Benchmarking Graph Conformal Prediction: Empirical Analysis, Scalability, and Theoretical Insights

Anonymous authors
Paper under double-blind review

Abstract

Conformal prediction has become increasingly popular for quantifying the uncertainty associated with machine learning models. Recent work in graph uncertainty quantification has built upon this approach for conformal graph prediction. The nascent nature of these explorations has led to conflicting choices for implementations, baselines, and method evaluation. In this work, we analyze the design choices made in the literature and discuss the tradeoffs associated with existing methods. Building on the existing implementations for existing methods, we introduce techniques to scale existing methods to large-scale graph datasets without sacrificing performance. Our theoretical and empirical results justify our recommendations for future scholarship in graph conformal prediction.

1 Introduction

Modern machine learning models trained on losses based on point predictions are prone to be overconfident in their predictions (Guo et al., 2017). The Conformal Prediction (CP) framework (Vovk et al., 2005) provides a mechanism for generating statistically sound post hoc prediction sets (or intervals, in case of continuous outcomes) with coverage guarantees under mild assumptions. The usual assumption made in CP is that data are exchangeable, i.e., the joint distribution of the data is invariant to permutations of the data points. CP's guarantees are distribution-free and can be added post hoc to arbitrary black-box predictor scores, making them ideal candidates for quantifying uncertainty in complex models, such as neural networks.

Network-structured data such as social networks, transportation networks, and biological networks are ubiquitous in modern data science applications. Graph Neural Networks (GNNs) have been developed to model vector representations of network-structured data and be effective in a variety of tasks such as node classification, link prediction, and graph classification (Hamilton, 2020; Wu et al., 2022). Uncertainty quantification approaches built for independent and identically distributed (iid) data cannot directly be applied to graph data, as the network structure introduces possible dependencies between the data points. However, recent work (Clarkson, 2023; H. Zargarbashi et al., 2023; Huang et al., 2023) has demonstrated that in specific settings, CP can be applied to graph data to generate statistically sound prediction sets for the node classification task. Variations of CP include full CP (Vovk et al., 2005), which has significant computational cost as the score function requires recomputation with replacement for each data point within the calibration set, cross-conformal prediction (Vovk, 2015)/CV+/Jackknife+ (Barber et al., 2021), and split (or inductive) conformal prediction (Papadopoulos et al., 2002; Papadopoulos, 2008). Prior work on graphs has mainly focused on the Split-CP setting due to its computational efficiency and distribution-free guarantees with black-box models, so we also focus on Split-CP for our work.

We undertake a comprehensive analysis of the choices made by existing Split-CP work to understand the trade-offs associated with various design choices. Our study provides a deeper theoretical understanding of some design choices and offers empirical insights and intuition on when and how to evaluate CP for graph data. In addition, we create a Python library that implements different design variations, which can help standardize practices in implementing and evaluating CP for graph data.

2 Conformal Prediction

Conformal prediction is used to quantify the uncertainty of a model by providing prediction sets/intervals with coverage guarantees. We will focus on conformal prediction in the classification setting. Given a calibration dataset $\mathcal{D}_{\text{calib}} = \{(\boldsymbol{x}_i, y_i)\}_{i=1}^{n}$, where $\boldsymbol{x}_i \in \mathcal{X} = \mathbb{R}^d$ and $y_i \in \mathcal{Y} = \{1, \ldots, K\}$, conformal prediction can be used to construct a prediction set C such that

$$\Pr\left[y_{n+1} \in C(\boldsymbol{x}_{n+1})\right] \geq 1 - \alpha$$

where $1 - \alpha \in [0, 1]$ is a user-specified coverage level. The only assumption required for the coverage guarantee is that $\mathcal{D}_{\text{calib}} \cup \{(\boldsymbol{x}_{n+1}, y_{n+1})\}$ is exchangeable. The following theorem provides a general recipe for constructing a prediction set with the coverage guarantee.

Theorem 2.1 ((Vovk et al., 2005; Angelopoulos & Bates, 2021)). *Suppose $\{(\boldsymbol{x}_i, y_i)\}_{i=1}^{n+1}$ are exchangeable, $s : \mathcal{X} \times \mathcal{Y} \to \mathbb{R}$ is a score function measuring the non-conformity of $(\boldsymbol{x}, y)$, with higher scores indicating lower conformity, and a target $\alpha \in [0, 1]$. Let $\hat{q}(\alpha) = \text{Quantile}\left(\frac{\lceil (n+1)(1-\alpha) \rceil}{n}; \{s(\boldsymbol{x}_i, y_i)\}_{i=1}^{n}\right)$. Define $C_\alpha(X) = \{y \in \mathcal{Y} : s(\boldsymbol{x}, y) \leq \hat{q}(\alpha)\}$. Then,*

$$1 - \alpha + \frac{1}{n+1} > \Pr\left[y_{n+1} \in C_\alpha(\boldsymbol{x}_{n+1})\right] \geq 1 - \alpha \tag{1}$$

The function s is the non-conformity score function, and it measures the degree of non-agreement between the input $\boldsymbol{x}$ and the label y, given exchangeability with the calibration data $\mathcal{D}_{\text{calib}}$, i.e., larger scores indicate worse agreement between $\boldsymbol{x}$ and y. While Theorem 2.1 does not place any restrictions on the choice of s, this choice can have a significant impact on the size of the prediction set.

The setup of theorem 2.1 is called Split-CP, as the score function remains fixed for the calibration split. In other versions of CP, the score function is usually more expensive as it maps $\mathcal{X}^k \times \mathcal{Y}^k \to \mathbb{R}$, for some $k \in \mathbb{N}$ which varies between n for full conformal prediction and smaller values for cross-conformal prediction and CV+/Jackknife+. When applying Split-CP, the dataset is partitioned into $\mathcal{D} = \mathcal{D}_{\text{train}} \cup \mathcal{D}_{\text{valid}} \cup \mathcal{D}_{\text{calib}} \cup \mathcal{D}_{\text{test}}$.

2.1 Conformal Prediction in Graphs

The usual tasks of interest in graph data are node classification, link prediction, and graph classification. In this work, we focus on node classification and its extensions to conformal prediction. Consider an attributed homogeneous graph $\mathcal{G} = (\mathcal{V}, \mathcal{E}, \boldsymbol{X})$, where $\mathcal{V}$ is the set of nodes, $\mathcal{E}$ is the set of edges, and $\boldsymbol{X}$ is the set of node attributes. Let $\boldsymbol{A}$ denote the adjacency matrix for the graph. Further, let $\mathcal{Y} = \{1, \ldots, K\}$ denote the set of class labels associated with the nodes. For $v \in \mathcal{V}$, $\boldsymbol{x}_v \in \mathcal{X} = \mathbb{R}^d$ denotes its features and $y_v \in \mathcal{Y}$ denotes its true class label. The task of node classification is to learn a model that predicts the label for each node given node features and the adjacency matrix, i.e. $(\boldsymbol{X}, \boldsymbol{A}, v) \mapsto y_v$. Corresponding to the CP partitions, we denote the nodes in the training set as $\mathcal{V}_{\text{train}}$, validation set as $\mathcal{V}_{\text{valid}}$, calibration set as $\mathcal{V}_{\text{calib}}$, and test set as $\mathcal{V}_{\text{test}}$. We denote $\mathcal{V}_{\text{dev}} = \mathcal{V}_{\text{train}} \cup \mathcal{V}_{\text{valid}}$ as the development set of the base model (non-conformalized). Note that labels are available only for nodes in $\mathcal{V}_{\text{train}}$, $\mathcal{V}_{\text{valid}}$, and $\mathcal{V}_{\text{calib}}$, and must be predicted for nodes in $\mathcal{V}_{\text{test}}$. The model cycle will involve four phases, viz. training, validation, calibration, and testing. Next, we discuss the different settings for node classification in graphs and the applicability of conformal prediction.

Transductive setting In this setting, the model has access to the fixed graph $\mathcal{G}$ during the model cycle. The nodes used in the model cycle are split into $\mathcal{V}_{\text{test}}$, $\mathcal{V}_{\text{valid}}$, and $\mathcal{V}_{\text{calib}} \cup \mathcal{V}_{\text{test}}$. The specific $\mathcal{V}_{\text{calib}}$ and $\mathcal{V}_{\text{test}}$ are randomly sampled from $\mathcal{V}_{\text{calib}} \cup \mathcal{V}_{\text{test}}$. This is the setting considered by H. Zargarbashi et al. (2023) and Huang et al. (2023). Note that the labels for $\mathcal{V}_{\text{calib}}$ are not available for training and validation of the base model, though all the neighborhood information of $\mathcal{G}$ and the features $\boldsymbol{x}_v$ and labels y_v, for $v \in \mathcal{V}_{\text{dev}}$ are available. During the calibration phase, the $(\boldsymbol{x}_v, y_v)$ for $v \in \mathcal{V}_{\text{calib}}$ and all the neighborhood information are used to compute the non-conformity scores. This split ensures that the base model cannot distinguish between the calibration and test nodes, and hence exchangeability holds for $v \in \mathcal{V}_{\text{calib}} \cup \mathcal{V}_{\text{test}}$. In line with previous work, we focus on the transductive setting. The following theorem states that in the transductive

Table 1: Summary statistics for datasets. Predefined splits from original source were noted.

Dataset	Nodes	Edges	Classes	Features	# Train	# Valid	# Test
Amazon_Computers	13,752	491,722	10	767	-	-	-
Cora	19,793	126,842	70	8,710	-	-	-
Coauthor_CS	18,333	163,788	15	6,805	-	-	-
Flickr	89,250	899,756	7	500	44,625	22,312	22,313
ogbn-arxiv	169,343	1,166,243	40	128	90,941	29,799	48,603

setting, a score model trained on the calibration set will generate scores exchangeable with the test set, thus allowing the use of conformal prediction.

Theorem 2.2 ((H. Zargarbashi et al., 2023; Huang et al., 2023)). *Let $\mathcal{G} = (\mathcal{V}, \mathcal{E}, \boldsymbol{X})$ be an attributed graph, and $\mathcal{V}_{calib} \cup \mathcal{V}_{test}$ be exchangeable. Let $F : \mathcal{X}^{|V|} \to \Delta^{|V|\times K}$ be any permutation equivariant model on the graph (e.g., GNNs). Define $F(G) = \Pi \in \Delta^{|V|\times K}$ be the output probability matrix for a model trained on only $\mathcal{V}_d$. Then any score function $s(v, y) = s(\Pi_v, y, \mathcal{G})$ is exchangeable for all $v \in \mathcal{V}_{calib} \cup \mathcal{V}_{test}$.*

The intuition for this theorem is that if the GNN does not depend on the order of the nodes in the graph then the outputs of the GNN will also be exchangeable. This holds for most standard GNNs. The formal proof for this theorem is available in (H. Zargarbashi et al., 2023; Huang et al., 2023). This theorem paves the way for using conformal prediction for transductive node classification.

For the following sections, we will assume that the base model $\hat{\pi} : \mathcal{X} \to \Delta_{\mathcal{Y}}$, where $\Delta_{\mathcal{Y}}$ is the probability simplex over the elements of $\mathcal{Y}$, is learned using the training and validation sets $\mathcal{D}_{\text{train}} \cup \mathcal{D}_{\text{valid}}$. The calibration set $\mathcal{D}_{\text{calib}}$ is used to determine the $\hat{q}(\alpha)$ from Theorem 2.1 and the test set $\mathcal{D}_{\text{test}}$ is the set for which we want to construct our prediction sets. In general, the scores need not lie over a simplex; they can be in $\mathbb{R}^K$. However, this greatly simplifies the exposition for the following sections and is the standard practice in prior work.

3 Empirical Analysis and Insights

Datasets and Methods Table 1 contains a representative set of datasets of varying sizes (i.e., number of nodes/edges) and the number of classes evaluated in this section. The Appendix contains the list of all the datasets used in this study. For these datasets, we used the version provided by the Deep Graph Library (Wang et al., 2019) and Open Graph Benchmark (Hu et al., 2020).

We compare CP methods, including TPS (Sadinle et al., 2019), APS (Romano et al., 2020), and RAPS (Angelopoulos et al., 2021) and describe the setups appropriate for their use in graph settings. For graph-focused CP methods, we include CFGNN (Huang et al., 2023), DAPS (H. Zargarbashi et al., 2023), and NAPS (Clarkson, 2023). To adapt NAPS to the transductive setting, we build on H. Zargarbashi et al. (2023) by using APS as a baseline score and compute a weighted quantile of scores from the set of k-hop nodes - of a given test node - that intersect the calibration nodes. (Further discussion in Appendix).

Metrics For evaluation, we used the following standard metrics (Shafer & Vovk, 2008): (i) **Coverage** the proportion of test instances for which the true label is contained in the prediction set. (ii) **Efficiency (or Prediction Efficiency)** the average size of the prediction set. and to measure adaptability, we used (iii) **Label (or Class) Stratified Coverage** (Sadinle et al., 2019) the mean of coverage for each class.

Dataset Splits and Training There are several methods of partitioning $\mathcal{D}$ into $\mathcal{D}_{\text{train}}$, $\mathcal{D}_{\text{valid}}$, $\mathcal{D}_{\text{calib}}$, and $\mathcal{D}_{\text{test}}$. Two methods used in existing works on graph conformal prediction for node classification are (1) **Full-Split (FS) Partitioning** (Huang et al., 2023) The data is split such that each subset of the partition adheres to a size constraint based on $\mathcal{D}$. For example, in CFGNN (Huang et al., 2023) the authors split the datasets in their experiments randomly, satisfying a 20%/10%/35%/35% split into $\mathcal{D}_{\text{train}}/\mathcal{D}_{\text{valid}}/\mathcal{D}_{\text{calib}}/\mathcal{D}_{\text{test}}$. Note that the overall percentage of data for which we do provide labels (in either the development or calibration

set) is a large proportion (65%) of the full dataset. This splitting scheme is ideal for non-conformity score models with numerous trainable (or tunable) parameters, as it allows for the use of a large amount of data for training the calibration score model. We explore the following splitting schemes under FS partitioning: $(\mathcal{D}_{\text{train}}, \mathcal{D}_{\text{valid}}, \mathcal{D}_{\text{calib}}, \mathcal{D}_{\text{test}}) = (0.2, 0.1, 0.35, 0.35)$, $(0.2, 0.2, 0.3, 0.3)$, $(0.3, 0.1, 0.3, 0.3)$, and $(0.3, 0.2, 0.25, 0.25)$. (2) **Label-Count (LC) Sample Partitioning** (H. Zargarbashi et al., 2023) The data is split to ensure an equal number of samples for each class label is present in $\mathcal{D}_{\text{train}}$, $\mathcal{D}_{\text{valid}}$, and $\mathcal{D}_{\text{calib}}$. The remaining nodes are $\mathcal{D}_{\text{test}}$. Such a setting is common in settings where only a small proportion of training/labeled nodes are available (e.g., semi-supervised learning). Intuitively, this setting is ideal for methods that do not have many parameters to train. We explore setting the number of samples per class to 10, 20, 40, and 80. Note that we assign nodes of each class sequentially, so it is feasible in this setup to have some classes having no representative samples in some data subset. If the dataset has **predefined splits** (e.g., Flickr, ogbn-arxiv), in addition to the described splitting rules, we ensure that $\mathcal{D}_{\text{train}}$ and $\mathcal{D}_{\text{valid}}$ come solely from the training and validation splits, while $\mathcal{D}_{\text{calib}} \cup \mathcal{D}_{\text{test}}$ come from the test split.

3.1. On TPS and adaptability Threshold Prediction Sets (TPS) (Sadinle et al., 2019) is a simple technique for generating conformal prediction sets. The score function $s(\boldsymbol{x}, y) = 1 - \hat{\pi}(\boldsymbol{x})_y$ directly maps the probability from the base model for the correct class into a non-conformity score. The score is higher if the model has a lower probability assigned to the correct class, indicating the label conforms less with the model. A $1-\alpha$ (approximate) quantile creates a probability inclusion threshold for this score over the calibration set, ensures coverage, and can be shown to generate prediction sets with the best-expected efficiency (Sadinle et al., 2019). However, the TPS score has been known to undercover hard examples and overcover easy ones (Angelopoulos et al., 2021; H. Zargarbashi et al., 2023) to achieve this efficiency. By overcovering easy examples, TPS can still maintain the overall coverage guarantee without having to correctly account for coverage over harder examples.

We note that this discrepancy is claimed to occur as the TPS scores are not 'adaptive', and consider only one dimension of the score for each calibration sample. However, Sadinle et al. (2019) also proposed a classwise control version of TPS. Instead of defining a single threshold for all classes, they separately compute the threshold for each class for a corresponding α. Thus, we define classwise quantile thresholds as

$$\hat{q}(\alpha, y_j) = \text{Quantile}\left(\frac{\lceil (n+1)(1-\alpha) \rceil}{n}; \{s(\boldsymbol{x}_i, y_i)\, i = 1, \ldots, n, y_i = y_j\}\right)$$

and the corresponding prediction sets as

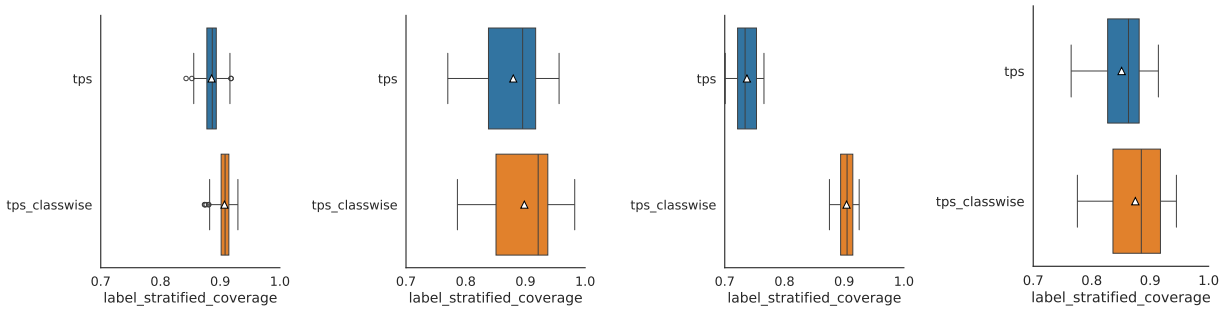
$$C_{\text{TPS}}(\boldsymbol{x}) = \{y \in \mathcal{Y} : s(\boldsymbol{x}, y) \leq \hat{q}(\alpha, y)\}$$

Note that this version would provide coverage for each class label, making it more 'adaptive'. The version defined by Sadinle et al. (2019) allows controlling α_y for each class, though we set $\alpha_y = \alpha$ for class-adaptability. The trade-off with the adaptive version is we have fewer calibration samples for each quantile threshold dimension, which may lead to higher variance in the distribution of coverage (Vovk, 2012). We call this variation of TPS, TPS-Classwise, and consider it in our baselines for comparison. From Figure 1, we see that using classwise TPS successfully provides label stratified coverage in both the FS and LC split settings.

We further observe that the dataset's label distribution impacts the difference in adaptability between TPS and TPS-Classwise. In Figure 1, we find an increase in the adaptability of TPS when using LC split instead of FS split for ogbn-arxiv compared to Amazon_Computers. The shift in performance is exacerbated due to many classes having little representation, as seen in Figure 2. The difference in adapatability, can be seen between other non-conformity scores in a more extreme setting in Figure D for the ogbn-products dataset, which has several classes with minimal representation.

Thus, TPS-Classwise is a good candidate for an adaptive baseline in place of TPS, at least for the datasets we studied.

3.2. On APS and randomization The most popular baseline in work on graph conformal prediction is Adaptive Prediction Sets (APS). Romano et al. (2020) introduce APS by defining an optimal prediction set construction mechanism by first considering an oracle-specified probability and then generalizing it to predictive



(a) Amazon_Computers (b) ogbn-arxiv

Figure 1: We set the target coverage rate $\alpha = 0.1$. The boxplots present the Label Stratified Coverage for Amazon_Computers (a) and ogbn-arxiv (b) for both the FS split (left) and LC split (right). We want the means (white triangle) to be around $1 - \alpha = 0.9$. For Labeled Stratified Coverage, TPS-Classwise can provide comparable performance to TPS.

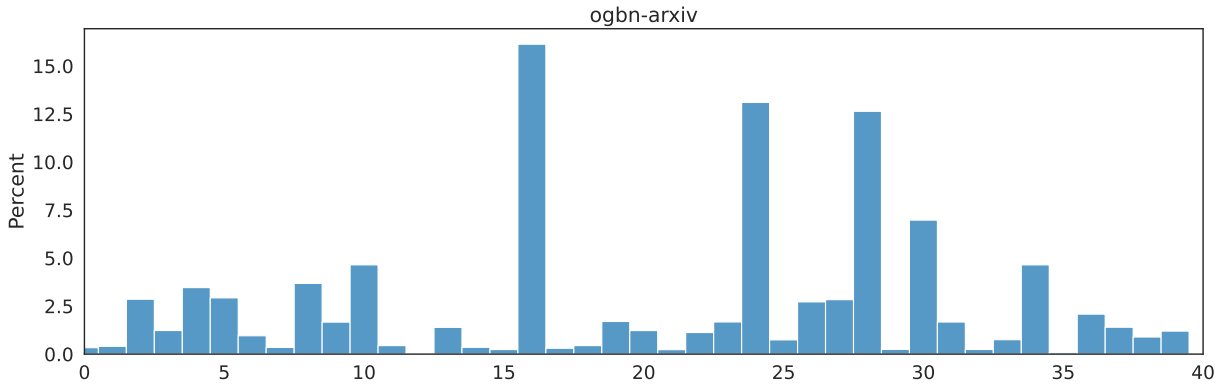


Figure 2: Label Distribution for ogbn-arxiv

probabilities. Consider a probability prediction function that estimates $\widehat{\Pr}[Y = y|X_{test} = \boldsymbol{x}] = \hat{\pi}_y(\boldsymbol{x})$ for each $y \in \mathcal{Y} = \{1, \ldots, K\}$. Assume that $\hat{\pi}$ are all distinct – for ease of defining rank. Suppose the rank of the true class amongst the sorted $\hat{\pi}$ is r_y, i.e., $\sum_{i=1}^{K} \mathbf{1}[\hat{\pi}_i(\boldsymbol{x}) \geq \hat{\pi}_y] = r_y$. Following the APS definitions, considering a uniform random variable $u \sim U(0, 1)$, it is possible to derive a randomized non-conformity score (see Appendix A, for derivations),

$$A(\boldsymbol{x}, y, u; \hat{\pi}) = \left[\sum_{i=1}^{r_y} \hat{\pi}_{(i)}(\boldsymbol{x})\right] - u\hat{\pi}_y \tag{2}$$

Instead, if a deterministic approach is used to define the conformal score instead (i.e., the randomized term is not included), then we could just add the probabilities until the true class is included:

$$\tilde{A}(\boldsymbol{x}, y; \hat{\pi}) = \left[\sum_{i=1}^{r_y} \hat{\pi}_{(i)}(\boldsymbol{x})\right] \tag{3}$$

The version of APS without randomization still provides the same conditional coverage guarantees and has a simpler exposition as the prediction sets are constructed by greedily including the classes until the true label is included. Thus, this version is implemented in the popular monographs on conformal prediction by Angelopoulos & Bates (2021). However, the lack of randomization may sacrifice (prediction) efficiency.

This modification affects the quantile threshold computation during the calibration phase and the prediction set construction during the test phase.

Formally, let $A(\boldsymbol{x}, y)$ be any non-conformity score function, and let q_A be the CP threshold, i.e. $1-\alpha \leq \Pr[A(\boldsymbol{x}_{n+1}, y_{n+1}) < q_A] < 1-\alpha+\frac{1}{n+1}$. Observe that $\Pr[A(\boldsymbol{x}_{n+1}, y_{n+1}) < q_A] \iff \Pr[y_{n+1} \in C_\alpha(\boldsymbol{x}_{n+1})]$. Let α_c^A be the significance level of incorrect labels being in the prediction set, i.e. $1-\alpha_c^A \leq \Pr[A(\boldsymbol{x}_{n+1}, y'_{n+1}) < q_A] < 1-\alpha_c^A+\frac{1}{n+1}$ for $y'_{n+1} \in \{1, 2, \ldots, K\} \setminus \{y_{n+1}\}$. Let $\tilde{A}$ be another score function and define $\alpha_c^{\tilde{A}}$ similarly. To determine when A is more efficient then $\tilde{A}$, we can use Theorem 3.1.

Theorem 3.1. *If $\alpha_c^A - \alpha_c^{\tilde{A}} \geq \frac{2}{n+1}$ then score function A produces a more efficient prediction set than $\tilde{A}$. Formally, $\mathbb{E}\left[|C_{\tilde{A}}(\boldsymbol{x}_{n+1})| - |C_A(\boldsymbol{x}_{n+1})|\right] \geq 0$*

Proof. Consider the case with only two class labels, i.e. $K = 2$.

Then, we have

$$\begin{aligned}
\mathbb{E}\left[|C_A^{n+1}|\right] &= \mathbb{E}\left[\sum_{i=1,2} \mathbf{1}[i \in C_A^{n+1}]\right] \\
&= \mathbb{E}\left[\mathbf{1}[y_{n+1} \in C_A^{n+1}]\right] + \mathbb{E}\left[\mathbf{1}[y'_{n+1} \in C_A^{n+1}]\right] \\
&= \Pr[y_{n+1} \in C_A^{n+1}] + \Pr[y'_{n+1} \in C_A^{n+1}] \\
&\leq 1-\alpha+1-\alpha_c^A+\frac{2}{n+1} && \text{(Exchangeability, Theorem 2.1)}
\end{aligned}$$

From a similar argument, we can show that

$$\mathbb{E}\left[|C_{\tilde{A}}^{n+1}|\right] \geq 1-\alpha+1-\alpha_c^{\tilde{A}}$$

Thus,

$$\begin{aligned}
\mathbb{E}\left[|C_{\tilde{A}}^{n+1}| - |C_A^{n+1}|\right] &\geq 1-\alpha+1-\alpha_c^{\tilde{A}} - \left(1-\alpha+1-\alpha_c^A+\frac{2}{n+1}\right) \\
&= \alpha_c^A - \alpha_c^{\tilde{A}} - \frac{2}{n+1}
\end{aligned}$$

which is equivalent to our assumption, and this completes the proof.

For K classes,

$$\begin{aligned}
\mathbb{E}\left[\left|C_A^{n+1}\right|\right] &= \Pr[y_i \in C_A^{n+1}] + \sum_{y'_i} \Pr[y'_i \in C_A^{n+1}] \\
&= \Pr[y_i \in C_A^{n+1}] + (K-1)\Pr[y'_i \in C_A^{n+1}]
\end{aligned}$$

Thus,

$$\begin{aligned}
\mathbb{E}\left[\left|C_A^{n+1}\right|\right] &\leq 1-\alpha+\frac{1}{n+1} + (K-1)\left(1-\alpha_c^A+\frac{1}{n+1}\right) \\
&= 1-\alpha+(K-1)\left(1-\alpha_c^A\right)+\frac{K}{n+1}
\end{aligned}$$

and

$$\mathbb{E}\left[\left|C_A^{n+1}\right|\right] \geq 1-\alpha+(K-1)\left(1-\alpha_c^A\right)$$

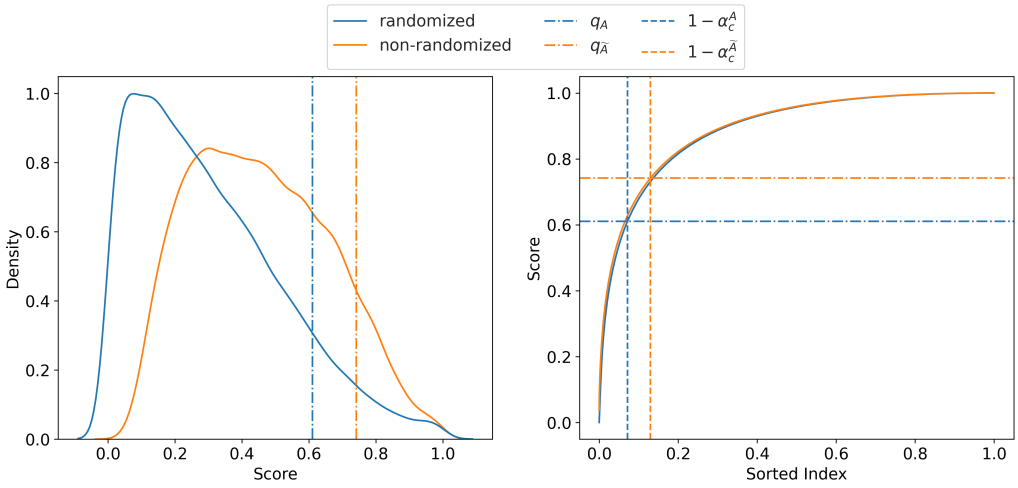


Figure 3: Scores for the Cora dataset using the randomized and non-randomized versions of APS. In the left plot shows, the vertical lines show the shift in the quantiles for A and $\tilde{A}$ considering the correct class with a 0.9 coverage. Right plot shows the shift in $1-\alpha_c$ for A and $\tilde{A}$ using scores for the incorrect classes. However, the shift is not as significant as the shift for correct labels as seen in the left plot. This demonstrates the scenario that is a condition for Theorem 3.1.

similar bounds can be derived for $\mathbb{E}\left[\left|C_{\tilde{A}}^{n+1}\right|\right]$. Thus,

$$\begin{aligned}
\mathbb{E}\left[|C_{\tilde{A}}^{n+1}| - |C_A^{n+1}|\right] &\geq (K-1)\left(\alpha_c^A - \alpha_c^{\tilde{A}}\right) - \frac{K}{n+1} \\
&\geq (K-1)\left(\alpha_c^A - \alpha_c^{\tilde{A}} - \frac{K}{(K-1)(n+1)}\right) \\
&> (K-1)\left(\alpha_c^A - \alpha_c^{\tilde{A}} - \frac{2}{n+1}\right) \geq 0, \qquad \text{Since } \alpha_c^A - \alpha_c^{\tilde{A}} \geq \frac{2}{n+1}
\end{aligned}$$

completing the proof for the general case. □

Applying Theorem 3.1 to randomized APS (A) and deterministic APS ($\tilde{A}$), intuitively, as each score in A gets shifted by a small $u\pi$ term to the left, q_A would be lower than $q_{\tilde{A}}$. Thus, the significance levels we would search for in the complementary scores $1-\alpha_c^A$ would be less than $1-\alpha_c^{\tilde{A}}$. $1-\alpha_c^A < 1-\alpha_c^{\tilde{A}} \implies \alpha_c^A - \alpha_c^{\tilde{A}} > 0$. If the shift is sufficiently large, then the randomized prediction set will be more efficient than the non-randomized one. In Figure 3, we show what this looks like for a practical example over the Cora dataset. In Figure 3 (right), the normalized sorted index at which the lower threshold q_A is reached when considering the incorrect classes is lower, i.e., $1-\alpha_c^A$ is lower, and hence α_c^A is higher. As a part of the proof, we show dependences on $\frac{1}{n+1}$ and $(K-1)$, which indicates that the improvements would be more pronounced for larger $\mathcal{D}_{\text{calib}}$ and a larger number of classes.

Figure 4 provides box plots that compare the efficiency of randomized and non-randomized versions of APS across different datasets. We observe that for each split type,the randomized version consistently provides a more efficient prediction set. This effect is most pronounced for a dataset with many potential classes in the FS split, which matches with the intuition from Theorem 3.1 described above. Overall, the empirical results show that the effect of randomized APS is more pronounced for larger values of K.

3.3. Conformalized GNN CFGNN (Huang et al., 2023) is a recent GNN-based approach for conformal prediction. The underlying observation of this approach is that inefficiencies are correlated between nodes having similar neighborhood topology. With this intuition, during the calibration stage, a second GNN is trained to correct the scores from the base model to optimize the efficiency of the prediction sets. This

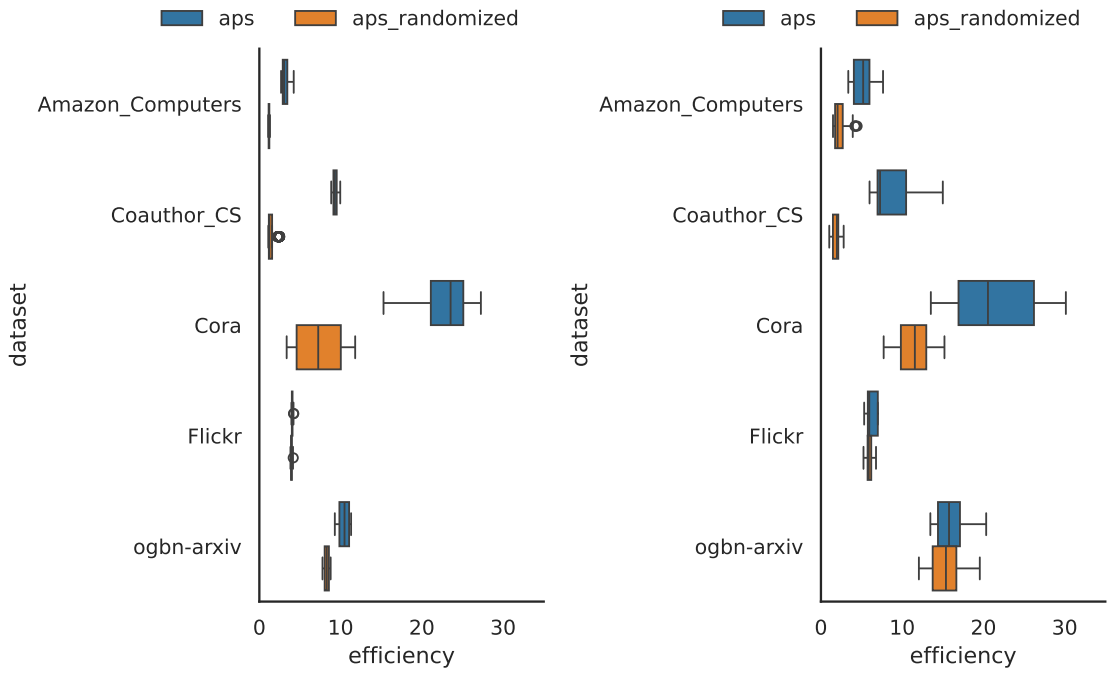


Figure 4: We set the target coverage rate $\alpha = 0.1$. Box plots depicting the efficiencies of APS and Randomized APS across different datasets and multiple runs in both the FS split (left) and LC split (right). Using randomization (the lower box plot for each dataset) consistently improves over the non-randomized version as the efficiencies are distributed around smaller values.

is feasible as all the steps of the conformal prediction framework (i.e., non-conformity score computation, quantile computation, thresholding) can be expressed as differentiable operations (Stutz et al., 2021). Thus, the second GNN can be trained using an efficiency-based loss function, which Huang et al. (2023) proposes. More details on CFGNN are available in the Appendix.

Impact of Inefficiency Loss: The choice of conformal loss during calibration and test plays a vital role in determining the overall performance of GNN-based conformal prediction. To illustrate, we replicate an experiment by Huang et al. (2023) who use TPS for its inefficiency loss during the calibration stage and non-randomized APS when constructing the final prediction sets and show a significant improvement in efficiency over the baseline (Figure 5 right). However, if instead randomized APS loss is used (Figure 5 left), we observe that the baseline is competitive across various coverage thresholds. It is worth noting that CFGNN appears robust to this choice, although the gains in efficiency are not as dramatic in the randomized setting. We also note that the confidence bars are narrower in the randomized setting. Similar results were observed on other datasets (see Appendix D).

Based on these insights, we implemented an improved version of CFGNN, which uses APS with randomization for *both* training and evaluation, labeled as 'cfgnn_aps.' The original implementation is labeled as 'cfgnn_orig'. Our library implementation of CFGNN allows for either TPS or APS to be used for training and evaluation and is extensible to other conformal prediction methods.

We compare the efficiency of 'cfgnn_aps,' 'cfgnn_orig,' and 'aps_randomized' in Figure 6. Note that 'cfgnn_aps' improves upon or matches the efficiency of 'cfgnn_orig' and can even improve upon 'aps_randomized.' We also benchmark CFGNN using LC splits in Figure 6 (originally CFGNN was evaluated using FS splits) One observes that CFGNN learned under this setting is quite brittle. One potential reason is that the LC split style doesn't provide enough data to train the second GNN. Adapting CFGNN to work in this setting is a potential area for future work.

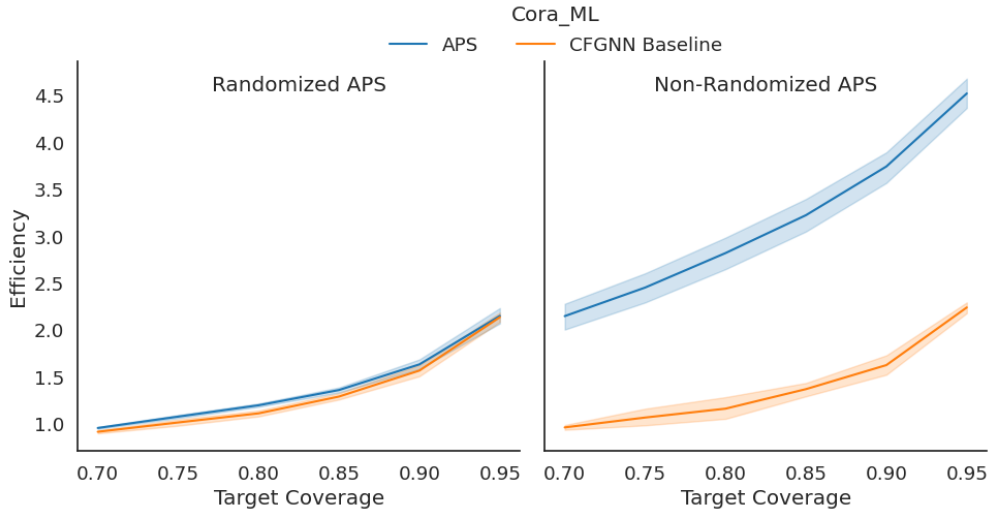


Figure 5: The plot on the right replicates an experiment (Huang et al., 2023) plotting efficiency over various coverage rates for the Cora_ML dataset (a subset of the Cora dataset) for both CFGNN and a baseline model. The plot on the left uses APS with randomization when constructing the final prediction sets. These plots illustrate the benefits of using randomization on baseline performance.

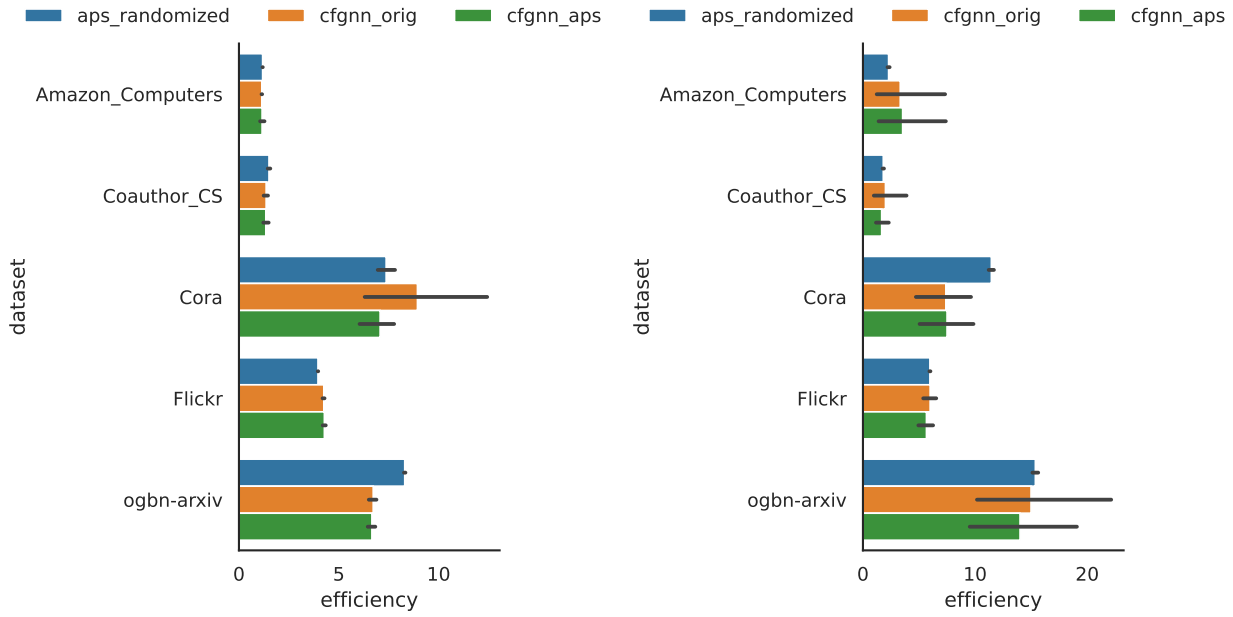


Figure 6: Bar charts denoting efficiency for 'cfgnn_aps', 'cfgnn_orig', and 'aps_randomized' for the FS split (left) and LC split (right) at $\alpha = 0.1$. We see that 'cfgnn_aps' improves or matches efficiency in most cases.

Scaling CFGNN In the original CFGNN implementation, full batch training was used. While this approach has merits, it also poses challenges, particularly when dealing with larger graphs. The need for a more scalable solution was evident, leading us to the modifications we have implemented. We implemented a batched version of CFGNN to ensure it can be used for larger graphs (e.g., ogbn-arxiv). To scale CFGNN for even larger graphs, we cache the outputs from the base model to be treated as features for the CFGNN training rather than having to sample neighbors for both the base model and CFGNN, significantly speeding up the computation in both training and evaluation for CFGNN.

Table 2: Impact of different CFGNN implementations starting from the baseline, then batching, and then both caching and batching combined. Used the **best** CFGNN architecture (w.r.t. validation efficiency) for each dataset. We run 5 trials for each setup and report a 95% confidence interval.

(a) Impact on Runtime

dataset(↓) / method(→)	original	batching	batching+caching
Amazon_Computers	664.98 ± 10.90	205.29 ± 3.96	73.85 ± 1.56
Cora	1378.01 ± 13.35	203.61 ± 2.52	73.60 ± 1.53
Coauthor_CS	638.56 ± 6.14	88.49 ± 1.14	31.67 ± 0.53
Flickr	868.87 ± 9.98	567.39 ± 6.50	56.71 ± 1.30
ogbn-arxiv	410.91 ± 8.29	373.19 ± 3.64	111.38 ± 1.95

(b) Impact on Efficiency

dataset(↓) / method(→)	baseline	batching	batching+caching
Amazon_Computers	1.29 ± 0.05	1.15 ± 0.01	1.14 ± 0.01
Cora	6.81 ± 1.07	8.34 ± 1.12	7.96 ± 0.59
Coauthor_CS	1.10 ± 0.01	1.15 ± 0.01	1.14 ± 0.01
Flickr	4.23 ± 0.07	4.23 ± 0.04	4.24 ± 0.03
ogbn-arxiv	7.07 ± 0.05	7.28 ± 0.06	6.91 ± 0.01

We compare three variations of the CFGNN implementation to demonstrate the impact of batching and caching on the runtime. Across all comparisons, we use the FS split, with 20%/20% assigned to train/valid sets and 35% to the calibration dataset. We use the best base GNN (w.r.t accuracy) and best CFGNN (w.r.t efficiency) architecture using hyperparameter tuning.

The baseline implementation follows the setup by Huang et al. (2023), where the CFGNN is trained with full batch gradient descent for 1000 epochs. Our improved implementation, which uses mini-batch gradient descent, achieves comparable efficiency in only 50 epochs without any batch size tuning (we set the batch size to 64 for consistent comparison) as shown in Table 2b. Finally, we add caching of the output probabilities from the base GNN to the batched implementation, which further reduces the runtime. Table 2 compares the batching and the combined batching + caching improvements. We note that our implementation can achieve improvements ranging from 3.69× (ogbn-arxiv) to 20.16× (Coauthor_CS) in runtime over the baseline implementation. [1]

3.4. Overall comparison of Graph Conformal Prediction We analyze the efficiency of the methods across different datasets. TPS is consistently the most efficient method for each dataset, regardless of the train/validation/calibration split. However, this often comes at a cost to adaptability, as shown in Figure 7 for the Flickr dataset. On the other hand, as previously noted, TPS-Classwise provides label adaptability. This particular dataset comes at the cost of efficiency, though this is not the case with all datasets. A new method we propose is Diffused Classwise TPS (DTPS), where we apply the diffusion operator from H. Zargarbashi et al. (2023) on top of the 'adaptive' TPS-Classwise, discussed earlier. As with other datasets, DTPS also provides label-stratified coverage for Flickr. More discussions on DTPS and Diffusion Adaptive Sets (DAPS) (H. Zargarbashi et al., 2023) can be found in the Appendix. A balance between efficiency and label-stratified coverage is achieved using the NAPS (Clarkson, 2023) scoring function. Before computing a quantile, NAPS uses neighbor distances to assign weights to non-conformity scores. In the original paper and our implementation, NAPS has a hyperparameter k, which is the maximum distance considered when assigning a weight, and everything beyond that is given zero weight. More details on NAPS and its variations can be found in Appendix B. Appendix D has similar plots for the other datasets.

[1] For each dataset and split type, we provide the configuration files for the best corresponding base GNN and CFGNN architectures in the attached code.

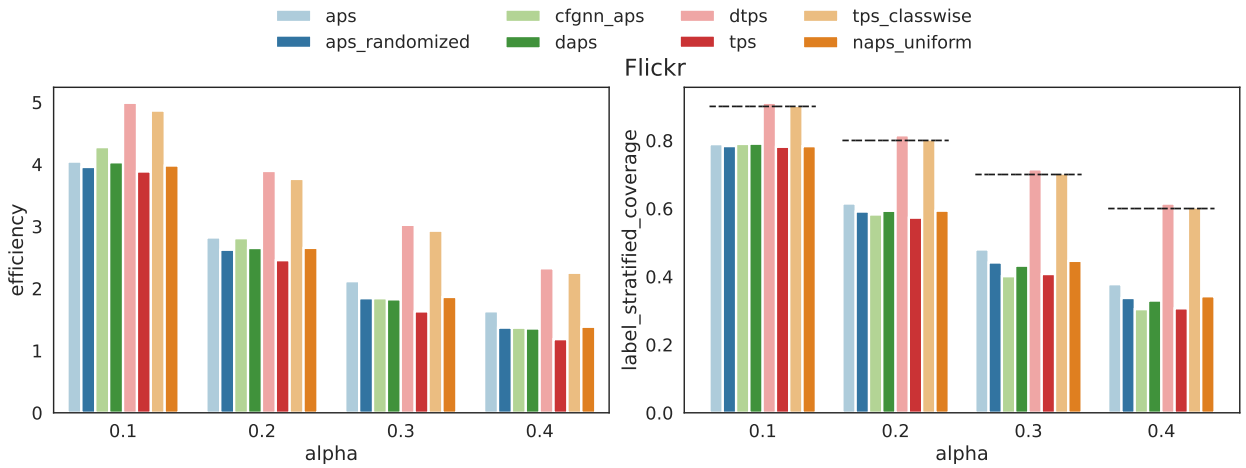


Figure 7: Efficiency (left) and Label Stratified Coverage (right) for all the conformal methods for the Flickr dataset. This uses the FS split style with multiple values for α. The dashed black line indicates the desired label stratified coverage. Other NAPS variants such as exponential and hyperbolic are discussed in the Appendix.

4 Related Works

4.1 Conformal Prediction for Graphs

In this work, we focus on the prominent methods of graph conformal prediction for the node classification task under a transductive setting. These included standard conformal prediction methods like TPS and APS and graph-specific methods like DAPS and CFGNN. There are works that consider other graph-based scenarios and tasks. Within node classification, we can also consider the inductive setting, where nodes/edges arrive in a sequence, thus violating the exchangeability assumption. One such method is Neighborhood Adaptive Prediction Sets (NAPS), which we consider a transductive variation in this work. Beyond node classification on static graphs, there is some work in link prediction (Zhao et al., 2024; Marandon, 2024) as well as with dynamic graphs (Davis et al., 2024). These settings are newly studied or infrequently studied due to additional assumptions being required for meaningful analysis. Thus, we omit these settings from our study and leave them as future work.

4.2 Uncertainty Quantification for Graphs

Beyond conformal prediction, many works propose ways to construct model-agnostic uncertainty estimates for classification tasks (Abdar et al., 2021; Guo et al., 2017; Kull et al., 2019). There also is work in methods specific to GNNs (Hsu et al., 2022; Wang et al., 2021) that leverage network principles such as homophily. However, these methods can fail to provide the desired coverage guarantees compared to conformal prediction-based methods. For an empirical comparison of popular UQ methods for GNN, please reference Huang et al. (2023).

5 Concluding Remarks

We present a comprehensive benchmarking study of conformal prediction for node classification. We provide novel insights related to design choices that impact efficiency, adaptability, and scalability. Along the way, we offer a new theoretical rationale for the importance of randomization and discuss some novel methodological improvements and directions for future work. One future direction pertains to the space of fairness auditing. Several works have dealt with the auditing fairness of ML models through measuring uncertainty in fairness definitions (Ghosh et al., 2021; Maneriker et al., 2023; Yan & Zhang, 2022), but they rely on the assumption

of IID. While conformal prediction works with the notion of *miscoverage*, more relevant notions of error can be considered using the generalized framework of conformal risk control (Angelopoulos et al., 2024).

References

Moloud Abdar, Farhad Pourpanah, Sadiq Hussain, Dana Rezazadegan, Li Liu, Mohammad Ghavamzadeh, Paul Fieguth, Xiaochun Cao, Abbas Khosravi, U. Rajendra Acharya, Vladimir Makarenkov, and Saeid Nahavandi. A review of uncertainty quantification in deep learning: Techniques, applications and challenges. *Inf. Fusion*, 76(C):243–297, December 2021. ISSN 1566-2535. doi: 10.1016/j.inffus.2021.05.008. URL https://doi.org/10.1016/j.inffus.2021.05.008.

Anastasios N Angelopoulos and Stephen Bates. A gentle introduction to conformal prediction and distribution-free uncertainty quantification. *arXiv preprint arXiv:2107.07511*, 2021.

Anastasios N Angelopoulos, Stephen Bates, et al. Conformal prediction: A gentle introduction. *Foundations and Trends® in Machine Learning*, 16(4):494–591, 2023.

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

Anastasios Nikolas Angelopoulos, Stephen Bates, Adam Fisch, Lihua Lei, and Tal Schuster. Conformal risk control. In *The Twelfth International Conference on Learning Representations*, 2024.

Rina Foygel Barber, Emmanuel J Candès, Aaditya Ramdas, and Ryan J Tibshirani. Predictive inference with the jackknife+. *The Annals of Statistics*, 49(1):486–507, 2021.

Rina Foygel Barber, Emmanuel J. Candès, Aaditya Ramdas, and Ryan J. Tibshirani. Conformal prediction beyond exchangeability. *The Annals of Statistics*, 51(2):816 – 845, 2023. doi: 10.1214/23-AOS2276. URL https://doi.org/10.1214/23-AOS2276.

Jase Clarkson. Distribution free prediction sets for node classification. In Andreas Krause, Emma Brunskill, Kyunghyun Cho, Barbara Engelhardt, Sivan Sabato, and Jonathan Scarlett (eds.), *Proceedings of the 40th International Conference on Machine Learning*, volume 202 of *Proceedings of Machine Learning Research*, pp. 6268–6278. PMLR, 23–29 Jul 2023. URL https://proceedings.mlr.press/v202/clarkson23a.html.

Ed Davis, Ian Gallagher, Daniel John Lawson, and Patrick Rubin-Delanchy. Valid conformal prediction for dynamic gnns, 2024. URL https://arxiv.org/abs/2405.19230.

Bishwamittra Ghosh, Debabrota Basu, and Kuldeep S Meel. Justicia: A stochastic sat approach to formally verify fairness. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 35, pp. 7554–7563, 2021.

Chuan Guo, Geoff Pleiss, Yu Sun, and Kilian Q Weinberger. On calibration of modern neural networks. In *International conference on machine learning*, pp. 1321–1330. PMLR, 2017.

Soroush H. Zargarbashi, Simone Antonelli, and Aleksandar Bojchevski. Conformal prediction sets for graph neural networks. In Andreas Krause, Emma Brunskill, Kyunghyun Cho, Barbara Engelhardt, Sivan Sabato, and Jonathan Scarlett (eds.), *Proceedings of the 40th International Conference on Machine Learning*, volume 202 of *Proceedings of Machine Learning Research*, pp. 12292–12318. PMLR, 23–29 Jul 2023. URL https://proceedings.mlr.press/v202/h-zargarbashi23a.html.

Aric A. Hagberg, Daniel A. Schult, and Pieter J. Swart. Exploring network structure, dynamics, and function using networkx. In Gaël Varoquaux, Travis Vaught, and Jarrod Millman (eds.), *Proceedings of the 7th Python in Science Conference*, pp. 11 – 15, Pasadena, CA USA, 2008.

William L Hamilton. *Graph representation learning*. Morgan & Claypool Publishers, 2020.

Hans Hao-Hsun Hsu, Yuesong Shen, Christian Tomani, and Daniel Cremers. What makes graph neural networks miscalibrated? In S. Koyejo, S. Mohamed, A. Agarwal, D. Belgrave, K. Cho, and A. Oh (eds.), *Advances in Neural Information Processing Systems*, volume 35, pp. 13775–13786. Curran Associates, Inc., 2022. URL https://proceedings.neurips.cc/paper_files/paper/2022/file/5975754c7650dfee0682e06e1fec0522-Paper-Conference.pdf.

Weihua Hu, Matthias Fey, Marinka Zitnik, Yuxiao Dong, Hongyu Ren, Bowen Liu, Michele Catasta, and Jure Leskovec. Open graph benchmark: Datasets for machine learning on graphs. *arXiv preprint arXiv:2005.00687*, 2020.

Kexin Huang, Ying Jin, Emmanuel Candes, and Jure Leskovec. Uncertainty quantification over graph with conformalized graph neural networks. 36:26699–26721, 2023. URL https://proceedings.neurips.cc/paper_files/paper/2023/file/54a1495b06c4ee2f07184afb9a37abda-Paper-Conference.pdf.

Meelis Kull, Miquel Perello Nieto, Markus Kängsepp, Telmo Silva Filho, Hao Song, and Peter Flach. Beyond temperature scaling: Obtaining well-calibrated multi-class probabilities with dirichlet calibration. In H. Wallach, H. Larochelle, A. Beygelzimer, F. d'Alché-Buc, E. Fox, and R. Garnett (eds.), *Advances in Neural Information Processing Systems*, volume 32. Curran Associates, Inc., 2019. URL https://proceedings.neurips.cc/paper_files/paper/2019/file/8ca01ea920679a0fe3728441494041b9-Paper.pdf.

Richard Liaw, Eric Liang, Robert Nishihara, Philipp Moritz, Joseph E Gonzalez, and Ion Stoica. Tune: A research platform for distributed model selection and training. *arXiv preprint arXiv:1807.05118*, 2018.

Clemence Magnien, Matthieu Latapy, and Michel Habib. Fast computation of empirically tight bounds for the diameter of massive graphs, 2009.

Pranav Maneriker, Codi Burley, and Srinivasan Parthasarathy. Online fairness auditing through iterative refinement. In *Proceedings of the 29th ACM SIGKDD Conference on Knowledge Discovery and Data Mining*, KDD '23, pp. 1665–1676, New York, NY, USA, 2023. Association for Computing Machinery. ISBN 9798400701030. doi: 10.1145/3580305.3599454. URL https://doi.org/10.1145/3580305.3599454.

Ariane Marandon. Conformal link prediction for false discovery rate control, 2024. URL https://arxiv.org/abs/2306.14693.

Julian McAuley, Christopher Targett, Qinfeng Shi, and Anton Van Den Hengel. Image-based recommendations on styles and substitutes. In *Proceedings of the 38th international ACM SIGIR conference on research and development in information retrieval*, pp. 43–52, 2015.

M. Newman. *Networks*. OUP Oxford, 2018. ISBN 9780192527493. URL https://books.google.com/books?id=YdZjDwAAQBAJ.

Jerzy Neyman and Egon Sharpe Pearson. On the problem of the most efficient tests of statistical hypotheses. *Philosophical Transactions of the Royal Society of London. Series A, Containing Papers of a Mathematical or Physical Character*, 231(694–706):289–337, February 1933. ISSN 2053-9258. doi: 10.1098/rsta.1933.0009. URL http://dx.doi.org/10.1098/rsta.1933.0009.

Harris Papadopoulos. Inductive conformal prediction: Theory and application to neural networks. In *Tools in artificial intelligence*. Citeseer, 2008.

Harris Papadopoulos, Vladimir Vovk, and Alex Gammerman. Qualified predictions for large data sets in the case of pattern recognition. In M. Wani, H. Arabnia, K. Cios, K. Hafeez, and G. Kendall (eds.), *Proceedings of the International Conference on Machine Learning and Applications*, pp. 159–163. CSREA Press, 2002. Proceedings of the International Conference on Machine Learning and Applications, CSREA Press, Las Vegas, NV, pages 159-163, 2002.

Hongbin Pei, Bingzhe Wei, Kevin Chen-Chuan Chang, Yu Lei, and Bo Yang. Geom-gcn: Geometric graph convolutional networks, 2020.

Yaniv Romano, Matteo Sesia, and Emmanuel Candes. Classification with valid and adaptive coverage. *Advances in Neural Information Processing Systems*, 33:3581–3591, 2020.

Mauricio Sadinle, Jing Lei, and Larry Wasserman. Least ambiguous set-valued classifiers with bounded error levels. *Journal of the American Statistical Association*, 114(525):223–234, 2019.

Glenn Shafer and Vladimir Vovk. A tutorial on conformal prediction. *Journal of Machine Learning Research*, 9(3), 2008.

Oleksandr Shchur, Maximilian Mumme, Aleksandar Bojchevski, and Stephan Günnemann. Pitfalls of graph neural network evaluation. *Relational Representation Learning Workshop, NeurIPS 2018*, 2018.

David Stutz, Krishnamurthy Dj Dvijotham, Ali Taylan Cemgil, and Arnaud Doucet. Learning optimal conformal classifiers. In *International Conference on Learning Representations*, 2021.

Vladimir Vovk. Conditional validity of inductive conformal predictors. In *Asian conference on machine learning*, pp. 475–490. PMLR, 2012.

Vladimir Vovk. Cross-conformal predictors. *Annals of Mathematics and Artificial Intelligence*, 74:9–28, 2015.

Vladimir Vovk, Alexander Gammerman, and Glenn Shafer. *Algorithmic learning in a random world*, volume 29. Springer, 2005.

Kuansan Wang, Zhihong Shen, Chiyuan Huang, Chieh-Han Wu, Yuxiao Dong, and Anshul Kanakia. Microsoft Academic Graph: When experts are not enough. *Quantitative Science Studies*, 1(1):396–413, 02 2020. ISSN 2641-3337. doi: 10.1162/qss_a_00021. URL https://doi.org/10.1162/qss_a_00021.

Minjie Wang, Da Zheng, Zihao Ye, Quan Gan, Mufei Li, Xiang Song, Jinjing Zhou, Chao Ma, Lingfan Yu, Yu Gai, Tianjun Xiao, Tong He, George Karypis, Jinyang Li, and Zheng Zhang. Deep graph library: A graph-centric, highly-performant package for graph neural networks. *arXiv preprint arXiv:1909.01315*, 2019.

Xiao Wang, Hongrui Liu, Chuan Shi, and Cheng Yang. Be confident! towards trustworthy graph neural networks via confidence calibration. In M. Ranzato, A. Beygelzimer, Y. Dauphin, P.S. Liang, and J. Wortman Vaughan (eds.), *Advances in Neural Information Processing Systems*, volume 34, pp. 23768–23779. Curran Associates, Inc., 2021. URL https://proceedings.neurips.cc/paper_files/paper/2021/file/c7a9f13a6c0940277d46706c7ca32601-Paper.pdf.

Lingfei Wu, Peng Cui, Jian Pei, Liang Zhao, and Le Song. Graph neural networks. *Graph Neural Networks: Foundations, Frontiers, and Applications*, pp. 27–37, 2022.

Tom Yan and Chicheng Zhang. Active fairness auditing. In *International Conference on Machine Learning*, pp. 24929–24962. PMLR, 2022.

Zhilin Yang, William Cohen, and Ruslan Salakhudinov. Revisiting semi-supervised learning with graph embeddings. In *International conference on machine learning*, pp. 40–48. PMLR, 2016.

Hanqing Zeng, Hongkuan Zhou, Ajitesh Srivastava, Rajgopal Kannan, and Viktor Prasanna. Graphsaint: Graph sampling based inductive learning method, 2020.

Tianyi Zhao, Jian Kang, and Lu Cheng. Conformalized link prediction on graph neural networks. In *Proceedings of the 30th ACM SIGKDD Conference on Knowledge Discovery and Data Mining*, KDD '24, pp. 4490–4499, New York, NY, USA, 2024. Association for Computing Machinery. ISBN 9798400704901. doi: 10.1145/3637528.3672061. URL https://doi.org/10.1145/3637528.3672061.

A Optimal τ for APS

The most popular baseline in work on graph conformal prediction is Adaptive Prediction Sets (APS). (Romano et al., 2020) introduce APS by defining an optimal prediction set construction mechanism under oracle probability. Suppose we estimate a prediction function $\hat{f}$ that correctly models the oracle probability $\Pr[Y = y|X_{test} = \boldsymbol{x}] = \pi_y(\boldsymbol{x})$ for each $y \in \mathcal{Y} = \{1, \ldots, K\}$ Let $\pi_{(1)}(\boldsymbol{x}), \ldots, \pi_{(K)}(\boldsymbol{x})$ be the sorted probabilities in descending order. For any $\tau \in [0, 1]$, define the generalized conditional quantile funciton at τ as

$$L(x; \pi, \tau) = \min \left\{ k \in \{1, \ldots, K\}, \sum_{j=1}^{k} \pi_{(j)}(\boldsymbol{x}) \geq \tau \right\} \tag{4}$$

The corresponding prediction set, $C_\alpha^{\text{or}}(\boldsymbol{x})$, is constructed as

$$C_\alpha^{\text{or}+}(\boldsymbol{x}) = \{y \in \mathcal{Y} : \pi_y(\boldsymbol{x}) \geq \pi_{(L(\boldsymbol{x};\pi,1-\alpha))}(\boldsymbol{x})\}$$

where or indicates the usage of the oracle probability. Further, they define tighter prediction sets in a randomized fashion using an additional uniform random variable $u \sim \text{Uniform}(0, 1)$ as a parameter to construct a generalized inverse. This idea draws upon the idea of uniformly most powerful tests in the Neyman-Pearson lemma for level-α sets (Neyman & Pearson, 1933). Define

$$S(\boldsymbol{x}, u; \pi, \tau) = \begin{cases} \{y \in \mathcal{Y} : \pi_y(\boldsymbol{x}) > \pi_{(L(\boldsymbol{x};\pi,\tau))}(\boldsymbol{x})\} & u < V(\boldsymbol{x}; \pi, \tau) \\ \{y \in \mathcal{Y} : \pi_y(\boldsymbol{x}) \geq \pi_{(L(\boldsymbol{x};\pi,\tau))}(\boldsymbol{x})\} & \text{otherwise} \end{cases} \tag{5}$$

i.e., the class at the $L(\boldsymbol{x}; \pi, \tau)$ rank is included in the prediction set with probability $1 - V(\boldsymbol{x}; \pi, \tau)$, where

$$V(\boldsymbol{x}; \pi, \tau) = \frac{1}{\pi_{(L(\boldsymbol{x};\pi,\tau))}(x)} \left\{ \left[\sum_{j=1}^{L(\boldsymbol{x};\pi,\tau)} \pi_{(j)}(x) \right] - \tau \right\}$$

The corresponding randomized prediction sets are $C_\alpha^{\text{or}}(\boldsymbol{x}) = S(\boldsymbol{x}, U; \pi, 1-\alpha)$, $U \sim U(0, 1)$ Note that in general, the coverage guarantees provided in conformal prediction hold only in expectation over the randomness in $(\boldsymbol{x}_i, y_i), i = 1, \ldots, n + 1$. The randomized prediction sets continue to provide the guarantee with additional randomness over u_i. To make this work for a non-oracle probability $\hat{\pi}(\boldsymbol{x})$, they define a non-conformity score A

$$A(\boldsymbol{x}, y, u; \hat{\pi}) = \min\{\tau \in [0, 1] : y \in S(\boldsymbol{x}, u; \hat{\pi}, \tau)\} \tag{6}$$

Assume that $\hat{\pi}$ are all distinct - for ease of defining rank. Suppose the rank of the true class amongst the sorted $\hat{\pi}$ be r_y, i.e., $\sum_{i=1}^{K} \mathbf{1}[\hat{\pi}_i(\boldsymbol{x}) \geq \hat{\pi}_y] = r_y$ Solving for τ as a function of $\hat{\pi}$ (see Appendix A, for proof),

$$A(\boldsymbol{x}, y, u; \hat{\pi}) = \left[\sum_{i=1}^{r_y} \hat{\pi}_{(i)}(\boldsymbol{x}) \right] - u\hat{\pi}_y \tag{7}$$

Instead, if a deterministic set is used to define the conformal score instead (i.e., the randomized set construction is not carried out), then we could just add the probabilities until the true class is included:

$$\tilde{A}(\boldsymbol{x}, y; \hat{\pi}) = \left[\sum_{i=1}^{r_y} \hat{\pi}_{(i)}(\boldsymbol{x}) \right] \tag{8}$$

This version of APS still provides the same conditional coverage guarantees and has a simpler exposition as the prediction sets are constructed by greedily including the classes until the true label is included. Thus, this version is provided as the implementation in the popular monographs on conformal prediction

by (Angelopoulos & Bates, 2021; Angelopoulos et al., 2023). However, the lack of randomization may sacrifice efficiency. This modification of the score function affects both the quantile threshold computation during the calibration phase and the prediction set during the test phase. We will now show the conditions that impact the efficiency more formally.

For simplicity, assume that the probabilities are distinct.

From the definition of A equation 6

$$A(\boldsymbol{x}, y, u; \hat{\pi}) = \min\{\tau \in [0,1] : y \in S(\boldsymbol{x}, u; \hat{\pi}, \tau)\}$$

Define

$$\Sigma_{\hat{\pi}}(\boldsymbol{x}, m) = \sum_{i=1}^{m} \hat{\pi}_{(i)}(\boldsymbol{x})$$

From the definition of $S(\boldsymbol{x}, u; \hat{\pi}, \tau)$ from equation 5, conisder the following cases:

Case 1: $\tau = \Sigma_{\hat{\pi}}(\boldsymbol{x}, r_y)$, then $L(x; \hat{\pi}, \tau) = y$ and thus, $V(\boldsymbol{x}; \pi, \tau) = 0$. Thus $\Pr[u > V(\boldsymbol{x}; \pi, \tau)] = 1$ and hence, $P[y \in S(\boldsymbol{x}, u; \hat{\pi}, \tau)] = 1$.

Case 2: $\tau = \Sigma_{\hat{\pi}}(\boldsymbol{x}, r_y - 1)$, then $y \notin S(\boldsymbol{x}, u, \hat{\pi}, \tau)$ in either case, since only classes with $\hat{\pi}_i(\boldsymbol{x}) > \hat{\pi}_y(\boldsymbol{x})$ could be included.

Case 3: $\tau = \Sigma_{\hat{\pi}}(\boldsymbol{x}, r_y) - \varepsilon\hat{\pi}_y$. Then we have $L(x; \hat{\pi}, \tau) = y$ again, and

$$\begin{aligned} V(\boldsymbol{x}; \pi, \tau) &= \frac{1}{\hat{\pi}_y(\boldsymbol{x})} \left\{ \left[\sum_{j=1}^{r_y} \hat{\pi}_{(j)}(\boldsymbol{x}) \right] - \tau \right\} \\ &= \frac{1}{\hat{\pi}_y(\boldsymbol{x})} \left\{ \left[\sum_{j=1}^{r_y} \hat{\pi}_{(j)}(\boldsymbol{x}) \right] - (\Sigma_{\hat{\pi}}(\boldsymbol{x}, r_y) - \varepsilon\hat{\pi}_y) \right\} \\ &= \varepsilon \end{aligned}$$

For y to be included in $S(\boldsymbol{x}, u; \hat{\pi}, \tau)$, we would require that $u \geq V(\boldsymbol{x}; \pi, \tau)$, i.e., $u \geq \varepsilon$. We want the minimal τ, which is equivalent to maximizing ε. Thus, $\tau = \Sigma_{\hat{\pi}}(\boldsymbol{x}, r_y) - u\hat{\pi}_y$ is the required solution.

A.1 Non-randomized set

The inclusion criterion for the score given the threshold τ is $\tilde{A}(\boldsymbol{x}, y; \hat{pi}) \leq \tau$

To include the currct label y_i while minimizing the chosen threshold τ, we would require $\tau = \sum_{j=1}^{r_{y_i}} \hat{\pi}_{(j)}(\boldsymbol{x})$

B Method Details and Innovations

B.1 Notes on Transductive NAPS

Neighborhood Adaptive Prediction Sets (NAPS) constructs prediction sets under relaxed exchangeability (or non-exchangeability) assumptions (Barber et al., 2023) and was initially implemented for the inductive setting (Clarkson, 2023). However, NAPS can also be used in the transductive setting (H. Zargarbashi et al., 2023). To compute scores for $\mathcal{D}_{\text{calib}}$ nodes, NAPS uses APS. Using these scores, Equation 9 is used to compute a *weighted quantile* for the score threshold to be used when constructing the prediction sets. The weighted quantile is defined by placing a point mass, δ_{s_i}, for each calibration point's score, s_i, as well as a point mass at $_\infty$ to represent the test point, v_{n+1}'s, score. The reason for the $\delta_{+\infty}$ point mass is because the score for v_{n+1} is unknown, and potentially unbounded, due to non-exchangeability.

$$\hat{q}_{n+1}^{\text{NAPS}} = \text{Quantile}\left(1-\alpha, \left[\sum_{i\in\mathcal{D}_{\text{calib}}} \tilde{w}_i \cdot \delta_{s_i}\right] + \tilde{w}_{n+1}\cdot\delta_{+\infty}\right) \tag{9}$$

For NAPS to produce viable prediction sets, the weights, $w_i \in [0,1]$, for the calibration nodes must be chosen in a data-independent fashion, i.e., they cannot leverage the associated node features (Barber et al., 2023). NAPS leverages the graph structure to assign these weights, assigning non-zero weights to nodes within a k-hop neighborhood, $\mathcal{N}_{n+1}^k$, of a test node v_{n+1}. For nodes that are in $\mathcal{N}_{n+1}^k$, let d_i be the distance from the node to v_{n+1} to $v_i \in \mathcal{V}_{\text{calib}}$. The three implemented weight functions are *uniform*: $w_u(d_i) = 1$, *hyperbolic*: $w_h(d_i) = \frac{1}{d_i}$, and *exponential*: $w_e(d_i) = 2^{-d_i}$. Nodes that are not in the $\mathcal{N}_{n+1}^k$ have zero weight. The weights are then normalized, $\tilde{w}_i$, such that $\sum_{i\in\mathcal{D}_{\text{calib}}} \tilde{w}_i + \tilde{w}_{n+1} = 1$ (Barber et al., 2023).

NAPS Implementation NAPS is computationally expensive in terms of time and memory since the k-hop intersection is computed for each test node. To allow for scalability, our implementation of NAPS, shown in Algorithms 1 and 2, uses batching to ensure sufficient memory is available and uses sparse-tensor multiplication to reduce memory and time costs.

To ensure scalability for large graphs, all the computations until the quantile computation step were done via sparse tensors. Algorithm 2 illustrates how the distance to each calibration node in the k-hop neighborhood can be computed via sparse tensor primitives.

Algorithm 1 NAPS Quantile Implementation

procedure NAPS_QUANTILE($w, k, \mathcal{D}_{\text{calib}}, \mathcal{D}_{\text{test}}, \mathcal{D}, \mathcal{S}_{\text{calib}}, b, \alpha$)
  $\{\mathcal{B}_1, \mathcal{B}_2, \ldots, \mathcal{B}_b\} \leftarrow$ SPLIT($\mathcal{D}_{\text{test}}, b$) ▷ Split test nodes into b batches
  $q \leftarrow$ ZEROS($\mathcal{D}_{\text{test}}, 1$) ▷ $q \in \mathbb{R}^{|\mathcal{D}_{\text{test}}|\times 1}$
  for $\mathcal{B}_n \in \{\mathcal{B}_1, \mathcal{B}_2, \ldots, \mathcal{B}_b\}$ **do**
    k_hop $\leftarrow$ SPARSE_K_HOP($k, \mathcal{B}_n, \mathcal{D}_{\text{calib}}, \mathcal{D}$) ▷ k_hop $\in \mathbb{R}^{|\mathcal{B}_n|\times|\mathcal{D}_{\text{calib}}|}$
    weights $\leftarrow$ COMPUTE_WEIGHTS(w, k_hop) ▷ weights $\in \mathbb{R}^{|\mathcal{B}_n|\times|\mathcal{D}_{\text{calib}}|}$
    $q[\mathcal{B}_n] \leftarrow$ COMPUTE_QUANTILE($1-\alpha$, weights, $\mathcal{S}_{\text{calib}}$)
  end for
  return q ▷ Return the quantiles for each test node
end procedure

Parameter Analysis: Apart from the particular weighting function, the main parameter in the NAPS algorithm is the number of hops to consider, k. Figure B1 shows the trend between efficiency and coverage as we increase k from 1 to D, where D is a lower bound on the diameter of the largest strongly connected component of the dataset, computed using the NetworkX Hagberg et al. (2008). For each of the datasets, we observe that there is a value of k after which the efficiency does not improve, while still achieving the desired coverage. This behavior can suggest a heuristic akin to the 'elbow method' used in clustering analysis for determining the number of clusters for choosing the value of k in NAPS.

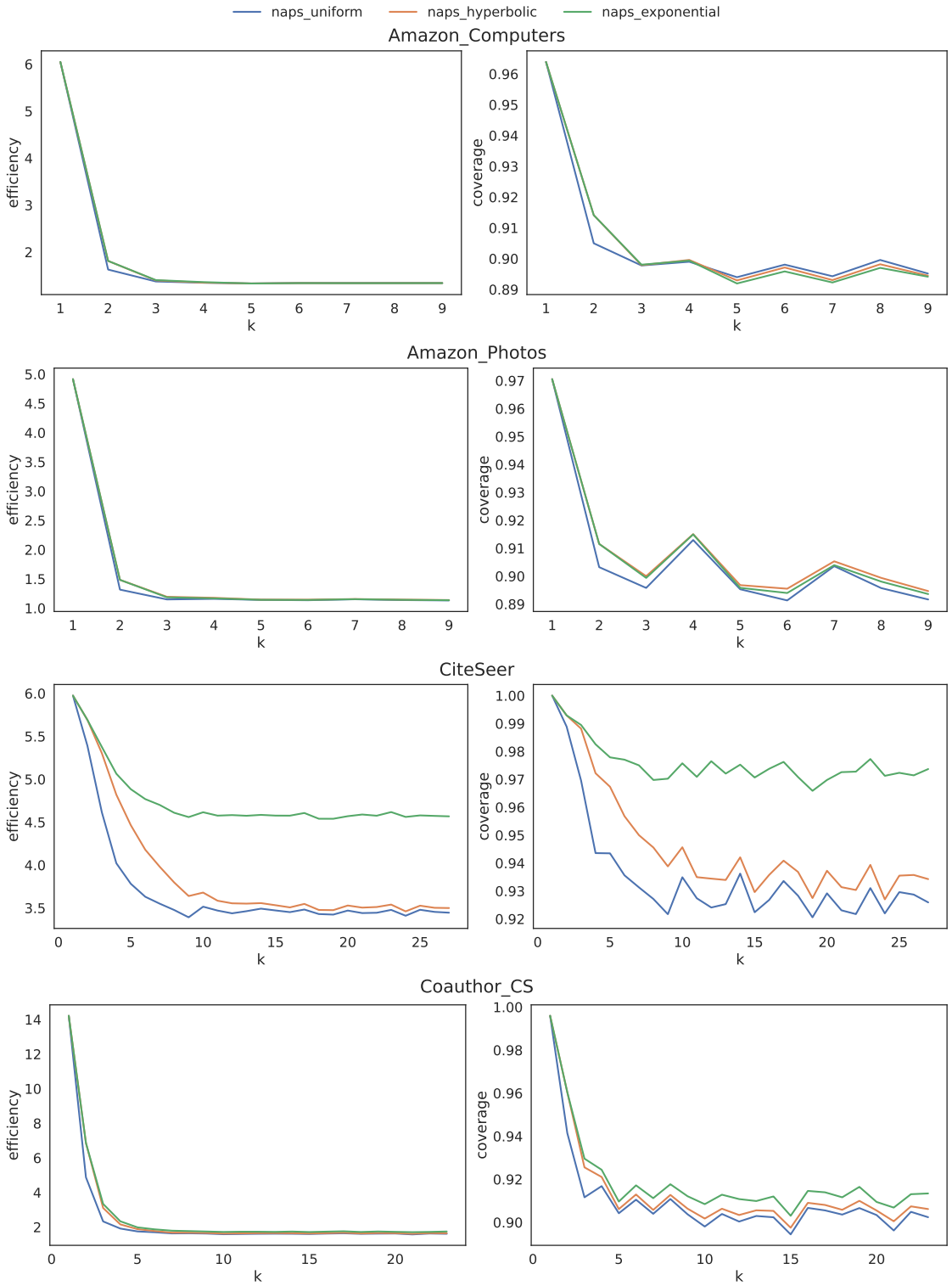


Figure B1: Plotting the Efficiency and Coverage when using NAPS for k from 1 to D. The above results are with FS split and $\alpha = 0.1$.

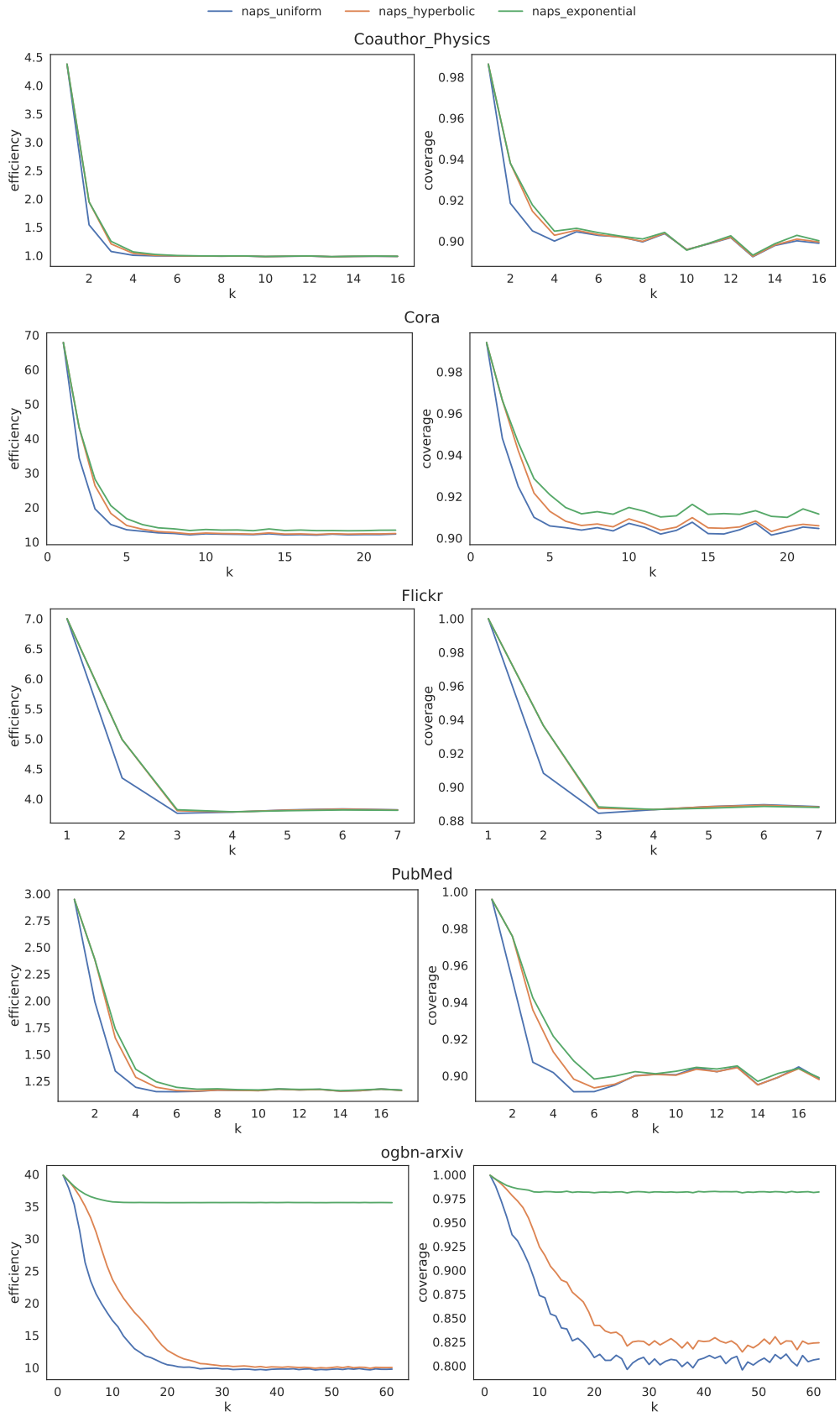


Figure B1: Plotting the Efficiency and Coverage when using NAPS for k from 1 to D. The above results are with FS split and $\alpha = 0.1$ (cont.). The ogbn-products dataset is omitted due to size and lack of data points.

Algorithm 2 Sparse K Hop Neighborhood Implementation

procedure SPARSE_K_HOP($k, \mathcal{B}, \mathcal{D}_{\text{calib}}, \mathcal{D}$)
  A $\leftarrow$ GET_ADJACENCY($\mathcal{D}$) ▷ Adjacency of $\mathcal{D}$, A $\in \mathbb{R}^{|\mathcal{D}|\times|\mathcal{D}|}$
  path_n $\leftarrow$ A[$\mathcal{B}$, :] ▷ path_n $\in \mathbb{R}^{|\mathcal{B}|\times|\mathcal{D}|}$
  k_hop $\leftarrow$ path_n[:, $\mathcal{D}_{\text{calib}}$] ▷ k_hop $\in \mathbb{R}^{|\mathcal{B}|\times|\mathcal{D}_{\text{calib}}|}$
  for $n \in \{2, 3, \ldots, k\}$ **do**
    path_n $\leftarrow$ (path_n)A
    neg_if_n $\leftarrow$ k_hop $-$ SGN(path_n[:, $\mathcal{D}_{\text{calib}}$]) ▷ negative value $\implies$ n hops away
    in_n_hop $\leftarrow$ (neg_if_n < 0) $\times\, n$ ▷ Nodes that are a min distance of n
    k_hop $\leftarrow$ k_hop + in_n_hop
  end for
  return k_hop ▷ $\forall_{i,j}$ **If** dist$(i, j) \leq k$ **then** k_hop$[i, j]$ = dist(i, j), **else** k_hop$[i, j] = 0$
end procedure

B.2 Diffusion Adaptive Prediction Sets (DAPS)

The Diffusion Adaptive Prediction Sets (DAPS) approach for conformal node classification on graphs was introduced by (H. Zargarbashi et al., 2023). The intuition behind DAPS follows the prevalence of homophily graphs, which suggests non-conformity scores for two connected nodes should be related. DAPS uses a diffusion step to capture this relationship and uses the non-conformity scores modified by diffusion to generate the prediction sets. Formally, suppose $s(v, y)$ is a point wise non-conformity score for a node v and label y (e.g., TPS or APS)

$$\hat{s}(v, y) = (1 - \lambda)s(v, y) + \frac{\lambda}{|\mathcal{N}_v|} \sum_{u \in \mathcal{N}_v} s(u, y)$$

where $\mathcal{N}_v$ is the 1-hop neighborhood of v and $\lambda \in [0, 1]$ is a hyperparameter controlling the diffusion.

H. Zargarbashi et al. (2023) use the APS score as the point-wise score in the diffusion process as it is adaptive and uniformly distributed in $[0, 1]$ under oracle probability. However, as we noted earlier, using classwise thresholds provides a mechanism to produce adaptive scores from TPS as well. Thus, we create DTPS, a variation of DAPS using TPS scores as the point-wise scores in the diffusion process.

We compare our proposed method of using diffusion on top of TPS-Classwise (DTPS) against DAPS, which was proposed by H. Zargarbashi et al. (2023). From Figure B2 (left), we see that DTPS can be competitive with DAPS in efficiency while providing better label stratified coverage. However, for some of the larger datasets (Cora, Flickr, ogbn-arxiv, ogbn-products), DTPS suffers from poorer efficiency compared to DAPS. This can be partially explained by the worse performance of the pre-diffusion TPS-classwise (Figure D2) which is forced to sacrifice efficiency on these datasets to achieve label-stratified coverage (Figure D1). However, when we control the number of samples per class with LC splits (Figure B2 right), we see that DTPS label stratified coverage deteriorates significantly compared to DAPS. Based on these results, we can conclude that DTPS is not a universally better method than DAPS, and its performance is sensitive to the calibration set size and the number of classes. It may be a viable candidate over DAPS in scenarios when there is a sufficiently large calibration set when TPS-classwise has competitive efficiency to TPS.

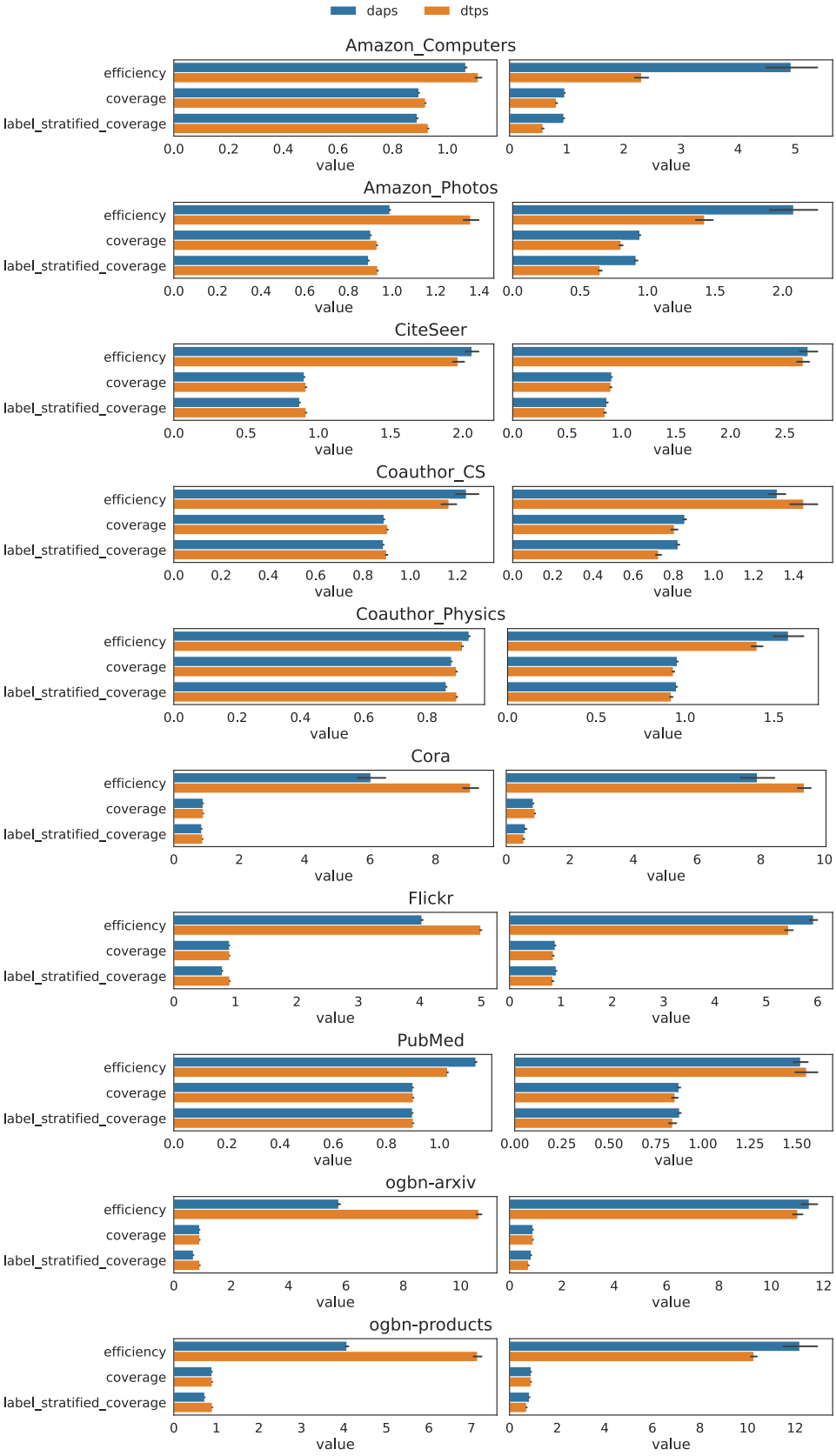


Figure B2: Bar charts denoting different metrics associated with DAPS and DTPS across all datasets for FS split (left) and LC split (right) at $\alpha = 0.1$.

C Datasets and Hyperparameter Tuning Details

C.1 Datasets

We selected datasets of varying sizes and origins to evaluate the performance of the graph conformal prediction methods. The first category of datasets is citation datasets, where the nodes are publications and the edges denote citation relationships. Nodes have features that are bag-of-words representations of the publication. The task is to predict the category of each publication. The citation networks we use are **CiteSeer** (Yang et al., 2016), CoraFull (Shchur et al., 2018), an extended version of the common Cora (Yang et al., 2016) citation network dataset, and **Pubmed** (Yang et al., 2016). The second category comes from the Amazon Co-Purchase graph (McAuley et al., 2015), where nodes represent goods, edges represent goods frequently bought together, and node features are bag-of-words representations of product reviews. The task is to predict the category of a good. We use the **Amazon_Photos** and **Amazon_Computers** datasets. The last category is co-authorship networks extracted from the Microsoft Academic Graph (Wang et al., 2020) used for KDD Cup'16. The nodes are authors, edges represent coauthorship, and node features represent paper keywords of the author's publications. The task is to predict the author's most active field of study. We use **Coauthor_CS** and **Coauthor_Physics** which both come from the Microsoft Academic Graph (Wang et al., 2020). Other datasets that were use include **Flickr** (Zeng et al., 2020), **ogbn-arxiv**, and **ogbn-products** Hu et al. (2020). These last three datasets have predefined splits for train/validation/test, shown in Table C2, which we used when constructing our train/validation/calibration/test splits for the different split styles. For all the chosen datasets, we used the version provided by the Deep Graph Library (Wang et al., 2019). DGL uses an Apache 2.0 license, and OGB uses an MIT license.

Table C1 presents summary statistics for each dataset. These include the average local clustering coefficient (Avg CC), global clustering coefficient (Global CC) (Newman, 2018), an approximate lower bound on the diameter (D) given by Magnien et al. (2009), node homophily ratio ($\hat{H}$) (Pei et al., 2020), and expected node homophily ratio (H_{rand}). Figure C1 shows the label distribution for each dataset.

Table C1: Summary statistics for all datasets evaluated.

Dataset	Nodes	Edges	Classes	Features	Avg CC	Global CC	D	$\hat{H}$	H_{rand}
CiteSeer	3,327	9,228	6	3,703	0.141	0.130	28	0.722	0.178
Amazon_Photos	7,650	238,163	8	745	0.404	0.177	10	0.836	0.165
Amazon_Computers	13,752	491,722	10	767	0.344	0.108	10	0.785	0.208
Cora	19,793	126,842	70	8,710	0.261	0.131	23	0.586	0.022
PubMed	19,717	88,651	3	500	0.060	0.054	17	0.792	0.357
Coauthor_CS	18,333	163,788	15	6,805	0.343	0.183	24	0.832	0.112
Coauthor_Physics	34,493	495,924	5	8,415	0.378	0.187	17	0.915	0.321
Flickr	89,250	899,756	7	500	0.033	0.004	8	0.322	0.267
ogbn-arxiv	169,343	1,166,243	40	128	0.118	0.115	62	0.567	0.077
ogbn-products	2,449,029	61,859,140	47	100	0.411	0.130	27	0.817	0.106

Table C2: Predefined splits from original source noted.

Dataset	# Train	# Valid	# Test
Flickr	44,625	22,312	22,313
ogbn-arxiv	90,941	29,799	48,603
ogbn-products	196,615	39,323	2,213,091

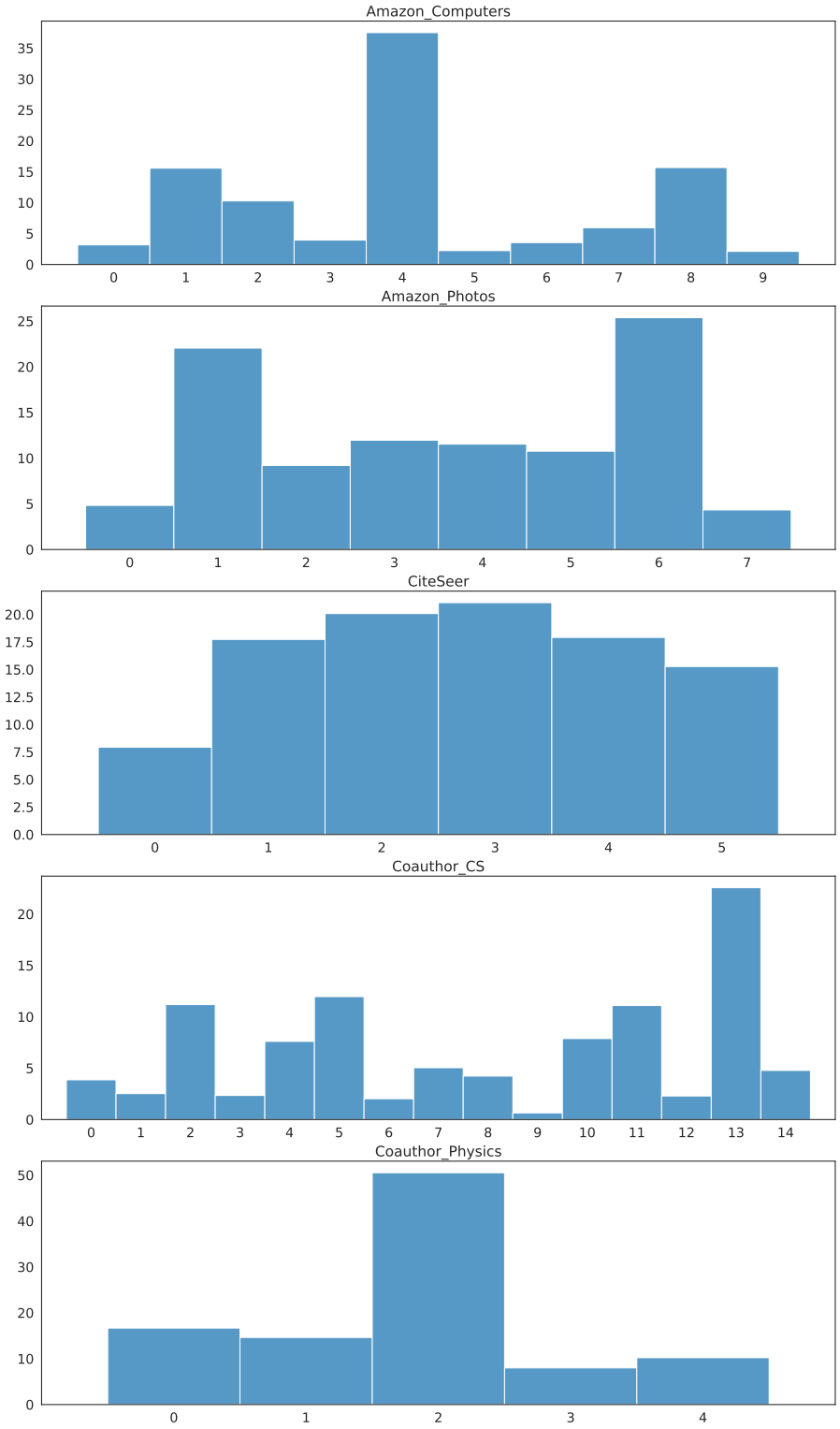


Figure C1: Plots of label distribution for each dataset. Each class for each dataset has at least one occurrence.

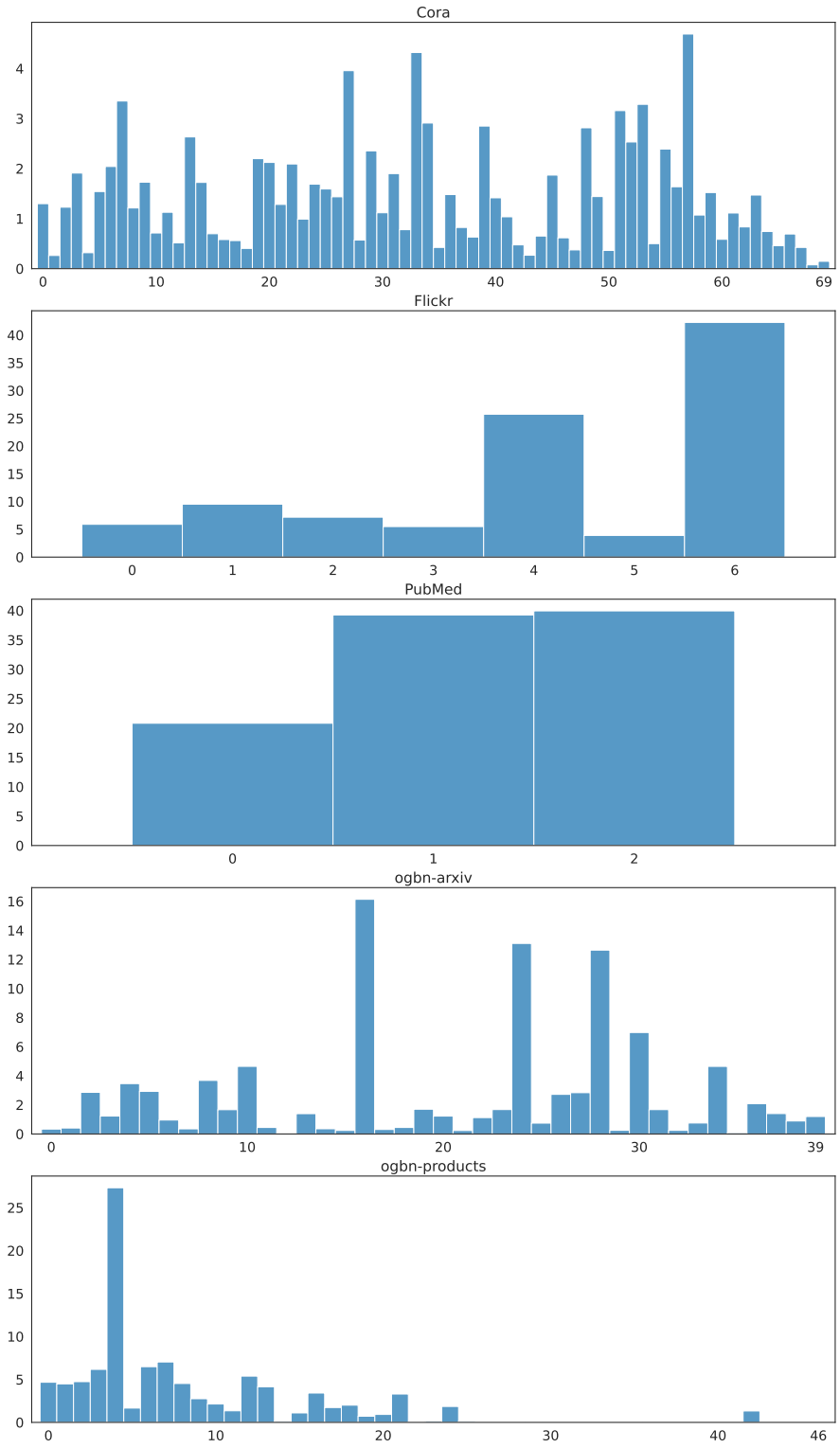


Figure C1: Plots of label distribution for each dataset. Each class for each dataset has at least one occurrence (cont.).

C.2 Hyperparameter Tuning

Hyperparameter tuning was done using Ray Tune (Liaw et al., 2018). The hyperparameters for the base model were tuned via random search using Table C3 for each model type (e.g. GCN, GAT, and GraphSAGE), for each dataset, except for the OGB (Hu et al., 2020) datasets, and each splitting scheme (FS vs LC and different value settings). For the OGB datasets, we took the hyperparameters and architectures from the corresponding leaderboard for all splitting schemes.

Once we got the best base model for each dataset and splitting scheme, we tune the hyperparameters for the CF-GNN model via random search using Table C4 for each model type (e.g. GCN, GAT, and GraphSAGE), for each dataset. For the FS splitting schemes, we set $\mathcal{D}_{\text{calib}} = \mathcal{D}_{\text{test}} = (1 - \mathcal{D}_{\text{train}} - \mathcal{D}_{\text{valid}})/2$. For the ogbn-products dataset, due to its size, we used a batch size of 512 for the CF-GNN training and also used a NeighborSampler with fanouts $[10, 10, 5]$ rather than a MultiLayerFullNeighborSampler.

[2]

All experiments with the ogbn-products datasets were run on a single A100 GPU while the remaining experiments for the other datasets were run on a single P100 GPU.

Table C3: Hyperparameter search space for the base GNN model for non-OGB datasets. The last two rows are layer-type specific for GAT and GraphSAGE, respectively.

Hyperparameter	Search Space
batch_size	64
lr	loguniform$(10^{-4}, 10^{-1})$
hidden_channels	$\{16, 32, 64, 128\}$
layers	$\{1, 2, 4\}$
dropout	uniform$(0.1, 0.8)$
heads	$\{2, 4, 8\}$
aggr_fn	{mean, gcn, pool, lstm}

Table C4: Hyperparameter search space for the CF-GNN model. The last two rows are layer-type specific for GAT and GraphSAGE, respectively.

Hyperparameter	Search Space
batch_size	64
lr	loguniform$(10^{-4}, 10^{-1})$
hidden_channels	$\{16, 32, 64, 128\}$
layers	$\{1, 2, 3, 4\}$
dropout	uniform$(0.1, 0.8)$
τ	loguniform$(10^{-3}, 10^{1})$
heads	$\{2, 4, 8\}$
aggr_fn	{mean, gcn, pool, lstm}

[2]For each dataset and split type, we provide the configuration files for the best corresponding base GNN and CFGNN architectures in the attached code.

D Additional Empirical Results, Analysis, and Insights

This section expands the figures and tables seen in the main body but for all datasets considered. Figure D1 compares TPS and TPS-Classwise for all the datasets. We observe that TPS-Classwise achieves the desired label stratified coverage of 0.9 while TPS doesn't necessarily for the FS split. For the LC split, we observe that TPS-Classwise slightly improves on TPS in terms of label-stratified coverage; however, neither method necessarily achieves the target label-stratified coverage. In Figure D2, we find TPS-Classwise generally is less efficient than TPS for FS and LC splitting. Figure D3 compares the label stratified coverage for APS with and without randomization. We observe that randomization does sacrifice the label-stratified coverage for both FS and LC splitting. Noticeably, the change in coverage is smaller for the LC split. Figure D4 compares the efficiency of APS with and without randomization at $\alpha = 0.1$. For both split types (FS & LC), we observe that the randomized version of APS produces more efficient prediction sets, in line with Theorem 3.1. The efficiency improvements come with a sacrifice in label stratified coverage since smaller prediction set sizes are preferred over covering every class, particularly if the classes are rare. To visualize this trade-off, we observe that in both Figure D1 and D3 the difference in label stratified coverage for ogbn-products with FS splitting is more extreme than with other datasets and with LC splitting. This is because ogbn-products has a lot of classes that have almost no representation (see Figure C1). Datasets that have near-uniform label distribution (e.g., PubMed, CiteSeer) – or when using LC split which controls for label counts – we observe label stratified coverage isn't sacrificed as much in the name of efficiency.

D.1 Conformalized GNN

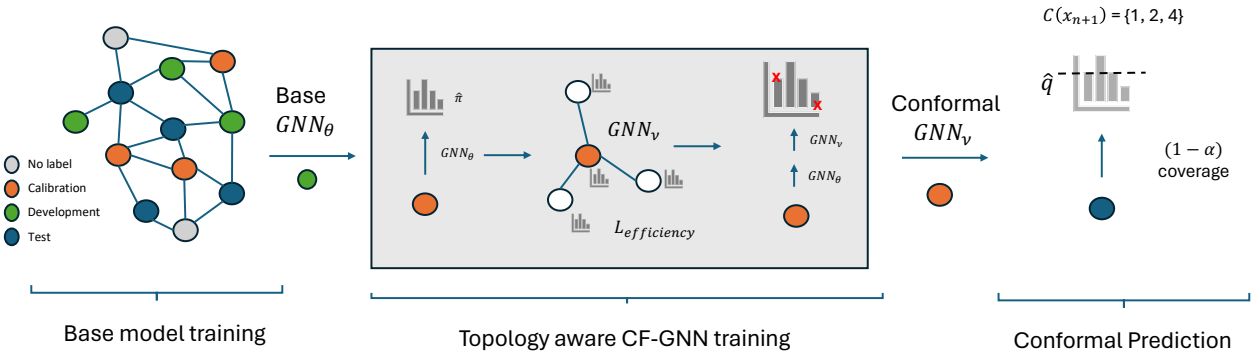


Figure D5: Procedure for training CFGNN. First (left), the base model is trained on the training set. Then, (middle) CFGNN is trained to maximize efficiency over the calibration set. Finally, (right) the non-conformity scores from the combined models are used to generate the prediction sets.

Section 3 introduces CFGNN (Huang et al., 2023) as a conformal prediction method specific for graphs. Figure D5 shows the end-to-end CFGNN procedure that is split into three steps. The first is to train a base GNN model, GNN_θ, on the development nodes, $\mathcal{V}_{\text{dev}}$, normally using a label-based loss function such as Cross Entropy Loss. Using the outputs of GNN_θ as inputs, a second GNN, GNN_φ is trained using an efficiency-based loss function proposed by Huang et al. (2023). Equation 10 presents the efficiency-based loss function for node classification, where σ is the sigmoid function and τ is a temperature hyperparameter. The calibration nodes, $\mathcal{V}_{\text{calib}}$, are split into two sets, $\mathcal{V}_{\text{cor-cal}}$ and $\mathcal{V}_{\text{cor-test}}$ for training and validation, respectively. The fully trained GNN_φ is then used for the conformal prediction and prediction set construction.

$$\begin{aligned} \hat{\eta} &= \text{DiffQuantile}(\{s(\boldsymbol{x}_i, y_i)\}, (1-\alpha)(1 + 1/|\mathcal{V}_{\text{cor-cal}}|)) \\ \mathcal{L}_{eff} &= \frac{1}{|\mathcal{V}_{\text{cor-cal}}|} \sum_{i \in \mathcal{V}_{\text{cor-cal}}} \sum_{k \in \mathcal{Y}} \sigma\left(\frac{s(\boldsymbol{x}_i, k) - \hat{\eta}}{\tau}\right) \end{aligned} \tag{10}$$

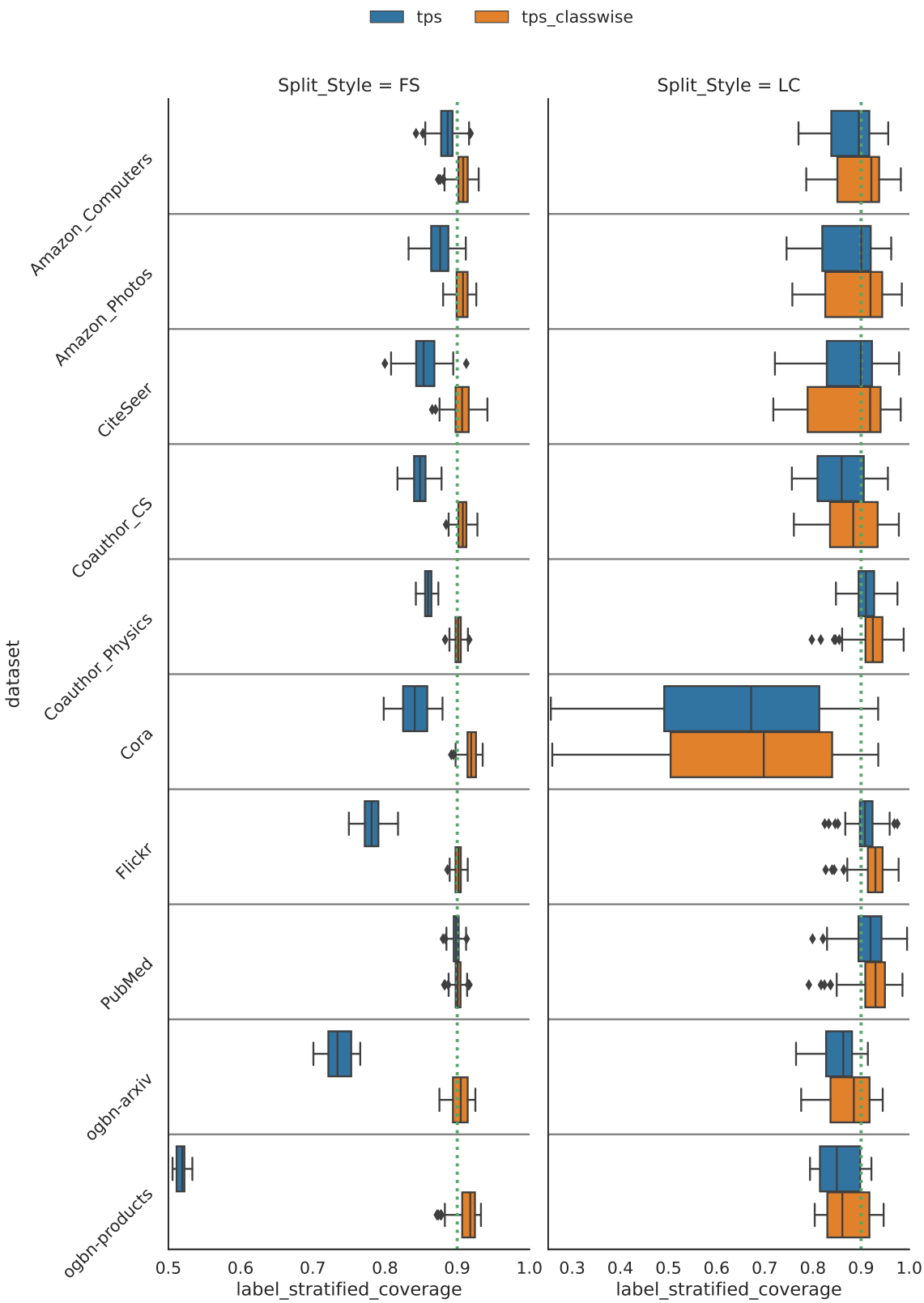


Figure D1: Plots for TPS vs TPS-Classwise for all the data sets at $1 - \alpha = 0.9$ coverage (green dotted line). TPS-Classwise, on average, meets label stratified coverage for FS split (left). For the LC split, the label stratified coverage slightly improves with TPS-Classwise but does not necessarily meet the $1 - \alpha$ coverage.

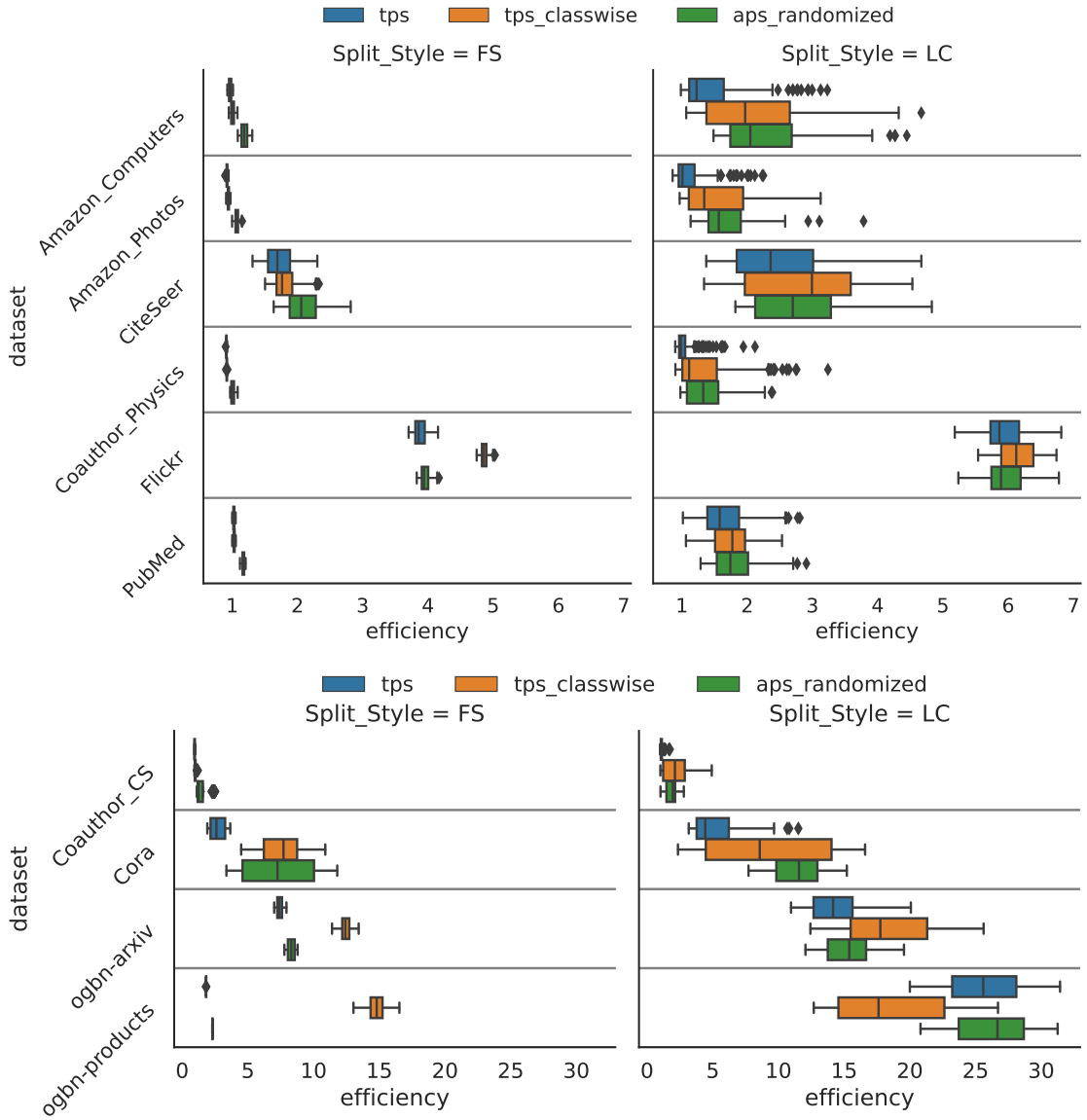


Figure D2: Plot for TPS vs TPS-classwise at $\alpha = 0.1$. For the FS split type, TPS-classwise becomes more inefficient compared to TPS for larger graph sizes but is competitive in other settings. To maintain label stratified coverage TPS-classwise may be forced to overcover certain classes at the cost of efficiency.

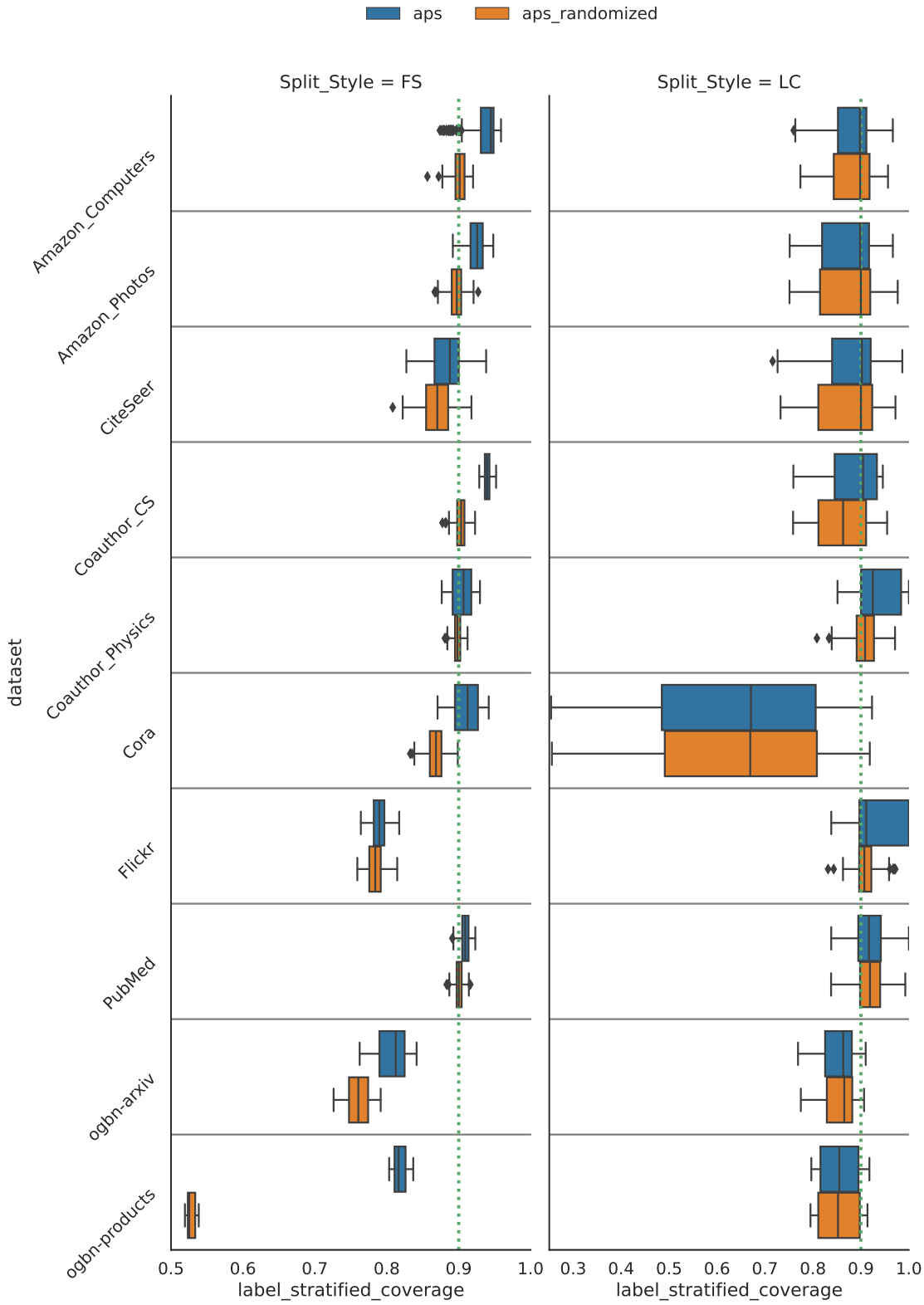


Figure D3: Plots for APS vs APS Randomized for all the data sets at $1 - \alpha = 0.9$ coverage (green dotted line). For the FS split type, APS-Randomized has a lower label stratified coverage. However, with the LC split type, the decrease in label-stratified coverage is not as significant.

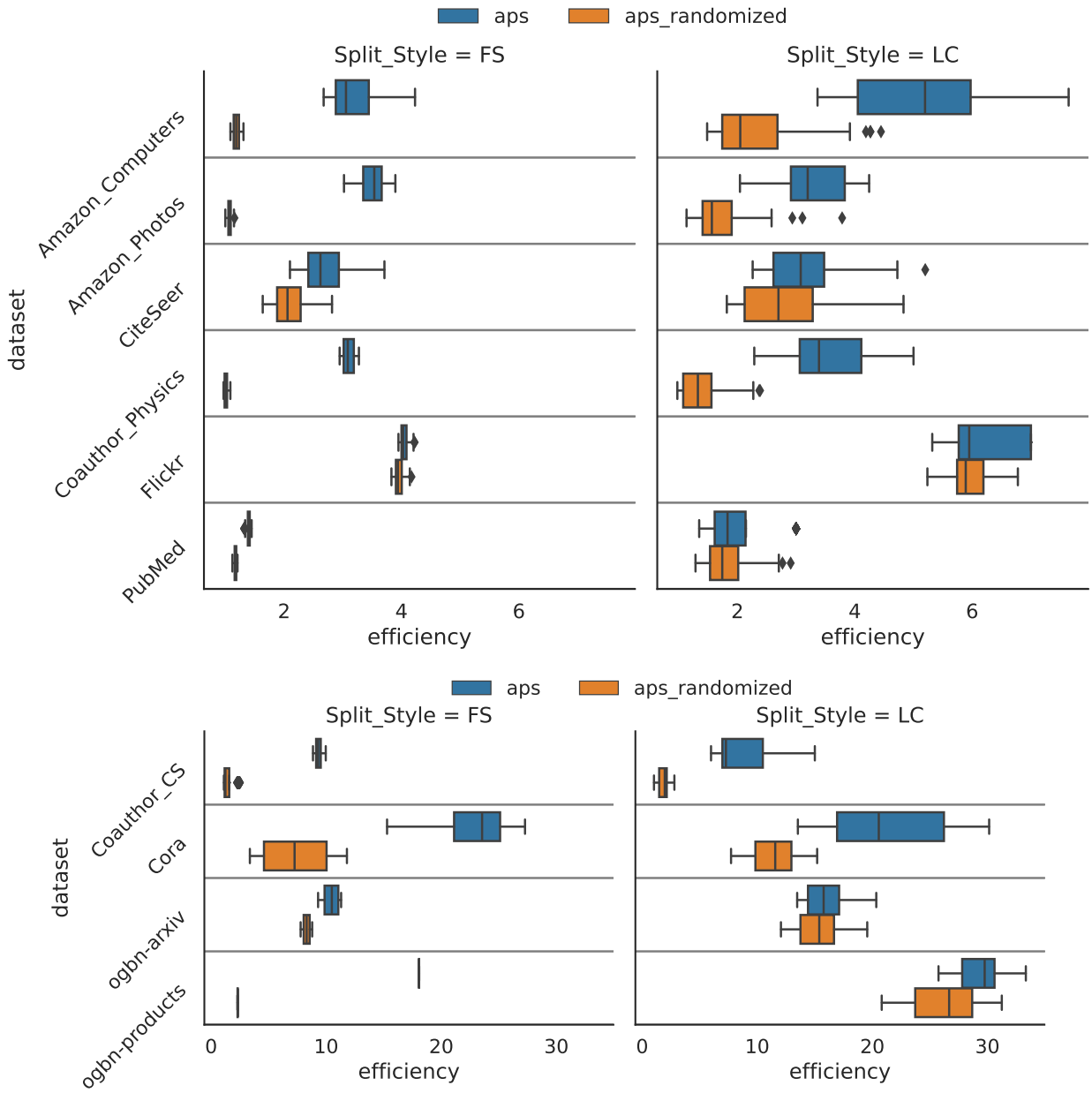
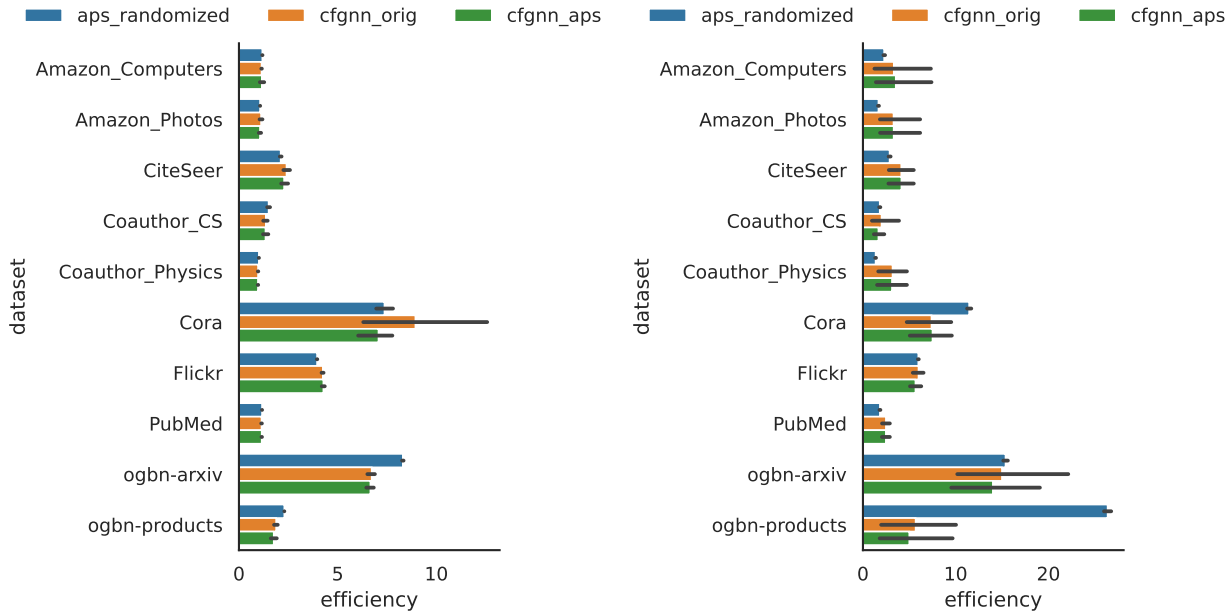


Figure D4: Plots for APS vs APS-randomized at $\alpha = 0.1$. For both split types (FS & LC), the randomized version of APS produces more efficient sets for all the datasets.

D.1.1 Further Discussions on CFGNNN

Impact of Inefficiency Loss: Figure D6 compares the efficiency of 'cfgnn_aps', 'cfgnn_orig,' and 'aps_randomized' for all the other datasets. For the FS split, we see that 'cfgnn_aps' improves upon 'cfgnn_orig' for all datasets and can improve upon 'aps_randomized' for most datasets. For the LC split, we see that CFGNN is still quite brittle for all datasets. For datasets with a larger number of classes, CFGNN can still improve upon 'aps_randomized' as seen with Cora and ogbn-products.



(a) FS split style

(b) LC split style

Figure D6: Bar charts denoting efficiency for 'cfgnn_aps', 'cfgnn_orig', and 'aps_randomized' for both split styles at $\alpha = 0.1$. We see that 'cfgnn_aps' improves or matches efficiency in most cases.

Scaling CFGNN: Expanding on Table 2 from Section 3, Table D1 present the runtime improvements (see Table D1a) as well as the efficiencies (see Table D1b) for each CFGNN implementation using the best CFGNN architecture found through hyperparameter tuning (see Section C4). We observe that our improved implementation achieves comparable efficiency to the original in only 50 epochs as opposed to 1000 used in the original.

Table D1: Impact of different CFGNN implementations starting from the baseline, then batching, and then both caching and batching combined. Used the **best** CFGNN architecture (w.r.t. validation efficiency) for each dataset. We run 5 trials for each setup and report a 95% confidence interval. OOM = Out of Memory

(a) Impact on Runtime

dataset(↓) / method(→)	original	batching	batching+caching
CiteSeer	379.28 ± 9.36	36.36 ± 0.97	27.45 ± 0.80
Amazon_Photos	496.36 ± 4.92	102.92 ± 0.95	54.44 ± 1.36
Amazon_Computers	664.98 ± 10.90	205.29 ± 3.96	73.85 ± 1.56
Cora	1378.01 ± 13.35	203.61 ± 2.52	73.60 ± 1.53
PubMed	571.60 ± 12.09	219.63 ± 5.23	109.68 ± 1.64
Coauthor_CS	638.56 ± 6.14	88.49 ± 1.14	31.67 ± 0.53
Coauthor_Physics	5942.30 ± 176.90	3918.38 ± 17.16	1585.26 ± 6.84
Flickr	868.87 ± 9.98	567.39 ± 6.50	56.71 ± 1.30
ogbn-arxiv	410.91 ± 8.29	373.19 ± 3.64	111.38 ± 1.95
ogbn-products	OOM	OOM	8709.16 ± 55.00

(b) Impact on Efficiency

dataset(↓) / method(→)	baseline	batching	batching+caching
CiteSeer	2.52 ± 0.27	2.45 ± 0.19	2.42 ± 0.37
Amazon_Photos	1.06 ± 0.01	1.06 ± 0.02	1.12 ± 0.02
Amazon_Computers	1.29 ± 0.05	1.15 ± 0.01	1.14 ± 0.01
Cora	6.81 ± 1.07	8.34 ± 1.12	7.96 ± 0.59
PubMed	1.17 ± 0.00	1.16 ± 0.00	1.17 ± 0.00
Coauthor_CS	1.10 ± 0.01	1.15 ± 0.01	1.14 ± 0.01
Coauthor_Physics	0.97 ± 0.01	1.01 ± 0.03	0.99 ± 0.01
Flickr	4.23 ± 0.07	4.23 ± 0.04	4.24 ± 0.03
ogbn-arxiv	7.07 ± 0.05	7.28 ± 0.06	6.91 ± 0.01
ogbn-products	OOM	OOM	1.86 ± 0.06

D.2 Overall Results

In Figure D7, we provide a plot of all the different methods discussed in this work for each dataset across different values of α. If applicable, for each method, we show the best-performing version, e.g., APS with randomization vs without and CFGNN with APS training ('cfgnn_aps') rather than TPS training ('cfgnn_orig'). We present the results for $k = 5$ for the different NAPS variations, since almost all datasets – except for CiteSeer and ogbn-arxiv – achieved their best efficiencies at or before that point.

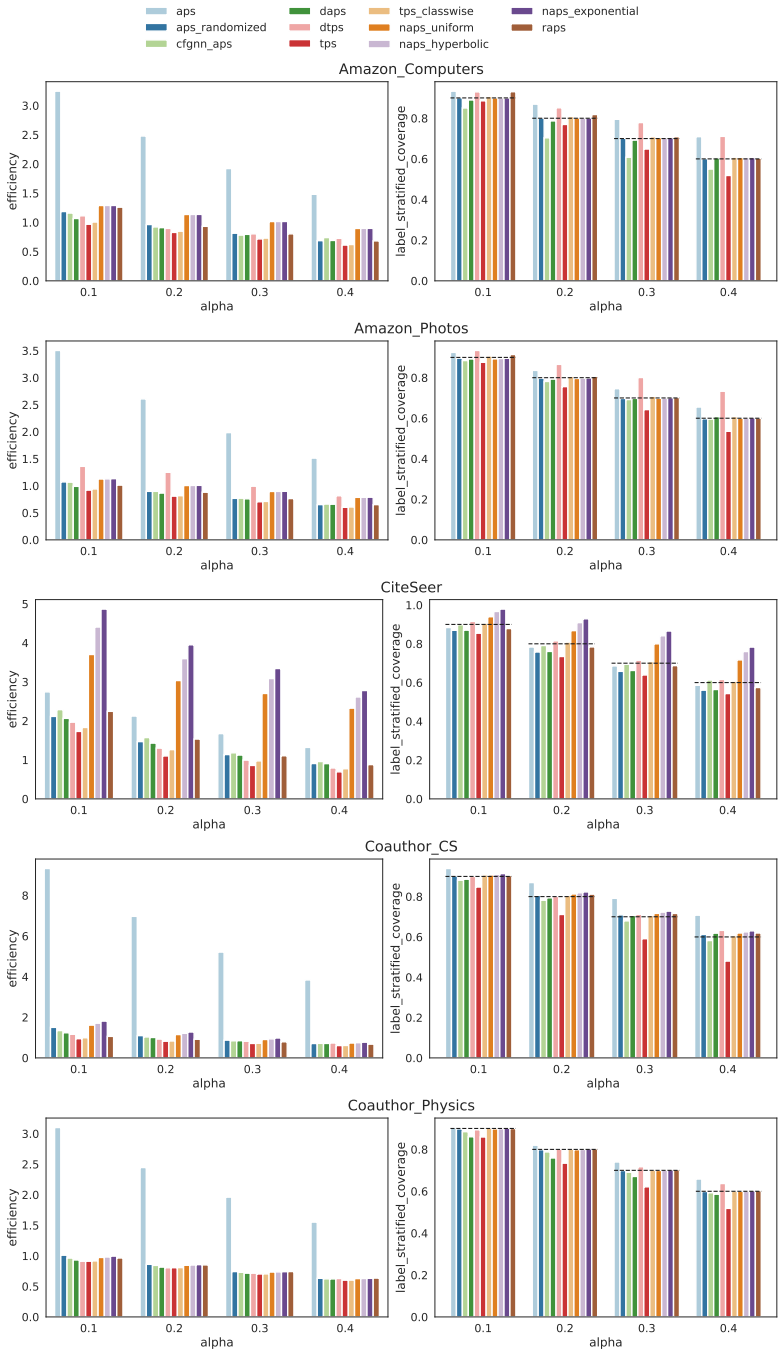


Figure D7: Plots for efficiency vs α for all the major methods (with best parameters) across all the datasets (cont.). Among the baseline methods, TPS consistently has the best efficiency. The results are for the FS split style. The dashed black line indicates the desired label stratified coverage.

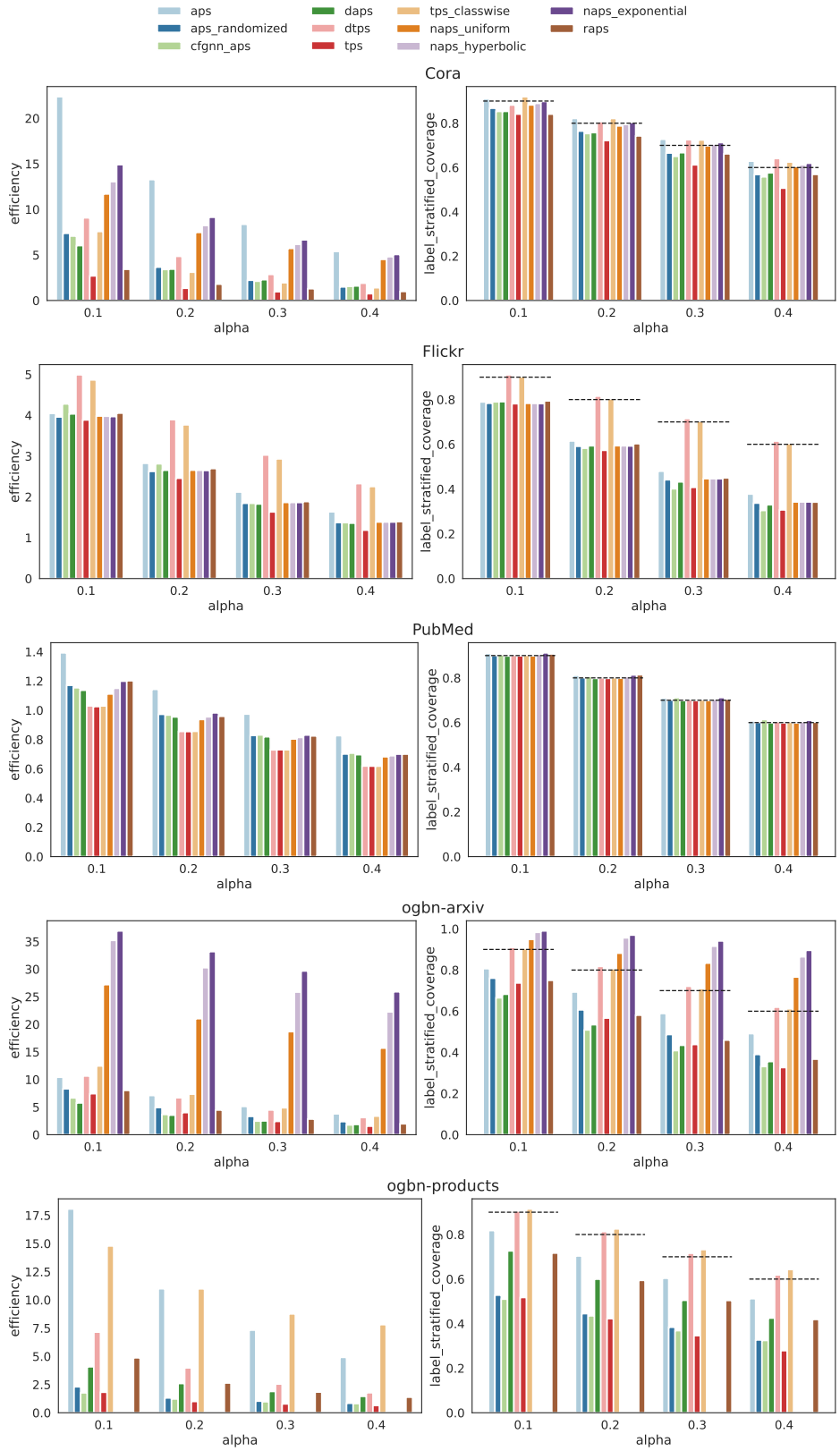


Figure D7: Plots for efficiency vs α for all the major methods (with best parameters) across all the datasets (cont.). Among the baseline methods, TPS consistently has the best efficiency. Results are for the FS split style. The dashed black line indicates the desired label stratified coverage. NAPS results for ogbn-products dataset are omitted due to size and lack of data points with existing hardware.