	0.000 ± 0.000	0.000 ± 0.000	0.002 ± 0.005	$\mathbf{0.016 \pm 0.029}$
	3	0.000 ± 0.000	0.000 ± 0.000	0.002 ± 0.005	$\mathbf{0.108 \pm 0.098}$
	4	0.000 ± 0.000	0.000 ± 0.000	0.002 ± 0.007	$\mathbf{0.276 \pm 0.132}$
	5	1.000 ± 0.000	1.000 ± 0.000	0.002 ± 0.007	$\mathbf{0.351 \pm 0.112}$

Table 16: Empirical FDR and power for the global r-out-of-d conjunction rule across varying values of the conjunction parameter r, with the dimension fixed at $d = 5$. The data-generating process follows simulation setting 1. The nominal FDR level is $q = 0.30$, multivariate linear regression is used, and the coordinate-wise target region is $R_k = (-\infty, c_{\text{low}}) \cup (c_{\text{high}}, \infty)$ for all $k = 1, \ldots, 5$. Results are reported as mean $\pm$ standard deviation. Boldface indicates the best-performing method among those that approximately control the FDR.

Metric	r	*Bonferroni*	*Omnibus*	*BinCS*	*PCCS*
FDR	1	0.129 ± 0.035	0.158 ± 0.023	0.338 ± 0.246	0.296 ± 0.031
	2	0.015 ± 0.026	0.101 ± 0.028	0.277 ± 0.254	0.301 ± 0.027
	3	0.001 ± 0.008	0.102 ± 0.037	0.289 ± 0.254	0.298 ± 0.031
	4	0.001 ± 0.007	0.114 ± 0.113	0.309 ± 0.251	0.299 ± 0.040
	5	0.498 ± 0.022	0.498 ± 0.022	0.314 ± 0.274	0.288 ± 0.054
Power	1	0.301 ± 0.082	0.444 ± 0.063	0.040 ± 0.049	$\mathbf{0.760 \pm 0.065}$
	2	0.035 ± 0.056	0.329 ± 0.079	0.027 ± 0.041	$\mathbf{0.762 \pm 0.052}$
	3	0.007 ± 0.025	0.270 ± 0.084	0.026 ± 0.037	$\mathbf{0.703 \pm 0.058}$
	4	0.001 ± 0.005	0.201 ± 0.218	0.024 ± 0.039	$\mathbf{0.547 \pm 0.080}$
	5	1.000 ± 0.000	1.000 ± 0.000	0.016 ± 0.029	$\mathbf{0.238 \pm 0.084}$

Table 17: Empirical FDR and power for the global r-out-of-d conjunction rule across varying values of the conjunction parameter r, with the dimension fixed at $d = 5$. The data-generating process follows simulation setting 1. The nominal FDR level is $q = 0.30$, multivariate linear regression is used, and the coordinate-wise target region is $R_k = (-\infty, c_{\text{low}}) \cup (c_{\text{med},1}, c_{\text{med},2}) \cup (c_{\text{high}}, \infty)$ for all $k = 1, \ldots, 5$. Results are reported as mean $\pm$ standard deviation. Boldface indicates the best-performing method among those that approximately control the FDR.

Metric	r	*Bonferroni*	*Omnibus*	*BinCS*	*PCCS*
FDR	1	0.174 ± 0.201	0.183 ± 0.082	0.318 ± 0.258	0.286 ± 0.118
	2	0.000 ± 0.004	0.036 ± 0.037	0.316 ± 0.288	0.301 ± 0.045
	3	0.000 ± 0.000	0.020 ± 0.039	0.262 ± 0.242	0.294 ± 0.037
	4	0.000 ± 0.000	0.001 ± 0.009	0.314 ± 0.260	0.302 ± 0.042
	5	0.005 ± 0.051	0.005 ± 0.051	0.283 ± 0.288	0.309 ± 0.121
Power	1	0.070 ± 0.079	0.180 ± 0.105	0.028 ± 0.042	$\mathbf{0.294 \pm 0.181}$
	2	0.001 ± 0.010	0.073 ± 0.058	0.020 ± 0.028	$\mathbf{0.472 \pm 0.160}$
	3	0.000 ± 0.000	0.040 ± 0.041	0.024 ± 0.037	$\mathbf{0.522 \pm 0.128}$
	4	0.000 ± 0.000	0.006 ± 0.009	0.023 ± 0.032	$\mathbf{0.395 \pm 0.130}$
	5	0.010 ± 0.100	0.010 ± 0.100	0.015 ± 0.026	$\mathbf{0.075 \pm 0.060}$

A.4 Details of Real Data Applications

A.4.1 Details of Multi-property Molecular Selection

The drug discovery dataset used in Section 6 was compiled from various public sources (Wenzel et al., 2019; Iwata et al., 2022; Kim et al., 2023; 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) and published by Bai et al. (2025). The raw dataset compiled from these ten sources contains missing entries, since not every compound was assayed for all d = 15 properties across the original studies. Following Bai et al. (2025), missing values were imputed using Chemprop (Yang et al., 2019; Heid et al., 2024), a message-passing neural network for molecular property prediction; the resulting fully imputed dataset is used directly in all our experiments. The dataset contains a $d = 15$-dimensional response and $N = 22{,}805$ data points. Table 18 lists the names, units, cutoffs, and descriptions of the responses.

For the block-level r_0-out-of-L rule on this ADMET dataset, we partition the $d = 15$ responses into a safety block (the hard-constraint block, $d_1 = 6$) and an ADME block (the soft-constraint block, $d_2 = 9$). The safety block ($d_1 = 6$) consists of hERG_binding, CYP2C8_inhibition, CYP2C9_inhibition, CYP2D6_inhibition, CYP3A4_inhibition, CL_total_human, all of which must satisfy their respective constraints ($r_1 = d_1 = 6$, a hard conjunction). The ADME block ($d_2 = 9$) consists of CL_microsome_human, CL_microsome_mouse, CL_microsome_rat, CL_total_dog, CL_total_monkey, CL_total_rat, Papp_Caco2, Pgp_human, LogD_pH_7.4, of which at least r_2 must meet their respective cutoffs (a soft constraint). This partition reflects real-world drug-screening priorities: the six safety-block variables (hERG, four CYP-inhibition assays, and human total clearance) are safety red lines and the direct human PK endpoint—hard constraints that must all hold—while the nine ADME-block variables are redundant, interchangeable ADME/developability proxies, of which only r_2 need meet their cutoffs. We emphasize that this partition is one reasonable instantiation guided by pharmacological practice rather than the unique or optimal choice; PCCS itself is agnostic to the specific block structure and accommodates any user-specified partition $\{G_\ell\}_{\ell=1}^{L}$, with the FDR guarantee holding for any such fixed partition chosen prior to seeing the test data.

Table 18: List of responses in the drug discovery dataset.

Name	Unit	Cutoff c_k	Description
CL_microsome_human	log 10 (mL/min/kg)	4	Intrinsic metabolic clearance in human liver microsomes. (Helps predict the rate of liver metabolism in humans through in vitro–in vivo correlation.)
CL_microsome_mouse	log 10 (mL/min/kg)	4	Intrinsic metabolic clearance in mouse liver microsomes.
CL_microsome_rat	log 10 (mL/min/kg)	4	Intrinsic metabolic clearance in rat liver microsomes.
CL_total_dog	log 10 (mL/min/kg)	0.5	Total body clearance measured in vivo in dogs.
CL_total_human	log 10 (mL/min/kg)	0	Total body clearance measured in vivo in humans. (Reflects drug exposure in this species; crucial for determining human dose regimens through translational modeling.)
CL_total_monkey	log 10 (mL/min/kg)	0.5	Total body clearance measured in vivo in monkeys.
CL_total_rat	log 10 (mL/min/kg)	1	Total body clearance measured in vivo in rats.

continued on next page

Table 18 – continued from previous page

Name	Unit	Cutoff c_k	Description
CYP2C8_inhibition	log 10 (nMolar IC50)	3.5	Inhibition potential against the human CYP2C8 enzyme. (Assesses the risk of drug-drug interactions (DDIs) with CYP2C8 substrates; important for safety assessment.)
CYP2C9_inhibition	log 10 (nMolar IC50)	3.5	Inhibition potential against the human CYP2C9 enzyme. (Assesses the risk of drug-drug interactions (DDIs) with CYP2C9 substrates such as warfarin; important for safety assessment.)
CYP2D6_inhibition	log 10 (nMolar IC50)	3.5	Inhibition potential against the human CYP2D6 enzyme. (Assesses the risk of drug-drug interactions (DDIs) with CYP2D6 substrates such as certain antidepressants and beta-blockers; important for safety assessment.)
CYP3A4_inhibition	log 10 (nMolar IC50)	3.5	Inhibition potential against the human CYP3A4 enzyme. (Assesses the risk of drug-drug interactions (DDIs) with CYP3A4 substrates; CYP3A4 is a very common metabolic pathway, making this important for safety assessment.)
Papp_Caco2	log 10 (10^{-6} cm/s)	0.8	Apparent permeability across a Caco-2 cell monolayer. (Indicates the potential for oral absorption by crossing the intestinal wall; a high Papp suggests good absorption is likely.)
Pgp_human	log 10 (efflux ratio)	-0.2	Interaction with the human P-glycoprotein (Pgp) efflux transporter. (Indicates whether the drug is pumped out of cells by Pgp, potentially limiting oral absorption and brain penetration.)
hERG_binding	log 10 (nMolar IC50)	3.5	Binding/inhibition of the hERG potassium channel. (Assesses the risk of cardiac toxicity such as QT prolongation and arrhythmias; a critical safety screen for which high binding is a major safety concern.)
LogD_pH-7.4	log 10 (M/M)	2	Logarithm of the distribution coefficient at pH 7.4. (Measures the drug's lipophilicity — fat vs. water solubility — under physiological conditions; influences absorption, distribution, membrane crossing, and CNS penetration.)

A.4.2 Details of Response Filtering

```
You are an expert annotator evaluating the quality of an AI
↪ assistant's response to a user prompt. Rate the response on the
↪ following five attributes, each on an integer scale from 0 to 4.
↪ Base your ratings ONLY on the definitions below.

[Prompt]
{prompt}

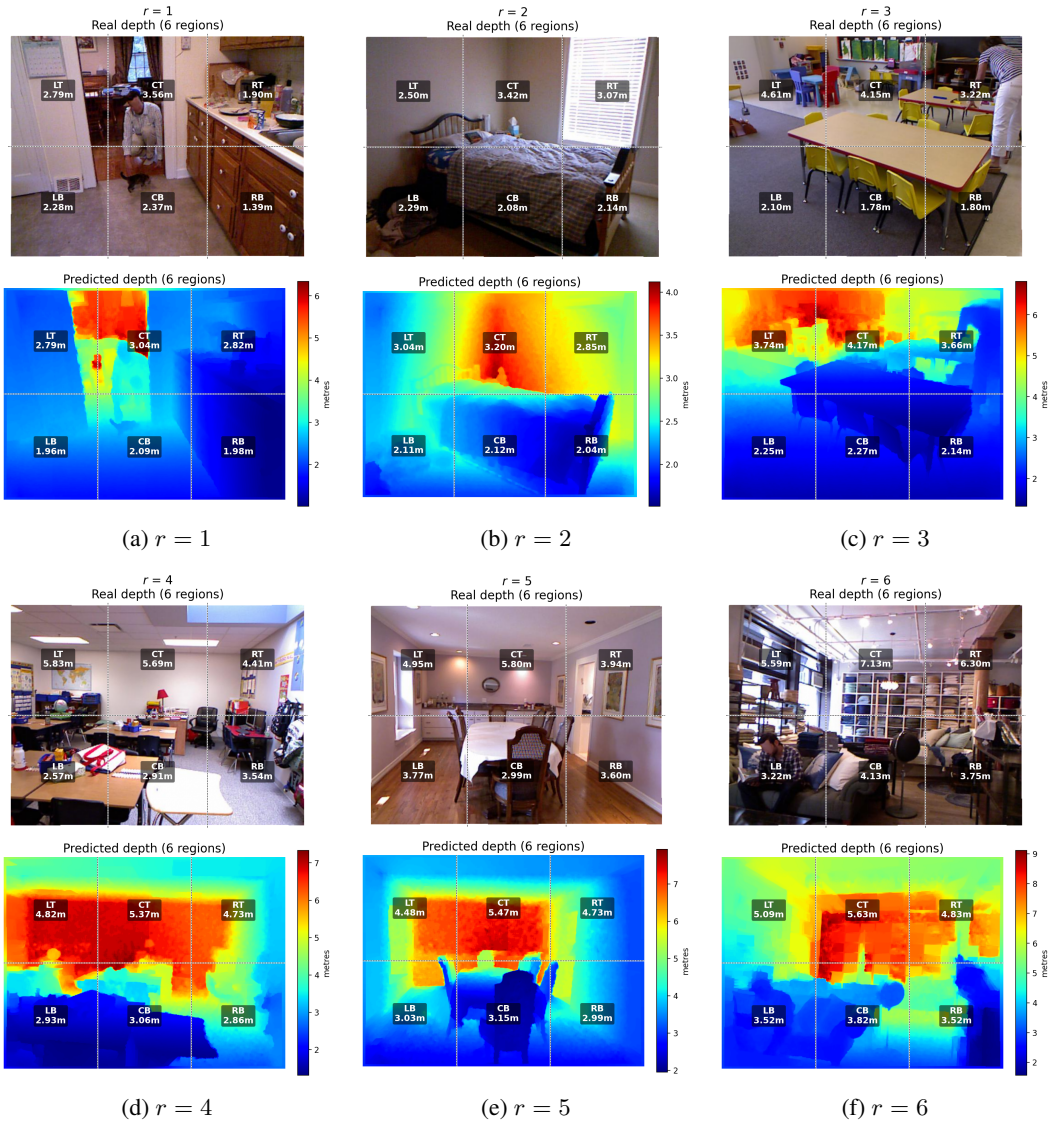
[Response]
{response}

Rate each attribute as follows:
1. Helpfulness (overall usefulness and alignment with what the
↪ prompt asks):
   4 = Extremely helpful, completely aligned with the spirit of the
↪ prompt.
   3 = Mostly helpful and aligned, with some room for improvement.
   2 = Partially helpful but misses the overall goal in some way.
   1 = Borderline unhelpful; mostly misses what the user wanted but
↪ still slightly usable.
   0 = Not useful at all; completely misses the user's intent.
2. Correctness (factual accuracy and completeness; no hallucinations
↪ ):
   4 = Completely correct and complete; no false or missing
↪ information.
   3 = Mostly accurate with a small amount of missing information.
   2 = A mix of correct and incorrect information, or minor
↪ omissions.
   1 = Some correct elements but mostly wrong or incomplete.
   0 = Completely incorrect, false, or hallucinated; irrelevant.
3. Coherence (clarity, self-consistency, logical flow):
   4 = Perfectly clear and self-consistent throughout.
   3 = Mostly coherent; one or two confusing spots.
   2 = A little unclear; some inconsistencies or hard-to-follow
↪ parts.
   1 = Mostly incoherent or unclear, with only some clear parts.
   0 = Completely incomprehensible; no sensible meaning can be
↪ discerned.
4. Complexity (simple vs. complex language required):
   4 = Expert: technical vocabulary a domain expert would use.
   3 = Advanced: sophisticated vocabulary; a college major could
↪ write it.
   2 = Intermediate: understandable to someone with a high-school
↪ education.
   1 = Simple: straightforward language; some schooling needed.
   0 = Basic: very easy language understandable by children.
5. Verbosity (succinct vs. verbose, relative to the prompt):
   4 = Verbose: lengthy/wordy with extra detail beyond what was
↪ requested.
   3 = Moderately long: on the longer side but not yet rambling.
   2 = Average length: adequate, neither wordy nor terse.
   1 = Pretty short: on the shorter side; some content could be
↪ trimmed.
   0 = Succinct: short, to the point, the most concise it can be.

Output ONLY a JSON object in exactly this format, with no extra text
↪ :
{"helpfulness": <0-4>, "correctness": <0-4>, "coherence": <0-4>, "
↪ complexity": <0-4>, "verbosity": <0-4>}
```

In NLP tasks, we apply PCCS to select high-quality large language model responses, curating reliable training data for subsequent alignment (e.g., preference optimization). We conduct experiments on the HelpSteer2 dataset (Wang et al., 2024a;b), which comprises 7,593 prompts, each paired with a chosen and a rejected response. We assess LLM responses across five attributes—*Helpfulness*, *Correctness*, *Coherence*, *Complexity*, and *Verbosity*—each scored on a scale from 0 to 4.

A.4.3 Details of Monocular Depth Estimation



(a) $r = 1$ (b) $r = 2$ (c) $r = 3$

(d) $r = 4$ (e) $r = 5$ (f) $r = 6$

Figure 4: Examples of NYU Depth V2 samples selected by *PCCS* under the r-out-of-6 regional depth criterion, with panels (a)–(f) corresponding to $r = 1, \ldots, 6$, respectively. ResNet-50 is used as the feature extractor. Within each panel, the top row shows the selected RGB image annotated with the ground-truth and predicted depths for the six regions (left-top, left-bottom, center-top, center-bottom, right-top, right-bottom), while the bottom row shows the corresponding predicted depth map. The target criterion requires the regional depth to exceed 3 meters.

We further evaluate *PCCS* on a computer vision task involving monocular depth estimation, where the goal is to select images in which the depths of at least r predefined spatial regions satisfy a target criterion. We conduct experiments on the NYU Depth V2 dataset (Silberman et al., 2012), which contains 1,449 RGB-D images of indoor scenes with corresponding depth maps captured using a

Microsoft Kinect sensor. Each image has a resolution of 640×480 pixels, and its depth values are measured in meters. We partition each depth map into six non-overlapping spatial regions by dividing it into left, center, and right thirds along the horizontal axis and top and bottom halves along the vertical axis, and define the response vector as $\mathbf{y} = (y_{\mathrm{LT}}, y_{\mathrm{LB}}, y_{\mathrm{CT}}, y_{\mathrm{CB}}, y_{\mathrm{RT}}, y_{\mathrm{RB}})$, where each component is the average ground-truth depth within the corresponding region (left-top, left-bottom, center-top, center-bottom, right-top, right-bottom). The dataset is randomly partitioned into training (74%), calibration (13%), and test (13%) sets. We use ResNet-50 (He et al., 2016) as the feature extractor and train a multi-output Ridge regression model to predict the six regional depths from each RGB image. For each region $k \in \{\mathrm{LT}, \mathrm{LB}, \mathrm{CT}, \mathrm{CB}, \mathrm{RT}, \mathrm{RB}\}$, we define the target region as $R_k = (3, +\infty)$. Therefore, under the r-out-of-6 conjunction rule, an image is desirable if the average ground-truth depth exceeds 3 meters in at least r of the six spatial regions, where $r \in \{2, 3, 4, 5, 6\}$. Figure 4 presents representative NYU Depth V2 samples selected by *PCCS* under these regional conjunction criteria.

Algorithm 1 Conjunction-Aware Score for the Global r-out-of-d Rule

function GLOBALSCORE($\mathbf{x}, \mathbf{y}, r, M$)
  for $k = 1, \ldots, d$ **do**
    Compute the signed margin $D_k\big(\widehat{f}_k(\mathbf{x}), R_k\big)$
  end for
  Let $D^{\downarrow}_{[r]}(\mathbf{x})$ be the r-th largest value among $\Big\{D_k\big(\widehat{f}_k(\mathbf{x}), R_k\big) : k \in [d]\Big\}$
  $r^*(\mathbf{y}) \leftarrow \sum_{k=1}^{d} \mathbb{1}\{y_k \in R_k\}$
  $V^{r/d}(\mathbf{x}, \mathbf{y}) \leftarrow M\mathbb{1}\{r^*(\mathbf{y}) \geq r\} - D^{\downarrow}_{[r]}(\mathbf{x})$
  return $V^{r/d}(\mathbf{x}, \mathbf{y})$
end function

Algorithm 2 Conjunction-Aware Score for the Block-Level r_0-out-of-L Rule

function BLOCKSCORE($\mathbf{x}, \mathbf{y}, \{G_\ell, r_\ell\}_{\ell=1}^{L}, r_0, M$)
  for $\ell = 1, \ldots, L$ **do**
    for $k \in G_\ell$ **do**
      Compute the signed margin $D_k\big(\widehat{f}_k(\mathbf{x}), R_k\big)$
    end for
    Let $D^{\downarrow}_{\ell,[r_\ell]}(\mathbf{x})$ be the r_ℓ-th largest value among $\Big\{D_k\big(\widehat{f}_k(\mathbf{x}), R_k\big) : k \in G_\ell\Big\}$
    $B_\ell(\mathbf{y}) \leftarrow \mathbb{1}\left\{\sum_{k \in G_\ell} \mathbb{1}\{y_k \in R_k\} \geq r_\ell\right\}$
  end for
  Let $D^{\downarrow}_{\mathrm{block},[r_0]}(\mathbf{x})$ be the r_0-th largest value among $\Big\{D^{\downarrow}_{\ell,[r_\ell]}(\mathbf{x}) : \ell \in [L]\Big\}$
  $V^{r_0/L}(\mathbf{x}, \mathbf{y}) \leftarrow M\mathbb{1}\left\{\sum_{\ell=1}^{L} B_\ell(\mathbf{y}) \geq r_0\right\} - D^{\downarrow}_{\mathrm{block},[r_0]}(\mathbf{x})$
  return $V^{r_0/L}(\mathbf{x}, \mathbf{y})$
end function

Algorithm 3 Partial Conjunction Conformal Selection (PCCS)

Input: Calibration data $\mathcal{D}_{\text{cal}} = \{(\mathbf{x}_i, \mathbf{y}_i)\}_{i=1}^{n}$; test inputs $\mathcal{D}_{\text{test}} = \{\mathbf{x}_{n+j}\}_{j=1}^{m}$; a conjunction rule $\mathcal{R}$; its associated oracle score V; its null-branch practical score $\widehat{V}$; nominal FDR level $q \in (0, 1)$

for $i = 1, \ldots, n$ **do**
  $V_i \leftarrow V(\mathbf{x}_i, \mathbf{y}_i)$
end for
for $j = 1, \ldots, m$ **do**
  $\widehat{V}_{n+j} \leftarrow \widehat{V}(\mathbf{x}_{n+j})$
  $p_j \leftarrow \frac{1+\sum_{i=1}^{n} \mathbb{1}\{V_i \le \widehat{V}_{n+j}\}}{n+1}$
end for
Sort the p-values as $p_{(1)} \le \cdots \le p_{(m)}$
$\eta^* \leftarrow \max\left(\left\{\eta \in [m] : p_{(\eta)} \le \frac{q\eta}{m}\right\} \cup \{0\}\right)$
if $\eta^* = 0$ **then**
  $\mathcal{S} \leftarrow \varnothing$
else
  $\mathcal{S} \leftarrow \left\{j \in [m] : p_j \le \frac{q\eta^*}{m}\right\}$
end if
return $\mathcal{S}$