# Conformalized Interval Arithmetic with Symmetric Calibration

**Rui Luo**[1*], **Zhixin Zhou**[2]

[1]City University of Hong Kong, Hong Kong SAR, China
[2]Alpha Benito Research, Los Angeles, USA
ruiluo@cityu.edu.hk, zzhou@alphabenito.com

**Abstract**

Uncertainty quantification is essential in decision-making, especially when joint distributions of random variables are involved. While conformal prediction provides distribution-free prediction sets with valid coverage guarantees, it traditionally focuses on single predictions. This paper introduces novel conformal prediction methods for estimating the sum or average of unknown labels over specific index sets. We develop conformal prediction intervals for single target to the prediction interval for sum of multiple targets. Under permutation invariant assumptions, we prove the validity of our proposed method. We also apply our algorithms on class average estimation and path cost prediction tasks, and we show that our method outperforms existing conformalized approaches as well as non-conformal approaches.

**Code** — https://github.com/luo-lorry/CIA
**Extended version** — https://arxiv.org/abs/2408.10939

## 1 Introduction

Conformal prediction (Vovk, Gammerman, and Shafer 2005) has gained popularity as a method for uncertainty quantification due to its validity guarantee. It plays a crucial role in modern decision-making processes, particularly in domains where understanding the joint distribution of random variables is essential. Under the exchangeability assumptions on the calibration and test samples, conformal prediction can provide prediction sets with valid coverage guarantees. Existing works have studied prediction intervals for single targets (Lei et al. 2018; Romano, Patterson, and Candes 2019; Sesia and Romano 2021; Luo and Zhou 2024). Given a black box algorithm to estimate the target label, current conformalized algorithms can provide prediction intervals for the target with reliable coverage probability. However, although numerous studies have developed the methodology of conformal prediction, they are limited to predicting a single label, leaving a gap in finding the prediction set for the sum of labels.

This gap is particularly relevant in applications such as transductive conformal prediction on traffic networks. For example, existing Graph Neural Network (GNN) methods can predict the label of each road, where the label can be considered as the cost of traversing that road. This problem has been studied in (Huang et al. 2024; Zargarbashi, Antonelli, and Bojchevski 2023; Zhao, Kang, and Cheng 2024; Luo and Colombo 2025). Conformalized GNN can output a prediction set of each edge's label, but the cost of a route, which is the sum of the labels of the edges on the route, cannot be directly obtained by applying conformal prediction to individual edges. This challenge arises from two main aspects. First, the sum or average of random variables involves the convolution of the density function, and simple interval arithmetic, such as adding up the lower and upper bounds, cannot provide a confidence interval with the desired coverage. Second, the coverage of each confidence interval is in a marginal sense, so the coverage of two labels from two confidence intervals are dependent events. Consequently, it is difficult to use the conformal prediction set for a single label to devise a prediction set for multiple labels. To address this critical gap in the current literature on conformal prediction and expand its applicability to a broader range of uncertainty quantification problems, we introduce the method *Conformal Interval Arithmetic (CIA)*, specifically designed to estimate the average or other symmetric functions of unknown labels over a certain index set, demonstrating the usefulness of our problem setting in many applications where people are interested in obtaining estimates about multiple labels.

Our main proposed method can be described as follows. The core idea is to establish a confidence interval using the exchangeability of the groups of indices. This method involves finding the absolute value of the difference between the sum of labels in a group and its prediction, or absolute residual, and using this as the score function for split conformal prediction. Assuming the groups of indices are exchangeable, we can provide prediction sets with valid marginal coverage. However, this approach cannot handle the case when the calibration samples and the test samples belong to different groups. To address this issue, we introduce a new split conformal prediction method called *symmetric calibration*, which creates a scenario in which the calibration set and the test set play symmetric roles. At the group level, each index within the same group has equal chance of being assigned to either a calibration sample or a test sample. This ensures that the residual from the sums of calibration samples and the sums of test samples have

*Corresponding author

identical distributions. By leveraging the exchangeability of the groups of samples in the calibration set, we can devise a conformal prediction set for the sum of residuals of calibration samples. Since both the calibration samples and the test samples share the same distribution, this conformal prediction set is also valid for the sum of the test samples.

In addition to the methodology of symmetric calibration and the validity guarantee discussed above, we will introduce some important extensions of our method to make it more widely applicable and efficient. First, symmetric calibration is easy to understand when the groups of indices are disjoint. We relax this condition to allow for small amounts of overlap between groups. Second, given prior knowledge of the number of unknown labels being summed, we can stratify the prediction into different levels. Finally, we can incorporate other conformal prediction methods, such as conformalized quantile regression (Romano, Patterson, and Candes 2019), to improve prediction efficiency.

To summarize the contribution of this paper:

1. We introduce the idea of conformalized interval arithmetic (CIA) to provide confidence interval for sum of multiple unknown labels.
2. We develop the technique of symmetric calibration to address the issue that unknown label can appear in different groups. The proposed method has valid coverage guarantee theoretically and empirically.
3. We apply our methods on datasets in (Sesia and Candès 2020), and show the reliability of our method. The code of our method will be published as open source.

## 2 Preliminary and Problem Setup

We will first provide an overview of conformal prediction in the literature. Then, we will study conformalized interval arithmetic (CIA) in a simple setting, where the calibration set and test set are exchangeable at the group level. In this case, we can use the residuals of the sums to perform split conformal prediction on the test group.

### 2.1 Split Conformal Prediction

Conformal prediction has been applied to the following types of problems. Suppose a dataset with an index set $\mathcal{I}$ is partitioned into training, calibration, and test sets, with index sets $\mathcal{I}_{\text{train}}$, $\mathcal{I}_{\text{cal}}$, and $\mathcal{I}_{\text{test}}$, respectively. Firstly, we apply a black-box model to the training data $\{(x_i, y_i)\}_{i\in\mathcal{I}_{\text{train}}}$ and obtain a function $\widehat{f}$ such that $\widehat{y}_i = \widehat{f}(x_i)$. We define a conformity score function $s(x, y)$, which depends on $\widehat{f}$. A larger value of the score function indicates that $y$ is less likely to be a true label. A typical choice of the score function is

$$s_i = s(x_i, y_i) = |y_i - \widehat{y}_i|.$$

We will first focus on this choice and then extend it to other choices in later sections. For data in $\{(x_i, y_i)\}_{i\in\mathcal{I}_{\text{cal}}}$, we evaluate $s_i = |y_i - \widehat{y}_i|$ and let $Q_{1-\alpha}$ be the $\lceil(1+|\mathcal{I}_{\text{cal}}|)(1-\alpha)\rceil$-th smallest value of $|y_i - \widehat{y}_i|$ for $i \in \mathcal{I}_{\text{cal}}$[1]. Then, for $i \in \mathcal{I}_{\text{test}}$, the prediction set of $y_i$ is defined as

$$\mathcal{C}(i) = \{y : |y - \widehat{y}_i| \le Q_{1-\alpha}\} = [\widehat{y}_i - Q_{1-\alpha}, \widehat{y}_i + Q_{1-\alpha}].$$

Under the assumption of exchangeability of the samples in the calibration set and the test set, it can be shown that this conformal prediction set has coverage of at least $1 - \alpha$. In prediction tasks, the resulting prediction set $\mathcal{C}(i)$ is usually an interval. Our present work will focus on the generalization of these prediction intervals.

### 2.2 Problem Setup and a Naive Approach

In many applications, we are not just satisfied with the prediction set for a single label $y_i$. Unlike existing works, this paper aims to provide the prediction set of the sum of variables, i.e., a prediction set $\mathcal{C}(S)$ for $\sum_{i\in S} y_i$. We will apply conformal prediction techniques so that $1 - \alpha$ coverage is guaranteed under certain exchangeability assumption.

The choice of $S \subseteq \mathcal{I}_{\text{test}}$ can depend on fixed or random procedures. Some concrete examples will be considered in Section 7. We can start with a simple example. Let $S = \{i, j\} \subseteq \mathcal{I}_{\text{test}}$. Suppose a simple split conformal prediction provides prediction intervals $\mathbb{P}(Y_i \in [L_i, U_i]) \ge 1 - \alpha$ and $\mathbb{P}(Y_j \in [L_j, U_j]) \ge 1 - \alpha$, then the addition operation of interval arithmetic gives

$$\begin{aligned}&\mathbb{P}(Y_i + Y_j \in [L_i + L_j, U_i + U_j])\\ &\ge \mathbb{P}(Y_i \in [L_i, U_i] \text{ and } Y_j \in [L_j, U_j])\\ &\ge 1 - 2\alpha.\end{aligned}$$

Hence, the addition operation of interval arithmetic of $1 - \alpha$ confidence intervals cannot guarantee $1 - \alpha$ coverage. One possible correction is to first find $1 - \alpha/2$ confidence intervals for $Y_i$ and $Y_j$, then construct the prediction interval by adding the intervals. This can be generalized using Bonferroni correction to predict the sum of any number of unknown labels, without requiring assumptions about the dependencies between variables. However, this method clearly lacks efficiency. Our method will be compared with Bonferroni correction in empirical experiments in Section 7.

## 3 Conformal Interval Arithmetic

We first study a simple case, in which the problem can be reduced to existing conformal prediction techniques. Suppose we have disjoint index sets $S_1, S_2, \ldots, S_K, S_{K+1}$ and suppose we observe samples $(x_i, y_i)$ for $i \in S_1 \cup S_2 \cup \cdots \cup S_K$, and $x_j$ for $j \in S_{K+1}$ as test samples. We can then calculate the residual as the score function on the set level:

$$s_k := s(z_k) := \left|\sum_{i\in S_k} (y_i - \widehat{y}_i)\right|, k \in [K]$$

where $K = \{1, \ldots, K\}$ and $z_k = (x_i, y_i)_{i\in S_k}$ denotes a group of pairs of $(x, y)$. The following procedure follows similarly from the conformal prediction from the previous section. Let $Q_{1-\alpha}$ be the $\lceil(1+K)(1-\alpha)\rceil$-th smallest value of $s_k, k \in [K]$. We define the prediction set $\mathcal{C}(S_{K+1})$ as

$$\left[\sum_{i\in S_{K+1}} \widehat{y}_i - Q_{1-\alpha}, \sum_{i\in S_{K+1}} \widehat{y}_i + Q_{1-\alpha}\right]. \quad (1)$$

[1] If $\lceil(1+|\mathcal{I}_{\text{cal}}|)(1-\alpha)\rceil > |\mathcal{I}_{\text{cal}}|$, then $Q_{1-\alpha} = \infty$. Here, $Q_{1-\alpha}$ represents the quantile of the scores. The meaning of $Q_{1-\alpha}$ may vary depending on how the scores are defined in different sections.

This procedure provides conformal prediction for the sum of unknown labels and is called Conformal Interval Arithmetic (CIA) in this paper. Assuming group exchangeability, it can be guaranteed that the coverage probability is valid.

**Proposition 1.** *Suppose that $Z_k$ is group exchangeable in the sense of for any permutation $\pi$ on $[K+1]$:*

$$\mathbb{P}(Z_1, \ldots, Z_{K+1}) = \mathbb{P}(Z_{\pi(1)}, \ldots, Z_{\pi(K+1)}), \qquad (2)$$

*then the prediction set $\mathcal{C}(S_{K+1})$ in* (1) *satisfies*

$$\mathbb{P}\left(\sum_{i \in S_{K+1}} y_i \in \mathcal{C}(S_{K+1})\right) \geq 1 - \alpha.$$

The idea of conformal interval arithmetic can be applied to the simple case where the test sample with unknown labels belongs only to the same group, i.e., $S_{K+1}$, and $y_i$ for $i \in S_1, \ldots, S_K$ must be fully observed. However, this assumption may not hold for many scenarios, such as cost estimation of edges in a route, as discussed in the introduction. Consider a traffic network where we randomly sample two nodes and find the shortest path between them. If a sufficient amount of edge weights are not observed, then only a few routes will contain all edges with observed weights. Let $S_1, \ldots, S_K$ be the set of edges on a path, determined by the network structure and edge costs. For unobserved weights, they can be estimated by their predictions. In this method, every shortest path can contain edges with unobserved weights, and edges can belong to multiple paths. These issues make the simple case method inapplicable.

## 4 Symmetric Calibration

Let us formally state the problem. We aim to provide a prediction set for $\sum_{i \in S_k} y_i$. Since $y_i$ is not observed only for $i \in \mathcal{I}$, it is equivalent to finding the prediction set for $\sum_{i \in S_k \cap \mathcal{I}_{\text{test}}} y_i$. In the previous section, we considered a simple case where $S_{K+1} = \mathcal{I}_{\text{test}}$. Here, we will consider a general case where every $S_k$ can contain indices of unobserved labels. To tackle this problem, we introduce the method of *symmetric calibration*.

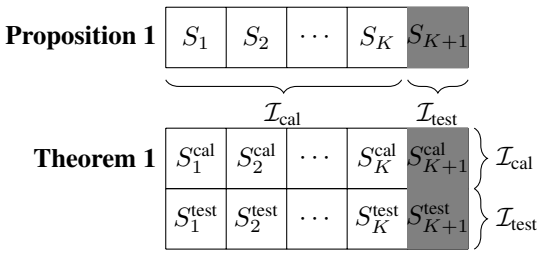


Figure 1: Diagram illustrating the partition of data into calibration and test sets for Proposition 1 and Theorem 1. The sums within each of $S_1, \ldots, S_K, S_{K+1}$ (**Top**) serve as the calibration data to predict the sum within $S_{K+1}$, mirroring the calibration row for the theorem's diagram (**Bottom**). As $S_{K+1}^{\text{cal}}$ and $S_{K+1}^{\text{test}}$ are exchangeable, their sums have the same distribution. Therefore, the prediction set (6) can be derived similarly to (5).

Before training the model $\widehat{f}$, we first hold out a calibration set that has the same size as the test set, i.e., $|\mathcal{I}_{\text{cal}}| = |\mathcal{I}_{\text{test}}|$. Note that we assume the number of observed labels must be greater than the number of unobserved labels. Now, we focus on the calibration and test samples. Let us assume

$$S_1 \cup \cdots \cup S_K \cup S_{K+1} = \mathcal{I}_{\text{cal}} \cup \mathcal{I}_{\text{test}},$$

and let us define

$$S_k^{\text{cal}} := S_k \cap \mathcal{I}_{\text{cal}} \quad \text{and} \quad S_k^{\text{test}} = S_k \cap \mathcal{I}_{\text{test}}. \qquad (3)$$

Our task becomes finding the prediction set for $\sum_{i \in S_k^{\text{test}}} y_i$ for every fixed $k \in [K+1]$. To simplify the presentation, let us focus on the prediction set of $\sum_{i \in S_{K+1}^{\text{test}}} y_i$. Other prediction sets and the coverage guarantee in the theorems can be obtained in the same manner. Using conformal interval arithmetic, we define the score function similar to (3):

$$s_k^{\text{cal}} = \left| \sum_{i \in S_k^{\text{cal}}} (y_i - \widehat{y}_i) \right| \qquad (4)$$

Let $Q_{1-\alpha}$ be the $\lceil (1+K)(1-\alpha) \rceil$-th smallest value of $s_k^{\text{cal}}, k = 1, \ldots, K$. Then we define the prediction set for $\sum_{i \in S_k^{\text{test}}} y_i$, denoted by $\mathcal{C}(S_k^{\text{test}})$:

$$\left[ \sum_{i \in S_k^{\text{test}}} \widehat{y}_i - Q_{1-\alpha}, \sum_{i \in S_k^{\text{test}}} \widehat{y}_i + Q_{1-\alpha} \right]. \qquad (5)$$

The above procedure is summarized in Algorithm 1.

Algorithm 1: CIA with Symmetric Calibration

1: **Input:** labeled samples $\{(x_i, y_i)\}_{i \in \mathcal{I} \setminus \mathcal{I}_{\text{test}}}$, unlabeled samples $\{x_i\}_{i \in \mathcal{I}_{\text{test}}}$, a black box algorithm $\mathcal{B}$, disjoint index sets $S_1, \ldots, S_K \subseteq \mathcal{I}$.
2: **Output:** Prediction sets for $\sum_{i \in S_{K+1}^{\text{test}}} y_i$.
3: Hold out a calibration set $\mathcal{I}_{\text{cal}}$ from $\mathcal{I} \setminus \mathcal{I}_{\text{test}}$ such that $\mathcal{I}_{\text{cal}}$ has the same size as $\mathcal{I}_{\text{test}}$.
4: $\widehat{f} \leftarrow \mathcal{B}(\{(x_i, y_i)\}_{i \in \mathcal{I} \setminus (\mathcal{I}_{\text{cal}} \cup \mathcal{I}_{\text{test}})}$.
5: $\widehat{y}_i \leftarrow \widehat{f}(x_i)$ for $i \in \mathcal{I}_{\text{cal}} \cup \mathcal{I}_{\text{test}}$.
6: $s_k^{\text{cal}} = \left| \sum_{i \in S_k^{\text{cal}}} (y_i - \widehat{f}(x_i)) \right|$ for $k \in [K]$
7: $Q_{1-\alpha} \leftarrow \lceil (1+K)(1-\alpha) \rceil$-th smallest value of $\{s_k^{\text{cal}} : k \in [K]\} \cup \{\infty\}$.
8: $\mathcal{C}(S_{K+1}^{\text{test}}) \leftarrow \left[ \sum_{i \in S_{K+1}^{\text{test}}} \widehat{y}_i - Q_{1-\alpha}, \sum_{i \in S_{K+1}^{\text{test}}} \widehat{y}_i + Q_{1-\alpha} \right]$.
9: **Output:** $\mathcal{C}(S_{K+1}^{\text{test}})$.

We will analyze the coverage probability of $\mathcal{C}(S_{K+1}^{\text{test}})$. In this section, we assume that $S_1, \ldots, S_K, S_{K+1}$ are disjoint. We will introduce the reasoning behind this in two steps. Firstly, under the assumption of group exchangeability, the tuple of score functions $(s_1^{\text{cal}}, \ldots, s_K^{\text{cal}}, s_{K+1}^{\text{cal}})$ is exchangeable. $s_k^{\text{cal}}$ has a uniform rank over $[K+1]$, so $s_{K+1}^{\text{cal}} \leq Q_{1-\alpha}$ with probability at least $1 - \alpha$. Equivalently, this suggests

that the interval

$$\left[\sum_{i\in S_{K+1}^{\text{cal}}} \widehat{y}_i - Q_{1-\alpha}, \sum_{i\in S_{K+1}^{\text{cal}}} \widehat{y}_i + Q_{1-\alpha}\right] \tag{6}$$

covers $\sum_{i\in S_{K+1}^{\text{cal}}} y_i$ with probability at least $1-\alpha$. This result comes from Proposition 1. However, we are not interested in the sum over $S_{K+1}^{\text{cal}}$. Instead, we want to show that the sum over $S_{K+1}^{\text{test}}$ has the same property. We need the following assumption: the index is randomly assigned to the calibration set or test set with equal probability:

$$\mathbb{P}(i \in \mathcal{I}_{\text{cal}}) = \mathbb{P}(i \in \mathcal{I}_{\text{test}}) = 0.5 \tag{7}$$

for all $i \in \mathcal{I}\backslash\mathcal{I}_{\text{train}}$ independently. In other words, $\mathcal{I}_{\text{cal}}$ has a uniform distribution on the power set $2^{\mathcal{I}\backslash\mathcal{I}_{\text{train}}}$. We summarize the above argument in the following theorem.

**Theorem 1.** *Suppose*

*(a) $S_1, \ldots, S_K, S_{K+1}$ are disjoint.*
*(b) The groups are exchangeable in the sense of* (2)*.*
*(c) The labels in $\mathcal{I}_{cal} \cup \mathcal{I}_{test}$ are assigned to $\mathcal{I}_{cal}$ and $\mathcal{I}_{test}$ with equiprobability independently.*

*Then the coverage probability satisfies*

$$\mathbb{P}\left(\sum_{i\in S_{K+1}^{test}} y_i \in \mathcal{C}(S_{K+1}^{test})\right) \geq 1-\alpha.$$

**Remark.** *The random splitting assumption* (7) *is slightly different from the procedure of holding out a calibration set $\mathcal{I}_{cal}$ with the same size as $\mathcal{I}_{test}$ in Algorithm 1. Theorem 1 suggests that it only requires the calibration set and test set have roughly the same size. Suppose $\mathcal{I}_{cal}$ and $\mathcal{I}_{test}$ are partitions of $\mathcal{I}_{cal} \cup \mathcal{I}_{test}$ with equal size, this complicates the proof of the theorem, as we can only swap $S_k^{cal}$ and $S_k^{test}$ when they have the same size.*

# 5 Extensions

In this section, we will extend the applicability of the proposed method to handle overlapping subsets. Furthermore, we will introduce a stratified technique and utilize Conformal Quantile Regression (CQR) (Romano, Patterson, and Candes 2019) to enhance the efficiency of our approach.

## 5.1 Overlapping Subsets

Algorithm 1 can be applied to overlapping subsets $S_1, \ldots, S_K, S_{K+1}$, but the conclusion of Theorem 1 is valid only if the subsets are disjoint. There are applications where we must assume that the subsets overlap. We will study these applications in Section 7.2. Theoretically, overlapping subsets affect the validity of the method. The following theorem shows that as long as each subset overlaps with only a small portion of the other subsets, the effect on the coverage probability is acceptable.

**Theorem 2.** *We still assume condition (b) and (c) in Theorem 1, and suppose that*

$$\max_{\ell\in[K+1]} \frac{1}{K+1}\sum_{k\neq\ell} \mathbb{1}\{S_k \cap S_\ell \neq \emptyset\} = \delta,$$

*Then for fixed $k \in [K+1]$, the coverage probability satisfies*

$$\mathbb{P}\left(\sum_{i\in S_{K+1}^{test}} y_i \in \mathcal{C}(S_{K+1}^{test})\right) \geq 1-\alpha-\delta.$$

## 5.2 Stratified Conformal Prediction

Given the known number of unknown labels in each test set, we can leverage this information to enhance the efficiency of conformal prediction. For instance, consider a case where $|S_k^{\text{test}}|$ is relatively small, such as $|S_k^{\text{test}}| = 1$. In this case, it is more appropriate to compare the residual with other residuals that have a smaller number of unknown labels in the calibration set. By incorporating this idea, we can modify Algorithm 1 and derive a stratified version of the algorithm that takes advantage of this additional information to improve its efficiency.

Algorithm 2: Stratified CIA with Symmetric Calibration

1: **Input:** labels and predictions $\{(y_i, \widehat{y}_i)\}_{i\in\mathcal{I}\backslash\mathcal{I}_{\text{train}}}$, index sets $S_1, \ldots, S_K, S_{K+1} \subseteq \mathcal{I}$. Partition of positive integers $C_1, C_2, \ldots$
2: **Output:** Prediction sets for $\sum_{i\in S_{K+1}^{\text{test}}} y_i$.
3: $s_k^{\text{cal}} \leftarrow \left|\sum_{i\in S_k^{\text{cal}}}(y_i - \widehat{y}_i)\right|$ for $k \in [K+1]$.
4: Find $C_j$ such that $|S_k^{\text{test}}| \in C_j$.
5: $n_j \leftarrow |\{k \in [K] : |S_k^{\text{cal}}| \in C_j\}|$.
6: $Q_{1-\alpha}^j \leftarrow \lceil(1+n_j)(1-\alpha)\rceil$-th smallest value of $\{s_k^{\text{cal}} : |S_k^{\text{cal}}| \in C_j, k \in [K]\} \cup \{\infty\}$.
7: $\mathcal{C}(S_{K+1}^{\text{test}}) \leftarrow \left[\sum_{i\in S_{K+1}^{\text{test}}} \widehat{y}_i - Q_{1-\alpha}^j, \sum_{i\in S_{K+1}^{\text{test}}} \widehat{y}_i + Q_{1-\alpha}^j\right]$.
8: **Output:** $\mathcal{C}(S_{K+1}^{\text{test}})$.

In practice, the stratified approach can often reduce the size of the prediction set. However, if the size of $C_j$ is too small, it may negatively impact the validity of the method. To mitigate this issue, it is advisable to set lower bounds for $n_j$ when defining the sets $C_j$. This ensures that each stratum has a sufficient number of samples to maintain the desired coverage probability while still benefiting from the increased efficiency provided by the stratified technique.

## 5.3 Conformal Quantile Regression

From the theorems, we can observe that the form of the score function can be modified. For instance, we can adopt Conformal Quantile Regression (CQR) (Romano, Patterson, and Candes 2019). The score function can be defined as

$$s_k^{\text{cal}} = \max\left\{\sum_{i\in S_k^{\text{cal}}} (\widehat{y}_i^{\alpha/2} - y_i), \sum_{i\in S_k^{\text{cal}}} (y_i - \widehat{y}_i^{1-\alpha/2})\right\}$$

for $k \in [K]$, where $\widehat{y}_i^{\alpha/2}$ and $\widehat{y}_i^{1-\alpha/2}$ are the outputs of quantile regression at quantiles $\alpha/2$ and $1-\alpha/2$, respectively, to predict $y_i$. We can obtain $Q_{1-\alpha}$ in Algorithm 1 using the

score functions of CQR. The prediction set $\mathcal{C}(S_k^{\text{test}})$ is then given by

$$\left[\sum_{i\in S_k^{\text{test}}} \widehat{y}_i^{\alpha/2} - Q_{1-\alpha}, \sum_{i\in S_k^{\text{test}}} \widehat{y}_i^{1-\alpha/2} + Q_{1-\alpha}\right].$$

CQR can also be employed in Algorithm 2, with similar modifications required in Step 4 and Step 7 of the algorithm.

In CQR, if $\widehat{y}_i^{\alpha/2}$ and $\widehat{y}_i^{1-\alpha/2}$ are the true quantiles of the distribution of $Y_i$, then CQR can achieve conditional $1-\alpha$ coverage. However, after combining with CIA, we do not have conditional validity because the sum of quantiles is not necessarily the quantile of the sum. Nevertheless, the upper and lower quantiles can still help in measuring the variation of the estimation and improving the efficiency of the conformal prediction.

# 6 Related Work

In this section, we will explore the relationships and distinctions between the present work and prior research. Although we maintain that the ideas of Conformal Interval Arithmetic (CIA), symmetric calibration, and stratified conformal prediction are original and groundbreaking, and that they are unparalleled among existing techniques for addressing the issues presented in this paper, there are studies in the literature that are relevant to these concepts.

**Multiple Test.** Conformal prediction has been applied to multiple testing (Fisch et al. 2021; Jin and Candès 2023; Bates et al. 2023; Bashari et al. 2024; Angelopoulos et al. 2024). While our present work also deal with multiple random variables, CIA provides only one confidence interval for the sum of these random variables, so our method essentially differs from these multiple test-based approaches.

**Half Sampling.** In Theorem 1, the conclusion can be restated as follows: the $(1-\alpha)$-th quantile of the score function values in the calibration set is larger than a proportion of $1-\alpha$ of the score function values in the test set. At a higher level, this is a comparison between the $(1-\alpha)$-th quantile of half of the samples and the entire sample set. In this regard, symmetric calibration can be linked to the balanced half-sample method (Kish and Frankel 1970; Shao and Wu 1992), which was popular from 1970 to 2000. The balanced half-sample method shares the same idea as symmetric calibration, as it compares the statistic from the half-sample with the entire sample. It is possible that further theory about symmetric calibration can be developed by building upon classical half-sample theory.

**Fair Conformal Prediction.** Fair conformal prediction (Lu et al. 2022; Liu et al. 2022; Wang et al. 2024) and group-conditional conformal prediction (Romano et al. 2020; Feldman, Bates, and Romano 2021; Melki et al. 2023; Plassier et al. 2024) aim to equalize coverage across various groups characterized by group attributes or auxiliary data. This idea is related to the stratified conformal prediction discussed in this paper. The procedures in these methods are very similar. However, the motivation behind stratified conformal prediction is to improve the efficiency of the conformal prediction, while group-conditional conformal prediction is designed to ensure fairness among different groups.

**Robust Route Planning.** As mentioned in the introduction, our method can be applied to path cost prediction. We will further discuss this application in Section 7.2. In (Sun, Liu, and Li 2023; Patel, Rayan, and Tewari 2024), the authors introduce Conformal-Predict-Then-Optimize (CPO) for linear programming by predicting edge weights and minimizing upper bounds of path costs for optimal robust solutions. While our method can generate conformal predictions for algorithm-selected paths, choosing the path with the smallest upper bound depends on the score function, compromising the guarantee of $1-\alpha$ coverage. Additionally, applying CPO to our problem is inefficient, as it only provides conformal predictions for the joint distribution of edge weights. We are also aware of recent works (Bertsimas, Gupta, and Kallus 2018; Chassein, Dokka, and Goerigk 2019; Johnstone and Cox 2021; Stanton, Maddox, and Wilson 2023; Sadana et al. 2024) which address decision making under uncertainty and methods which tackle adversary data (Gendler et al. 2021; Bastani et al. 2022; Ghosh et al. 2023) and distribution shifts (Tibshirani et al. 2019; Prinster et al. 2024; Zhao et al. 2024) in conformal prediction.

# 7 Application and Experiment

In this section, we will introduce two main applications of the proposed methods. Then we will analyze the empirical performance of our algorithms on public datasets.

## 7.1 Application 1: Group Average Prediction

Suppose that in a dataset, $x_{ij}$ is the $j$th feature of the $i$th sample, and suppose that this feature is categorical. We can then define the subsets as

$$S_k = \{i \in \mathcal{I}_{\text{cal}} \cup \mathcal{I}_{\text{test}} : x_{ij} = k\}$$

To predict the average value in class $k$, we can equivalently predict $\sum_{i\in S_k^{\text{test}}} y_i$, since $S_k^{\text{test}}$ is the set of indices corresponding to the unknown labels in class $k$.

**Datasets**

1. *Bike Sharing (Fanaee-T 2013):* This dataset is used to investigate the factors influencing bike rental demand. The grouping features are SEASON, WORKINGDAY, and WEATHER, which have 272 unique value combinations out of 10886 samples.
2. *Community Crime (Redmond 2009):* This dataset is used to predict the per capita violent crime rate of a community using its demographic features. The grouping features are STATE and COUNTY, which have 350 unique value combinations out of 1994 samples.
3. *Medical Expenditure Panel Survey (AfHRaQ 2021):* This dataset is used to predict the utilization of medical services based on various features. The grouping features are AGE, RACE, and MARRIAGE, which have 302, 302, and 301 unique value combinations out of 15785, 17541,

and 15656 samples for the years 2019, 2020, and 2021, respectively.

Following similar procedures in (Sesia and Candès 2020), we standardize the response variables $Y$ for all datasets. We use $70\%$ of data for training the quantile regression model and $30\%$ for calibration and testing.

### 7.2 Application 2: Path Cost Prediction

As discussed in the introduction, our method has a valuable application in route planning problems. We consider a transductive setting in edge regression problems, where $\mathcal{I}$ represents the set of edge indices. In the experiment, we assume that there are edges with unknown labels, and we denote the set of these edges as $\mathcal{I}_{\text{test}}$. Similar to Algorithm 1, we can hold out a set of edges with known labels as the calibration set $\mathcal{I}_{\text{cal}}$. Utilizing node and/or edge features, the network structure, and the observed edge labels (costs), we train a GNN to predict the labels of the calibration set and test set. The transductive setting of GNN is introduced in (Huang et al. 2024), where they consider the prediction of node labels. However, more recent work (Zhao, Kang, and Cheng 2024; Luo and Colombo 2025) which focuses on edge prediction is more suitable for this particular application set.

While the choice of $S_k$'s can be quite general, the most interesting case in the traffic network data is when we use the edges on the shortest path as the set of edges $S_k$. First, we randomly sample two nodes $(s, t)$ as the source and target nodes, respectively. Then, we use $y_i$ for $i \in \mathcal{I}_{\text{train}}$ and $\widehat{y}_i$ for $i \in \mathcal{I}_{\text{cal}} \cup \mathcal{I}_{\text{test}}$ as the cost of each edge. Next, we find the shortest path using Dijkstra's algorithm. By sampling $K$ pairs of $(s, t)$, the indices of edges on the path form an index set $S_{(s,t)}$. Given the index sets and the predictions $\widehat{y}_i$ for calibration samples and test samples, we can apply Algorithm 1 or Algorithm 2 to obtain the conformal prediction.

In this setting, the index sets can overlap, meaning that different paths can contain the same edge. However, in a large network without bottlenecks, two shortest paths with random sources and targets share the same edge with a small probability. Therefore, Theorem 2 can guarantee that the coverage of the prediction set is very close to $1 - \alpha$.

**Dataset:** *Road Traffic in Anaheim and Chicago (Bar-Gera, Stabler, and Sall 2023).* This dataset is used to predict the traffic flow along certain roads based on the traffic flow along other roads. The predicted traffic flow is then utilized as edge cost for shortest path planning. This scenario corresponds to the situation of subsets with overlaps, where the edges (roads) in different subsets (shortest paths) may overlap with each other. We collect 2000 shortest paths by randomly sampling $(s, t)$ pairs. The Anaheim dataset consists of 413 nodes and 858 edges, and the Chicago dataset consists of 541 nodes and 2150 edges. We adopt a similar procedure from (Jia and Benson 2020; Huang et al. 2024), and allocate 50%, 10%, and 40% for training, validation, and calibration and testing.

### 7.3 Baseline

Despite of the significance of our method in application, there are few method can provide valid confidence interval for finding the intervals. We introduce three possible methods for comparison.

**Group Sampling Conformal Prediction.** This method corresponds to the approach described in Section 3. To predict $\sum_{i \in S_k^{\text{test}}} y_i$, we sample $K+1$ groups of samples from the calibration sets, where each group contains $|S_k^{\text{test}}|$ samples. Let $S_1, \ldots, S_K, S_{K+1}$ be the index sets of these groups. The prediction set is then given by (1).

It is important to note that in this method, the indices of $S_1, \ldots, S_K, S_{K+1}$ may not belong to the same class, which means that this approach may not be appropriate in all cases. However, the method is always applicable, regardless of the class composition of the sampled groups.

**Normal Confidence Interval.** This approach assumes that $\widehat{y}_i - y_i$ follows an i.i.d. $N(0, \sigma^2)$ distribution. The predicted common variance is estimated as

$$\widehat{\sigma}^2 = \frac{1}{|\mathcal{I}_{\text{cal}}| - 1} \sum_{i \in \mathcal{I}_{\text{cal}}} (\widehat{y}_i - y_i)^2.$$

Then, the prediction set for $\sum_{i \in S_k^{\text{test}}} y_i$ is given by

$$\left[ \sum_{i \in S_k^{\text{test}}} \widehat{y}_i + z_{\frac{\alpha}{2}} \sqrt{|S_k^{\text{test}}|} \widehat{\sigma}, \sum_{i \in S_k^{\text{test}}} \widehat{y}_i + z_{1-\frac{\alpha}{2}} \sqrt{|S_k^{\text{test}}|} \widehat{\sigma} \right],$$

where $z_{\alpha/2}$ and $z_{1-\alpha/2}$ are the $\alpha/2$ and $(1-\alpha/2)$-th quantile of the standard normal distribution. It is important to note that this approach relies on the assumption of normality, which is unlikely to hold in most real-world datasets.

**Bonferroni correction.** The Bonferroni correction has been employed in various conformal methods (Fisch et al. 2021; Jin and Candès 2023; Cleaveland et al. 2024). As mentioned in Section 2.2, it is also applicable to our problem. While the Bonferroni correction is a valid approach, it often lacks efficiency in the context of our problem.

### 7.4 Results

We split the calibration and test sets equally as indicated by (7), repeating this process 100 times to record the average results and the standard deviation. Figure 2 presents the results for constructing prediction sets for subsets without overlaps, specifically for the *Bike Sharing*, *Community Crime*, and *Medical Expenditure Panel Survey* datasets. We compare the baseline methods with Algorithm 2 using CQR, as described in Section 5.3. In these figures, we refer to our method as "CIA." The baseline methods from Section 7.3 are labeled as "Group," "Normal," and "Bonferroni," respectively. Additional results will appear in the appendix.

For these five datasets, the left plots present the desired miscoverage rate $\alpha$ versus the true coverage rate. The closer the curve aligns with the line $1 - \alpha$, the easier it is for the method to achieve the desired coverage. Our proposed method, CIA, is very close to the line $1 - \alpha$. There is a small gap in the result of the Anaheim dataset. This result aligns with Theorem 2. Compared to the Chicago dataset, Anaheim is a smaller traffic network. The probability of two shortest paths overlapping each other can affect the validity. This

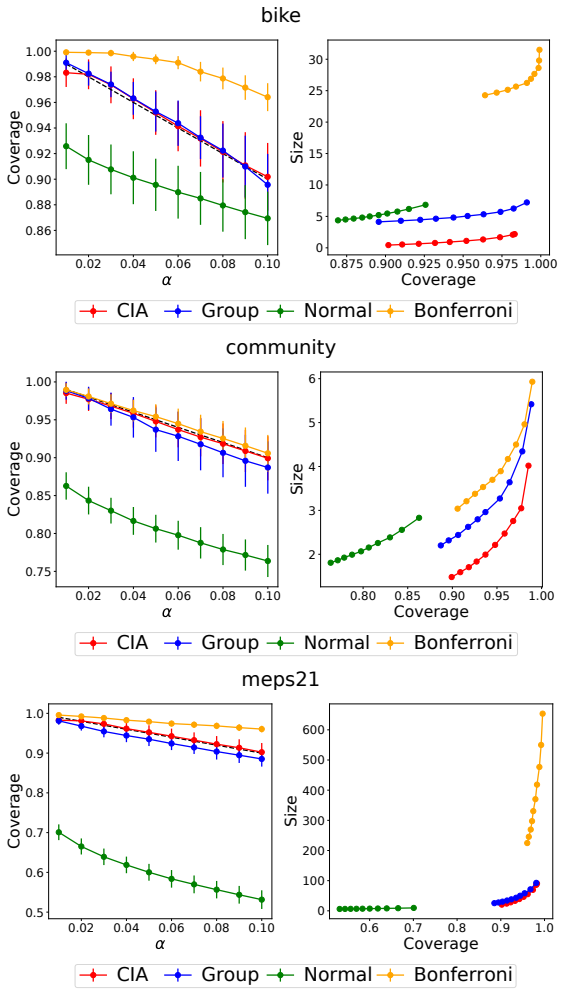


Figure 2: Results for constructing prediction sets for subsets without overlaps (Section 7.1). CIA achieves guaranteed $1-\alpha$ coverage as well as optimal efficiency on various datasets.

issue is much less obvious in a larger traffic network like Chicago. The quantity $\delta$ from the theorem, along with the coverage gap, is provided in the appendix. Recall that the group-sample conformal method is proposed in Section 3. This is also a conformal prediction method, so its validity is also close to the desired line $1-\alpha$ in the results of Figure 2. However, the validity is much less reliable in the results of Figure 3. This is because edge weights from random samples are not exchangeable with the edge weights on the shortest path. The method of Bonferroni correction has good validity in the example of the community dataset. The reason is that under the current setting of calibration and test set ratio, each subset contains one or two samples, which is very close to the case of classical conformal prediction. In other datasets, Bonferroni correction has too high coverage. For normal confidence intervals, since the assumption is not true in general, the coverage is much lower than the desired one. In conclusion, our proposed method has the most reliable validity in all datasets that we have studied.

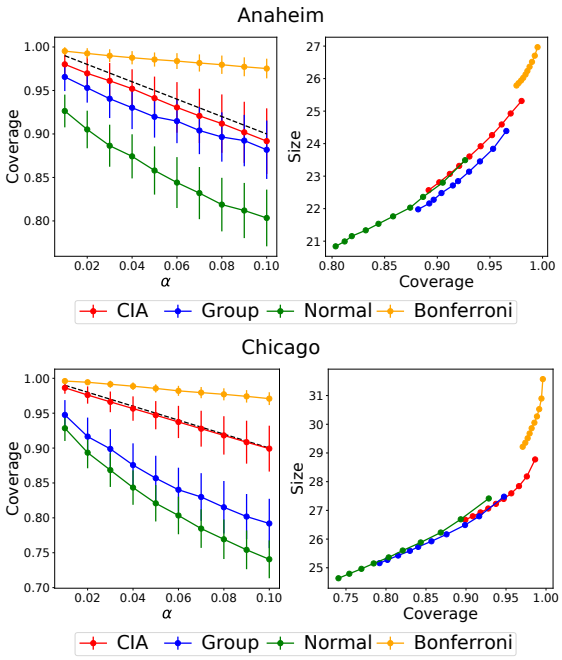


Figure 3: Results for constructing prediction sets for subsets with overlaps (Section 7.2). CIA achieves close to $1-\alpha$ coverage and optimal efficiency on two road traffic datasets.

Under the choices of $\alpha$ on the left plots, the right plots compare the coverage versus the prediction set size. The lower the curve, the more efficient and informative the prediction set provided by the method. In the plots of Figure 2, CIA is the most efficient in all datasets. In the plots of Figure 3, CIA has similar efficiency to the Group and Normal methods. In conclusion, CIA is the most efficient method with valid coverage.

## 8 Conclusion

In this paper, we have introduced conformalized interval arithmetic with symmetric calibration, to tackle the challenges of conformal prediction for interdependent data. Our approach offers methodological contributions and theoretical guarantees for achieving the desired coverage under reasonable exchangeability assumptions. Extensive experiments on real-world datasets demonstrate the validity and efficiency of our method, outperforming others even when they fail to ensure valid coverage. We hope our core ideas and techniques will foster the development of more robust conformal prediction methods across various domains.

## Acknowledgments

The work described in this paper was partially supported by grants from City University of Hong Kong (Project No.9610639, 6000864) and a grant from Chengdu Municipal Office of Philosophy and Social Science (Project No.2024BS013). We thank the reviewers for their insightful suggestions. We also appreciate the helpful discussions with Andro Sabashvili about Theorem 1.

## References


AfHRaQ. 2021. Medical expenditure panel survey. https://meps.ahrq.gov/mepsweb/data_stats/data_overview.jsp.

Angelopoulos, A. N.; Bates, S.; Fisch, A.; Lei, L.; and Schuster, T. 2024. Conformal Risk Control. In *The Twelfth International Conference on Learning Representations*.

Bar-Gera, H.; Stabler, B.; and Sall, E. 2023. Transportation networks for research core team. https://github.com/bstabler/TransportationNetworks. Accessed: 2023-09-10.

Bashari, M.; Epstein, A.; Romano, Y.; and Sesia, M. 2024. Derandomized novelty detection with FDR control via conformal e-values. *Advances in Neural Information Processing Systems*, 36.

Bastani, O.; Gupta, V.; Jung, C.; Noarov, G.; Ramalingam, R.; and Roth, A. 2022. Practical adversarial multivalid conformal prediction. *Advances in Neural Information Processing Systems*, 35: 29362–29373.

Bates, S.; Candès, E.; Lei, L.; Romano, Y.; and Sesia, M. 2023. Testing for outliers with conformal p-values. *The Annals of Statistics*, 51(1): 149–178.

Bertsimas, D.; Gupta, V.; and Kallus, N. 2018. Data-driven robust optimization. *Mathematical Programming*, 167: 235–292.

Chassein, A.; Dokka, T.; and Goerigk, M. 2019. Algorithms and uncertainty sets for data-driven robust shortest path problems. *European Journal of Operational Research*, 274(2): 671–686.

Cleaveland, M.; Lee, I.; Pappas, G. J.; and Lindemann, L. 2024. Conformal prediction regions for time series using linear complementarity programming. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 38, 20984–20992.

Fanaee-T, H. 2013. Bike Sharing Dataset. UCI Machine Learning Repository. DOI: https://doi.org/10.24432/C5W894.

Feldman, S.; Bates, S.; and Romano, Y. 2021. Improving conditional coverage via orthogonal quantile regression. *Advances in neural information processing systems*, 34: 2060–2071.

Fisch, A.; Schuster, T.; Jaakkola, T. S.; and Barzilay, R. 2021. Efficient Conformal Prediction via Cascaded Inference with Expanded Admission. In *International Conference on Learning Representations*.

Gendler, A.; Weng, T.-W.; Daniel, L.; and Romano, Y. 2021. Adversarially robust conformal prediction. In *International Conference on Learning Representations*.

Ghosh, S.; Shi, Y.; Belkhouja, T.; Yan, Y.; Doppa, J.; and Jones, B. 2023. Probabilistically robust conformal prediction. In *Uncertainty in Artificial Intelligence*, 681–690. PMLR.

Huang, K.; Jin, Y.; Candes, E.; and Leskovec, J. 2024. Uncertainty quantification over graph with conformalized graph neural networks. *Advances in Neural Information Processing Systems*, 36.

Jia, J.; and Benson, A. R. 2020. Residual correlation in graph neural network regression. In *Proceedings of the 26th ACM SIGKDD international conference on knowledge discovery & data mining*, 588–598.

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

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

Kish, L.; and Frankel, M. R. 1970. Balanced repeated replications for standard errors. *Journal of the American Statistical Association*, 65(331): 1071–1094.

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

Liu, M.; Ding, L.; Yu, D.; Liu, W.; Kong, L.; and Jiang, B. 2022. Conformalized fairness via quantile regression. *Advances in Neural Information Processing Systems*, 35: 11561–11572.

Lu, C.; Lemay, A.; Chang, K.; Höbel, K.; and Kalpathy-Cramer, J. 2022. Fair conformal predictors for applications in medical imaging. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 36, 12008–12016.

Luo, R.; and Colombo, N. 2025. Conformal load prediction with transductive graph autoencoders. *Machine Learning*, 114(3): 1–22.

Luo, R.; and Zhou, Z. 2024. Conformal Thresholded Intervals for Efficient Regression. *arXiv preprint arXiv:2407.14495*.

Melki, P.; Bombrun, L.; Diallo, B.; Dias, J.; and Da Costa, J.-P. 2023. Group-Conditional Conformal Prediction via Quantile Regression Calibration for Crop and Weed Classification. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 614–623.

Patel, Y. P.; Rayan, S.; and Tewari, A. 2024. Conformal contextual robust optimization. In *International Conference on Artificial Intelligence and Statistics*, 2485–2493. PMLR.

Plassier, V.; Fishkov, A.; Panov, M.; and Moulines, E. 2024. Conditionally valid Probabilistic Conformal Prediction. *arXiv preprint arXiv:2407.01794*.

Prinster, D.; Stanton, S. D.; Liu, A.; and Saria, S. 2024. Conformal Validity Guarantees Exist for Any Data Distribution (and How to Find Them). In *Forty-first International Conference on Machine Learning*.

Redmond, M. 2009. Communities and Crime. UCI Machine Learning Repository. DOI: https://doi.org/10.24432/C53W3X.

Romano, Y.; Barber, R. F.; Sabatti, C.; and Candès, E. 2020. With malice toward none: Assessing uncertainty via equalized coverage. *Harvard Data Science Review*, 2(2): 4.

Romano, Y.; Patterson, E.; and Candes, E. 2019. Conformalized quantile regression. *Advances in neural information processing systems*, 32.

Sadana, U.; Chenreddy, A.; Delage, E.; Forel, A.; Frejinger, E.; and Vidal, T. 2024. A survey of contextual optimization methods for decision-making under uncertainty. *European Journal of Operational Research*.

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

Sesia, M.; and Romano, Y. 2021. Conformal prediction using conditional histograms. *Advances in Neural Information Processing Systems*, 34: 6304–6315.

Shao, J.; and Wu, C. 1992. Asymptotic properties of the balanced repeated replication method for sample quantiles. *The Annals of Statistics*, 1571–1593.

Stanton, S.; Maddox, W.; and Wilson, A. G. 2023. Bayesian optimization with conformal prediction sets. In *International Conference on Artificial Intelligence and Statistics*, 959–986. PMLR.

Sun, C.; Liu, L.; and Li, X. 2023. Predict-then-calibrate: A new perspective of robust contextual lp. *Advances in Neural Information Processing Systems*, 36: 17713–17741.

Tibshirani, R. J.; Foygel Barber, R.; Candes, E.; and Ramdas, A. 2019. Conformal prediction under covariate shift. *Advances in neural information processing systems*, 32.

Vovk, V.; Gammerman, A.; and Shafer, G. 2005. *Algorithmic learning in a random world*, volume 29. Springer.

Wang, F.; Cheng, L.; Guo, R.; Liu, K.; and Yu, P. S. 2024. Equal opportunity of coverage in fair regression. *Advances in Neural Information Processing Systems*, 36.

Zargarbashi, S. H.; Antonelli, S.; and Bojchevski, A. 2023. Conformal prediction sets for graph neural networks. In *International Conference on Machine Learning*, 12292–12318. PMLR.

Zhao, T.; Kang, J.; and Cheng, L. 2024. Conformalized Link Prediction on Graph Neural Networks. *arXiv preprint arXiv:2406.18763*.

Zhao, Y.; Hoxha, B.; Fainekos, G.; Deshmukh, J. V.; and Lindemann, L. 2024. Robust conformal prediction for stl runtime verification under distribution shift. In *2024 ACM/IEEE 15th International Conference on Cyber-Physical Systems (ICCPS)*, 169–179. IEEE.