TRADING COVERAGE FOR PRECISION: CONFORMAL PREDICTION WITH LIMITED FALSE DISCOVERIES

Anonymous authors
Paper under double-blind review

ABSTRACT

In this paper, we develop a new approach to conformal prediction in which we aim to output a precise set of promising prediction candidates that is guaranteed to contain a limited number of incorrect answers. Standard conformal prediction provides the ability to adapt to model uncertainty by constructing a calibrated candidate set in place of a single prediction, with guarantees that the set contains the correct answer with high probability. In order to obey this coverage property, however, conformal sets can often become inundated with noisy candidates—which can render them unhelpful in practice. This is particularly relevant to large-scale settings where the cost (monetary or otherwise) of false positives is substantial, such as for in-silico screening for drug discovery, where any positively identified molecular compound is then manufactured and tested. We propose to trade coverage for precision by enforcing that the presence of incorrect candidates in the predicted conformal sets (i.e., the total number of false discoveries) is bounded according to a user-specified tolerance. Subject to this constraint, our algorithm then optimizes for a generalized notion of set coverage (i.e., the true discovery rate) that allows for any number of true answers for a given query (including zero). We demonstrate the effectiveness of this approach across a number of classification tasks in natural language processing, computer vision, and computational chemistry.

1 INTRODUCTION

For many classification problems, returning a set of plausible responses instead of single prediction is a useful way of representing uncertainty (Gammerman and Vovk, 2007; Lei, 2014; Bates et al., 2020). Aligned with this goal, conformal prediction (Vovk et al., 2005) is an increasingly popular method for creating confident prediction sets that provably contain the correct answer with high probability. Unfortunately, these guarantees do not come for free; in order to achieve proper coverage on difficult tasks, conformal predictors are often unable to rule out an overwhelming number of candidates—making their prediction sets large and inefficient (Angelopoulos et al., 2021c). This can make conformal predictors unusable in settings in which the cost of false positives is substantial.

Consider the example of in-silico screening for drug discovery (see Figure 1). In-silico screening uses computational tools to identify potentially viable molecular compounds for a particular purpose of interest. For instance, Stokes et al. (2020) recently performed a high-throughput screen of compounds from the multi-million molecule ZINC15 (Sterling and Irwin, 2015) database using models predictive of *E. coli* inhibition in order to find a new antibiotic, and were successful. Clearly, guaranteeing high hit rates for this task with conformal prediction is an attractive proposition. The catch, however, is that each returned candidate must be verified experimentally—and too many false positives can quickly consume the available budget (e.g., of time, funding, or other limited resources). This is especially relevant when a valid answer—in this case, an effective drug—might not even exist.

In this paper, we are interested in creating confident prediction sets that trade off coverage for precision in order to provably constrain the total *number of false discoveries* (FD). In other words, we shift the focus of our performance guarantee to be on limiting the number of incorrect answers in our outputs (e.g., to $\leq k$ on average), with the understanding that we will potentially fail to recover some proportion of the true answers—i.e., we will obtain a lower *true discovery rate* (TDR). Concretely, suppose we have been given n (potentially) multi-label classification examples $(X_i, Z_i) \in \mathcal{X} \times 2^{\mathcal{Y}}$, $i = 1, \ldots n$ as calibration data, that have been drawn exchangeably from some underlying distribution P_{XZ}. Here we consider the generalized scenario where each observation X_i may have any number

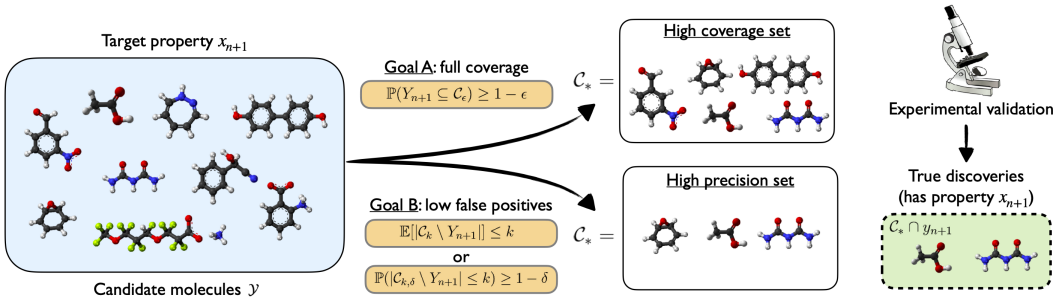


Figure 1: A demonstration of our proposed approach to trading off standard coverage guarantees ("Goal A") for rigorous limits on the total number of false positives included in the output $\mathcal{C}_{k,\delta}$ ("Goal B"). Predictions made using our framework may miss some true positives, but will remain relatively noise-free across their other predictions. For in-silico screening for drug discovery, bounding the number of false positives is critical when balancing a budget for experimental validation.

of corresponding correct labels (including zero, in the case of having no answer at all). That is, the response Z_i can be any subset of the full label space $\mathcal{Y}$.[1] Let $X_{n+1} \in \mathcal{X}$ be a new exchangeable test example for which we would like to predict $Z_{n+1} \subseteq \mathcal{Y}$. Our goal is to construct a set predictor $\mathcal{C}_k(X_{n+1})$ that maximizes recall with respect to the true Z_{n+1}, but more importantly has an expected false discovery budget that is controlled at a user-defined tolerance level $k \in \mathbb{R}_{>0}$, i.e.,

$$\text{maximize } \mathbb{E}\left[\frac{|\mathcal{C}_k(X_{n+1}) \cap Z_{n+1}|}{\max(|Z_{n+1}|, 1)}\right] \quad \text{s.t.} \quad \mathbb{E}\Big[|\mathcal{C}_k(X_{n+1}) \setminus Z_{n+1}|\Big] \leq k. \tag{1}$$

As an additional alternative to bounding the expected number of false discoveries to k, we also seek a set predictor $\mathcal{C}_{k,\delta}$ that has more direct control of the probability of *exceeding* k false discoveries:

$$\text{maximize } \mathbb{E}\left[\frac{|\mathcal{C}_{k,\delta}(X_{n+1}) \cap Z_{n+1}|}{\max(|Z_{n+1}|, 1)}\right] \quad \text{s.t.} \quad \mathbb{P}\Big(|\mathcal{C}_{k,\delta}(X_{n+1}) \setminus Z_{n+1}| \leq k\Big) \geq 1 - \delta, \tag{2}$$

where $\delta \in (0, 1)$ is another user-defined tolerance level. Both constructions define different, but useful, operating conditions; the first is more straightforward (e.g., for the general practitioner), while the second offers a finer, two-parameter level of control. Note that both are well-known concepts in the context of multiple hypothesis testing—i.e., the PFER and k-FWER (Spjøtvoll, 1972; Romano and Wolf, 2007), though the setting (multi-label classification) and available tools (conformal prediction) considered here differ substantially (see §2). Here, and in the rest of the main paper, both constraints are marginal over the choice of calibration and testing data points (X_i, Z_i), $i = 1, \ldots, n+1$; though enforcing (relatively stronger) calibration-conditional constraints only requires minor modifcation.

In order to achieve the desired levels of false discovery control, we present an approach that is based on *set classification*, combined with a form of conformal calibration. Specifically, we use a set nonconformity measure $\mathcal{F}\colon \mathcal{X} \times 2^{\mathcal{Y}} \to \mathbb{R}$ to score candidate output sets, $\mathcal{S} \in 2^{\mathcal{Y}}$. We learn this function from separate multi-label classification training data. Intuitively, a *high* nonconformity score should reflect the confidence that the candidate set might contain a *high* number of false discoveries, and vice-versa. To empirically maximize the true discovery rate, we then return the largest set among all candidates that have nonconformity scores below a threshold that we set such that our false positive constraints are satisfied. Candidate sets are explored greedily with a best-first strategy that adds top-ranked individual labels to a growing, nested output set $\mathcal{S}$ until our nonconformity threshold is met. This both allows us to find efficient solutions for large label spaces $\mathcal{Y}$, and to leverage prior theory for calibrating expectations of monotonic losses for set predictors (Angelopoulos et al., 2021b).

Contributions. In summary, our main results are as follows:

- A theoretical reframing of conformal prediction that provides rigorous false positive control;
- A simple and effective strategy for constructing valid output sets with high true discovery rates;
- A demonstration of the practical utility of our framework across a range of classification tasks.

[1]Note that the classification setting typically considered by conformal prediction, where the correct class is assumed to both exist and be unique, is a special case of this paradigm (i.e., the answer set is always of size one).

2 RELATED WORK

Uncertainty estimation. In recent years, there has been a growing interest in estimating model uncertainty. A large body of work focuses on calibrating the model posterior, $p_\theta(\hat{y}_{n+1}|x_{n+1})$, such that the accuracy, $y_{n+1} = \hat{y}_{n+1}$, is equal to the estimated probability (Niculescu-Mizil and Caruana, 2005; Kuleshov et al., 2018; Kumar et al., 2019). In theory, these estimates could be used to create prediction sets with few false discoveries, but they are not always guaranteed to be accurate (Guo et al., 2017; Ashukha et al., 2020; Hirschfeld et al., 2020). In a similar vein, Bayesian formalisms underlie several popular approaches to quantifying predictive uncertainty via computing the posterior predictive distribution over model parameters (Neal, 1996; Graves, 2011; Hernández-Lobato and Adams, 2015; Gal and Ghahramani, 2016). Like calibration, however, the quality of these methods can vary depending on the suitability of the presumed prior, as well as on any approximation error.

Conformal prediction. As introduced in §1, conformal prediction (Vovk et al., 2005) provides a finite-sample, distribution-free method for obtaining prediction sets $\mathcal{C}$ with guarantees on the event $\mathbf{1}\{Y_{n+1} \in \mathcal{C}(X_{n+1})\}$. Most efforts in CP focus on improving the predictive efficiency, $\mathbb{E}[|\mathcal{C}(X_{n+1})|]$, of the conformal sets (Vovk et al., 2016; Sadinle et al., 2019; Romano et al., 2020; Angelopoulos et al., 2021c; Fisch et al., 2021; Hoff, 2021). As coverage is guaranteed by design, improving efficiency will naturally lead to more precise sets with fewer false positives—but not to a specifiable level. Cauchois et al. (2021) develop a conformal approach to multi-label classification that can guarantee that the prediction set *only* contains true labels (i.e., FD = 0), but does not offer fine-grained control. Most relevant to our work, Bates et al. (2020) gives a flexible framework for controlling the risk, $\mathbb{E}[\mathcal{L}(Y, \mathcal{T}(X))]$, of a set-valued predictor $\mathcal{T}$ with an arbitrary loss function $\mathcal{L}$—as long the loss respects a monotonic *nesting* property, $\mathcal{S} \subset \mathcal{S}' \Rightarrow \mathcal{L}(\mathcal{S}) \geq \mathcal{L}(\mathcal{S}')$, for any two prediction sets $\mathcal{S}$ and $\mathcal{S}'$. The calibration strategy we use here for marginal expectations is based on the recent extension in Angelopoulos et al. (2021b). Concurrent to this work, Angelopoulos et al. (2021a) proposed methods to rigorously control non-monotonic losses, including the related false discovery *rate* (FDR), which normalizes the number of false discoveries over the size of the prediction set. However, as most of our target applications have relatively few true positives, FDR control can be somewhat volatile and lead to many empty predictions (making controlling total false discoveries a more natural fit).

Multiple testing. Controlling the number of false discoveries over a collection of hypothesis tests is a well-studied problem in statistics (Dunn, 1961; Benjamini and Hochberg, 1995; Lehmann and Romano, 2005; Romano and Wolf, 2007, *etc.*). Recently, FDR control has also been studied for outlier detection in a conformal inference setting (Bates et al., 2021). Most statistical approaches to false discovery control operate over p-values for each hypothesis test that have specific dependency structures (e.g., independent or positively dependent), or otherwise use more conservative corrections. Though similar, our multi-label classification setting is slightly different from standard multiple testing in that there is (1) an unknown dependency structure between candidate labels for the same query, but also (2) an extra layer of exchangeability over the $n + 1$ queries. Our approach is able to ignore (1) by leveraging (2) within a conformal calibration framework to obtain desirable guarantees.

Selective classification. Our work also bears some relation to *selective classification* (El-Yaniv and Wiener, 2010), where models have the option to abstain from answering. In particular, Geifman and El-Yaniv (2017) propose a strategy for finding classifiers with specific selective risks (i.e., the expected accuracy over *answered* examples). In our setting, this is analogous to controlling false discoveries at $k \approx 0$. If uncertain, the model would have to "abstain" by outputting an empty set.

3 BACKGROUND

We begin with a brief review of conformal prediction (see Shafer and Vovk, 2008). Here, and in the rest of the paper, upper-case letters (X) denote random variables; lower-case letters (x) denote scalars, and script letters ($\mathcal{X}$) denote sets, unless otherwise specified. Proofs are deferred to the appendix.

Given a new example x, for every candidate label $y \in \mathcal{Y}$, *single label* conformal prediction (where y_{n+1} is a scalar) either accepts or rejects the null hypothesis that the pairing (x, y) is correct. The test statistic for this test is a *nonconformity measure*, $\mathcal{M}\left((x, y), \mathcal{D}\right)$, where $\mathcal{D}$ is a dataset of exchangeable, labeled examples. Informally, a lower value of $\mathcal{M}$ reflects that point (x, y) "conforms" to $\mathcal{D}$, whereas a higher value of $\mathcal{M}$ reflects that (x, y) is atypical relative to $\mathcal{D}$. A practical choice for $\mathcal{M}$ is model-based likelihood, e.g., $-\log p_\theta(y|x)$, where θ is a model fit to $\mathcal{D}$. For conformal prediction to work, is important that $\mathcal{M}$ preserves the exchangeability of its inputs: i.e., $\mathcal{M}$ should be symmetric with respect to permutations. In order to avoid retraining $\mathcal{M}$ every time a new candidate label y is

considered, "split" conformal prediction (Papadopoulos, 2008) uses a proper training set $\mathcal{D}_{\text{train}}$ to learn a fixed $\mathcal{M}$. This preserves exchangeability of the calibration and test points without further modification—and is a popular and computationally efficient strategy (which we follow in this work).

To construct a prediction set for a new test point x_{n+1}, the conformal classifier outputs all y for which the null hypothesis (that pairing (x, y) is correct) is not rejected. This is achieved by comparing the scores of the test candidates to the scores computed over the first n calibration examples.

Theorem 3.1 (Split CP, Vovk et al. (2005); Papadopoulos (2008))**.** *Assume that examples $(X_i, Y_i) \in \mathcal{X} \times \mathcal{Y}$, $i = 1, \ldots, n+1$ are drawn exchangeably from a distribution P_{XY}. For a fixed nonconformity measure $\mathcal{M}$, let random variable $V_i = \mathcal{M}(X_i, Y_i)$ be the nonconformity score of (X_i, Y_i). For $\epsilon \in (0, 1)$, define the conformal set (based on the first n examples) at $x \in \mathcal{X}$ as*

$$\mathcal{C}_\epsilon(x) := \Big\{ y \in \mathcal{Y} \colon \mathcal{M}(x_{n+1}, y) \leq \text{Quantile}(1 - \epsilon;\, V_{1:n} \cup \{\infty\}) \Big\}. \tag{3}$$

Then $\mathcal{C}_\epsilon(X_{n+})$ satisfies $\mathbb{P}(Y_{n+1} \in \mathcal{C}_\epsilon(X_{n+1})) \geq 1 - \epsilon$.

Remark 3.2. Cauchois et al. (2021) recently extended conformal prediction to the multi-label case to obtain guarantees of the form $\mathcal{P}(\mathcal{C}_\epsilon^{\text{inner}}(X_{n+1}) \subseteq Z_{n+1} \subseteq \mathcal{C}_\epsilon^{\text{outer}}(X_{n+1})) \geq 1 - \epsilon$. Their approach follows a modified version of the algorithm for the single label case.

The motivation for our work is quickly evident from Eq. (3): if we are unable to reject most candidates based on their nonconformity scores $\mathcal{M}(x_{n+1}, y)$, then $\mathcal{C}_\epsilon(x)$ can contain many false positives.

4 Conformal prediction with false discovery control

We now introduce our strategy for trading off *coverage* for *precision* by imposing constraints on the number of false positives that are contained in our output sets. To briefly remind the reader of our setting, we assume that we have been given n exchangeable multi-label classification examples, $(X_i, Z_i) \in \mathcal{X} \times 2^{\mathcal{Y}}$, $i = 1, \ldots n$ as calibration data, that are drawn from a distribution P_{XZ}. The response Z_i is treated as a generalized set of correct labels for input X_i, and is a subset of $\mathcal{Y}$, where $\mathcal{Y}$ is finite. For a prediction $\mathcal{C}(x_{n+1}) \subseteq \mathcal{Y}$ evaluated at a point $x_{n+1} \in \mathcal{X}$ with true label set $\mathcal{Z}_{n+1} \subseteq \mathcal{Y}$, we define the *true discovery proportion* (TDP) as the ratio of correct labels that are recovered:

$$\text{TDP}(\mathcal{Z}_{n+1}, \mathcal{C}(x_{n+1})) := \frac{|\mathcal{C}(x_{n+1}) \cap \mathcal{Z}_{n+1}|}{\max(|\mathcal{Z}_{n+1}|, 1)} \qquad \text{(Note that TDR} := \mathbb{E}[\text{TDP}]), \tag{4}$$

and the number of *false discoveries* (FD) as the total count of incorrect labels in $\mathcal{C}(x_{n+1})$:

$$\text{FD}(\mathcal{Z}_{n+1}, \mathcal{C}(x_{n+1})) := |\mathcal{C}(x_{n+1}) \setminus \mathcal{Z}_{n+1}|. \tag{5}$$

Our goal, as stated in §1, is to maximize the expected TDP (in standard classification, this would equate to accuracy; or power in statistics), while constraining the FD in either of two ways:

Definition 4.1 (k-FD validity)**.** *A conformal classifier $\mathcal{C}_k$ evaluated at test point X_{n+1} is k-FD valid if the expectation of its FD satisfies $\mathbb{E}[\text{FD}(Z_{n+1}, \mathcal{C}_k(X_{n+1}))] \leq k$.*

Definition 4.2 ((k, δ)-FDP validity)**.** *A conformal classifier $\mathcal{C}_{k,\delta}$ evaluated at test point X_{n+1} is (k, δ)-FD valid if the CDF of its FD satisfies $\mathbb{P}(\text{FD}(Z_{n+1}, \mathcal{C}_{k,\delta}(X_{n+1})) \leq k) \geq 1 - \delta$.*

4.1 An oracle set predictor

To motivate our approach, imagine an *oracle* with access to $P_{Z|X}$, the conditional distribution of the multi-label set Z given the input X. Given this information, for any input $x \in \mathcal{X}$ and candidate set $\mathcal{S} \in 2^{\mathcal{Y}}$, such an oracle would be able to exactly calculate both the expectation and the conditional distribution of the number of false (and true) discoveries in $\mathcal{S}$ given x. In order to maximize the TDR while meeting k-FD and (k, δ)-FD validity, it could then yield the following set predictions:

$$\mathcal{C}_k^{\text{oracle}}(x_{n+1}) := \operatorname*{arg\,max}_{\mathcal{S} \subseteq 2^{\mathcal{Y}}} \big\{ \mathbb{E}[\text{TDP}(\mathcal{S}) \mid x_{n+1}] \colon \mathbb{E}[\text{FD}(Z_{n+1}, \mathcal{S}) \mid x_{n+1}] \leq k \big\}, \text{ and} \tag{6}$$

$$\mathcal{C}_{k,\delta}^{\text{oracle}}(x_{n+1}) := \operatorname*{arg\,max}_{\mathcal{S} \subseteq 2^{\mathcal{Y}}} \big\{ \mathbb{E}[\text{TDP}(\mathcal{S}) \mid x_{n+1}] \colon \mathbb{P}(\text{FD}(Z_{n+1}, \mathcal{S}) \mid x_{n+1}] > k) < \delta \big\}, \tag{7}$$

where ties in the $\arg\max$ are broken by set size (smaller is better). This oracle has the advantage of not only being marginally valid as required, but also being *conditionally* valid given the input X_{n+1}.

Algorithm 1 Conformal prediction with false discovery control (for in-silico screening).

Definitions: $x_{\mathrm{n}+1}$ is the test point, $\mathcal{D}_{\mathrm{train}}$ is a proper training set, $\mathcal{D}_{\mathrm{cal}}$ is an exchangeable calibration set, k is the FD tolerance, and B is a parameter for considering only the top individually ranked candidates $y \in \mathcal{Y}$.

function PREDICT(x_{n+1}, $\mathcal{D}_{\mathrm{train}}$, $\mathcal{D}_{\mathrm{cal}}$, k, $B = |\mathcal{Y}|$)
  $\mathcal{D}^{(1)}_{\mathrm{train}}, \mathcal{D}^{(2)}_{\mathrm{train}} \leftarrow$ SPLIT ($\mathcal{D}_{\mathrm{train}}$)
  # Use part of the training set to learn individual label likelihood $p_\theta(y_i \in \mathcal{Z} \mid x)$.
  $p_\theta(y_i \in \mathcal{Z} \mid x) \leftarrow$ TRAIN(MPNN, $\mathcal{D}^{(1)}_{\mathrm{train}}$)
  # Use the other (smaller) part to learn set function $\mathcal{F}(x, \mathcal{S})$.
  $\mathcal{F}(x, \mathcal{S}) \leftarrow$ TRAIN(DeepSets, $\mathcal{D}^{(2)}_{\mathrm{train}}$)
  # Compile calibration data.
  $\mathcal{T}_{\mathrm{cal}} = \varnothing$
  FOR $(x_i, \mathcal{Z}_i) \in \mathcal{D}_{\mathrm{cal}}$ **DO**
    # Rank candidates $y \in \mathcal{Y}$ by top p_θ up to bound B.
    $\{\mathcal{S}_1, \ldots, \mathcal{S}_B\} = \{\mathrm{SORT}(\mathcal{Y}, p_\theta(y \in \mathcal{Z}_i \mid x_i))_{1:j} \colon j \in \{1, \ldots, B\}\}$
    # Implicitly define MFD *based on this ordering.*
    $\mathrm{MFD}(x_i, \mathcal{Z}_i) \leftarrow$ INIT$(x_i, \mathcal{Z}_i, \mathcal{F}, \{\mathcal{S}_1, \ldots, \mathcal{S}_B\})$
    $\mathcal{T}_{\mathrm{cal}} \leftarrow \mathcal{T}_{\mathrm{cal}} \cup \{\mathrm{MFD}(x_i, \mathcal{Z}_i)\}$
  # Find threshold t_k on the calibration set using Eq. (11); return largest set that passes.
  $t_k \leftarrow$ CALIBRATE$(\mathcal{T}_{\mathrm{cal}}, B, k)$
  RETURN $\arg\max_{\mathcal{S} \in \{\mathcal{S}_1 \ldots, \mathcal{S}_B\}} \{|\mathcal{S}| \colon \tilde{\mathcal{F}}(x_{n+1}, S) \leq t_k\}$

4.2 Approximating the oracle using Deep Sets

Of course, computing such an oracle is not pactical, as $P_{Z|X}$ is not known. Instead, let $\mathcal{F}$ be a set function $\mathcal{F} \colon \mathcal{X} \times 2^{\mathcal{Y}} \to \mathbb{R}$ that generates a score for the candidate set $\mathcal{S}$ given x_{n+1}. Importantly, $\mathcal{F}$ can be any function (as we will later show), but to best replicate the oracle's behavior for k-FD and (k, δ)-FD valid predictions, $\mathcal{F}$ should be a good estimator of the distribution of false positives in $\mathcal{S}$. We develop a simple approach to modeling and learning $\mathcal{F}$ via a *deep sets* model (Zaheer et al., 2017). Let $\{\phi(x, y_1), \ldots, \phi(x, y_l)\}$ featurize a candidate set $\mathcal{S} \subseteq \mathcal{Y}$, where $\phi(x, y_i) \in \mathbb{R}^d$ is the representation of an individual label $y_i \in \mathcal{S}$ computed as a function of (x, y_i).

In practice, we find that taking $\phi(x, y_i)$ to be a (fixed) scalar estimate $p_\theta(y_i \in \mathcal{Z} \mid x)$—the marginal likelihood of y_i being a correct label for x—performs reasonably well and is simple to implement. This means that the features to the set predictor model need only be defined by the (uncalibrated) confidence thresholds of any backbone model. For example, in our chemistry we use a directed MPNN (Yang et al., 2019) that independently classifies molecules for being active/inactive for the target property. $\phi(x, y_i)$ is then simply a sigmoidal output. Given ϕ, the deep sets model is defined as

$$\Psi(x, \mathcal{S}) := \mathrm{softmax}\left(\mathrm{dec}\left(\sum_{y_i \in \mathcal{S}} \mathrm{enc}(\phi(x, y_i); \theta_1)); \theta_2\right)\right), \tag{8}$$

where $\mathrm{enc}(\cdot)$ and $\mathrm{dec}(\cdot)$ are two neural models parameterized by θ_1 and θ_2, respectively, and $\mathrm{softmax}(\cdot)$ is taken over the total possible number of false positives, $\{0, \ldots, |\mathcal{Y}|\}$. Ψ is trained to predict the number of false positives in $\mathcal{S}$ via cross entropy, using labeled sets sampled from auxiliary training data. We then compute $\mathcal{F}_k$ and $\mathcal{F}_{k,\delta}$ (for either k-FD or (k, δ)-FD validity) as

$$\mathcal{F}_k(x, \mathcal{S}) := \sum_{i=0}^{|\mathcal{S}|} i \cdot \Psi(x, \mathcal{S})_i \qquad \text{and} \qquad \mathcal{F}_{k,\delta}(x, \mathcal{S}) := 1 - \sum_{i=0}^{\min(k, |\mathcal{S}|)} \Psi(x, \mathcal{S})_i, \tag{9}$$

where $\Psi(x, \mathcal{S})_i$ denotes the ith index of the $\mathrm{softmax}$ (i.e., the estimated probability that FD = i). In the next section we will refer only to $\mathcal{F}$, with the understanding it differs based on the validity goal.

4.3 Identifying valid candidate sets

Although our set predictor $\mathcal{F}$ is trained to model either the expected FD or its CDF, it is not necessarily accurate. If $\mathcal{F}$ were simply substituted into Eq. (6) or Eq. (7), it would not produce set predictions that are guaranteed to be valid. To remedy this, we must calibrate a threshold for rejecting or accepting candidate sets based on $\mathcal{F}$. To do so, we first greedily identify a sequence of *nested* candidate sets, $\mathcal{S}_1 \subset \mathcal{S}_2 \subset \ldots \subset \mathcal{Y}$, by ranking individual labels $y \in \mathcal{Y}$ by their estimated likelihoods of being true

positives, i.e., $p_\theta(y \in \mathcal{Z} \mid x)$, and including them one by one. Each of these candidate sets is then given a score by $\mathcal{F}$, which is taken to be a *nonconformity score*. We manually enforce that the scores produce a ranking consistent with their size (though this is rarely necessary). Specifically, for set S_i we compute its adjusted nonconformity score $\tilde{\mathcal{F}}(x, \mathcal{S}_i)$ as $\tilde{\mathcal{F}}(x, \mathcal{S}_i) := \max\{\mathcal{F}(x, \mathcal{S}_1), \ldots, \mathcal{F}(x, \mathcal{S}_i)\}$.

Let $\mathrm{MFD}(x, \mathcal{Z})(t) := \max\left\{\mathrm{FD}(\mathcal{Z}, \mathcal{S}_i) \colon \tilde{\mathcal{F}}(x, \mathcal{S}_i) \leq t\right\}$ denote the maximum number of false discoveries over all nested sets S_i for an input x that have nonconformity scores less than t. If this thresholded set is empty, then the MFD is defined to be 0. Since the sets are ordered by size, this quantity is guaranteed to be monotonically non-decreasing in t, i.e.,

$$t \leq t' \implies \mathrm{MFD}(x, \mathcal{Z})(t) \leq \mathrm{MFD}(x, \mathcal{Z})(t') \tag{10}$$

Using this property, we can calibrate t to create a set filter with controlled MFD, as formalized next:

Theorem 4.3 (FD-CP)**.** *Assume that examples $(X_i, Z_i) \in \mathcal{X} \times 2^{\mathcal{Y}}$, $i = 1, \ldots, n+1$ are drawn exchangeably from a distribution P_{XY}. For each example i, let $S_{i,j}$, $j = 1, \ldots, |\mathcal{Y}|$, where $|\mathcal{Y}| < B$, be nested candidate sets with monotonically non-decreasing, lower-bounded nonconformity scores $\tilde{\mathcal{F}}(X_i, S_{i,j})$. For $k \in \mathbb{R}_{>0}$ and $\delta \in (0, 1)$ define the random variables T_k and $T_{k,\delta}$ as*

$$T_k := \sup\left\{t \in \mathbb{R} \colon \frac{1}{n+1}\Big(B + \sum_{i=1}^{n} \mathrm{MFD}(X_i, Z_i)(t)\Big) \leq k\right\} \text{ and} \tag{11}$$

$$T_{k,\delta} := \sup\left\{t \in \mathbb{R} \colon \frac{1}{n+1}\sum_{i=1}^{n} \mathbf{1}\{\mathrm{MFD}(X_i, Z_i)(t) \leq k\} \geq 1 - \delta\right\}. \tag{12}$$

Then we have that for any $j \in \{1, \ldots, |\mathcal{Y}|\}$, $\mathbb{E}[\mathrm{FD}(Z_{n+1}, S_{n+1,j}) \mid \tilde{\mathcal{F}}(X_i, \mathcal{S}_{n+1,j}) \leq T_k] \leq k$, and $\mathbb{P}(\mathrm{FD}(Z_{n+1}, S_{n+1,j}) \leq k \mid \tilde{\mathcal{F}}(X_i, \mathcal{S}_{n+1,j}) \leq T_{k,\delta}) \geq 1 - \delta$.

The proof, given in Appendix A, uses the results from marginal RCPS (Angelopoulos et al., 2021b).

Remark 4.4. T_k may be very conservative if $|\mathcal{Y}|$ is very large, to the point where $T_k = -\infty$ always if $|\mathcal{Y}| > k * (n+1)$. To account for this, it can be beneficial to truncate the considered label space $\mathcal{Y}$ for an example x_i to only the top $m \ll k * (n+1)$ individual candidates, $\{y_1, \ldots, y_m\} \in \mathcal{Y}^m$. For example, for text generation tasks (like NMT) $\mathcal{Y}$ is infinite, but we can restrict our predictions to a subset of the top m beach search candidates (where m can still be large to a reasonable degree).

Remark 4.5. Note that no constraints are put on the underlying $\tilde{\mathcal{F}}$ in Theorem 4.3 (it is model-agnostic). If, however, $\tilde{\mathcal{F}}$ is a good estimator of $\mathrm{FD}(Z, \mathcal{S}) \mid X$, then T_k and $T_{k,\delta}$ are likely to identify sets that are approximately valid conditional on X_{n+1}, which we show empirically in §6.

4.4 Selecting the final output set

The main result of Theorem 4.3 is that, using the calibrated nonconformity threshold T_k or $T_{k,\delta} = t^*$, we can construct a collection of sets that are simultaneously valid. In other words, we are free to select *any* set contained in the filtered set of candidates $\{S_i \colon \tilde{\mathcal{F}}(x, S_i) \leq t^*\}$ as a valid output. Ideally, we would be able to follow the the oracle strategy in returning the smallest set with the highest true positive proportion. This would make our predictions *efficient*, in the sense that we are not including more false positives than necessary (even if the total number is still $\leq k$). A reasonable choice is to then choose $\mathcal{S}_{i^*}$ where $i^* = \arg\max_i |\mathcal{S}_i| - \mathcal{F}(x, \mathcal{S}_i)$; but this can be sub-optimal if $\mathcal{F}$ is not accurate. As a robust and straightforward alternative, we simply take the *largest* set in our filtered set of candidates as our final output. That is, for an input $x \in \mathcal{X}$ and calibrated threshold $t^* \in \mathbb{R}$

$$\mathcal{C}_*(x, t^*) := \underset{\mathcal{S} \in \{\mathcal{S}_1 \ldots, \mathcal{S}_{|\mathcal{Y}|}\}}{\arg\max} \left\{|\mathcal{S}| \colon \tilde{\mathcal{F}}(x, S) \leq t^*\right\} \tag{13}$$

where $\mathcal{C}_*$ is either $\mathcal{C}_k$ or $\mathcal{C}_{k,\delta}$ (i.e., they are constructed the same way, but for different thresholds t^*), and where we define $\arg\max$ to return $\varnothing$ if the set $\{\mathcal{S} \in \{\mathcal{S}_1, \ldots, \mathcal{S}_{|\mathcal{Y}|}\} \colon \tilde{\mathcal{F}}(x, S) \leq t^*\}$ is empty.

Proposition 4.6 (Greedy FD-CP)**.** *As defined in Eq. (13), the set $\mathcal{C}_k(X_{n+1}, T_k)$ is k-FD valid, and the set $\mathcal{C}_{k,\delta}(X_{n+1}, T_{k,\delta})$ is (k, δ)-FD valid. Furthermore, among sets $\mathcal{S}_{n+1,j} \in \{\mathcal{S}_{n+1,1}, \ldots, \mathcal{S}_{n+1,|\mathcal{Y}|}\}$ that have nonconformity scores $\leq T_k$ or $T_{k,\delta}$, the set $\mathcal{C}_*(X_{n+1}, T_*)$ maximizes the TDP.*

The proof follows from Theorem 3.1, and from the fact that the TDP is monotonically non-decreasing in set size when $\mathcal{S}_{n+1,j}$ are nested. Example pseudo-code for our procedure (specifically as would be applied to in-silico screening with the goal of k-FD control) is given in Algorithm 1.

5 EXPERIMENTAL SETUP

We empirically evaluate our method on three different tasks covering computational chemistry, computer vision, and natural language processing. In this section, we briefly outline each task and its associated modeling choices. We also describe our implementation, evaluation, and baselines. For ease of experimentation (and performance reasons, see Remark 4.4), we choose to cap the number of considered labels *per example* to 100 for each task (but the full label space $\mathcal{Y}$ can be much larger).

5.1 TASKS

In-silico screening for drug discovery. As introduced in §1, the goal of in-silico screening is to identify potentially effective drugs to manufacture and test. We use the ChEMBL database (Mayr et al., 2018) to screen molecules for combinatorial constraint satisfaction, where given a constraint such as "*has property A but not property B*", we want to identify the subset of molecules from a given set of candidates that have the desired attributes. We partition the dataset both by molecules and property combinations, so that at test time the model must make predictions on combinations it has never been tested on before, over a pool of molecules that it has never seen before. Scores for individual candidate molecules are obtained via combining independent property assignment probabilities predicted by an ensemble of directed MPNNs (Yang et al., 2019) using a naive independence assumption. Our training, validation, and test sets consist of $4.4k$, 797, and $5k$ screening examples, respectively. We reuse the training set to learn both the base MPNN model and the deep sets model $\mathcal{F}$.

Object detection. The goal of object detection is to place bounding boxes around all objects in an image (of which there may be many). We use the MS-COCO dataset (Lin et al., 2014), a large scale object detection dataset with images of complex everyday scenes containing 80 object categories. We extract typed bounding box candidates (i.e., tuples of both location *and* category) using an EfficientDet model (Tan et al., 2020) with non-maximum suppression. True positives are defined as boxes that have an intersection over union (IoU) > 0.5 with a gold annotation of the same type. Since the official test set is hidden, we use the $5k$ examples from the validation set and randomly partition them into $1/1/3k$ sized splits for training $\mathcal{F}$, validation, and testing.

Entity extraction. In entity extraction, we are interested in identifying all named entities that appear in a sentence x of length l, $x = \{w_1, \ldots, w_l\}$. We report results on the CoNLL NER dataset (Tjong Kim Sang and De Meulder, 2003), where we use the PURE span-based entity extraction model of Zhong and Chen (2021) to predict scores for all $\mathcal{O}(l^2)$ candidate spans. We consider exact span predictions of the correct category to be true positives, and all others to be false positives. Not all sentences are guaranteed to have entities (in fact, many do not contain any). We extract $1k$ examples for training $\mathcal{F}$ out of the $3.3k$ validation set, and report test results on the original $3.5k$ test set.

5.2 EVALUATION

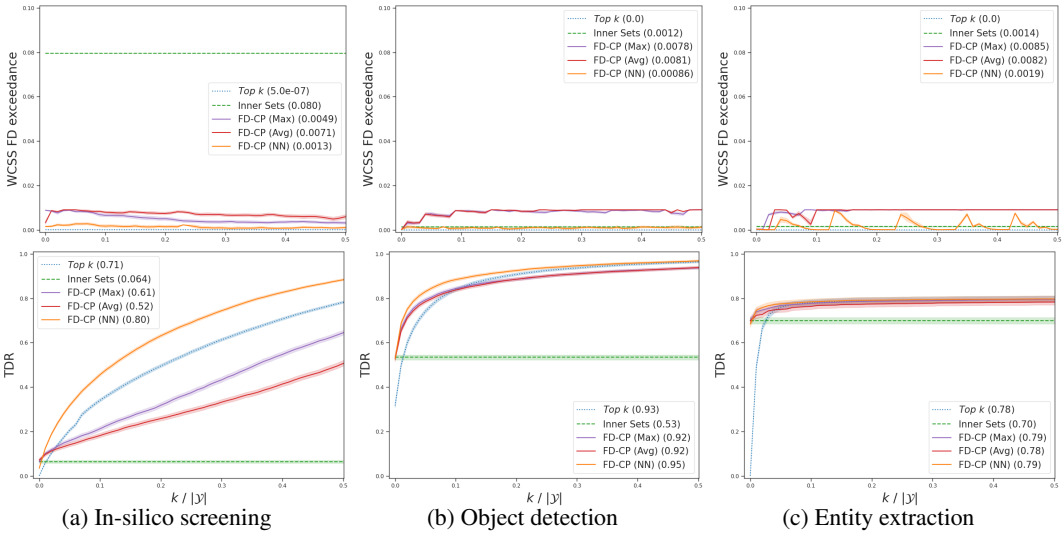
We use a proper training, validation, and test set for each task. We use the training set to learn all models, i.e., $p_\theta(y_i \in \mathcal{Z} \mid x)$ and $\mathcal{F}$. We perform model selection on the validation set, and report final numbers as the average over 1000 random trials on the test set, where in each trial we partition the data into 80% calibration ($x_{1:n}$) and 20% prediction points (x_{n+1}). To compare aggregate performance across tolerance levels, we plot each metric as a function of $k/|\mathcal{Y}|$ (up to $k/|\mathcal{Y}| = 0.5$), and compute the AUC. Shaded regions show the 16-84th percentiles across trials. In addition to TDR and FD metrics, inspired by Angelopoulos et al. (2021c), we compute the *size-stratified k-FD violation*:

$$\text{SSFD}_k(\mathcal{C}, \{\mathcal{A}\}_{j=1}^a) := \sup_j \max\left(\widehat{\mathbb{E}}[\text{FD}(Z_{n+1}, \mathcal{C}_k(X_{n+1})) \mid |\mathcal{C}_k(X_{n+1})| \in \mathcal{A}_j] - k, 0\right) \quad (14)$$

where $\{\mathcal{A}\}_{j=1}^a$ forms a partition of $\{1, \ldots, |\mathcal{Y}|\}$, and $\widehat{\mathbb{E}}$ denotes the empirical average over our test trials. $\text{SSFD}_{k,\delta}$ for (k, δ)-FD violation is defined similarly when compared to the target $1 - \delta$ level. As argued in Angelopoulos et al. (2021c), a lower size-stratified violation suggests that a classifier has better conditional coverage. As marginal validity is theoretically guaranteed by our procedure (and holds empirically in our experiments), we focus on this metric in our results.

5.3 BASELINES

For all experiments, we compare our conformalized deep sets model (NN), to the following baselines:



(a) In-silico screening (b) Object detection (c) Entity extraction

Figure 2: (k, δ)-FD results as a function of $k/|\mathcal{Y}|$ for $\delta = 0.1$. The top row plots size-stratified (k, δ)-FD violation (lower is better). The bottom row plots the TDR (higher is better). We see that compared to the other baselines, our conformal deep sets approach (NN) has the best (or close to the best) TDR AUC across tasks, while having the lowest (or close to the lowest) $\mathrm{SSFD}_{k,\delta}$ violation.

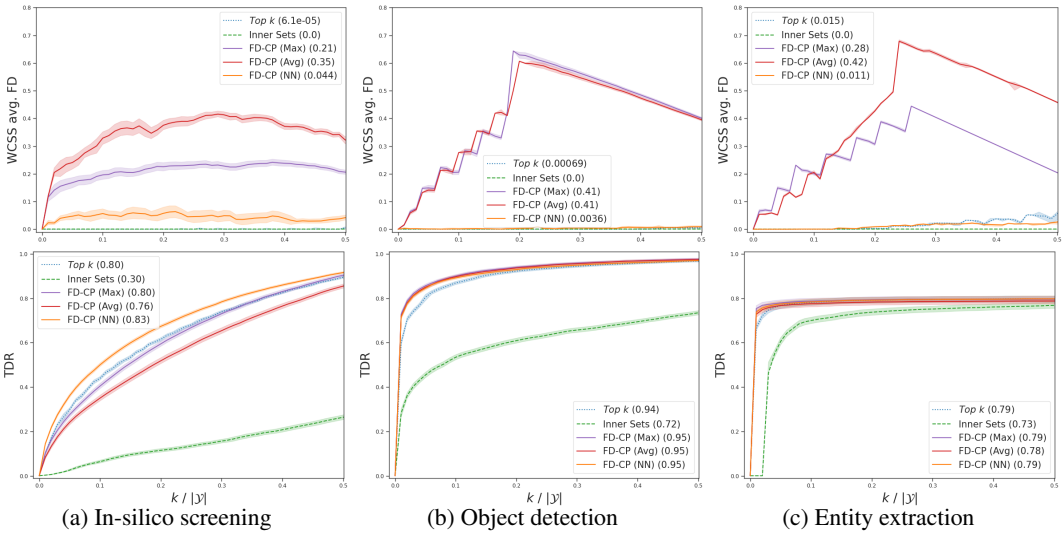
- **Top k.** We naively take the top k' fixed predictions for any x_{n+1}, where k' is found using average performance on the calibration set (without any correction factors, so it is *not* generally guaranteed to be valid). Note that k' can be (and mostly is) different than the user-specified k for FD.
- **Inner Sets.** We use the method of Cauchois et al. (2021) to bound $\mathbb{P}(\mathcal{C}_\epsilon(X_{n+1}) \subseteq Z_{n+1}) \geq 1 - \epsilon$ at level $\epsilon = k/|\mathcal{Y}|$ for k-FD control and at level $\epsilon = \delta$ for (k, δ)-FD control. It is straightforward to show that these levels of ϵ conservatively achieve k-FD and (k, δ)-FD control. For simplicity, we use a one-sided version of the direct inner/outer method described in Cauchois et al. (2021), and do not use conformalized quantile regression for computing thresholds.
- **Independent scoring (max).** We take $\mathcal{F}(x, \mathcal{S})$ to be the maximum individual label uncertainty in $\mathcal{S}$, $\max\{1 - p_\theta(y_i \in \mathcal{Z} \mid x) \colon y_i \in \mathcal{S}\}$. This is equivalent to choosing labels independently. The full model is calibrated using the same FD-CP algorithm (it is a drop-in replacement for the NN).
- **Average scoring (avg).** We take $\mathcal{F}(x, \mathcal{S})$ to be the average individual label uncertainty in $\mathcal{S}$, $|\mathcal{S}|^{-1} \sum_{y_i \in \mathcal{S}} 1 - p_\theta(y_i \in \mathcal{Z} \mid x)$. We calibrate $p_\theta(y_i \in \mathcal{Z} \mid z)$ using Platt scaling. As with the max score, the full model is calibrated using the same FD-CP algorithm.

6 Experimental Results

In the following, we present our main results for conformal prediction with false discovery control. We evaluate both k-FD and (k, δ)-FD approaches. In all of the following experiments, we report results on the test set, using configurations selected based on validation set performance.

Constraining false discoveries. The top row of Figure 2 and 3 shows the size-stratified violation (Eq. 14) for (k, δ)-FD and k-FD, respectively. Our deep set NN is able to predict the FD of a set better than the max or average baselines, and therefore, the model has better conditional coverage, meaning fewer excessive false discoveries. Other than the in-silico screening with (k, δ)-FD where the Inner Sets method is very conservative, the Top k and the Inner Sets baselines allow very small violations. However, this result by itself is not impressive—as simply always abstaining (returning an empty set) will lead to zero SSFD. When also considering the TDR (bottom row in the figures), we see that our FD-CP model is superior (often considerably so).

Maximizing true discovery rates. Table 1 summarizes the empirical results for the in-silico screening task with several representative configurations. We see that our FD-CP (NN) method provides significantly higher true discovery rates while meeting the specified false discovery bounds. For example, allowing no more than 5 false discoveries leads to a TDR of 36.1% with k-FD. In



(a) In-silico screening (b) Object detection (c) Entity extraction

Figure 3: k-FD results as a function of $k/|\mathcal{Y}|$. The top row plots size-stratified k-FD violation (lower is better). The bottom row plots the TDR (higher is better). As the number of allowed expected false positives grows, our methods quickly achieve high power. In particular, consistent with Figure 2, our conformal deep sets approach (NN) has the best (or close to the best) TDR AUC across tasks, while having the lowest (or close to the lowest) SSFD_k violation as compared to other baselines.

	Top k		**Inner Sets**		**FD-CP (Max)**		**FD-CP (Avg)**		**FD-CP (NN)**	
k-FD:	*Avg.* FD	*TDR*	*Avg.* FD	*TDR*	*Avg.* FD	*TDR*	*Avg.* FD	*TDR*	*Avg.* FD	*TDR*
$k = 5$	4.59	29.8	0.14	2.5	4.98	27.5	4.99	24.1	4.98	**36.1**
$k = 15$	14.47	53.4	0.88	9.5	14.98	50.7	14.98	43.8	14.99	**59.9**
$k = 25$	24.51	68.0	1.49	13.4	24.98	66.8	24.98	58.9	24.99	**73.2**
$k = 35$	34.54	78.2	2.45	18.4	34.97	78.4	35.96	71.2	34.99	**82.5**
(k, δ)-*FD with* $1 - \delta = 0.9$:	FD $\leq k$	*TDR*	FD $\leq k$	*TDR*	FD $\leq k$	*TDR*	FD $\leq k$	*TDR*	FD $\leq k$	*TDR*
$k = 5$	100.0	20.5	96.6	6.36	90.0	15.8	90.0	14.0	90.0	**31.6**
$k = 15$	94.7	42.4	99.5	6.36	90.0	26.7	90.0	22.3	90.0	**55.3**
$k = 25$	96.6	55.7	100.0	6.36	90.0	37.4	90.0	29.3	90.0	**69.0**
$k = 35$	97.5	66.2	100.0	6.36	90.0	49.1	90.0	36.9	90.0	**79.0**

Table 1: Results for the in-silico screening task over the ChEMBL dataset. TDR and FD $\leq k$ are expressed as percents (i.e., $\times 100$). We see that our FD-CP methods always remain valid, while using the Inner Sets formulation is conservative (as expected). Applying FD-CP calibration with our deep sets function $\mathcal{F}$ (NN) yields classifiers with substantially higher TDR across tolerance levels k.

comparison, the TDR of Top k is only 29.8%. The advantage of the deep set neural network for estimating the number of false discoveries is most pronounced in the challenging in-silico task. The two other tasks are overall easier, as suggested by their higher TDR (Figures 2 and 3). Yet, our FD-CP (NN) consistently provides high TDR compared to all baselines for both k-FD and (k, δ)-FD.

7 Conclusion

Conformal prediction, in its standard formulation, already grants theoretical performance guarantees that can be a critical in many machine learning applications. Naively applying CP, however, can yield disappointing results. Even if the target coverage is indeed upheld, the predicted sets may be too large, include too many false positives to be practical. In this paper, we proposed a method for trading coverage guarantees in favor of precision guarantees, where we enforce a limit to the number of false positives that are contained in our prediction sets. Our results show that our method yields classifiers that (1) still achieve strong true discovery rates compared to their coverage-seeking counterparts, and (2) provide meaningful output sets with effectively controlled total false discovery counts.

ETHICS STATEMENT

Our FP control methods are general and can be applied to many applications and on top of any model for computing nonconformity scores. It's worth noting that any undesirable biases exhibited by underlying models can still propagate to the prediction sets of our methods. While our methods provide marginal performance guarantees, we recommend any application to perform controlled evaluation across target populations to ensure fairness.

REPRODUCIBILITY STATEMENT

All datasets used in this paper are publicly available (see §5.1). Also, we use publicly available models for computing the nonconformity scores. For drug discovery, we use the ensemble of directed MPNNs from ChemProp (https://github.com/chemprop/chemprop). For object detection, we use tf_efficientdet_d2 from https://github.com/rwightman/efficientdet-pytorch. For entity extraction, we train the PURE entity model (https://github.com/princeton-nlp/PURE) on the ConLL03 dataset using a context window of 100, batch size 32, albert-base-v2 encoding model, and the suggested learning rate of $1e-5$, and task learning rate $5e-4$. Default values were used for all other parameters. The results in Section 6 are all based on the experimental setting described in Section 5. We will release our code for running these experiments and for reproducing all plots and tables.

REFERENCES

Anastasios Nikolas Angelopoulos, Stephen Bates, Emmanuel J. Candès, Michael I. Jordan, and Lihua Lei. Learn then test: Calibrating predictive algorithms to achieve risk control. *ArXiv preprint: 2110.01052*, 2021a.

Anastasios Nikolas Angelopoulos, Stephen Bates, and Lihua Lei. Risk control in expectation: A proof and counter example, 2021b. URL http://angelopoulos.ai/publications/downloads/rce_note.pdf.

Anastasios Nikolas Angelopoulos, Stephen Bates, Jitendra Malik, and Michael I. Jordan. Uncertainty sets for image classifiers using conformal prediction. In *International Conference on Learning Representations (ICLR)*, 2021c.

Arsenii Ashukha, Alexander Lyzhov, Dmitry Molchanov, and Dmitry Vetrov. Pitfalls of in-domain uncertainty estimation and ensembling in deep learning. In *International Conference on Learning Representations (ICLR)*, 2020.

Stephen Bates, Anastasios Nikolas Angelopoulos, Lihua Lei, Jitendra Malik, and Michael I. Jordan. Distribution free, risk controlling prediction sets. *ArXiv preprint: 2101.02703*, 2020.

Stephen Bates, Emmanuel Candès, Lihua Lei, Yaniv Romano, and Matteo Sesia. Testing for outliers with conformal p-values. *ArXiv preprint: 2104.08279*, 2021.

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 (Statistical Methodology)*, 1995.

Maxime Cauchois, Suyash Gupta, and John C. Duchi. Knowing what you know: valid and validated confidence sets in multiclass and multilabel prediction. *Journal of Machine Learning Research (JMLR)*, 2021.

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

Ran El-Yaniv and Yair Wiener. On the foundations of noise-free selective classification. *Journal of Machine Learning Research (JMLR)*, 2010.

Adam Fisch, Tal Schuster, Tommi Jaakkola, and Regina Barzilay. Efficient conformal prediction via cascaded inference with expanded admission. In *International Conference on Learning Representations (ICLR)*, 2021.

Yarin Gal and Zoubin Ghahramani. Dropout as a bayesian approximation: Representing model uncertainty in deep learning. In *International Conference on Machine Learning (ICML)*, 2016.

Alexander Gammerman and Vladimir Vovk. Hedging Predictions in Machine Learning: The Second Computer Journal Lecture . *The Computer Journal*, 50(2):151–163, Feb. 2007.

Yonatan Geifman and Ran El-Yaniv. Selective classification for deep neural networks. In *Advances in Neural Information Processing Systems (NeurIPS)*, pages 4878–4887. 2017.

Alex Graves. Practical variational inference for neural networks. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2011.

Chuan Guo, Geoff Pleiss, Yu Sun, and Kilian Q. Weinberger. On calibration of modern neural networks. In *International Conference on Machine Learning (ICML)*, 2017.

José Miguel Hernández-Lobato and Ryan P. Adams. Probabilistic backpropagation for scalable learning of bayesian neural networks. In *International Conference on Machine Learning (ICML)*, 2015.

Lior Hirschfeld, Kyle Swanson, Kevin Yang, Regina Barzilay, and Connor W. Coley. Uncertainty quantification using neural networks for molecular property prediction. *ArXiv preprint: 2005.10036*, 2020.

Peter Hoff. Bayes-optimal prediction with frequentist coverage control. *ArXiv preprint: 2105.14045*, 2021.

Volodymyr Kuleshov, Nathan Fenner, and Stefano Ermon. Accurate uncertainties for deep learning using calibrated regression. In *International Conference on Machine Learning (ICML)*, 2018.

Ananya Kumar, Percy Liang, and Tengyu Ma. Verified uncertainty calibration. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2019.

E. L. Lehmann and Joseph P. Romano. Generalizations of the familywise error rate. *The Annals of Statistics*, 33(3):1138 – 1154, 2005. doi: 10.1214/009053605000000084. URL https://doi.org/10.1214/009053605000000084.

Jing Lei. Classification with confidence. *Biometrika*, 101(4), 2014.

Tsung-Yi Lin, Michael Maire, Serge J. Belongie, James Hays, Pietro Perona, Deva Ramanan, Piotr Dollár, and C. Lawrence Zitnick. Microsoft coco: Common objects in context. In *ECCV*, 2014.

Andreas Mayr, Günter Klambauer, Thomas Unterthiner, Marvin Steijaert, Joerg Wegner, Hugo Ceulemans, Djork-Arné Clevert, and Sepp Hochreiter. Large-scale comparison of machine learning methods for drug target prediction on chembl. *Chemical Science*, 9, 06 2018. doi: 10.1039/C8SC00148K.

Radford M. Neal. *Bayesian Learning for Neural Networks*. Springer-Verlag, 1996. ISBN 0387947248.

Alexandru Niculescu-Mizil and Rich Caruana. Predicting good probabilities with supervised learning. In *International Conference on Machine Learning (ICML)*, 2005.

Harris Papadopoulos. Inductive conformal prediction: Theory and application to neural networks. In *Tools in Artificial Intelligence*, chapter 18. IntechOpen, Rijeka, 2008.

Joseph Romano and Michael Wolf. Control of generalized error rates in multiple testing. *Annals of Statistics*, 2007.

Yaniv Romano, Matteo Sesia, and Emmanuel J. Candès. Classification with valid and adaptive coverage. In *Advances in Neural Information Processing (NeurIPS)*, 2020.

Mauricio Sadinle, Jing Lei, and Larry Wasserman. Least ambiguous set-valued classifiers with bounded error levels. *Journal of American Statistical Association*, 2019.

Glenn Shafer and Vladimir Vovk. A tutorial on conformal prediction. *Journal of Machine Learning Research (JMLR)*, 9:371–421, June 2008.

Emil Spjøtvoll. On the Optimality of Some Multiple Comparison Procedures. *The Annals of Mathematical Statistics*, 43(2):398 – 411, 1972. doi: 10.1214/aoms/1177692621. URL https://doi.org/10.1214/aoms/1177692621.

Teague Sterling and John J. Irwin. Zinc 15 – ligand discovery for everyone. *Journal of Chemical Information and Modeling*, 55(11):2324–2337, 11 2015.

Jonathan Stokes, Kevin Yang, Kyle Swanson, Wengong Jin, Andres Cubillos-Ruiz, Nina Donghia, Craig MacNair, Shawn French, Lindsey Carfrae, Zohar Bloom-Ackerman, Victoria Tran, Anush Chiappino-Pepe, Ahmed Badran, Ian Andrews, Emma Chory, George Church, Eric Brown, Tommi Jaakkola, Regina Barzilay, and James Collins. A deep learning approach to antibiotic discovery. *Cell*, 180:688–702.e13, 02 2020.

Mingxing Tan, Ruoming Pang, and Quoc V. Le. Efficientdet: Scalable and efficient object detection. *2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 10778–10787, 2020.

Erik F. Tjong Kim Sang and Fien De Meulder. Introduction to the CoNLL-2003 shared task: Language-independent named entity recognition. In *Proceedings of the Seventh Conference on Natural Language Learning at HLT-NAACL 2003*, pages 142–147, 2003. URL https://aclanthology.org/W03-0419.

Vladimir Vovk, Alex Gammerman, and Glenn Shafer. *Algorithmic Learning in a Random World*. Springer-Verlag, Berlin, Heidelberg, 2005.

Vladimir Vovk, Valentina Fedorova, Ilia Nouretdinov, and Alexander Gammerman. Criteria of efficiency for conformal prediction. In *International Symposium on Conformal and Probabilistic Prediction with Applications - Volume 9653*, 2016.

Kevin Yang, Kyle Swanson, Wengong Jin, Connor Coley, Philipp Eiden, Hua Gao, Angel Guzman-Perez, Timothy Hopper, Brian Kelley, Miriam Mathea, Andrew Palmer, Volker Settels, Tommi Jaakkola, Klavs Jensen, and Regina Barzilay. Analyzing learned molecular representations for property prediction. *Journal of Chemical Information and Modeling*, 59(8):3370–3388, 2019.

Manzil Zaheer, Satwik Kottur, Siamak Ravanbakhsh, Barnabas Poczos, Russ R Salakhutdinov, and Alexander J Smola. Deep sets. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2017.

Zexuan Zhong and Danqi Chen. A frustratingly easy approach for entity and relation extraction. In *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 50–61, Online, June 2021. Association for Computational Linguistics. doi: 10.18653/v1/2021.naacl-main.5. URL https://aclanthology.org/2021.naacl-main.5.

Appendix

A PROOFS

A.1 PROOF OF THEOREM 3.1

Proof. This is a well-known result; we prove it here for completeness. Since the nonconformity scores V_i are constructed symmetrically, then

$$((X_1, Y_1), \ldots, (X_{n+1}, Y_{n+1})) \stackrel{d}{=} ((X_{\sigma(1)}, Y_{\sigma(1)}), \ldots, (X_{\sigma(n+1)}, Y_{\sigma(n+1)}))$$
$$\iff (V_1, \ldots, V_{n+1}) \stackrel{d}{=} (V_{\sigma(1)}, \ldots, V_{\sigma(n+1)})$$

for all permutations $(\sigma(1), \ldots \sigma(n+1))$. Therefore, if $\{(X_i, Y_i)\}_{i=1}^{n+1}$ are exchangeable, then so too are their nonconformal scores $\{V_i\}_{i=1}^{n+1}$.

By the construction of $\mathcal{C}$, we have

$$Y_{n+1} \in \mathcal{C}_k(X_{n+1}) \iff V_{n+1} \leq \text{Quantile}(1-\epsilon, V_{1:n} \cup \{\infty\})$$

This implies that V_{n+1} is ranked among the $\lfloor (1-\epsilon) \cdot (n+1) \rfloor$ smallest of $V_1, \ldots, V_n, \infty$. Since V_i are exchangeable, this happens with probability at least $1-\epsilon$. □

A.2 PROOF OF THEOREM 4.3

Our proof will use Theorem 1 of Angelopoulos et al. (2021a), which we state here:

Theorem A.1 (Marginal RCPS)**.** *Let $L_i \colon \mathbb{R} \to \mathbb{R}$, $i = 1, \ldots, n+1$ be exchangeable loss functions, where $L_i(t)$ is non-increasing in t. Also, take $g \colon \mathbb{R} \to \mathbb{R}$ where $g(x)$ is non-decreasing in x. Further assume that $g \circ L_i$ is right-continuous, and*

$$\inf_t g(L_i(t)) < \gamma, \quad \sup_t g(L_i(t)) \leq B < \infty \text{ almost surely.} \tag{15}$$

For any $\gamma \geq 0$, define the random variable $T(\gamma, g)$ as

$$T(\gamma; g) := \inf \left\{ t \colon \frac{1}{n+1} \sum_{i=1}^{n} g(L_i(t)) \leq \gamma \right\}. \tag{16}$$

Then $\mathbb{E}[g \circ L_{n+1}(T(\gamma; g))] \leq \gamma + \frac{B}{n+1}$.

Corollary A.2 (Marginal RCPS, adjusted)**.** *Under the same setting as in Theorem A.1,*

$$\mathbb{E}[g \circ L_{n+1}(\tilde{T}(\gamma; g))] \leq \gamma, \tag{17}$$

where

$$\tilde{T}(\gamma; g) = \inf \left\{ t \colon \frac{1}{n+1} \left(B + \sum_{i=1}^{n} g(L_i(t)) \right) \leq \gamma \right\}. \tag{18}$$

Following their analysis, we give a second corollary for lower bounding the "reward" of function R_{n+1}, where $R_1, \ldots, R_{n+1}$ are now non-decreasing functions.

Corollary A.3 (Marginal RCPS, monotonic non-decreasing case)**.** *Similar to the setting in Theorem A.1, let $R_i \colon \mathbb{R} \to \mathbb{R}$, $i = 1, \ldots, n+1$ be exchangeable reward functions, where $R_i(t)$ is non-decreasing in t. Also, take $g \colon \mathbb{R} \to \mathbb{R}$ where $g(x)$ is non-decreasing in x. Further assume that $g \circ R_i$ is right-continuous, and*

$$\inf_t g(R_i(t)) \geq 0, \quad \sup_t g(R_i(t)) > C \geq \gamma \text{ almost surely.} \tag{19}$$

For any $\gamma \leq 0$, define the random variable $T(\gamma, g)$ as

$$T(\gamma; g) := \inf \left\{ t \colon \frac{1}{n+1} \sum_{i=1}^{n} g(R_i(t)) \geq \gamma \right\}. \tag{20}$$

where we define $\inf \varnothing = \infty$. Then $\mathbb{E}[g \circ R_{n+1}(T(\gamma; g))] \geq \gamma$.

Proof. Let

$$T'(\gamma; g) := \inf \left\{ t \colon \frac{1}{n+1} \sum_{i=1}^{n+1} g(R_i(t)) \geq \gamma \right\}, \tag{21}$$

Since $\inf_t g(R_i(t)) \geq 0$, $\sup_t g(R_i(t)) > C \geq \gamma$, $T'(\gamma; g)$ and $T(\gamma, g)$ are both well-defined almost surely. Since $\inf_t g(R_i(t)) \geq 0$,

$$\frac{1}{n+1} \sum_{i=1}^{n} g(R_i(t)) \geq \gamma \rightarrow \frac{1}{n+1} \sum_{i=1}^{n+1} g(R_i(t)) \geq \gamma. \tag{22}$$

Thus, $T'(\gamma; g) \leq T(\gamma; g)$. Since $g(R_i(t))$ is non-decreasing in t,

$$\mathbb{E}[g \circ R_{n+1}(T(\gamma; g))] \geq \mathbb{E}[g \circ R_{n+1}(T'(\gamma; g))] \tag{23}$$

Let E_f be the unordered set (bag) of $\{R_1, \ldots, R_{n+1}\}$. Then $T'(\gamma; g)$ is a function of E_f, and is a constant conditional on E_f. Exchangeability of R_i and right-continuity of $g \circ R_i$ imply

$$\mathbb{E}[g \circ R_{n+1}(T'(\gamma; g)) \mid E_f] = \frac{1}{n+1} \sum_{i=1}^{n+1} g(R_i(t)) \geq \gamma. \tag{24}$$

The proof is completed by Eq. (23). □

We now prove our main theorem.

Proof. *w.l.o.g.* let us the flip signs of our non-conformity measures such that

$$\overline{\mathrm{MFD}}(x, \mathcal{Z}, t) := \max \left\{ \mathrm{FD}(\mathcal{Z}, \mathcal{S}_i) \colon -\tilde{\mathcal{F}}(x, \mathcal{S}_i) \geq t \right\}, \tag{25}$$

and T_k, $T_{k,\delta}$ are instead defined via $\inf$.

We are given that the sets $\mathcal{S}$ are nested, and it is clear from its definition that FD is non-decreasing for $\mathcal{S}_i \subset \mathcal{S}_j$. Since the ranking of $-\tilde{\mathcal{F}}(x, \mathcal{S})$ is consistent with that of $-|\mathcal{S}|$ and monotonically non-increasing, $\overline{\mathrm{MFD}}$ is therefore also monotonically non-increasing in t.

We prove T_k first. Since $\mathcal{Y}$ is finite by assumption, we have that $\sup_t \overline{\mathrm{MFD}}(x, \mathcal{Z}, t) \leq |\mathcal{Y}| < \infty$. Furthermore, since $-\tilde{\mathcal{F}}(X_i, S_{i,j})$ is upper-bounded by assumption, we have that $\inf_t \overline{\mathrm{MFD}}(x, \mathcal{Z}, t) = 0 < k \in \mathbb{R}_{>0}$. Finally, as evident from its definition in Eq. (25), $\overline{\mathrm{MFD}}(x, \mathcal{Z}, t)$ is right-continuous.Let $L_i(t) = \overline{\mathrm{MFD}}(X_i, Z_i, t)$ and $g(x) = x$. We can then directly apply Corollary A.2 to obtain

$$\mathbb{E}[\mathrm{MFD}(X_{n+1}, Z_{n+1}, T_k)] \leq k. \tag{26}$$

The proof for T_k is completed by the fact that $\mathrm{MFD}(X_{n+1}, Z_{n+1}, T_k) \geq \mathrm{FD}(Z_{n+1}, \mathcal{S}_{n+1,j})$ conditioned on $\mathcal{S}_{n+1,j} \leq T_k$.

We proceed similarly for $T_{k,\delta}$. Let $L_i(t) = \mathbf{1}\{\overline{\mathrm{MFD}}(x, \mathcal{Z}, t) \leq k\}$. Since $\overline{\mathrm{MFD}}$ is non-increasing and right-continuous, $L_i(t)$ is non-decreasing and right-continuous. Let $\gamma = 1 - \delta \in (0, 1)$. By the same argument as before about $\overline{\mathrm{MFD}}$ and from the fact that $L_i(t) \in \{0, 1\}$, we can apply Corollary A.3 to get

$$\mathbb{E}[\mathbf{1}\{\mathrm{MFD}(X_{n+1}, Z_{n+1}, T_{k,\delta}) \leq k\}] = \mathbb{P}(\mathrm{MFD}(X_{n+1}, Z_{n+1}, T_{k,\delta}) \leq k) \geq 1 - \delta. \tag{27}$$

This gives $\mathbb{P}(\mathrm{MFD}(X_{n+1}, Z_{n+1}, T_{k,\delta}) \leq k) \geq 1 - \delta$. The proof for $T_{k,\delta}$ is then completed by the fact that $\mathrm{MFD}(X_{n+1}, Z_{n+1}, T_{k,\delta}) \geq \mathrm{FD}(Z_{n+1}, \mathcal{S}_{n+1,j})$ conditioned on $\mathcal{S}_{n+1,j} \leq T_{k,\delta}$. □

A.3 Proof of Proposition 4.6

Proof. By definition, we have $\mathcal{F}(X_{n+1}, \mathcal{C}_*(X_{n+1}, T_*)) \leq T_*$. By Theorem 4.3, we can then conclude that

$$\mathbb{E}[\mathrm{FD}(Z_{n+1}, \mathcal{C}_*(X_{n+1}, T_*))] \leq k, \tag{28}$$

and

$$\mathbb{P}(\mathrm{FD}(Z_{n+1}, \mathcal{C}_*(X_{n+1}, T_*)) \leq k) \leq T_{k,\delta}) \geq 1 - \delta. \tag{29}$$

Finally, if $S \subset S'$ then $y_i \in S \Rightarrow y_i \in S'$ for any $y_i \in \mathcal{Z}$. Since the candidate sets are nested,

$$|S| < |S'| \Longrightarrow \mathrm{TDP}(\mathcal{Z}, S) \leq \mathrm{TDP}(\mathcal{Z}, S') \tag{30}$$

Thus picking the biggest candidate set maximizes the TDP. □