Journal of the Royal Statistical Society Series B: Statistical Methodology, 2025, **00**, 1–21
https://doi.org/10.1093/jrsssb/qkaf016

ROYAL STATISTICAL SOCIETY
DATA | EVIDENCE | DECISIONS
Journal of the Statistics Society
Series B
Statistical Methodology
B

Original Article

Confidence on the focal: conformal prediction with selection-conditional coverage

Ying Jin[1,†] **and Zhimei Ren**[2,†]

[1]Data Science Initiative & Department of Health Care Policy, Harvard University, Boston, MA 02138, USA
[2]Department of Statistics and Data Science, University of Pennsylvania, 265 South 37th street, Room 317, Philadelphia, PA 19104, USA

Address for correspondence: Zhimei Ren, Department of Statistics and Data Science, University of Pennsylvania, 265 South 37th Street, Room 317, Philadelphia, PA 19104, USA. Email: zren@wharton.upenn.edu

Abstract

Conformal prediction builds marginally valid prediction intervals that cover the unknown outcome of a randomly drawn test point with a prescribed probability. However, in practice, data-driven methods are often used to identify specific test unit(s) of interest, requiring uncertainty quantification tailored to these focal units. In such cases, marginally valid conformal prediction intervals may fail to provide valid coverage for the focal unit(s) due to selection bias. This article presents a general framework for constructing a prediction set with finite-sample exact coverage, conditional on the unit being selected by a given procedure. The general form of our method accommodates arbitrary selection rules that are invariant to the permutation of the calibration units and generalizes Mondrian Conformal Prediction to multiple test units and non-equivariant classifiers. We also work out computationally efficient implementation of our framework for a number of realistic selection rules, including top-K selection, optimization-based selection, selection based on conformal *p*-values, and selection based on properties of preliminary conformal prediction sets. The performance of our methods is demonstrated via applications in drug discovery and health risk prediction.

Keywords: conformal inference, predictive inference, selective inference, uncertainty quantification

1 Introduction

Conformal prediction is a versatile framework for quantifying the uncertainty of any black-box prediction model, by issuing a prediction set that covers the unknown outcome with a prescribed probability. Formally, suppose the task is to predict an outcome $Y \in \mathcal{Y}$ based on features $X \in \mathcal{X}$. Given a set of calibration data $\{(X_i, Y_i)\}_{i=1}^n$ and the features of a new test point X_{n+1}, conformal prediction builds upon a given prediction model and delivers a prediction set $\widehat{C}_{\alpha,n+1} \subseteq \mathcal{Y}$ at level $\alpha \in (0, 1)$, which obeys

$$\mathbb{P}\left(Y_{n+1} \in \widehat{C}_{\alpha,n+1}\right) \geq 1 - \alpha, \tag{1}$$

as long as $\{(X_i, Y_i)\}_{i=1}^{n+1}$ are exchangeable (e.g. when they are i.i.d. samples). The probability in (1) is over both the calibration data and the test point (Lei et al., 2018; Vovk et al., 2005).

With this finite-sample, distribution-free guarantee, the conformal prediction set $\widehat{C}_{\alpha,n+1}$ describes a range of plausible values the unknown outcome Y_{n+1} may take, thereby expected to inform downstream decision-making based on the black-box prediction model. With such a promise, methods for constructing *marginally valid*—in the sense of (1)—prediction sets have been developed for various problems; see e.g. Angelopoulos and Bates (2021) for a review.

† Author names listed alphabetically

Received: March 31, 2024. **Revised:** March 1, 2025. **Accepted:** March 7, 2025

In many downstream applications, however, people are often only interested in a *selective* subset of units. For example, practitioners may only act upon a unit if it exhibits an interesting property (Levitskaya, 2023; Olsson et al., 2022; Sokol et al., 2024), or focus only on a subset of test units picked by a complicated data-dependent process such as resource optimization (Castro & Petrovic, 2012; Gocgun & Puterman, 2014; Kemper et al., 2014; Svensson et al., 2018). It would be misleading for the practitioners if the prediction sets fail to deliver the promised coverage guarantee for the selected unit(s). Let us discuss a few applications where such cases may arise.

- In *drug discovery*, an important task is to predict the binding affinity of a drug candidate to a disease target, which informs subsequent drug prioritization (Laghuvarapu et al., 2024). Among many drug candidates, scientists may only focus on those with highest predicted affinities, or those selected by a false discovery rate (FDR)-controlling procedure (Jin & Candès, 2023), or whose prediction sets only cover high values (Svensson et al., 2017), or those optimizing resource usage (Svensson et al., 2018). It may lead to a waste of resources if the prediction sets for the selected drugs fail to cover the actual binding affinities with an exceedingly high chance.
- In *business decision-making*, companies may take different inventory decisions based on whether a conformal prediction set suggests a strong demand or a weak demand (Levitskaya, 2023). Similarly, it will be problematic if a strong-demand prediction cannot cover with at least $1-\alpha$ of the time.
- In *disease diagnosis*, Olsson et al. (2022) suggest human intervention if the prediction set for a disease status is too large (this implicitly declares confidence in small prediction sets; see similar ideas in Ren et al., 2023; Sokol et al., 2024). However, it would be concerning if, with more than a chance α, the small-sized prediction sets—'approved' as confident—miss the true disease status.
- In *healthcare management*, patients may be sent to different healthcare categories based a program that optimizes some performance measure (such as waiting time) subject to certain constraints, such as budget, capacity, or fairness (Castro & Petrovic, 2012; Gocgun & Puterman, 2014; Kemper et al., 2014). The subset of patients in each category is therefore data-dependent.

In all these examples, it is highly desirable that a prediction set should cover the unknown outcome for a unit *of interest* with a prescribed probability. This motivates a stronger, selection-conditional guarantee. Supposing there are $m \geq 1$ test units $\mathcal{D}_{\text{test}} = \{X_{n+j}\}_{j=1}^{m}$, we aim for

$$\mathbb{P}\Big(Y_{n+j} \in \widehat{C}_{\alpha,n+j} \,|\, j \in \mathcal{S}(\mathcal{D}_{\text{calib}}, \mathcal{D}_{\text{test}})\Big) \geq 1-\alpha, \tag{2}$$

where $\mathcal{D}_{\text{calib}} = \{(X_i, Y_i)\}_{i=1}^{n}$ is the calibration data, and $\mathcal{S}(\cdot\,,\,\cdot)$ is a data-driven process to decide the units of interest, which maps the calibration and test data to a subset of $[m] := \{1, \ldots, m\}$. Our target is similar in spirit to post-selection inference (Lee et al., 2016; Tibshirani et al., 2018), but we consider predictive inference settings and develop distinct techniques; see online supplementary material Section S1.1 for more discussion. Throughout, we focus on settings where the prediction sets are constructed after $\mathcal{S}(\mathcal{D}_{\text{calib}}, \mathcal{D}_{\text{test}})$ is determined. By separating the selection process from the (post hoc) uncertainty quantification step, we leave the freedom of defining selection rules to the practitioners.

A prediction set with marginal validity (1) does not necessarily cover an unknown outcome *conditional on* being of interest as in (2). Such a selection issue has recently been raised in the literature of predictive inference: through analysis of a real drug discovery dataset, Jin and Candès (2023, Section 1) demonstrate that more than 30% of seemingly promising prediction sets (in the sense that $\widehat{C}_{\alpha,n+j} = \{\text{active}\}$) miss the actual outcomes when the nominal marginal miscoverage rate is $\alpha = 0.01$. We shall see more examples of this issue in our numerical experiments with various selection processes. New techniques are therefore needed for constructing prediction sets achieving (2).

1.1 Exchangeability via reference sets

Accounting for selection in conformal prediction is a delicate task since it breaks exchangeability. The core of conformal prediction is to leverage the exchangeability among the calibration and test data, such that their 'prediction residuals', referred to as *nonconformity scores*, are comparable in distribution (see Section 3.1 for more details). The calibration scores therefore inform the magnitude of uncertainty in a new test point (Vovk et al., 2005). However, given a selection event, the calibration data are no longer exchangeable with the test point, leading to violation of the coverage guarantee mentioned above. In general, the selection-conditional distributions of these scores are complex since the selection event can be highly data-dependent.

Such a challenge motivates our new framework, named *J*Oint *M*ondrian Conformal *I*nference (JOMI), which builds prediction sets that achieve selection-conditional coverage, with (2) as a special case. As visualized in Figure 1, our key idea is to find a 'reference set'—a data-dependent subset of calibration data that remain exchangeable with respect to the new test point conditional on the selection event. This reference set thus provides calibrated quantification of uncertainty for a selected unit. The mechanism we devise accommodates arbitrary selection rules $\mathcal{S}(\mathcal{D}_{\text{calib}}, \mathcal{D}_{\text{test}})$ that are invariant to permutations of data in $\mathcal{D}_{\text{calib}}$ and can be computed efficiently for a wide range of commonly used selection rules.

We also note that the selection-conditional coverage (2) may be implied by other stronger notions such as conditional coverage: $\mathbb{P}(Y_{n+1} \in \widehat{C}_{\alpha,n+1} \mid X_{n+1} = x) \geq 1 - \alpha$ for $\mathbb{P}$-almost all $x \in \mathcal{X}$. Conditional coverage implies (2) if $\{(X_i, Y_i)\}_{i=1}^{n+1}$ are mutually independent and the selection rule only depends on test features. However, it is not achievable by finite-length prediction intervals without distributional assumptions (Barber et al., 2021; Vovk et al., 2005), and practical selection rules may also depend on other information. Given these considerations, we may view (2) as a 'relaxed' version of conditional coverage that is both achievable and relevant for practical use.

1.2 Preview of contributions

In Section 3, we introduce the general formulation of JOMI, including the construction of reference set and its use in deriving a selection-conditionally valid prediction set. We prove that under exchangeability, the prediction set $\widehat{C}_{\alpha,n+j}$ produced by JOMI covers Y_{n+j} with probability at least $1 - \alpha$ conditional on a selection event. The selection event can be 'test unit j is selected', which leads to (2); it can also be more granular, such as 'test unit j is selected, and there are k selected test units'. This framework is valid for any selection rule $\mathcal{S}(\mathcal{D}_{\text{calib}}, \mathcal{D}_{\text{test}})$ that is permutation-invariant to $\mathcal{D}_{\text{calib}}$, without any modelling assumptions on the data generating process.

In Section 4, we study the computational aspect of JOMI. We show that when $|\mathcal{Y}| < \infty$, our method can be computed with a worst-case complexity of $O(|\mathcal{Y}|mn)$ times that of the selection rule. Moreover, for general continuous outcome space $\mathcal{Y}$, we work out efficient implementation of JOMI for a number of selection processes that may be of practical interest, including:

- *Covariate-dependent selection*. When the selection rule does not involve $\{Y_i\}_{i=1}^n$, the computation complexity of our generic method is (at most) $O(mn)$ times that of the selection rule. This implies efficient implementation for a wide range of problems, including various forms of top-K selection (for which the computation can be further reduced to $O(m + n)$) and selection based on complicated constrained optimization programs. As a special case, we recover the method of Bao et al. (2024) for top-K selection among test data and provide valid solutions to other ranking-based selection methods attempted in their work.
- *Conformal p-value-based selection*. We derive efficient implementation for a class of selection rules based on thresholding conformal p-values with 'stopping time-type' cutoffs. For instance, our prediction set is finite-sample exactly valid for units selected by the Conformal Selection method (Jin & Candès, 2023), which is studied in Bao et al. (2024) with approximate FCR guarantee.
- *Selection based on preliminary conformal prediction sets*. In addition, we derive a general, efficient instantiation when selecting units whose marginal prediction sets demonstrate certain interesting properties, such as being of a short length, or having a lower bound above some threshold. Our method can be useful for re-calibrating the uncertainty quantification of such seemingly promising units.

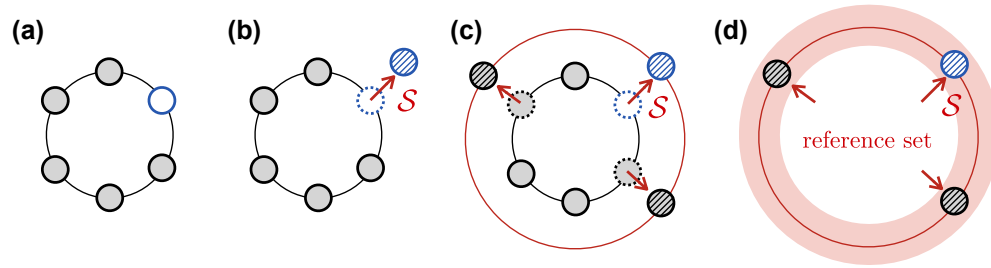

Figure 1. Visualization of the intuition behind the reference set. (a) Marginally, the calibration data (shaded) are exchangeable with respect to the test point (no shade). (b) The calibration data are not exchangeable with respect to the test point (shaded slash) given a selection event. (c) We find calibration data which, when posited as a 'test point', would lead to the same selection event. (d) The reference set consists of calibration data that are exchangeable with respect to the test point given selection, and we use them to construct JOMI prediction sets.

Finally, we demonstrate the application of our methods via several realistic selection rules that may occur in drug discovery (Section 5) and health risk prediction (Section 6). Our results show that marginal prediction sets may undercover or overcover the selected units, while our methods always achieve the promised coverage guarantees.

Due to space constraints, we delegate a comprehensive literature review to online supplementary material Section S1.1. We end this section with some useful notations.

Notations. For a positive integer $n \in \mathbb{N}^+$, write $[n] := \{1, \ldots, n\}$. We write the data pair as $Z = (X, Y)$, so that the calibration data are $\mathcal{D}_{\text{calib}} = \{Z_i\}_{i=1}^n$. For any $j \in [m]$, we define the augmented calibration set as $\mathcal{D}_j = \mathcal{D}_{\text{calib}} \cup \{Z_{n+j}\}$ and the remaining test set as $\mathcal{D}_j^c = \mathcal{D}_{\text{test}} \backslash \{X_{n+j}\}$. The unordered set of $\mathcal{D}_j$ is denoted $[\mathcal{D}_j] = [Z_1, \ldots, Z_n, Z_{n+j}]$, which provides the order statistics.

2 Problem setup

Following the split conformal prediction framework (Lei et al., 2018; Vovk et al., 2005), we build our prediction sets based on a prediction model fitted on a training set $\mathcal{D}_{\text{train}}$, assuming $\mathcal{D}_{\text{train}}$ is independent of the calibration and test data. In what follows, we always condition on $\mathcal{D}_{\text{train}}$, thereby treating the fitted models as fixed. In this section, we formally introduce the selection-conditional coverage guarantees and compare them to other related notions in the literature.

2.1 Selection-conditional coverage

Recall that $j \in [m]$ is a focal unit if $j \in \widehat{\mathcal{S}}$, where $\widehat{\mathcal{S}} = \mathcal{S}(\mathcal{D}_{\text{calib}}, \mathcal{D}_{\text{test}}) \subseteq [m]$ is obtained from a selection rule $\mathcal{S}$ that depends on the observed data. Its potential dependence on $\mathcal{D}_{\text{train}}$ is clear since we treat $\mathcal{D}_{\text{train}}$ as fixed. Without loss of generality, we posit that $\mathcal{S}$ is a deterministic function; when the selection rule is randomized, one can condition on its randomness and follow our framework.

For each test unit j, we wish to construct a prediction set $\widehat{C}_{\alpha,n+j} \subseteq \mathcal{Y}$ such that

$$\mathbb{P}\Big(Y_{n+j} \in \widehat{C}_{\alpha,n+j} \,|\, j \in \widehat{\mathcal{S}},\ \widehat{\mathcal{S}} \in \mathfrak{S}\Big) \geq 1 - \alpha, \tag{3}$$

where $\alpha \in (0, 1)$ is the confidence level, and $\mathfrak{S} \subseteq 2^{[m]}$ is some pre-specified collection of subsets of $[m]$. We call $\mathfrak{S}$ the *selection taxonomy* in what follows. Different choices of $\mathfrak{S}$ lead to a spectrum of granularity in the conditioning event. For instance, taking $\mathfrak{S} = 2^{[m]}$ puts no restrictions on the selection set, giving rise to the guarantee (2) introduced in the beginning (coverage conditional on a unit being selected). Taking $\mathfrak{S} = \{S \subseteq [m] : |S| = r\}$ for some $r \leq m$ achieves coverage conditional on selecting a specific number of units. Finally, taking $\mathfrak{S} = \{S_0\}$ for some $S_0 \subseteq [m]$ achieves coverage conditional on selecting a specific set; this is similar to the coverage guarantee conditional on a selected model in Lee et al. (2016) for high-dimensional parameter inference.

In words, the selection-conditional coverage guarantee (3) can be interpreted as follows: imagine there are infinitely many independent realizations of $\mathcal{D}_{\text{calib}}$ and $\mathcal{D}_{\text{test}}$; among those realizations where the selection event of interest happens, the prediction set $\widehat{C}_{\alpha,n+j}$ will cover the true outcome for at least $1-\alpha$ fraction of times. Here, $\mathfrak{S}$ is introduced to provide practitioners with the language to specify the granularity of the conditioning event, allowing them to tailor the guarantee to their specific needs. Moreover, it allows us to properly describe the relationship between selection-conditional coverage and FCR control, as we will see below.

2.2 Relations among notions of selective coverage

Before introducing our methods, we take a moment to compare different notions of selective coverage. Readers who are more interested in our methodology may skip the remaining of this section.

Our first observation is that (3) implies (2) under appropriate conditions. The proof of the next proposition is in online supplementary material Section S4.1.

Proposition 1 Suppose a family of prediction sets $\{\widehat{C}^{(\ell)}_{\alpha,n+j}\}_{\ell\in\mathcal{L}}$ satisfy $\mathbb{P}(Y_{n+j}\in\widehat{C}^{(\ell)}_{\alpha,n+j}\,|\,j\in\widehat{\mathcal{S}},\ \widehat{\mathcal{S}}\in\mathfrak{S}_\ell)\geq 1-\alpha$ for a set of disjoint taxonomies $\{\mathfrak{S}_\ell\}_{\ell\in\mathcal{L}}$ such that $\cup_{\ell\in\mathcal{L}}\mathfrak{S}_\ell=2^{[m]}$. Define the prediction set $\widehat{C}_{\alpha,n+j}=\widehat{C}^{(\ell)}_{\alpha,n+j}$ when $j\in\widehat{\mathcal{S}}$ and $\widehat{\mathcal{S}}\in\mathfrak{S}_\ell$. Then $\mathbb{P}(Y_{n+j}\in\widehat{C}_{\alpha,n+j}\,|\,j\in\widehat{\mathcal{S}})\geq 1-\alpha$.

To distinguish the selection-conditional coverage in (3) and that in (2), we refer to (3) as *strong* selection-conditional coverage and (2) as *weak* selection-conditional coverage.

Another widely used post-selection guarantee is the false coverage rate (FCR) (Benjamini & Yekutieli, 2005), defined as the expected proportion of selected units missed by the prediction set:

$$\text{FCR}:=\mathbb{E}\left[\frac{\sum_{j\in[m]}\mathbb{1}\{j\in\widehat{\mathcal{S}},\ Y_{n+j}\notin\widehat{C}_{\alpha,n+j}\}}{|\widehat{\mathcal{S}}|\vee 1}\right], \tag{4}$$

where $a\vee b=\max\{a,b\}$ for any $a,b\in\mathbb{R}$. Previous works (Bao et al., 2024; Gazin et al., 2025; Weinstein & Ramdas, 2020) mainly focus on constructing prediction sets with FCR control. However, with $m=1$, it holds that $\text{FCR}=\mathbb{P}(j\in\widehat{\mathcal{S}},\ Y_{n+j}\notin\widehat{C}_\alpha(X_{n+j}))\leq\alpha$ for any marginally valid prediction set. Thus, FCR does not always address the selection issue.

The following proposition shows that strong selection-conditional coverage implies FCR control with a proper choice of selection taxonomy, whose proof is in online supplementary material Section S4.2.

Proposition 2 Suppose a family of prediction sets $\{\widehat{C}^{(\ell)}_{\alpha,n+j}\}_{\ell\in\mathcal{L}}$ satisfy $\mathbb{P}(Y_{n+j}\in\widehat{C}^{(\ell)}_{\alpha,n+j}\,|\,j\in\widehat{\mathcal{S}},\ \widehat{\mathcal{S}}\in\mathfrak{S}_\ell)\geq 1-\alpha$ for a set of disjoint selection taxonomies $\{\mathfrak{S}_\ell\}_{\ell\in\mathcal{L}}$ such that $\cup_{\ell\in\mathcal{L}}\mathfrak{S}_\ell=2^{[m]}$, and for each $\ell\in\mathcal{L}$, $\mathfrak{S}_\ell\subseteq\{S\subseteq[m]:|S|=r(\ell)\}$ for some $0\leq r(\ell)\leq m$. Define the prediction set $\widehat{C}_{\alpha,n+j}=\widehat{C}^{(\ell)}_{\alpha,n+j}$ when $j\in\widehat{\mathcal{S}}$ and $\widehat{\mathcal{S}}\in\mathfrak{S}_\ell$. Then its FCR (4) is upper bounded by $\alpha\cdot\mathbb{P}(\widehat{\mathcal{S}}\neq\emptyset)\leq\alpha$.

The selection taxonomies in Proposition 2 require the selection set to be of a specific size, without other conditions on its form. This can be automatically satisfied by some selection rules such as top-K selection. Extending Proposition 2 to non-exact scenarios yields useful insights in practice. For instance, consider the family of taxonomies with $\mathfrak{S}_\ell\subseteq\{S\subseteq[m]:r(\ell)-\delta\leq|S|\leq\delta+r(\ell)\}$ for some small $\delta\in\mathbb{N}^+$. When the selection size is stable, weak selection-conditional coverage implies $\mathbb{P}(Y_{n+j}\in\widehat{C}_{\alpha,n+j}\,|\,j\in\widehat{\mathcal{S}},\ \widehat{\mathcal{S}}\in\mathfrak{S}_\ell)\approx 1-\alpha$, which leads to $\text{FCR}\approx\alpha\cdot\mathbb{P}(\widehat{\mathcal{S}}\neq\emptyset)$. This approximate result helps explain several phenomena in our numerical experiments: first, the FCR is usually controlled empirically by our method; second, we sometimes observe significantly lower FCR than the selection-conditional coverage when $\mathbb{P}(\widehat{\mathcal{S}}=\emptyset)$ is moderately large. As we shall see later, a version of our proposed method achieves coverage guarantees conditional on being selected and the selection size, which requires checking a slightly more complex condition in the reference set

construction when the selection size varies. The computation of such prediction sets can be done efficiently for all the instances provided in the article, with the corresponding worst-case computational complexity explicitly stated in each section. One potential concern, though, is that the additional condition on the selection size may reduce the number of calibration data points in the reference set, especially when the selection step is highly variable. This could potentially lead to wider and/or unstable prediction intervals.

As a side note, the weak selection-conditional coverage does not necessarily imply FCR control, although in some special cases both can be true (see, e.g. Bao et al., 2024 for such examples). We put this as a proposition below, with a counterexample given in online supplementary material Section S4.3.

Proposition 3 There exists an instance and prediction sets $\{\widehat{C}_{\alpha,n+j} : j \in \widehat{\mathcal{S}}\}$ that satisfy the weak selection-conditional coverage at level α but violate the FCR control at level α.

We end this section with a remark on the interpretation of selection-conditional coverage and FCR control.

Remark 1 (Interpretation of selection-conditional coverage and FCR control). As shown by Proposition 2, the guarantee in (3) is stronger than FCR control (for a proper choice of the selection taxonomy). In fact, the former has often been used as a device to derive the latter in the literature (e.g. Bao et al., 2024; Weinstein et al., 2013).

In terms of interpretation, there are two main differences between selection-conditional coverage and FCR. First, the guarantee of FCR is averaged over all the selection events, including the case of empty selection set where the false coverage proportion is by definition zero. Therefore, even with FCR control, one could still suffer from a high false coverage proportion when the selection set is non-empty as long as this is compensated by the cases where the selection set is empty. On the other hand, selection-conditional coverage delivers guarantees conditioning on selection events of interest, which prevents the aforementioned undesired situation. Second, FCR control provides a guarantee that is averaged over all the selected units: it could be possible that for some selected units, the coverage is much lower than the nominal level, and for others the coverage is much higher, so that the average coverage over all the selected units is controlled at the nominal level. In contrast, selection-conditional coverage provides guarantees specific to each selected units.

Finally, we note that the distinction between the two concepts can be asymptotically negligible in special cases. We refer to Lemma S1 in the online supplementary material for such an instance.

3 JOMI: a unified framework

3.1 Warm-up: split conformal prediction

To warm up, we briefly summarize the split conformal prediction (SCP) method (Lei et al., 2018; Vovk et al., 2005), and how it achieves finite-sample coverage under exchangeability.

SCP starts with a nonconformity score function $V : \mathcal{X} \times \mathcal{Y} \mapsto \mathbb{R}$ determined by $\mathcal{D}_{train}$, so that $V(x, y)$ informs how well a hypothetical value $y \in \mathcal{Y}$ conforms to a machine prediction. For instance, one may set $V(x, y) = |y - \widehat{\mu}(x)|$, where $\widehat{\mu}(x)$ is regression function fitted on $\mathcal{D}_{train}$. Other popular choices include *conformalized quantile regression* (CQR, Romano et al., 2019) for regression and *adaptive prediction sets* (APS, Romano et al., 2020) for classification.

Compute $V_i = V(X_i, Y_i)$ for $i \in [n]$. The split conformal prediction set for test unit $j \in [m]$ is

$$\widehat{C}^{\mathrm{SCP}}_{\alpha,n+j} = \left\{y : V(X_{n+j}, y) \le \mathrm{Quantile}\left(1-\alpha; \{V_i\}_{i=1}^{n} \cup \{+\infty\}\right)\right\},$$

where $\mathrm{Quantile}(1-\alpha; \cdot)$ is the $(1-\alpha)$th empirical quantile of the set in the second argument. When $(Z_1, \ldots, Z_n, Z_{n+j})$ are exchangeable, $\widehat{C}^{\mathrm{SCP}}_{\alpha,n+j}$ achieves (1) (Vovk et al., 2005).

It helps motivate our approach to see the ideas behind the validity of $\widehat{C}^{\text{SCP}}_{\alpha,n+j}$. In words, $\widehat{C}^{\text{SCP}}_{\alpha,n+j}$ finds hypothesized values of y that make $V(X_{n+j}, y)$ look similar to calibration scores. Note that

$$\mathbb{P}\left(Y_{n+j} \in \widehat{C}^{\text{SCP}}_{\alpha,n+j}\right) = \mathbb{P}\left(V_{n+j} \leq \text{Quantile}\left(1-\alpha; \{V_i\}_{i=1}^n \cup \{V_{n+j}\}\right)\right), \tag{5}$$

where $V_{n+j} = V(X_{n+j}, Y_{n+j})$. Recall the unordered set $[\mathcal{D}_j] = [Z_1, \ldots, Z_n, Z_{n+j}]$ where $Z_i = (X_i, Y_i)$. Conditional on the event $\{[\mathcal{D}_j] = [z_1, \ldots, z_{n+j}]\}$, the only randomness is in the ordering of $(Z_1, \ldots, Z_n, Z_{n+j})$; due to exchangeability, the probability of Z_{n+j} taking on each value in $z_1, \ldots, z_n, z_{n+j}$ is equal. Therefore, conditional on $[\mathcal{D}_j] = [z_1, \ldots, z_{n+j}]$, the chance of V_{n+j} being no greater than the $(1-\alpha)$th quantile of $[v_1, \ldots, v_n, v_{n+j}]$, where $v_i = v(z_i)$, is at least $1-\alpha$, i.e.

$$\mathbb{P}\left(V_{n+j} \leq \text{Quantile}\left(1-\alpha; \{v_i\}_{i=1}^n \cup \{v_{n+j}\}\right) \mid [\mathcal{D}_j] = [z_1, \ldots, z_{n+j}]\right) \geq 1-\alpha. \tag{6}$$

This leads to (5) by the tower property. Therefore, inverting the criterion on the right-hand side of (5) gives a valid prediction set for Y_{n+j}.

The reason why vanilla SCP may fail to achieve selection-conditional coverage like (3) is that, conditional on the selection event, the data points $\{Z_1, \ldots, Z_n, Z_{n+j}\}$ are no longer exchangeable. In other words, in (6), it is unclear how V_{n+j} is distributed over $v_1, \ldots, v_{n+j}$ if we additionally condition on the selection event. As such, a natural remedy is to find a subset of calibration data that are 'exchangeable' with respect to the test point conditioning on the selection event, and leverage their scores to calibrate the prediction of the test unit. We introduce our methods below.

3.2 Conformal inference via a reference set

Fix a test unit $j \in [m]$. Recall that $Z_{n+j}(y) = (X_{n+j}, y)$ is the imputed test point with a hypothesized response $y \in \mathcal{Y}$. The core of our method is to find calibration points that are exchangeable with respect to the test point conditional on the selection set. At a high level, they are calibration units $i \in [n]$ that are 'indistinguishable' with $j \in [m]$, in the sense that treating $Z_{n+j}(y)$ as a calibration point and Z_i as a test point results in the same selection event.

We formalize this idea via a 'swap' operation. For any calibration unit $i \in [n]$, we define the 'swapped' calibration data $\mathcal{D}^{\text{swap(i,j)}}_{\text{calib}}(y)$ and the swapped test data $\mathcal{D}^{\text{swap(i,j)}}_{\text{test}}$ as follows:

$$\begin{aligned}
\mathcal{D}^{\text{swap}(i,j)}_{\text{calib}}(y) &= (Z_1^{\text{swap}(i,j)}(y), Z_2^{\text{swap}(i,j)}(y), \ldots, Z_n^{\text{swap}(i,j)}(y)), \\
\mathcal{D}^{\text{swap}(i,j)}_{\text{test}} &= (X_{n+1}^{\text{swap}(i,j)}, X_{n+2}^{\text{swap}(i,j)}, \ldots, X_{n+m}^{\text{swap}(i,j)}),
\end{aligned}$$

where for $k \in [n]$ and $\ell \in [m]$,

$$Z_k^{\text{swap}(i,j)}(y) = \begin{cases} Z_{n+j}(y) & k = i, \\ Z_k & k \neq i. \end{cases} \quad X_{n+\ell}^{\text{swap}(i,j)} = \begin{cases} X_i & \ell = j, \\ X_{n+\ell} & \ell \neq j. \end{cases}$$

That is, $\mathcal{D}^{\text{swap(i,j)}}_{\text{calib}}(y)$ and $\mathcal{D}^{\text{swap(i,j)}}_{\text{test}}$ are the calibration and test data if we treat Z_i as the jth test point, and $Z_{n+j}(y)$ as the ith calibration point. Figure 2 is an illustration of the swap operation.

Applying the same selection rule $\mathcal{S}$ to the swapped data, we define the swapped selection set with the hypothesized y as

$$\widehat{\mathcal{S}}^{\text{swap}(i,j)}(y) = \mathcal{S}\left(\mathcal{D}^{\text{swap}(i,j)}_{\text{calib}}(y), \mathcal{D}^{\text{swap}(i,j)}_{\text{test}}\right).$$

Then, we define the 'reference set' for achieving (3) with taxonomy $\mathfrak{S}$ as

$$\widehat{\mathcal{R}}_{n+j}(y) = \left\{i \in [n] : j \in \widehat{\mathcal{S}}^{\text{swap}(i,j)}(y), \quad \text{and} \quad \widehat{\mathcal{S}}^{\text{swap}(i,j)}(y) \in \mathfrak{S}\right\}.$$

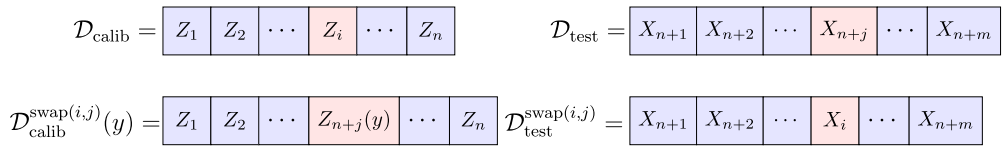

Figure 2. Graphical illustration of $\mathcal{D}_{\text{calib}}^{\text{swap}(i,j)}(y)$ and $\mathcal{D}_{\text{test}}^{\text{swap}(i,j)}$.

In words, the reference set is the collection of calibration points $i \in [n]$ such that, after swapping unit i and $n+j$, the (posited) test point j (which is the original unit i) remains in the focal set and the focal set remains in $\mathfrak{S}$. We use the notation $\widehat{\mathcal{R}}_{n+j}(y)$ to emphasize that $\widehat{\mathcal{R}}_{n+j}(\cdot)$ is a data-dependent mapping from $\mathcal{Y}$ to the power set of $[n]$.

With all the preparation, we define our prediction set for Y_{n+j} as

$$\widehat{C}_{\alpha,n+j} = \left\{ y \in \mathcal{Y} : V(X_{n+j}, y) \le \text{Quantile}\left(1-\alpha; \{V_i\}_{i \in \widehat{\mathcal{R}}_{n+j}(y)} \cup \{+\infty\}\right) \right\}. \tag{7}$$

As we shall show shortly, our prediction set (7) achieves near-exact coverage when the nonconformity score is continuous and the reference set is of moderate size. In addition, we can achieve exact coverage by introducing extra randomness:

$$\widehat{C}_{\alpha,n+j}^{\text{rand}} = \left\{ y : \frac{\sum_{i \in \widehat{\mathcal{R}}_{n+j}(y)} \mathbb{1}\{V(X_{n+j}, y) < V_i\} + U_j \cdot (1 + \sum_{i \in \widehat{\mathcal{R}}_{n+j}(y)} \mathbb{1}\{V(X_{n+j}, y) = V_i\})}{1 + |\widehat{\mathcal{R}}_{n+j}(y)|} \le 1-\alpha \right\}, \tag{8}$$

where $U_1, \ldots, U_m$ are i.i.d. random variables drawn from Unif[0, 1] independent of the data. In online supplementary material Sections S1.2 and S1.3, we discuss in detail the connection of our method to Mondrian conformal prediction (MCP) and comparison with the BY procedure (Benjamini & Yekutieli, 2005). While the BY procedure is a natural and heuristic solution to post-selection inference, it suffers from over-conservativeness and non-adaptivity to the selection event.

3.3 Theoretical guarantees

Theorem 1 confirms the conditional validity of our prediction sets $\widehat{C}_{\alpha,n+j}$ in (7) and $\widehat{C}_{\alpha,n+j}^{\text{rand}}$ in (8).

Theorem 1 Suppose $\mathcal{S}(\mathcal{D}_{\text{calib}}, \mathcal{D}_{\text{test}})$ is invariant to permutations of $\mathcal{D}_{\text{calib}}$, and that $\{Z_i\}_{i=1}^n \cup \{Z_{n+j}\}$ are exchangeable conditional on $\{X_{n+\ell}\}_{\ell \in [m]\setminus\{j\}}$ for any $j \in [m]$. Then, for any selection taxonomy $\mathfrak{S}$, the following statements hold.

(a) $\widehat{C}_{\alpha,n+j}$ defined in (7) obeys

$$\mathbb{P}\left(Y_{n+j} \in \widehat{C}_{\alpha,n+j} \,|\, j \in \widehat{\mathcal{S}},\ \widehat{\mathcal{S}} \in \mathfrak{S}\right) \ge 1-\alpha. \tag{9}$$

Furthermore, if ties among $V_1, \ldots, V_n, V_{n+j}$ occur with probability zero, then

$$\mathbb{P}\left(Y_{n+j} \in \widehat{C}_{\alpha,n+j} \,|\, j \in \widehat{\mathcal{S}},\ \widehat{\mathcal{S}} \in \mathfrak{S}\right) \le 1-\alpha + \mathbb{E}\left[\frac{1}{1+|\widehat{\mathcal{R}}_{n+j}(Y_{n+j})|}\right].$$

(b) The randomized prediction set $\widehat{C}^{\text{rand}}_{\alpha,n+j}$ defined in (8) satisfies

$$\mathbb{P}(Y_{n+j} \in \widehat{C}^{\text{rand}}_{\alpha,n+j} \,|\, j \in \widehat{\mathcal{S}},\ \widehat{\mathcal{S}} \in \mathfrak{S}) = 1-\alpha.$$

We defer the detailed proof of Theorem 1 to online supplementary material Section S4.4 and provide some intuition here. Similar to the ideas of SCP in Section 3.1, the key fact we rely on is that, plugging in the true value $y = Y_{n+j}$, data in the reference set and the new test point are still exchangeable conditional on the selection event. To be specific, to prove (9), we are to show a stronger result:

$$\mathbb{P}\left(V_{n+j} \le \text{Quantile}\left(1-\alpha; \{V_i\}_{i\in\widehat{\mathcal{R}}_{n+j}(Y_{n+j})} \cup \{V_{n+j}\}\right) \,\middle|\, j \in \widehat{\mathcal{S}},\ \widehat{\mathcal{S}} \in \mathfrak{S},\ [\mathcal{D}_j],\ \mathcal{D}_j^c\right) \ge 1-\alpha, \tag{10}$$

where we recall $[\mathcal{D}_j] = [Z_1, \ldots, Z_n, Z_{n+j}]$ and $\mathcal{D}_j^c = \mathcal{D}_{\text{test}} \backslash \{X_{n+j}\}$. For any fixed values $z_1, \ldots, z_n, z_{n+j}$, once given the unordered values $[\mathcal{D}_j] = [d_j] = [z_1, \ldots, z_n, z_{n+j}]$ and the values of other test points $\{X_{n+\ell}\}_{\ell\neq j}$, the only randomness is in the ordering of $Z_1, \ldots, Z_{n+j}$ among $[z_1, \ldots, z_n, z_{n+j}]$. Meanwhile, we show that our reference set $\widehat{\mathcal{R}}_{n+j}(\cdot)$ is constructed in a delicate way such that the (unordered set of) scores $\{V_i\}_{i\in\widehat{\mathcal{R}}_{n+j}(Y_{n+j})} \cup \{V_{n+j}\}$ is fully determined by $[z_1, \ldots, z_n, z_{n+j}]$. That is, $\widehat{R}^+_{n+j} := [V_i : i \in \widehat{\mathcal{R}}_{n+j}(Y_{n+j}) \cup \{n+j\}]$ is fully determined given $[d_j]$. Then, (10) reduces to

$$\mathbb{P}\left(V_{n+j} \le \text{Quantile}\left(1-\alpha; \widehat{R}^+_{n+j}\right) \,\middle|\, j \in \widehat{\mathcal{S}},\ \widehat{\mathcal{S}} \in \mathfrak{S},\ [\mathcal{D}_j],\ \mathcal{D}_j^c\right) \ge 1-\alpha,$$

Finally, by the exchangeability of $Z_1, \ldots, Z_n, Z_{n+j}$, the probability of V_{n+j} taking on any value in $\widehat{R}^+_{n+j}$ is equal given $[\mathcal{D}_j]$ and $\mathcal{D}_j^c$, leading to the validity in Theorem 1 via Bayes' rule.

4 Computationally tractable instances

So far, we have presented a general framework for constructing valid prediction sets conditional on general selection events. However, computing the prediction sets according to their definition requires looping over all possible values of $y \in \mathcal{Y}$, which can be computationally intractable. When $|\mathcal{Y}|$ is finite (and relatively small), our proposed method can be efficiently implemented according to its definition; the corresponding computational complexity is at most $O(|\mathcal{Y}|mn)$ times the complexity of the selection rule.

In this section, we instantiate our general procedure beyond the small $|\mathcal{Y}|$ setting with concrete examples where special structures enable efficient computation. We focus on three classes of selection rules that can be of practical interest: selection using only the covariates, selection based on conformal p-values, and selection based on conformal prediction sets. When practitioners are willing to slightly modify the selection rule to improve computation efficiency, they may refer to online supplementary material Section S5.4 where we discuss extensions to simplify the construction of prediction sets by further splitting the calibration data.

4.1 Covariate-dependent selection rules

We first consider covariate-dependent selection rules, i.e. when $\mathcal{S}(\mathcal{D}_{\text{calib}}, \mathcal{D}_{\text{test}})$ is only a function of $\{X_i\}_{i=1}^{n+m}$. Under such rules, the reference set no longer depends on y; we shall suppress the dependence on y and write $\widehat{\mathcal{R}}_{n+j}(y) \equiv \widehat{\mathcal{R}}_{n+j}$ throughout this subsection.

Here, $\widehat{\mathcal{R}}_{n+j}$ can be efficiently computed by looping over $i \in [n]$. The complete procedure for constructing $\widehat{C}_{\alpha,n+j}$ and $\widehat{C}^{\text{rand}}_{\alpha,n+j}$ with an arbitrary covariate-dependent selection rule is summarized in Algorithm 1. Its overall computational complexity is $O(mn \cdot C_{\mathcal{S}})$, with $C_{\mathcal{S}}$ being the complexity of the selection process.

By Theorem 1, the output of Algorithm 1 is valid as long as the selection rule does not rely on the ordering of the calibration points. This includes many commonly used selection rules:

Algorithm 1 JOMI for arbitrary covariate-dependent selection rules

Input: Calibration data $\mathcal{D}_{\text{calib}}$; test data $\mathcal{D}_{\text{test}}$; miscoverage level α; nonconformity score $V(\cdot,\cdot)$; selection rule $\mathcal{S}$; selection taxonomy $\mathfrak{S}$; form of prediction set $\in \{\texttt{dtm}, \texttt{rand}\}$.

Compute $\widehat{\mathcal{S}} = \mathcal{S}(\mathcal{D}_{\text{train}}, \mathcal{D}_{\text{test}})$.

for $j \in \widehat{\mathcal{S}}$ **do**

Initialize $\widehat{\mathcal{R}}_{n+j} = \emptyset$.

for $i = 1, \ldots, n$ **do**

$\widehat{\mathcal{R}}_{n+j} = \widehat{\mathcal{R}}_{n+j} \cup \{i\}$ if $j \in \widehat{\mathcal{S}}^{\text{swap}(i,j)}$ and $\widehat{\mathcal{S}}^{\text{swap}(i,j)} \in \mathfrak{S}$.

if form $=$ `dtm` **then**

$\widehat{C}_{\alpha,n+j} = \left\{ y \in \mathcal{Y} : V(X_{n+j}, y) \le \text{Quantile}(1-\alpha; \{V_i\}_{i \in \widehat{\mathcal{R}}_{n+j}} \cup \{+\infty\}) \right\}$.

if form $=$ `rand` **then**

Sample $U_j \sim \text{Unif}[0,1]$.

$$\widehat{C}_{\alpha,n+j} \leftarrow \left\{ y : \frac{\sum_{i \in \widehat{\mathcal{R}}_{n+j}} \mathbb{1}\{V(X_{n+j}, y) < V_i\} + U_j \cdot (1 + \sum_{i \in \widehat{\mathcal{R}}_{n+j}} \mathbb{1}\{V(X_{n+j}, y) = V_i\})}{1 + |\widehat{\mathcal{R}}_{n+j}|} \le 1 - \alpha \right\}$$

Output: $\{\widehat{C}_{\alpha,n+j}\}_{j \in \widehat{\mathcal{S}}}$.

1. *Top-K selection.* The K test units with the highest scores $S(X_i)$ are selected, where $\mathcal{S} : \mathcal{X} \to \mathbb{R}$ is a pre-trained score function. For example, $S(X_i)$ may be the predicted binding affinity for a drug with chemical structure X_i or the predicted health risk of a patient with features X_i, and the drug discovery process or clinical system admits a fixed number of new units.
2. *Selection based on joint quantiles.* A unit j is selected if its score $S(X_{n+j})$ surpasses the qth quantile of both the calibration *and* test scores $\{S(X_i)\}_{i=1}^{n+m}$. For instance, a scientist may be interested in toxicities of drug candidates in $\mathcal{D}_{\text{test}}$ where those in $\mathcal{D}_{\text{calib}}$ have been tested, but they only focus on drugs with highest predicted activities $S(X_i)$ in the entire library.
3. *Selection based on calibration quantiles.* A test unit j is selected if its score $S(X_{n+j})$ surpasses the qth quantile of calibration scores $\{S(X_i)\}_{i=1}^{n}$. This may happen when a doctor uses the predicted health risks of existing patients to determine a normal range, and picks test units anticipated to have a relatively extreme health risk.
4. *Selection by black-box optimization procedures.* A test unit j is selected by running an arbitrary (even black-box) optimization program that does not involve $\{Y_i\}_{i \in \mathcal{I}_{calib}}$. One such example is optimization under constraints: apart from the score $S(X_{n+j})$, each unit j is associated with a cost C_{n+j}, and $\widehat{\mathcal{S}}$ is the subset of test units that maximizes $\sum_{j \in \widehat{\mathcal{S}}} S(X_{n+j})$ subject to $\sum_{j \in \widehat{\mathcal{S}}} C_{n+j} \le c$, where c reflects the total budget. This may happen in healthcare management systems that optimize resources subject to constraints and send patients to different care categories, or in candidate screening for job interviews subject to budget and diversity constraints. More generally, in drug discovery, scientists may run a complex Bayesian optimization algorithm to select the next batch of drugs to evaluate (Pyzer-Knapp, 2018), which may be viewed as a black box process. No matter how complicated these optimization programs are, as long as they do not involve calibration labels, JOMI supports efficient uncertainty quantification afterwards. We will see some stylized examples in our numerical experiments.

We additionally show in online supplementary material Section S2 that we can further improve the computational efficiency of JOMI for rules (1)–(3) by deriving exact forms of the reference sets, where in each case the computation complexity is $O(\max\{m, n\})$. We also note that rules (1)–(2) were considered in Bao et al. (2024), where they propose methods that achieve (2) and FCR control. The prediction sets proposed therein coincide with ours, and our results imply that they in fact achieve the strong selection-conditional coverage in (3) for free.

4.2 Selection based on conformal p-values

The second class of selection rules we study concern selecting units whose outcomes satisfy certain conditions while controlling some type-I error. To this end, the test units are selected by

thresholding a class of conformal p-values, where each p-value is computed via contrasting a test point with the calibration data. As such, the selection rule can be complicated and asymmetric.

We will follow the framework of Jin and Candès (2023) to define the p-values and selection rules, who study the problem of discovering test units with large outcomes. Examples include selecting drugs with sufficiently high binding affinities, finding highly competent job candidates, and identifying patients who benefit from a treatment, etc. In these problems, predictions from machine learning models serve as proxies for the true outcomes of interest that are too expensive or impossible to evaluate, and the selection procedure leverages the power of predictions to select units with large outcomes while ensuring error control.

Given test points $\{X_{n+j}\}_{j\in[m]}$ and (potentially random) thresholds $c_{n+j} \in \mathbb{R}$, the goal is to select those $Y_{n+j} > c_{n+j}$ while controlling the number/fraction of false positives. The statistical evidence for detecting a large outcome is quantified by the so-called 'conformal p-values'.

Suppose the calibration data are $\{(X_i, Y_i, c_i)\}_{i=1}^{n}$, such that the tuples $\{X_i, Y_i, c_i\}_{i=1}^{n+m}$ are exchangeable. Assume access to a score function $S : \mathcal{X} \times \mathbb{R} \to \mathbb{R}$ such that $S(x, y)$ is non-increasing in y for any $x \in \mathcal{X}$. An example is $S(x, y) = \widehat{\mu}(x) - y$, where $\widehat{\mu}(x)$ is a point predictor trained on $\mathcal{D}_{\text{train}}$. We then compute $\widehat{S}_i = S(X_i, c_i)$ for $i \in [n + m]$ and define the conformal p-values[1]

$$p_j = \frac{1 + \sum_{i\in[n]} \mathbb{1}\{\widehat{S}_i \geq \widehat{S}_{n+j},\ Y_i \leq c_i\}}{n+1}, \quad j \in [m]. \tag{11}$$

Hereafter, we call S the *selection* score function. We can show that p_j is valid in the sense that $\mathbb{P}(p_j \leq t,\ \ Y_{n+j} \leq c_{n+j}) \leq t$, $\forall t \in [0, 1]$. That is, testing with p_j controls the type-I error in finding one large outcome, accounting for the randomness in the outcomes as well. We then select test units whose conformal p-values are below a threshold, the choice of which determines the type of error control guarantee. Some examples are given below.

1. *Fixed threshold*. We select test units whose conformal p-values (11) are below a fixed threshold $q \in (0, 1)$. This could happen when testing a single hypothesis, or testing multiple hypotheses with Bonferroni correction. The latter controls the family-wise error rate in finding large outcomes, which is useful in highly risk-sensitive settings such as disease diagnosis.
2. *Benjamini–Hochberg threshold*. We select test units whose conformal p-values (11) are below a data-dependent threshold given by the Benjamini–Hochberg (BH) procedure (Benjamini & Hochberg, 1995). This selection procedure is shown in Jin and Candès (2023) to control the FDR in detecting large outcomes, which is useful in exploratory screening such as drug discovery for ensuring efficient resource use in follow-up investigations.

For generality, we consider the selection rule in the form of $\widehat{\mathcal{S}}_{cp} = \{j \in [m] : \widehat{S}_{n+j} \geq \tau\}$, where τ is a stopping time adapted to the filtration $\{\sigma(\{A_s\}_{s\leq t}, \{B_s\}_{s\leq t})\}_{t\in\mathbb{R}}$:

$$A_s = 1 + \sum_{i\in[n]} \mathbb{1}\{\widehat{S}_i \geq s,\ Y_i \leq c_i\}, \quad B_s = \sum_{j\in[m]} \mathbb{1}\{\widehat{S}_{n+j} \geq s\}. \tag{12}$$

Equivalently, for any $t \in \mathbb{R}$, we can write $\mathbb{1}\{\tau \leq t\} = f_t(\{A_s\}_{s\leq t}, \{B_s\}_{s\leq t})$ for some function f_t. It can be shown that the two examples above are special cases of this general rule (these results can be found in Section S5.3 in the online supplementary material).

We now provide a general solution for constructing selection-conditional prediction sets corresponding to such rules. The challenge here is that all the p-values depend on each other, as they leverage the same set of calibration data, and/or the data-driven threshold determined by all p-values would further add to the intricacy. We note that the BH-based rule is studied in Bao et al. (2024) with approximate FCR control, while we are to provide an efficient solution with

[1] Our p-value slightly modify the definition in Jin and Candès (2023) for the ease of describing our prediction sets. Similar to the original ones, our p-values control the type-I error in detecting one large outcome. In addition, using our p-values in their original procedures maintains error control with improved power; we discuss these results in online supplementary material Section S5.2 for completeness.

exact coverage guarantee. The following proposition lays out the form of the reference set and its validity, with its proof delegated to Section S4.5 in the online supplementary material.

Proposition 4 Suppose the selection set is $\widehat{\mathcal{S}}_{cp} = \{j \in [m] : \widehat{S}_{n+j} \geq \tau\}$, where τ is determined by $\mathbb{1}\{\tau \leq t\} = f_t(\{A_s\}_{s\leq t}, \{B_s\}_{s\leq t})$, for some f_t and (A_s, B_s) defined in (12), $\forall t \in \mathbb{R}$. For any $\mathfrak{S} \in 2^{[m]}$ and $j \in [m]$ such that $j \in \widehat{\mathcal{S}}_{cp}$ and $\widehat{\mathcal{S}}_{cp} \in \mathfrak{S}$, the reference set can be simplified as

$$\begin{aligned}
\widehat{\mathcal{R}}^{\mathrm{cp}}_{n+j}(y) &= \mathbb{1}\{y \leq c_{n+j}\} \cdot \widehat{\mathcal{R}}^{\mathrm{cp},1}_{n+j} + \mathbb{1}\{y > c_{n+j}\} \cdot \widehat{\mathcal{R}}^{\mathrm{cp},0}_{n+j}, \quad \text{where} \\
\widehat{\mathcal{R}}^{\mathrm{cp},k}_{n+j} &= \{i \in [n] : Y_i \leq c_i, \widehat{S}_i \geq \tau^{(j)}(k, 0), \\
&\qquad \{\ell \in [m] : \widehat{S}^{\mathrm{swap}(i,j)}_{n+\ell} \geq \tau^{(j)}(k, 0)\} \in \mathfrak{S}\} \\
&\quad \cup \{i \in [n] : Y_i > c_i, \widehat{S}_i \geq \tau^{(j)}(k, 1), \\
&\qquad \{\ell \in [m] : \widehat{S}^{\mathrm{swap}(i,j)}_{n+\ell} \geq \tau^{(j)}(k, 1)\} \in \mathfrak{S}\}.
\end{aligned} \tag{13}$$

For $k, \ell \in \{0, 1\}$, the adjusted threshold $\tau^{(j)}(k, \ell)$ is given by

$$\begin{aligned}
&\mathbb{1}\{\tau^{(j)}(k, \ell) \leq t\} = f_t(\{A^{(j)}_s(k, \ell)\}_{s\leq t}, \{B^{(j)}_s\}_{s\leq t}), \text{ where} \\
&A^{(j)}_s(k, \ell) = \ell + \sum_{i\in[n]} \mathbb{1}\{\widehat{S}_i \geq s, Y_i \leq c_i\} + k \cdot \mathbb{1}\{\widehat{S}_{n+j} \geq s\}, \\
&B^{(j)}_s = 1 + \sum_{\ell\neq j} \mathbb{1}\{\widehat{S}_{n+\ell} \geq s\}.
\end{aligned} \tag{14}$$

Based on Proposition 4, the JOMI prediction set is given by

$$\widehat{C}_{\alpha,n+j} = \left\{y \in \mathcal{Y} : y > c_{n+j},\, V(X_{n+j}, y) \leq \widehat{q}_0\right\} \cup \left\{y \in \mathcal{Y} : y \leq c_{n+j},\, V(X_{n+j}, y) \leq \widehat{q}_1\right\},$$

where $\widehat{q}_k = \mathrm{Quantile}(1-\alpha; \{V_i : i \in \widehat{\mathcal{R}}^{cp,k}_{n+j}\} \cup \{\infty\})$ for $k = 0, 1$. See Algorithm 2 for a summary of the complete procedure, where we only present the deterministic version for simplicity. The overall computation complexity is at most $O(m(m+n)|\widehat{\mathcal{S}}|)$.

4.3 Selection based on conformal prediction sets

The final class of selection rules we study are based on the properties of (preliminary) prediction sets, usually constructed by running the vanilla SCP. Such use cases have appeared implicitly in many heuristic applications of conformal prediction. For example, practitioners may select units whose prediction intervals are shorter/longer than a threshold, which roughly indicates enough confidence (Sokol et al., 2024). People may also select units whose prediction sets entirely lie above a threshold, which roughly indicates a desired outcome (Svensson et al., 2017). Note that the original prediction intervals are no longer valid *conditional on being selected* (Jin & Candès, 2023), and thus using them for interpreting downstream uncertainty can be misleading. In this section, we apply our general framework to re-calibrate prediction sets for the units selected in such a way.

Formally, we consider two stages of prediction set construction. The one constructed in the first stage, called the *preliminary* prediction set, is used for determining the selection set $\widehat{\mathcal{S}}$. The one in the second stage, which we call the *selective* prediction set, is the one we are to build with JOMI. Following SCP in Section 3.1, we let $S(x, y)$ and $V(x, y)$ be the nonconformity score functions for the two stages, respectively. The $(1-\beta)$-level preliminary prediction set for the jth test unit is

$$\widehat{C}^{\mathrm{prelim}}_{\beta,n+j} = \left\{y \in \mathcal{Y} : S(X_{n+j}, y) \leq \eta\right\}, \tag{15}$$

where η is the $K := \lceil (1-\beta)(n+1) \rceil$th smallest element in $\{S(X_i, Y_i)\}_{i=1}^n$.

We consider any selection rule based on the preliminary prediction set $\widehat{C}^{\mathrm{prelim}}_{\beta,n+j}$. Note that by (15), given the first-stage score function $S(\cdot, \cdot)$, the form of $\widehat{C}^{\mathrm{prelim}}_{\beta,n+j}$ is fully determined by X_{n+j} and η.

Algorithm 2 JOMI for selection based on conformal *p*-values

Input: Calibration data $\mathcal{D}_{\text{calib}}$; test data $\mathcal{D}_{\text{test}}$; miscoverage level α; selection taxonomy $\mathfrak{S}$; selection rule $\mathcal{S}$; nonconformity score function $V(\cdot, \cdot)$.

Compute $\widehat{\mathcal{S}} = \mathcal{S}(\mathcal{D}_{\text{calib}}, \mathcal{D}_{\text{test}})$.

for $j \in \widehat{\mathcal{S}}$ **do**

Compute $\tau^{(j)}(k, \ell)$ as in (14), for $(k, \ell) \in \{0, 1\}$.

Compute $\widehat{\mathcal{R}}^{\text{cp},0}_{n+j}$ and $\widehat{\mathcal{R}}^{\text{cp},1}_{n+j}$ as (13).

Compute $\widehat{q}_k = \text{Quantile}(1-\alpha; \{V_i : i \in \widehat{\mathcal{R}}^{\text{cp},k}_{n+j}\} \cup \{\infty\})$, for $k = 0, 1$.

Compute $\widehat{C}_{\alpha,n+j} = \{y \in \mathcal{Y} : y > c_{n+j},\ V(X_{n+j}, y) \le \widehat{q}_0\} \cup \{y \in \mathcal{Y} : y \le c_{n+j},\ V(X_{n+j}, y) \le \widehat{q}_1\}$.

Output: $\{\widehat{C}_{\alpha,n+j}\}_{j \in \widehat{\mathcal{S}}}$.

We can thus express any selection rule through $\mathcal{L} : \mathcal{X} \times \mathbb{R} \mapsto \{0, 1\}$, where $\mathcal{L}(X_{n+j}, \eta) = 1$ means selecting the unit and $\mathcal{L}(X_{n+j}, \eta) = 0$ otherwise. An example is selecting based on prediction interval lengths: following Sokol et al. (2024), suppose that we use CQR (Romano et al., 2019) in the first stage, i.e. $S(x, y) = \max\{\widehat{q}_L(x) - y, y - \widehat{q}_U(x)\}$, where $\widehat{q}_L(x)$ and $\widehat{q}_U(x)$ are estimates of some lower and upper conditional quantiles. Selecting prediction intervals shorter than a threshold λ gives $\mathcal{L}(x, \eta) = \mathbb{1}\{\widehat{q}_U(x) - \widehat{q}_L(x) + 2\eta \le \lambda\}$. As another example, for a binary outcome Y, we might want to select units whose prediction set is a singleton, leading to $\mathcal{L}(x, \eta) = 1\{S(x, 1) \le \eta < S(x, 0) \text{ or } S(x, 0) \le \eta < S(x, 1)\}$.

Having determined the selection rule $\mathcal{L}$, the selection set is thus $\mathcal{S}(\mathcal{D}_{calib}, \mathcal{D}_{test}) = \widehat{\mathcal{S}}_{ps} = \{j \in [m] : \mathcal{L}(X_{n+j}, \eta) = 1\}$. We are to derive a computationally efficient but slightly conservative version of the JOMI prediction set, which nevertheless has tight coverage in all our numerical experiments (see Section 6). Define

$$\begin{aligned}\widehat{C}^{\text{ps}}_{\alpha,n+j} := &\{y : \eta^- \le S(X_{n+j}, y) \le \eta^+\} \cup \{y : V(X_{n+j}, y) \le q_{j,1} \text{ and } S(X_{n+j}, y) < \eta^-\} \\ &\cup \{y : V(X_{n+j}, y) \le q_{j,2} \text{ and } S(X_{n+j}, y) > \eta^+\},\end{aligned} \tag{16}$$

where η^+ and η^- are the $(K+1)$th and $(K-1)$th smallest element in $\{S_i\}_{i=1}^n$, respectively, and

$$\begin{aligned}q_{j,1} := \text{Quantile}\Big(1-\alpha; &\Big\{V_i : i \in [n], S_i \le \eta^-, \mathcal{L}(X_i, \eta) = 1, \{\ell \in [m] : \mathcal{L}(X^{\text{swap}(i,j)}_{n+\ell}, \eta) = 1\} \in \mathfrak{S}\Big\} \\ &\cup \Big\{V_i : S_i > \eta^-,\ \mathcal{L}(X_i, \eta^-) = 1, \{\ell \in [m] : \mathcal{L}(X^{\text{swap}(i,j)}_{n+\ell}, \eta^-) = 1\} \in \mathfrak{S}\Big\}\Big); \\ q_{j,2} := \text{Quantile}\Big(1-\alpha; &\Big\{V_i : i \in [n], S_i \le \eta, \mathcal{L}(X_i, \eta^+) = 1, \{\ell \in [m] : \mathcal{L}(X^{\text{swap}(i,j)}_{n+\ell}, \eta^+) = 1\} \in \mathfrak{S}\Big\} \\ &\cup \Big\{V_i : S_i > \eta,\ \mathcal{L}(X_i, \eta) = 1, \{\ell \in [m] : \mathcal{L}(X^{\text{swap}(i,j)}_{n+\ell}, \eta) = 1\} \in \mathfrak{S}\Big\}\Big).\end{aligned} \tag{17}$$

We prove the validity of $\widehat{C}^{ps}_{\alpha,n+j}$ in (16) below, whose proof is in online supplementary material Section S4.6.

Proposition 5 For any selection rule $\mathcal{L}$, any $j \in [m]$ and any selection taxonomy $\mathfrak{S}$ such that $j \in \widehat{\mathcal{S}}_{ps}$ and $\widehat{\mathcal{S}}_{ps} \in \mathfrak{S}$, $\widehat{C}^{ps}_{\alpha,n+j}$ is a superset of the JOMI prediction set $\widehat{C}_{\alpha,n+j}$ defined in (7), and

$$\widehat{C}^{\text{ps}}_{\alpha,n+j} \backslash \widehat{C}_{\alpha,n+j} \subseteq \{y \in \mathcal{Y} : \eta^- \le S(X_{n+j}, y) \le \eta^+\}.$$

By Proposition 5, the conservativeness of $\widehat{C}^{ps}_{\alpha,n+j}$ is quite limited, as η^- and η^+ are usually very close to each other. We also verify its tight empirical coverage in Section 6.

The procedure is summarized in Algorithm 3. For each j, the computation cost of $\widehat{C}^{ps}_{\alpha,n+j}$ is $O(m+n)$, and therefore the overall computation cost is $O(m(m+n))$.

Algorithm 3 JOMI for selection based on preliminary prediction sets

Input: Calibration data $\mathcal{D}_{\text{calib}}$; test data $\mathcal{D}_{\text{test}}$; selection taxonomy Ξ; selective miscoverage level α; first-stage miscoverage level β; selection rule $\mathcal{S}$; first-stage score function $S(\cdot, \cdot)$; nonconformity score function $V(\cdot, \cdot)$.

$\widehat{\mathcal{S}} = \mathcal{S}(\mathcal{D}_{\text{calib}}, \mathcal{D}_{\text{test}})$.

Compute η^+ as the $(K+1)$-th order statistic of $\{S_i\}_{i=1}^n$, where $K = \lceil (1-\beta)(n+1) \rceil$.

Compute η^+ as the $(K-1)$-th order statistic of $\{S_i\}_{i=1}^n$.

for $j \in \widehat{\mathcal{S}}$ **do**

  Compute $q_{j,1}$ and $q_{j,2}$ as in (17).

  Compute $\widehat{C}^{\text{ps}}_{\alpha,n+j}$ as in (16).

Output: $\{\widehat{C}_{\alpha,n+j}\}_{j \in \widehat{\mathcal{S}}}$.

Remark 2 We note that it is possible that the prediction interval produced by Algorithm 3 does not satisfy the constraints on the preliminary prediction sets. This is, however, a desired feature in our setting where the selection rule is given and the inference step is decoupled from the selection step; the selection-conditional prediction sets should be used as tools for informing further decisions/analysis. Otherwise, we may need to take an orthogonal strategy: change the selection algorithm, for which ideas from Jin and Candès (2023) and Gazin et al. (2025) may be useful.

5 Application to drug discovery

In drug discovery, powerful prediction machines are increasingly used to guide the search of promising drug candidates. For such high-stakes decisions, it is important to quantify the uncertainty in the predictions (Jin & Candès, 2023; Laghuvarapu et al., 2024; Svensson et al., 2017). Meanwhile, selection issues naturally arise as scientists may only focus on seemingly promising drugs.

In this section, we apply *JOMI* to several application scenarios in drug discovery with a selective nature. In some cases, *JOMI* yields shorter prediction intervals than vanilla conformal prediction when the latter is under-confident; in others, it makes the just right inflation of the prediction interval to provide exact selection-conditional coverage.

Application scenarios. In the main text, we focus on *drug property prediction* (DPP), a classification problem where the binary outcome indicates whether a drug candidate binds to a prespecified disease target, and the covariates are the (encoded) chemical structure of the drug compound. Due to limited space, we defer the results for several selection scenarios in drug-target-interaction prediction (DTI) to online supplementary material Section S3.4. DTI is a regression problem where each sample is a pair of drug and disease target. The outcome of interest is a real-valued variable indicating the binding affinity of that pair. The covariates are the (concatenated) encoded structure of both.

Selection rules. We consider three types of realistic selection rules $\mathcal{S}$:

1. *Covariate-dependent top-K selection*: selecting drugs with highest predicted binding affinities.
   (i) Top-K among test data. When the scientist has a fixed budget of investigating K drug candidates in the next phase, one may select K test samples with the largest $\widehat{\mu}(X_{n+j})$.
   (ii) Top-K among mixed data. When the scientist is to investigate other properties for promisingly active drug candidates in the next phase, they may select K units in $\mathcal{D}_{\text{calib}} \cup \mathcal{D}_{\text{test}}$ with the highest predicted affinities.
   (iii) Calibration-referenced selection. The scientist may use the $\mathcal{D}_{\text{calib}}$ as reference and select test samples whose predicted activities are greater than the Kth highest in $\{\widehat{\mu}(X_i)\}_{i \in \mathcal{I}_{\text{calib}}}$.
2. *Conformal selection*. The scientist might also obtain a subset of active drugs while controlling the FDR below some $q \in (0, 1)$. In this case, $\mathcal{S}(\mathcal{D}_{\text{calib}}, \mathcal{D}_{\text{test}})$ is the set of test drugs picked by Conformal Selection in Jin and Candès (2023) at FDR level $q \in (0, 1)$.

3. *Selection with constraints*. The cost of developing a drug is a variable C_{n+j}, and one wants to select as many drugs with the highest predictions $\widehat{\mu}(X_{n+j})$ as possible while ensuring the total cost is below some constant C, or maximize the total reward while controlling the costs.

Evaluation metrics. We evaluate the selection-conditional (mis)coverage via the consistent estimator $\widehat{\text{Miscov}} = \frac{\sum_{j=1}^{m} \widehat{P}(j \in \widehat{\mathcal{S}},\, Y_{n+j} \notin \widehat{C}_{\alpha,n+j})}{\sum_{j=1}^{m} \widehat{P}(j \in \widehat{\mathcal{S}})}$, where $\widehat{P}$ is the empirical probability over all repeats. The size of prediction sets is evaluated by averaging the cardinality of $\widehat{C}_{\alpha,n+j}$ for classification or length of $\widehat{C}_{\alpha,n+j}$ for regression over selected test units in all repeats. We also evaluate (but do not show in figures for brevity) the false coverage rate via $\widehat{\text{FCR}} = \widehat{E}\left[\frac{\sum_{j=1}^{m} \mathbb{1}\{j \in \widehat{\mathcal{S}},\, Y_{n+j} \notin \widehat{C}_{\alpha,n+j}\}}{1 \vee |\widehat{\mathcal{S}}|}\right]$, where $\widehat{E}$ is the empirical mean over all repeats.

Data and prediction models. We use the HIV screening data in the DeepPurpose library (Huang et al., 2020) with a total sample size of $n_{tot} = 41{,}127$. The numerical features $X \in \mathcal{X}$ are encoded by Extended-Connectivity FingerPrints (Rogers & Hahn, 2010, ECFP) which characterize topological properties of the drug candidates. A small-scale neural network $\widehat{\mu}: \mathcal{X} \to [0, 1]$ is trained on randomly sampled 20% of the entire dataset. Then, the remaining is randomly split into two equally sized folds $\mathcal{D}_{\text{calib}}$ and $\mathcal{D}_{\text{test}}$. The exchangeability among $\mathcal{D}_{\text{calib}}$ and $\mathcal{D}_{\text{test}}$ is thus satisfied. Sampling of the training data and model training is independently repeated 10 times; within each, we randomly split calibration/test data for 100 times. This leads to $N = 1{,}000$ runs in total.

5.1 Top-K selection

We first consider the selection rule (i) top-K with test data. For a trained predictor $\widehat{\mu}$, we select (a) the highest K values of $\widehat{\mu}(X_j)$ and (b) the lowest K values of $\widehat{\mu}(X_j)$ among all test data. The results for (ii) top-K selection with both test and calibration data, as well as (iii) calibration-referenced selection are deferred to online supplementary material Section S3.1, showing similar patterns.

For a confidence level $\alpha \in (0, 1)$, we apply the vanilla conformal prediction and JOMI (both deterministic and randomized) to the selected units at level α, with $\mathfrak{S} = 2^{[m]}$. Since $|\widehat{\mathcal{S}}|$ is fixed, selection-conditional coverage implies FCR control at level α due to Proposition 2. In addition, we include the BY procedure (Benjamini & Yekutieli, 2005) as a heuristic baseline. We vary $\alpha \in \{0.1, 0.2, \ldots, 0.9\}$ and $K \in \{20, 100, 1{,}000, 2{,}000, 5{,}000, 10{,}000, 15{,}000\}$, and set V as the APS score (Romano et al., 2020). The results with another binary score are in online supplementary material Section S3.2.

Figure 3 shows the empirical selection-conditional miscoverage in the left panel and the average prediction set size in the right panel, both at nominal levels $\alpha \in \{0.1, 0.8\}$. The solid (resp. dashed) lines show the results when we select units with the highest (resp. lowest) predicted affinities.

The orange curves (`Vanilla_CP`) show that vanilla conformal prediction with APS scores is over-confident for units with the lowest predicted affinities while under-confident for units with highest predicted affinities. In contrast, the purple (`JOMI`) and blue (`JOMI_rand`) curves both show valid coverage for our proposed methods. Also, using the APS score introduces visible gap between the actual coverage and $1 - \alpha$ for *JOMI* due to discretization, which is made exact by its randomized version. Finally, while `BY` has lower-than-nominal miscoverage which shows its validity as a heuristic method, it is overly conservative, leading to much larger prediction sets.

Interestingly, vanilla conformal prediction with the APS score yields zero-cardinality prediction sets for units with the lowest predicted affinities with $\alpha = 0.8$ (the orange dashed line in the right panel of Figure 3). This is because vanilla CP covers other test units with very high rate, and thus marginal coverage is guaranteed even with empty prediction sets for the selected units. Of course, this is worrying if one cares more about these units at the bottom.

The behaviour of these methods also depends on the choice of the nonconformity score V. online supplementary material Section S3.2 shows the results for selection rules (i)–(iii) with the binary score $V(x, y) = y(1 - \widehat{\mu}(x)) + (1 - y)\widehat{\mu}(x)$. With the binary score, *JOMI* consistently achieves valid coverage for selected units, and the coverage gap of *JOMI* due to discretization is less visible. For vanilla CP, opposite to the situations here, it is over-confident for units with the highest predicted affinities yet under-confident for units with the lowest predicted affinities.

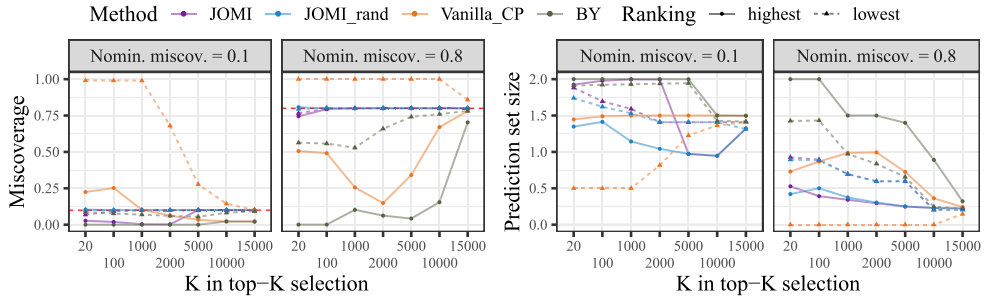

Figure 3. Empirical selection-conditional miscoverage (left) and prediction set size (right) in drug property prediction for vanilla conformal prediction (`Vanilla_CP`), BY correction (`BY`), JOMI and randomized JOMI (`JOMI_rand`) for test units whose $\widehat{\mu}(X_{n+j})$ are top-K (`highest`, solid line) or bottom-K (`lowest`, dashed line) among test units. The nominal miscoverage level $\alpha \in \{0.1, 0.8\}$. The results are averaged over $N = 1{,}000$ runs.

5.2 Conformal selection

We then consider conformal selection, where the focal test units are those believed to obey $Y = 1$ with false discovery rate control at level $q \in (0, 1)$. This problem was investigated in Bao et al. (2024) with no exact finite-sample coverage guarantees in theory (though their heuristic method performs reasonably in their empirical studies). We apply conformal selection (Jin & Candès, 2023) at FDR level $q \in \{0.2, \ldots, 0.9\}$ and $c_{n+j} \equiv 0.5$ to determine $\widehat{\mathcal{S}}$, and construct prediction intervals for units in $\widehat{\mathcal{S}}$ using both the APS score and the binary score. Experiments are repeated for $N = 1{,}000$ independent runs.

Figure 4 depicts $\widehat{Miscov}$ and prediction interval lengths with nominal miscoverage level $\alpha \in \{0.1, 0.8\}$ under various FDR levels q and two choices of nonconformity score V.

Vanilla CP is not calibrated for selected units: it is over-confident with both scores for $\alpha = 0.1$, while being over-confident with binary score and under-confident with APS score at $\alpha = 0.8$. In contrast, `JOMI` and `JOMI_rand` achieves valid selection-conditional coverage in all scenarios. There is some gap for `JOMI` with the APS score due to discretization, but not for `JOMI_rand` or the binary score. We observe an even lower empirical FCR than the conditional miscoverage for our methods (so they achieve valid FCR control); this is because the selection set can sometimes be empty for small values of FDR level q (recall Proposition 3). Compared with the heuristic methods of Bao et al. (2024), our method usually achieves smaller prediction set sizes whereas their method seems overly conservative. We conjecture that this is due to a more delicate choice of the reference set. Finally, `BY` is also overly conservative despite valid empirical coverage.

5.3 Selection with constraints

We now consider selecting units with the highest predicted binding affinities within a total budget of subsequent development. In this case, $\mathcal{S}(\mathcal{D}_{\text{calib}}, \mathcal{D}_{\text{test}}) = \{j \in [m] : \widehat{\mu}(X_{n+j}) \geq \bar{\mu}\}$, where $\bar{\mu} = \max\{\mu : \sum_{j=1}^{m} L_{n+j} \mathbb{1}\{\widehat{\mu}(X_{n+j}) \geq \mu\} \leq C\}$, and $\{L_{n+j}\}_{j \in [m]}$ are the costs.

We create semi-synthetic datasets since the original HIV data does not contain the cost information. Specifically, for each $i \in [n+m]$, we generate $L_i = \exp(3\widehat{\mu}(X_i)) + 2|\sin(\widehat{\mu}(X_i))| + \epsilon_i$, where $\widehat{\mu}(X_i)$ is the predicted binding affinity, and $\epsilon_i \sim \text{Exp}(1)$ are i.i.d. random variables that capture other cost-related information. Setting 20% of the data aside as the training set, we randomly sample the data without replacement so that $n = 2{,}500$ and $m = 2{,}500$.

The average miscoverage, prediction set size, and reference set size are reported in Figure 5. Interestingly, after adding cost constraints, vanilla conformal prediction is over-confident with the binary score and under-confident with the APS score. In contrast, our methods always yield near-exact coverage. From the right-most plot, we see that $|\widehat{\mathcal{R}}_j|$ is positively correlated with the number of selected test units. As usual, `BY` is overly conservative.

Finally, we report the empirical FCR in the three tasks in online supplementary material Section S3.3. Consistent with our theory in Section 2.2, selection-conditional coverage implies FCR control in top-K selection, and the empirical FCR is also close to the nominal coverage level under

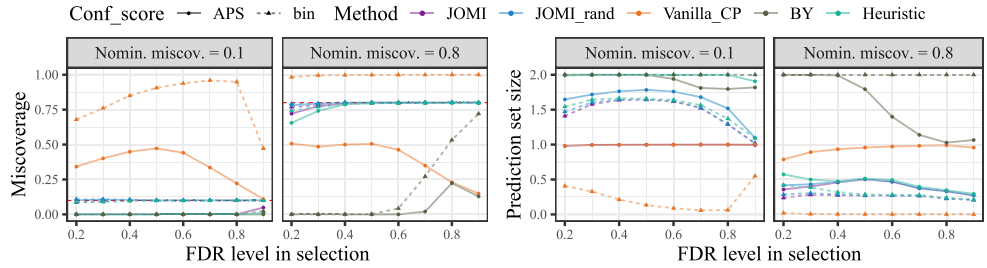

Figure 4. Empirical selection-conditional miscoverage (left) and prediction set size (right) in drug property prediction for test units selected by conformal selection at target FDR levels $q \in \{0.2, \ldots, 0.9\}$ when applying vanilla conformal prediction (`Vanilla_CP`), BY (`BY`), JOMI (`JOMI`), randomized JOMI (`JOMI_rand`), and the method of Bao et al. (2024) (`Heuristic`). Results with the APS score are in solid lines, while those with the binary score are in dashed lines. Details are otherwise the same as Figure 3.

conformal selection where the selection size is stable. However, in the constrained optimization task, we find that the empirical FCR of vanilla CP is much lower than the selection-conditional coverage. Similar to the ideas in Proposition 1, this is because the FCR is defined as zero when the selection set is empty. This shows the limitation of FCR as an error metric: even if it is controlled, on the 'interesting' event that the selection set is non-empty, the actual coverage can be lower than anticipated.

6 Application to health risk prediction

Prediction machines are also widely used in healthcare for guiding clinical decision-making. These decisions may come from complicated underlying processes such as clinical resource optimization (Ahmadi-Javid et al., 2017; Master et al., 2017) or preliminary uncertainty quantification (Olsson et al., 2022). Accounting for selections from complicated decision processes is necessary for reliable and informative uncertainty quantification. In this section, we demonstrate the application of our framework to health risk prediction settings. We will consider three selection rules:

1. *Selection with constraints*. Clinical decision makers may optimize a performance measure subject to certain constraints such as budget, capacity, or fairness (Castro & Petrovic, 2012; Gocgun & Puterman, 2014; Kemper et al., 2014), and send patients to different care categories. In Section 6.1, we consider a stylized example where we minimize the total predicted ICU stay subject to total cost budget, where the selected units can be viewed as patients sent to a certain category.
2. *Selecting small conformal prediction sets*. Based on preliminary conformal prediction, one may suggest human intervention for units with large set sizes (Olsson et al., 2022; Sokol et al., 2024) while leaving those with small prediction sets unattended. In Section 6.2, we recalibrate predictive inference for those whose preliminary prediction sets are shorter than a threshold.
3. *Selection based on upper prediction bounds*. Practitioners may also focus on units whose preliminary prediction intervals lie below a threshold (Levitskaya, 2023). In Section 6.3, we study selected test units whose preliminary prediction intervals have upper bounds below a threshold.

Among the above, rule (1) is covariate-dependent, while rules (2) and (3) can depend on the outcomes. All of them are efficiently tackled by the computation tricks in Section 4.

Our experiments use the ICU-stay data in the `MIMIC-IV` dataset (Johnson et al., 2023). Data (and features in X) are pre-processed using the pipeline provided by Gupta et al. (2022), with the outcome Y being the length of ICU stay. We use random forests in the `scikit-learn` Python package to train a point prediction model $\widehat{\mu}(\cdot)$ and two quantile regression models $\widehat{q}_{1-\beta/2}(\cdot)$ and $\widehat{q}_{\beta/2}(\cdot)$ using a holdout training set. Calibration and test data are randomly split with $n = 3,000$ and $m = 2,000$. As the discretization issue is minimal in this regression problem, we only present results for JOMI without randomization.

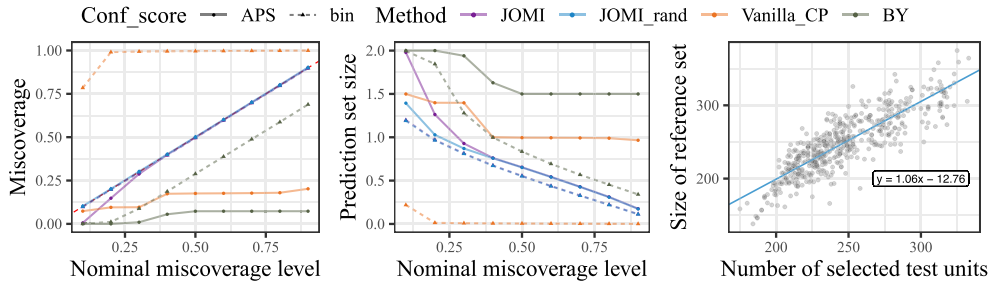

Figure 5. Empirical miscoverage rate (left), average length of prediction interval (middle), and scatter plots for averaged reference set size $\{|\widehat{\mathcal{R}}_j|\}_{j\in\widehat{\mathcal{S}}}$ vs. $|\widehat{\mathcal{S}}|$ (right), across $N = 500$ independent runs of `Vanilla_CP`, `BY`, `JOMI`, and `JOMI_rand`. The x-axis of left and middle plots is the nominal miscoverage level $\alpha \in \{0.1, 0.2, \ldots, 0.9\}$.

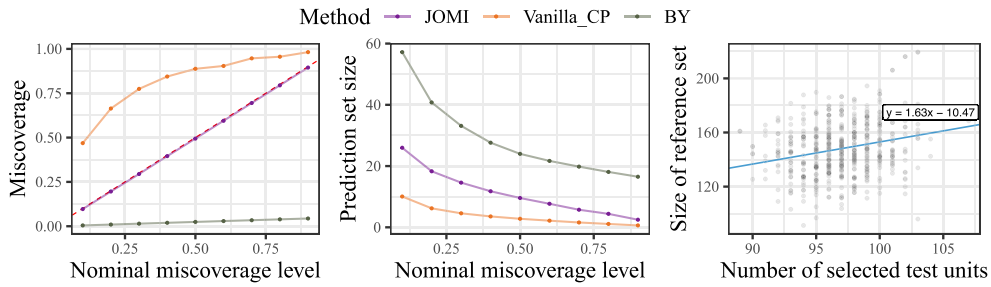

Figure 6. Empirical miscoverage rate (left), average length of prediction interval (middle), and scatter plots for the averaged reference set size vs. $|\widehat{\mathcal{S}}|$ (right), across $N = 500$ independent runs of `Vanilla_CP`, `BY`, `JOMI`, and `JOMI_rand`. The x-axis of left and middle plots is the nominal level $\alpha \in \{0.1, 0.2, \ldots, 0.9\}$.

6.1 Selection with constraints

We first study the case where test units are selected by minimizing the total predicted ICU stay time subject to a budget constraint. Formally, for each patient $i \in [m+n]$, we let $\widehat{\mu}(X_i)$ be its predicted ICU stay length, and $L_i > 0$ be the budget needed for them. Then, $\mathcal{S}(\mathcal{D}_{\text{calib}}, \mathcal{D}_{\text{test}})$ aims to solve the following optimization problem:

$$\begin{aligned} \underset{S\subseteq[m]}{\text{maximize}} \quad & \sum_{j\in S} \widehat{\mu}(X_{n+j}) \\ \text{subject to} \quad & \sum_{j\in S} L_{n+j} \le \bar{L}, \end{aligned} \tag{18}$$

where $\bar{L} = 200$ is a budget limit. Again, as the dataset does not come with drug development costs, we generate $L_i = \lceil \exp(3\widehat{\mu}(X_i)/\bar{\mu}) + |\sin(\widehat{\mu}(X_i))| + \epsilon_i - 1 + \varepsilon_i \rceil$, where $\bar{\mu} = \max_{i\in\mathcal{D}_{\text{train}}} |\widehat{\mu}(X_i)|$, and $\epsilon_i \sim \text{Exp}(1)$, $\varepsilon_i \sim \text{Unif}([0, 1])$ are independent random variables.

The optimization problem (18) is known as the Knapsack problem which is NP-hard. Nevertheless, there are efficient approximate solvers and we note that the validity of JOMI does not rely on exactness of the results; in our experiments, we use the Python package `mknapsack` (mknapsack, 2023). Existing methods such as Bao et al. (2024) cannot deal with such a complicated selection process. In contrast, our framework tackles this problem with a computation complexity that is polynomial in m, n, and the complexity of the subroutine $\mathcal{S}(\cdot, \cdot)$.

Figure 6 shows the empirical miscoverage, length of prediction interval, and sizes of the selection set and reference sets. While vanilla conformal prediction is over-confident and BY is overly conservative, our method achieves exact coverage for selected test units despite the complexity of the selection process. We also see a slightly positive correlation of $|\widehat{\mathcal{R}}_{n+j}|$ and $|\widehat{\mathcal{S}}|$.

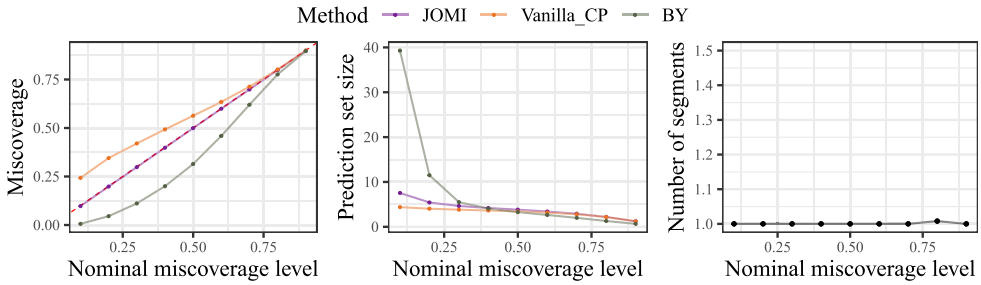

Figure 7. Empirical miscoverage rate (left), average length of prediction interval (middle), and average number of segments in $\widehat{C}^{\text{ps}}_{\alpha,n+j}$ (right), across $N = 500$ runs of `Vanilla_CP` and `JOMI` when test units with short preliminary prediction sets are selected. The x-axis is nominal levels $\alpha \in \{0.1, 0.2, \ldots, 0.9\}$.

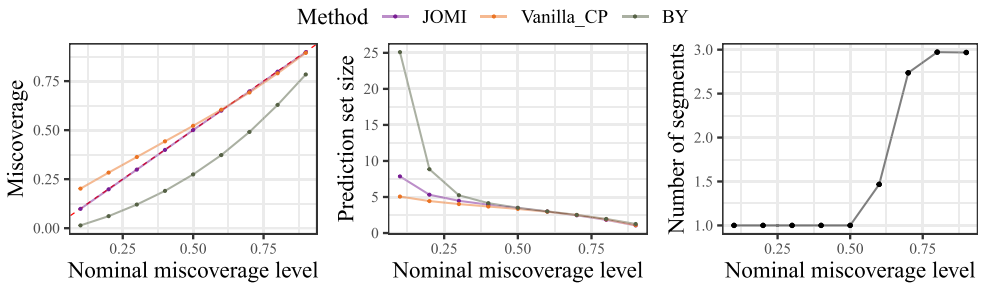

Figure 8. Empirical miscoverage rate (left), average length of prediction interval (middle), and average number of segments in $\widehat{C}^{\text{ps}}_{\alpha,n+j}$ (right), across $N = 500$ independent runs of `Vanilla_CP` and `JOMI`. The nominal miscoverage levels on the x-axis are $\alpha \in \{0.1, 0.2, \ldots, 0.9\}$.

6.2 Selecting small-sized prediction sets

We then consider the second selection rule, where we first build preliminary conformal prediction intervals $\widehat{C}^{\text{prelim}}_{\alpha,n+j}$ via the score function $S(x, y) = |y - \widehat{\mu}(x)|/\widehat{\sigma}(x)$ (Lei et al., 2018); both the point prediction function $\widehat{\mu}(\cdot)$ and the conditional standard deviation are estimated via random forests. We then select those test units with $|\widehat{C}^{\text{prelim}}_{\alpha,n+j}| \leq 5$, i.e. the upper and lower bounds of the preliminary prediction intervals are less than 5 days apart. This mimics the ideas in (Ren et al., 2023; Sokol et al., 2024) where small-sized prediction sets are 'certified' as confident.

After selection, we leverage the method in Section 4.3 to construct $\widehat{C}^{\text{ps}}_{\alpha,n+j}$ for all selected test units. Since $\widehat{C}^{\text{ps}}_{\alpha,n+j}$ is a superset of the exact output $\widehat{C}_{\alpha,n+j}$, we evaluate its empirical coverage to investigate whether it is over-conservative. Also, note from (16) that it is the union of three subsets; we also evaluate the number of disjoint segments in $\widehat{C}^{\text{ps}}_{\alpha,n+j}$.

The miscoverage and length of prediction sets are reported in the left and middle plots in Figure 7. We observe that selectively certifying short prediction intervals can lead to undercoverage (orange curve), while JOMI achieves exact coverage (purple curve) by inflating the prediction sets, meaning that the superset $\widehat{C}^{\text{ps}}_{\alpha,n+j}$ is effectively quite tight. The average number of disjoint segments in the right plot of Figure 7, which shows that $\widehat{C}^{\text{ps}}_{\alpha,n+j}$ is almost always one single interval.

6.3 Selecting prediction sets below a threshold

Finally, we study selection rule (iii) which is also based on a preliminary prediction set constructed in the same way as Section 6.2. We imagine that practitioners select a test unit $n + j$ if the upper bound of $\widehat{C}^{\text{pre}}_{\alpha,n+j}$ is below 6, i.e. it appears the patient will stay in ICU less than 6 days.

The miscoverage rate, length of prediction sets, and number of disjoint segments averaged over $N = 500$ independent runs of JOMI (Section 4.3) and vanilla conformal prediction are

summarized in Figure 8. We see that preliminary prediction sets with low upper bounds tend to under-cover for small α, while JOMI achieves exact coverage despite that we construct a superset of $\widehat{C}_{\alpha,n+j}$. However, JOMI may produce multiple segments for large values of α.

Finally, we present in online supplementary material Section S3.7 the tight empirical FCR control of JOMI in the above three tasks, since the size of the selection set is stable and nonzero.

Acknowledgments

The authors thank the anonymous reviewers for their constructive comments.

Conflict of interest: None declared.

Funding

Z.R. acknowledges support from the National Science Foundation under grant DMS-2413135. Y.J. was partially supported by the Harvard Data Science Wojcicki-Troper Postdoctoral Fellowship.

Data availability

The code for reproducing the numerical results in this article is available at https://github.com/ying531/JOMI-paper. The data underlying this article are available online, with references given at appropriate places in the article.

Supplementary material

Supplementary material is available online at *Journal of the Royal Statistical Society: Series B*.

References

Ahmadi-Javid A., Jalali Z., & Klassen K. J. (2017). Outpatient appointment systems in healthcare: A review of optimization studies. *European Journal of Operational Research*, *258*(1), 3–34. https://doi.org/10.1016/j.ejor.2016.06.064

Angelopoulos A. N., & Bates S. (2021). Conformal prediction: A gentle introduction. *Foundations and Trends® in Machine Learning*, *16*(4), 494–591. https://doi.org/10.1561/2200000101

Bao Y., Huo Y., Ren H., & Zou C. (2024). Selective conformal inference with false coverage-statement rate control. *Biometrika*, *111*(3), 727–742. https://doi.org/10.1093/biomet/asae010

Barber R. F., Candès E. J., Ramdas A., & Tibshirani R. J. (2021). The limits of distribution-free conditional predictive inference. *Information and Inference: A Journal of the IMA*, *10*(2), 455–482. https://doi.org/10.1093/imaiai/iaaa017

Benjamini Y., & Hochberg Y. (1995). 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. https://doi.org/10.1111/j.2517-6161.1995.tb02031.x

Benjamini Y., & Yekutieli D. (2005). False discovery rate–adjusted multiple confidence intervals for selected parameters. *Journal of the American Statistical Association*, *100*(469), 71–81. https://doi.org/10.1198/016214504000001907

Castro E., & Petrovic S. (2012). Combined mathematical programming and heuristics for a radiotherapy pretreatment scheduling problem. *Journal of Scheduling*, *15*(3), 333–346. https://doi.org/10.1007/s10951-011-0239-8

Gazin U., Heller R., Marandon A., & Roquain E. (2025). Selecting informative conformal prediction sets with false coverage rate control. *Journal of the Royal Statistical Society Series B: Statistical Methodology*. https://doi.org/10.1093/jrsssb/qkae120

Gocgun Y., & Puterman M. L. (2014). Dynamic scheduling with due dates and time windows: An application to chemotherapy patient appointment booking. *Health Care Management Science*, *17*(1), 60–76. https://doi.org/10.1007/s10729-013-9253-z

Gupta M., Gallamoza B., Cutrona N., Dhakal P., Poulain R., & Beheshti R. (2022). An extensive data processing pipeline for mimic-iv. In *Machine learning for health* (pp. 311–325). PMLR.

Huang K., Fu T., Glass L. M., Zitnik M., Xiao C., & Sun J. (2020). Deeppurpose: A deep learning library for drug–target interaction prediction. *Bioinformatics*, *36*(22–23), 5545–5547. https://doi.org/10.1093/bioinformatics/btaa1005

Jin Y., & Candès E. J. (2023). Selection by prediction with conformal *p*-values. *Journal of Machine Learning Research*, *24*(244), 1–41. https://dl.acm.org/doi/abs/10.5555/3648699.3648943

Johnson A. E., Bulgarelli L., Shen L., Gayles A., Shammout A., Horng S., Pollard T. J., Hao S., Moody B., Gow B., Lehman L.-W. H., Celi L. A., & Mark R. G. (2023). MIMIC-IV, a freely accessible electronic health record dataset. *Scientific Data*, *10*(1), 1. https://doi.org/10.1038/s41597-022-01899-x

Kemper B., Klaassen C. A., & Mandjes M. (2014). Optimized appointment scheduling. *European Journal of Operational Research*, *239*(1), 243–255. https://doi.org/10.1016/j.ejor.2014.05.027

Laghuvarapu S., Lin Z., & Sun J. (2024). Codrug: Conformal drug property prediction with density estimation under covariate shift. *Advances in Neural Information Processing Systems*, *36*, 37728–37747. https://openreview.net/forum?id=GgdFLb94Ld

Lee J. D., Sun D. L., Sun Y., & Taylor J. E. (2016). Exact post-selection inference, with application to the lasso. *Annals of Statistics*, *44*(3), 907–927. https://doi.org/10.1214/15-AOS1371

Lei J., G'Sell M., Rinaldo A., Tibshirani R. J., & Wasserman L. (2018). Distribution-free predictive inference for regression. *Journal of the American Statistical Association*, *113*(523), 1094–1111. https://doi.org/10.1080/01621459.2017.1307116

Levitskaya V. (2023). How to boost business decisions with conformal prediction and confidence. https://redfield.ai/conformal-prediction-for-business/

Master N., Zhou Z., Miller D., Scheinker D., Bambos N., & Glynn P. (2017). Improving predictions of pediatric surgical durations with supervised learning. *International Journal of Data Science and Analytics*, *4*(1), 35–52. https://doi.org/10.1007/s41060-017-0055-0

mknapsack (2023). Python package mknapsack 1.1.12.

Olsson H., Kartasalo K., Mulliqi N., Capuccini M., Ruusuvuori P., Samaratunga H., Delahunt B., Lindskog C., Janssen E. A., Blilie A., Egevad L., Spjuth O., & Eklund M. (2022). Estimating diagnostic uncertainty in artificial intelligence assisted pathology using conformal prediction. *Nature Communications*, *13*(1), 7761. https://doi.org/10.1038/s41467-022-34945-8

Pyzer-Knapp E. O. (2018). Bayesian optimization for accelerated drug discovery. *IBM Journal of Research and Development*, *62*(6), 2:1–2:7. https://doi.org/10.1147/JRD.2018.2881731

Ren A. Z., Dixit A., Bodrova A., Singh S., Tu S., Brown N., Xu P., Takayama L., Xia F., & Varley J. (2023). Robots That Ask For Help: Uncertainty Alignment for Large Language Model Planners. *Proceedings of the Conference on Robot Learning (CoRL)*. https://openreview.net/forum?id=4ZK8ODNyFXx

Rogers D., & Hahn M. (2010). Extended-connectivity fingerprints. *Journal of Chemical Information and Modeling*, *50*(5), 742–754. https://doi.org/10.1021/ci100050t

Romano Y., Patterson E., & Candes E. (2019). Conformalized quantile regression. *Advances in Neural Information Processing Systems*, *32*, 3543–3553. https://dl.acm.org/doi/10.5555/3454287.3454605

Romano Y., Sesia M., & Candes E. (2020). Classification with valid and adaptive coverage. *Advances in Neural Information Processing Systems*, *33*, 3581–3591. https://dl.acm.org/doi/10.5555/3495724.3496026

Sokol A., Moniz N., & Chawla N. (2024). 'Conformalized selective regression', arXiv, arXiv:2402.16300, preprint: not peer reviewed. https://doi.org/10.48550/arXiv.2402.16300

Svensson F., Afzal A. M., Norinder U., & Bender A. (2018). Maximizing gain in high-throughput screening using conformal prediction. *Journal of Cheminformatics*, *10*(1), 1–10. https://doi.org/10.1186/s13321-018-0260-4

Svensson F., Norinder U., & Bender A. (2017). Improving screening efficiency through iterative screening using docking and conformal prediction. *Journal of Chemical Information and Modeling*, *57*(3), 439–444. https://doi.org/10.1021/acs.jcim.6b00532

Tibshirani R. J., Rinaldo A., Tibshirani R., & Wasserman L. (2018). Uniform asymptotic inference and the bootstrap after model selection.

Vovk V., Gammerman A., & Shafer G. (2005). *Algorithmic learning in a random world* (Vol. 29). Springer.

Weinstein A., Fithian W., & Benjamini Y. (2013). Selection adjusted confidence intervals with more power to determine the sign. *Journal of the American Statistical Association*, *108*(501), 165–176. https://doi.org/10.1080/01621459.2012.737740

Weinstein A., & Ramdas A. (2020). Online control of the false coverage rate and false sign rate. In *International Conference on Machine Learning* (pp. 10193–10202). PMLR.