What If Without the Conformal Prediction?

Jing LIANG
Industrial Engineering and Decision Analytics
Hong Kong University of Science and Technology
Student id : 21266188
MATH 5472
jliangcd@connect.ust.hk

Abstract

Black-box predictive models are increasingly used in domains where reliable uncertainty quantification is essential. Conformal prediction provides a model-agnostic and distribution-free mechanism for constructing finite-sample valid prediction sets through score-based calibration under exchangeability. This paper examines what would be missing from the uncertainty quantification landscape in the absence of conformal prediction. The discussion outlines its core principles, representative variants, and theoretical structure, and contrasts these with several alternative approaches—including Bayesian inference, bootstrap resampling, and quantile regression. Empirical comparisons on synthetic and real datasets illustrate that these alternatives, while useful in specific settings, do not jointly offer finite-sample distribution-free validity and broad model compatibility. The analysis highlights the distinctive conceptual insights introduced by conformal prediction, as well as its practical value across diverse predictive applications.

1 Introduction

With the growing use of machine learning in risk-sensitive domains such as medical diagnosis and financial risk management, aggregate metrics such as average accuracy are no longer sufficient. Conventional learning algorithms such as neural networks, support vector machines, and decision trees typically provide only point predictions and do not offer calibrated measures of uncertainty. In practice, decision-makers require not only accurate predictions but also a principled assessment of their reliability.

Conformal prediction addresses this need by constructing, for each new input X_{n+1}, a prediction set

$$C_\alpha(X_{n+1}) \subseteq \mathcal{Y},$$

where $\mathcal{Y}$ denotes the label space and $\alpha \in (0, 1)$ is the target miscoverage level. The goal is to ensure the finite-sample, distribution-free validity guarantee

$$\mathbb{P}(Y_{n+1} \in C_\alpha(X_{n+1})) \geq 1 - \alpha,$$

for the unknown true label Y_{n+1}, assuming only that the data $(X_i, Y_i)_{i=1}^{n+1}$ are exchangeable.

To achieve this, conformal prediction defines a *nonconformity score*

$$V_i = V(X_i, Y_i), \qquad i = 1, \ldots, n,$$

which measures how atypical an example (X_i, Y_i) is relative to the others. A candidate label y for the new input X_{n+1} is evaluated via its test-time score $V(X_{n+1}, y)$ and included in the prediction set whenever it does not exceed the appropriate empirical quantile of the calibration scores.

If conformal prediction did not exist, what we would lose is not merely an algorithm but a critical meta-methodology: a systematic framework that decouples uncertainty quantification from model

design through the construction of nonconformity scores. The impact would extend beyond the disappearance of a single technique, altering the fundamental paradigm for developing uncertainty-aware predictive systems.

The source code for all experiments is publicly available at: `https://github.com/Revelyn-Jing/HKUST-MATH-5472/tree/main/Project`.

2 Background and Need

With the rapid development of artificial intelligence and its widespread deployment across society, both industry and government agencies are increasingly adopting AI-based decision-support systems and agentic task-execution systems. These systems are typically built upon predictive machine learning models, which generate predictions or recommended actions through supervised learning tasks[1].

At the same time, AI technologies are accelerating their penetration into high-risk domains such as medical diagnosis, biometrics, face recognition, nuclear fusion, and industrial diagnostics[2]. In these settings, model failures can lead to severe medical, ethical, economic, or safety consequences. As a result, society and regulatory bodies have raised expectations regarding the reliability, fairness, transparency, and accountability of AI systems, thereby driving the development of Trustworthy AI (TAI).

Within the TAI framework, uncertainty quantification (UQ) plays a central role in ensuring reliability. The goal of UQ is to characterize and evaluate various sources of uncertainty in predictive models, providing reliable and calibrated confidence information[3]. This enables decision-makers to understand the trust boundaries of model outputs and make robust decisions accordingly.

Traditional UQ approaches face structural limitations that constrain their effectiveness in high-risk applications. Bayesian methods, such as Bayesian neural networks, provide probabilistic predictions but rely heavily on strong prior assumptions and the correctness of the model specification; inference is also computationally expensive for large-scale models. Frequentist methods, such as bootstrapping or dropout-based UQ, estimate uncertainty through resampling or approximate inference but typically lack distribution-free, finite-sample guarantees. They also become unstable under distribution shift and incur high computational costs.

Consequently, traditional UQ methods generally fail to provide instance-wise, distribution-free guarantees of validity.

Conformal prediction, however, offers a systematic solution. It provides verifiable, finite-sample validity guarantees for each individual instance without requiring any distributional assumptions, enabling precise quantification of predictive uncertainty.

3 Conformal Prediction

3.1 Historical Development

The theoretical origins of conformal prediction were established in the late 1990s and early 2000s by Vladimir Vovk, Alex Gammerman, and Glenn Shafer[4, 5, 6]. Unlike traditional parametric statistics or Bayesian inference, the methodology arises from the frameworks of algorithmic randomness and game-theoretic probability[6, 7]. Its objective is to construct distribution-free predictive guarantees for machine learning that remain valid for finite samples. Within this paradigm, predictions produce sets of plausible labels rather than single point estimates, and the size of these sets adapts to model uncertainty, thereby providing a direct quantification of predictive reliability.

The foundational contribution to the field is the work that establishes the theoretical basis of conformal prediction [6]. A more recent overview is provided by the monograph "Conformal prediction for reliable machine learning: theory, adaptations and applications" [2].

Recent developments in conformal prediction have focused on three main areas: improving the validity concept, enhancing computational efficiency, and expanding application scope[8].

3.2 Conformal Prediction Framework

Vovk's work introduces full conformal prediction or transductive conformal prediction [6]. The central idea is that, given a training set and a new test instance, each possible label for the test instance is tentatively paired with the training data to compute a nonconformity score, and the label is included in the prediction set based on its rank among these scores. Because this procedure becomes computationally expensive as the dataset grows, subsequent research has proposed more efficient variants, most notably split conformal prediction[9], which is now widely used. This subsection will focus on split conformal prediction and its technical details, and illustrate the overall conformal framework.

3.2.1 Split Conformal Prediction

Recall the targert of conformal prediction : given i.i.d. data $(X_i, Y_i)_{i=1}^n$ with $X_i \in \mathcal{X}$, $Y_i \in \mathcal{Y}$ and a target miscoverage level $\alpha \in (0, 1)$, conformal prediction attempts to construct, for each new input X_{n+1}, a prediction set

$$C_\alpha(X_{n+1}) \subseteq \mathcal{Y}$$

such that the coverage rate guarantee

$$\mathbb{P}(Y_{n+1} \in C_\alpha(X_{n+1})) \geq 1 - \alpha$$

holds for any underlying distribution, assuming only exchangeability (i.i.d.) of $(X_i, Y_i)_{i=1}^{n+1}$.

Split conformal prediction achieves this by partitioning the data into two parts: a training subset for fitting an arbitrary base model $\hat{f}$, and a calibration subset for quantifying the model's residual uncertainty. The calibration samples provide an empirical distribution of nonconformity scores, whose appropriate quantile is then used to expand the model's raw prediction into a prediction set with controlled miscoverage. In this way, split CP (Conformal Prediction) converts any point predictor into a distribution-free, finite-sample valid prediction region.

Algorithm 1 Split Conformal Prediction

Require: Dataset $\{(X_i, Y_i)\}_{i=1}^n$; base learner $\mathcal{A}$; nonconformity function V_f; miscoverage level $\alpha \in (0, 1)$

Ensure: Mapping $x \mapsto C_\alpha(x)$

1: Randomly split the index set $\{1, \ldots, n\}$ into disjoint subsets: training indices I_{train} and calibration indices I_{cal}

2: Fit the prediction model on the training set:

$$f \leftarrow \mathcal{A}\big(\{(X_i, Y_i) : i \in I_{\text{train}}\}\big)$$

3: For each $i \in I_{\text{cal}}$, compute the nonconformity score:

$$V_i \leftarrow V_f(X_i, Y_i), \qquad i \in I_{\text{cal}}$$

4: Let $m \leftarrow |I_{\text{cal}}|$ and $k \leftarrow \lceil (1 - \alpha)(m + 1) \rceil$. Define $\widehat{Q}_{1-\alpha}$ as the k-th smallest value in $\{V_i : i \in I_{\text{cal}}\}$

5: For any new input $x \in \mathcal{X}$, define the prediction set:

$$C_\alpha(x) \leftarrow \big\{y \in \mathcal{Y} : V_f(x, y) \leq \widehat{Q}_{1-\alpha}\big\}$$

Moreover, the split conformal prediction algorithm comes with a rigorous theoretical coverage guarantee.

Theorem 1 (Conformal coverage guarantee for Split CP[10])**.** *Suppose the observations $(X_i, Y_i)_{i=1,\ldots,n}$ and the test pair $(X_{\text{test}}, Y_{\text{test}})$ are i.i.d. and let $\widehat{Q}_{1-\alpha}$ and $C_\alpha(X_{\text{test}})$ be defined as in Algorithm 1. Then the following coverage guarantee holds:*

$$\mathbb{P}(Y_{\text{test}} \in C_\alpha(X_{\text{test}})) \;\geq\; 1 - \alpha.$$

Notice that this theorem shows that the coverage guarantee of split CP is not asymptotic but holds non-asymptotically for any sample size.

Method	Objective	Core Idea
CV+ / Jackknife+ [11]	Improve statistical efficiency and avoid overly wide intervals from data splitting	Use cross-validation or leave-one-out residuals to obtain more stable error estimates and produce tighter prediction sets.
Weighted CP [12]	Achieve valid coverage under covariate shift or population heterogeneity	Rank scores using importance weights that reflect differences between the target and calibration distributions.
Distributional CP [13]	Handle complex, multimodal, or asymmetric conditional distributions	Construct scores based on estimated CDF values (e.g., $\|F(x, y) - 0.5\|$) to form high-density prediction regions.
CP for hierarchical structure data [14]	Handle data with group-level heterogeneity and hierarchical sampling structures	Replace standard exchangeability with hierarchical exchangeability and construct prediction sets using pooled CDF, double conformal, or subsampling-based aggregation across groups.
Adaptive Prediction Sets[15, 16]	Improve fairness of coverage across classes and sample difficulty	Use cumulative softmax probability to define an ordered score and adjust prediction set size according to sample difficulty.
Conformalized Quantile Regression[17]	Construct more adaptive regression intervals	Measure deviation from upper and lower quantiles and calibrate these to form intervals adaptive to conditional distribution shape.

3.2.2 Variancs and Shared Framework

In recent years, conformal prediction has attracted considerable attention and has advanced rapidly. In the following, we provide several representative variants and explain how the components that vary, as well as those that remain invariant, help reveal the essential structure of the conformal prediction framework.

From the table, we can see that although recent developments in conformal prediction have produced many variants, their innovations mainly focus on how the score is designed, what is calibrated, or how the exchangeability assumption is extended. These represent the variant component of the CP framework. Different methods modify the source of the scores, the geometric form of the scores, or model the underlying data structure and distribution shift to better adapt to new tasks, data structures, or distributional settings. These changes improve the applicability of CP in complex, high-dimensional, heteroscedastic, multimodal, and non-i.i.d. scenarios, allowing the framework to remain flexible across different use cases.

In contrast, all methods preserve the invariant component of conformal prediction. No matter how the method varies, the validity mechanism follows the same basic design: construct a score that measures the discrepancy between the prediction and the ground truth, compute its empirical distribution on the calibration set, and rely on exchangeability to obtain validity guarantees.

4 Key Insights and Benefit of Conformal Prediction

From Section 3.2.2, it is clear that although conformal prediction admits many variants, they all follow the same fundamental framework, whose core lies in the construction of the score function and the associated calibration mechanism.

The distribution-free property originates from the design of the score: model adequacy is entirely captured by the nonconformity measure $A(\cdot)$, while the coverage guarantee is provided solely by rank-based calibration under exchangeability. These two components are decoupled, allowing any black-box predictor to obtain formal validity without modifying its internal structure.

Abstractly, a score can be viewed as $1 -$ p-value[6]. Consequently, the remaining steps of conformal prediction amount to determining whether a new input behaves like a sample drawn from the empirical distribution of scores, and forming the prediction set accordingly. This procedure relies only on exchangeability to obtain marginal coverage guarantees, making conformal prediction one of the few methods that retain valid coverage even in finite-sample regimes.

In an era of increasingly complex models and highly specialized tasks, this structural and portable decoupling principle becomes especially important.

5 Alternative Methods

A wide range of uncertainty quantification techniques have been developed across statistics and machine learning. These approaches differ in their underlying assumptions, the form of guarantees they provide, and their suitability for various application settings. Table summarizes the main methodological families commonly used as competitors to conformal prediction.

Method Family	Key Idea	Representative Competitors
Classical parametric prediction intervals	Assume a fully specified parametric model, typically linear with Gaussian noise	Gaussian linear model (t-interval) [18]
Semi- / Non-parametric prediction bands	Assume a functional form for estimation, such as conditional means or quantiles, without making distributional assumptions	Quantile regression [19]
Bayesian framework	Characterize uncertainty through the posterior predictive distribution	Bayesian neural networks[20]; Empirical Bayesian [21]
PAC-learning	Provide worst-case upper bounds on prediction error over the hypothesis class	Vapnik–Chervonenkis bounds[22]; Littlestone–Warmuth bounds[23]
Hold-out and resampling-based methods	Estimate uncertainty empirically via data splitting or repeated resampling	cross-validation[24]; bootstrap[25]

None of these approaches simultaneously achieve the combination of distribution-free operation, finite-sample validity, and instance-wise guarantees that conformal prediction provides. Nevertheless, each method family has its own advantages and suitable application contexts. The remainder of this section outlines the several representative competitors and contrasts them with conformal prediction to highlight its distinct value.

5.1 Bayesian Framework

Bayesian methods can yield efficient and well-calibrated intervals when the likelihood and prior are correctly specified, and they naturally account for parameter uncertainty through the posterior. However, their guarantees depends on the prior.Under misspecification, Bayesian credible intervals often suffer from severe undercoverage despite appearing narrow, and the Bayesian framework provides no coverage guarantee.

In contrast, conformal prediction is agnostic to the data-generating mechanism and avoids reliance on priors or likelihood models. Its intervals remain finite-sample valid under the sole assumption of exchangeability, maintaining stable coverage even when the predictive model is misspecified. The trade-off is efficiency: conformal intervals are typically wider, particularly at high confidence levels or with limited calibration data, because they must guard against worst-case residual behavior.

In this section, the comparisons are organized across three settings. First, full Bayesian and empirical Bayesian approaches are examined in the context of ridge regression, contrasting full posterior integration with the empirical Bayes approximation that estimates prior hyperparameters from data. Second, the analysis is extended to a Bayesian neural network on the UCI ENB2012 Energy dataset[26], using an MLP backbone to isolate the effect of the Bayesian treatment itself. Together, these experiments provide a unified view of how Bayesian uncertainty behaves in both linear and nonlinear models, and how it compares with distribution-free conformal prediction.

5.1.1 Full Bayesian and Empirical Bayesian

This experiment based on the general setup of [27] and compare the interval performance of Full Bayesian Ridge Regression and Split Conformal Prediction using generated data. The input features $x \in [-10, 10]^5$ are independently sampled from a uniform distribution, and the weight vector w is drawn from a standard normal prior $\mathcal{N}(0, I)$. Labels are generated according to

$$y = x^\top w + \varepsilon, \qquad \varepsilon \sim \mathcal{N}(0, 1).$$

A single training–test split (100 training points and 100 test points) is fixed using a predetermined random seed, and we evaluate 20 significance levels $\alpha \in [0.01, 0.99]$ on this fixed dataset. Each (a, α) configuration is repeated 10 times: Bayesian Ridge uses the same dataset in all repetitions, whereas Split Conformal varies only the internal random seed that determines the calibration split.

Bayesian Ridge Regression applies

$$\Sigma = (X^\top X + aI)^{-1}, \qquad \mu = \Sigma X^\top y,$$

and constructs Gaussian predictive intervals via

$$y \mid \mathcal{D} \sim \mathcal{N}(x\mu,\ 1 + x\Sigma x)\,,$$

evaluating the effect of prior misspecification with $a \in \{1, 1000, 10000\}$.

Split Conformal Prediction uses the same Ridge predictor but randomly partitions the training data into 80% proper training and 20% calibration subsets, forming distribution-free intervals from the finite-sample quantile of the calibration residuals.

Across all confidence $(1 - \alpha)$, this experimentevaluate both the empirical coverage and the mean interval width, thereby comparing the sensitivity of Bayesian intervals to prior misspecification against the distribution-free robustness of Split Conformal Prediction.

In addition, since Full Bayesian methods are no longer the preferred choice after the introduction of Empirical Bayes, the experiment also incorporates an Empirical Bayesian variant to provide a more modern and practically relevant comparison.

As shown in figure 1, when the prior is correctly specified ($a = 1$), Bayesian Ridge Regression produces valid prediction intervals. As a increases and the prior becomes misspecified, the Bayesian intervals deteriorate sharply: their coverage collapses while the interval width remains almost unchanged. In contrast, the width of the Split Conformal intervals increases with the confidence level and maintains valid coverage across all settings.

A similar pattern appears in the empirical Bayesian variant in 2. With the correct prior, it achieves good coverage with intervals narrower than those of the full Bayesian method. Under severe prior misspecification, the empirical Bayesian intervals widen and display some degree of robustness, yet their coverage still remains below that of Split Conformal Prediction, can not reach the desired level.

5.1.2 Bayesian Neural Network

Recent years have introduced improved Bayesian neural network (BNN) methods [20]. Their core idea is simple: instead of learning a single set of network weights, a BNN places a prior over the weights and learns a posterior distribution. Sampling from this posterior gives multiple plausible models, allowing the prediction uncertainty to reflect both model uncertainty (epistemic) and data noise (aleatoric).

This experiment applies such a BNN to the UCI ENB2012 Energy dataset[26] , using the same MLP backbone as in the CP baseline.

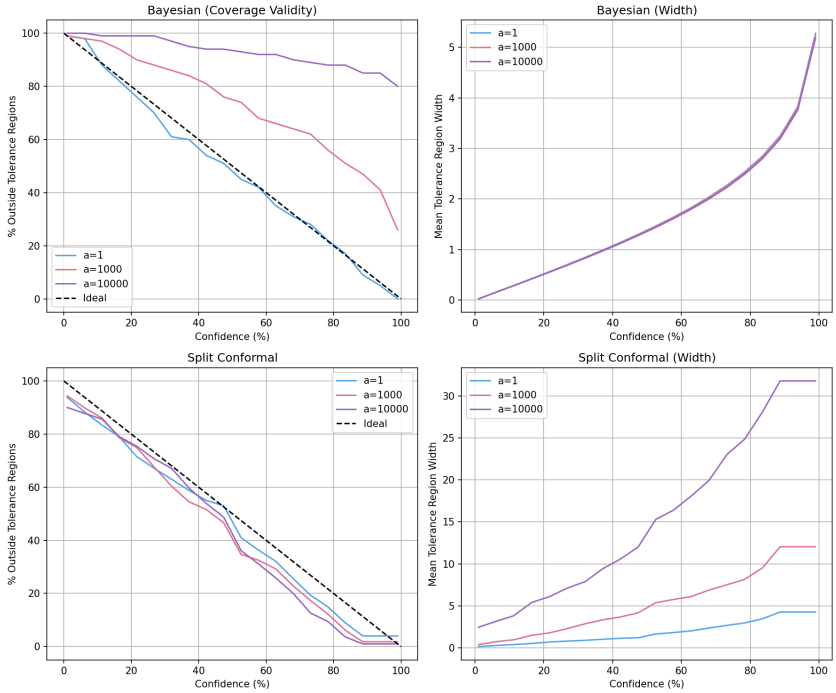

Figure 1: Full Bayesian RR and Split Conformal prediction for RR.

The UCI ENB2012 Energy dataset contains 768 simulated building designs, each described by eight physical and geometric features such as wall area, roof area, glazing ratio, and overall height. These features are used to predict two real-valued heating and cooling load indicators. The dataset is low-dimensional, noise-modest, and exhibits smooth nonlinear structure, making it a standard benchmark for regression and uncertainty quantification.

In this experiment, both the BNN and Split CP use the same base learner: an MLP with input dimension 8, one hidden layer of width 64 with ReLU. The BNN adopts a mean–field variational formulation (Bayes-by-Backprop style)[28], placing independent Gaussian priors on all weights and learning Gaussian posterior parameters via an ELBO objective with reparameterized sampling.

and a 1–dimensional output. The BNN employs Gaussian priors $w, b \sim \mathcal{N}(0, 1)$ and is trained for 100 epochs using Adam (10^{-3} learning rate, batch size 64). Posterior prediction is approximated with $T = 200$ Monte Carlo samples, yielding predictive mean μ and standard deviation σ (after inverting the response scaling). The $(1 - \alpha)$ interval is $\mu \pm z_{1-\alpha/2}\,\sigma$.

For Split CP, the same MLP is trained on 70% of the data, with the remaining 30% used for calibration. Absolute residuals are sorted, and the conformal radius for level α is taken at index $\lceil (n_{\text{cal}}+1)(1-\alpha) \rceil$. CP intervals are obtained by adding/subtracting this radius to the MLP predictions on the raw scale.

Figure 3 shows that, under the same MLP base model, BNN intervals deviate sharply from the ideal diagonal: the fraction of points outside the intervals is far above nominal, reflecting severely underestimated posterior variance and strong overconfidence. The intervals remain narrow across confidence levels, consistent with the observed undercoverage.

By contrast, Split Conformal closely follows the ideal line, with interval width increasing predictably with confidence. Although wider, these intervals achieve near-nominal coverage.

5.2 Boostrap Resampling

Bootstrap is a non-parametric resampling method used to assess the variability, bias, and confidence intervals of statistical estimators. It treats the observed data as an empirical distribution and repeatedly draws samples with replacement to mimic how the estimator would behave if the experiment

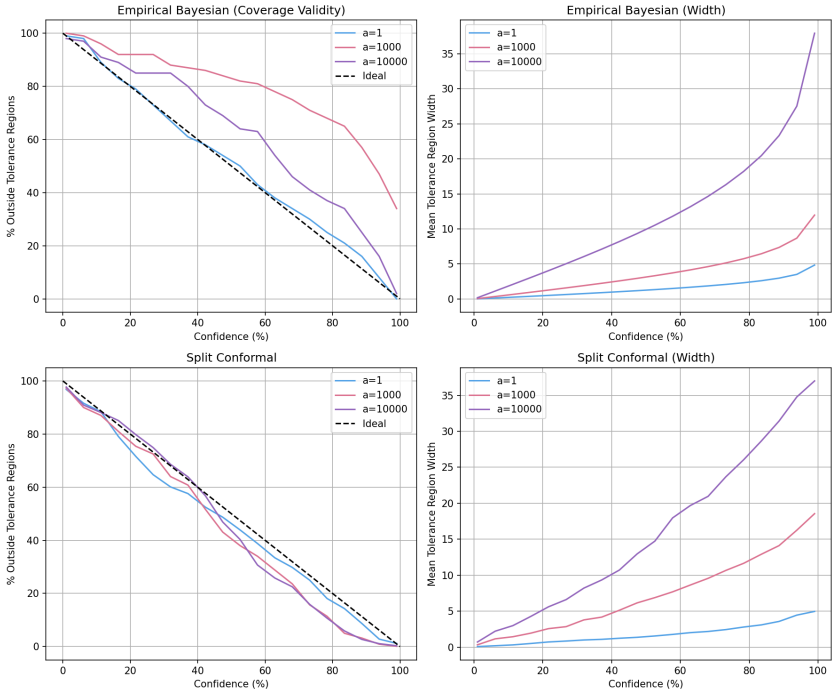

Figure 2: Empirical Bayesian RR and Split Conformal prediction for RR.

were replicated. This mechanism does not rely on explicit distributional assumptions, making it conceptually similar to conformal prediction in its distribution-free spirit.

However, bootstrap also has notable drawbacks. Classical bootstrap requires retraining the model and re-evaluating predictions for every resampled dataset; achieving stable estimates typically demands a large number of resamples, leading to significant computational cost.

The experiment evaluates the finite-sample behavior of Bootstrap prediction intervals and Split CP using the ENB2012 Heating Load dataset. The first eight features are used as inputs, and the heating load target is standardized for training stability. The dataset is randomly split into training (80%) and test (20%).

A common base learner is adopted for both methods: a two-layer MLP (64 hidden units, ReLU activations) trained for 200 epochs with Adam. Split CP trains a single model and uses calibration residuals to obtain a valid finite-sample quantile. In contrast, Bootstrap requires training (B) independent models on resampled datasets and aggregates their test predictions to form empirical $((\alpha/2, 1-\alpha/2))$ quantiles.

The experiment vary $(B \in 5, 10, 20, 100, 300, 600)$ to examine how computation time increases and how coverage behaves across resampling budgets. For each method, record empirical coverage, average interval length (after inverse scaling), and wall-clock computation time.

Table 1 reports the Bootstrap prediction interval results and the corresponding total computation times. Split Conformal Prediction, by contrast, requires only a single model fit and achieves a total runtime of approximately 0.68 s with empirical coverage 0.89. This setup isolates the computational burden of Bootstrap and assesses whether increasing B improves its coverage.

Bootstrap is not categorically inferior to conformal prediction. When the computational cost of increasing the resampling budget B is acceptable, bootstrap can produce smaller prediction sets, yielding more efficient predictions.

The experiment uses a misspecified regression setting. Covariates $X \in [-1, 1]^d$ are sampled i.i.d. from a uniform distribution, and the response is generated as

$$Y = X^\top \beta \ + \ 2X_1^2 \ + \ X_2^2 \ + \ \varepsilon, \qquad \varepsilon \sim \mathcal{N}(0, \sigma^2),$$

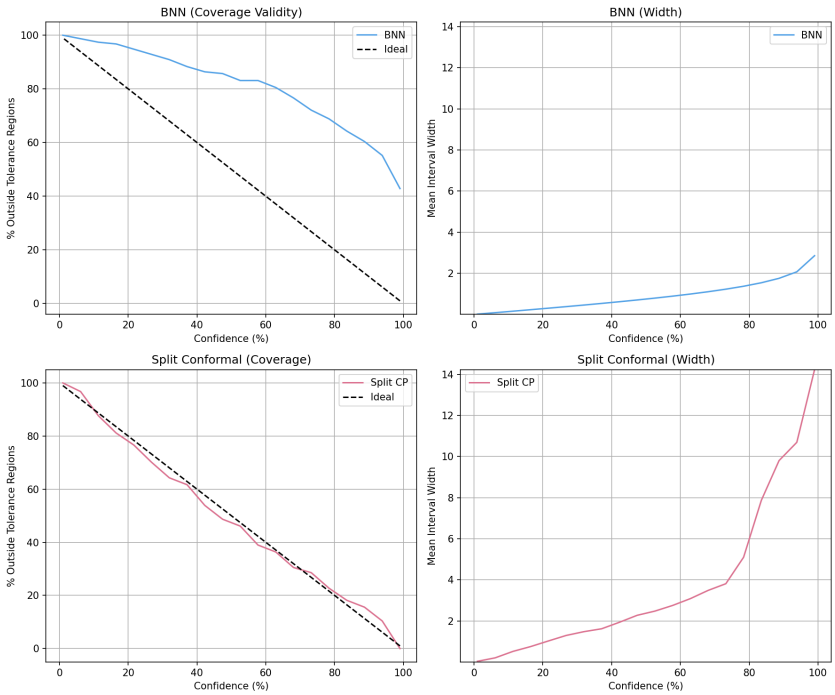

Figure 3: BNN and Split CP on UCI ENB2012.

B	Coverage	Avg. length	Total time (s)
5	0.448	2.09	3.00
10	0.604	2.37	6.06
20	0.604	2.79	12.03
100	0.688	3.10	61.36
300	0.747	3.23	181.74
600	0.753	3.23	359.54

Table 1: Bootstrap : larger B increases time but still lacks coverage.

so the true model is nonlinear while a linear model is fitted.

Split conformal prediction trains a linear regressor on 90% of the training data and uses the remaining 10% for calibration. The bootstrap method fits a linear model on the full training set and constructs predictive intervals by residual resampling with B bootstrap replicates.

Table 2 illustrates this trade-off. This aligns with the view that bootstrap is not categorically inferior to conformal prediction; when the computational cost of a sufficiently large resampling budget B is acceptable, bootstrap can produce smaller and more efficient prediction sets.bootstrap shows greater variability but yields consistently shorter intervals.

5.3 Quantile Regression

Quantile regression directly learns conditional quantile functions by minimizing asymmetrically weighted absolute deviations, capturing systematic variation across quantile levels [19]. This allows it to reveal heteroscedasticity, tail behavior, and distributional asymmetry, making it highly attractive in many economic and financial applications.

The previous experimental setup is retained, using the ENB2012 Heating Load dataset with the first eight features as predictors. Under this setting and with a shared MLP base learner, Split CP attains coverage 0.896 with an average length of 2.35, close to the nominal 90% level. Quantile regression, despite using the same architecture, yields lower coverage 0.799 and wider intervals 3.59.

Method	Coverage	Avg. Length
Split Conformal	0.927 ± 0.063	4.825 ± 1.145
Bootstrap	0.884 ± 0.342	3.911 ± 0.342

Table 2: Coverage and efficiency under model misspecification.

A key limitation relative to CP is that the quantile level τ is coupled with model training. Specifying different τ values (for example, 0.05, 0.5, 0.95) requires training a separate model for each. To obtain an interval $[q_{\alpha/2}(x),\ q_{1-\alpha/2}(x)]$, two quantile regression models must be fitted. This is inherent to the quantile-regression objective, as the loss function depends explicitly on τ, so each τ yields a different optimizer. In contrast, conformal prediction is more flexible: it does not require retraining the base model when the desired coverage level changes.

More importantly, quantile regression and conformal prediction are substantially complementary. Quantile regression directly fits the conditional lower and upper quantiles and therefore automatically captures conditional heteroscedasticity (interval width varying with x) and local density variation reflected in quantile spacing.

By contrast, CP—unless a specially designed score is used to encode local variability (e.g., Adaptive Prediction Sets [16]) learns only marginal information from residuals. Its validity guarantee is marginal, and the resulting intervals may lose efficiency on data with strong heteroscedasticity or asymmetric conditional distributions.

This motivates recent methods that combine the two approaches: quantile regression provides a flexible estimate of the conditional distribution, while conformalization adds finite-sample marginal validity, yielding interval estimators that are both adaptive and distribution-free.

Sesia and Candès [29] present a systematic comparison of methods that integrate conformal prediction with quantile regression. Their study examines three variants—CQR, CQR-m, and CQR-r—distinguished by their constructions of conformity scores, and establishes that all variants are asymptotically efficient, with prediction bands converging to the oracle conditional quantile interval under mild conditions.

Through experiments on synthetic and real datasets, the authors show that these procedures maintain finite-sample validity, with CQR typically producing the narrowest intervals. They also provide practical recommendations for data splitting, indicating that allocating approximately 70%–90% of the sample to quantile-regression training yields a favorable balance between adaptivity and stability.

6 Implications of a World Without Conformal Prediction

As discussed in Section 4, the core contribution of CP lies in its decoupling principle, whose conceptual influence extends beyond a single algorithm. Without conformal prediction, the field would also lose a clear illustration of this modular design idea, which is potentially useful not only for UQ but for other areas of machine learning as well.

This monograph by Balasubramanian, Ho and Vovk[2] surveys a large number of applications, most of which crucially exploit that conformal prediction provides provably valid and well-calibrated prediction regions under nothing more than an exchangeability assumption, outputs set-valued predictions with user-specified confidence levels whose empirical error frequencies match the nominal significance, and can be used as a model-agnostic wrapper around virtually any underlying classifier or regressor (including SVMs, neural networks, k-NN, ridge regression, etc.).

A particularly insightful observation, originally articulated by Vovk in an invited talk, concerns the potential role of conformal prediction in grounding large language models. The key argument is that contemporary LLMs lack an external notion of truth: much of what they produce reflects statistical regularities of language rather than factual grounding. From this viewpoint, conformal prediction provides a principled mechanism for injecting verifiable ground-truth signals into models whose training is otherwise purely text-driven.

This perspective has already inspired concrete methodological developments. For example, Cherian et al. [30] propose enhanced conformal procedures for evaluating and calibrating LLM outputs,

illustrating how CP can furnish validity guarantees even in settings where conventional ground truth is scarce or difficult to access.

Building on this perspective, the experiments in Section 5 show that, in the considered benchmarks, none of the competing methods (Bayesian credible intervals, bootstrap, or quantile regression) simultaneously attains such distribution-free finite-sample coverage and algorithm-agnostic deployability, so they cannot be regarded as full substitutes for conformal prediction.

Without conformal prediction, the UQ landscape would lack a universally applicable and theoretically rigorous framework; applying alternative methods would require researchers and practitioners to verify and tune model-specific assumptions and hyperparameters on a case-by-case basis, often without any distribution-free finite-sample guarantees. This would raise the technical barrier for deploying reliable UQ across domains and could slow the adoption of trustworthy machine-learning systems, while increasing the risk of calibration failures in real applications.

7 Conclusion

Conformal prediction offers more than a set of algorithms; it provides a structural template for uncertainty quantification that separates model construction from validity calibration. This decoupling principle enables a wide range of predictors to be equipped with distribution-free, finite-sample guarantees through a unified mechanism based on score ranking under exchangeability.

The comparative analysis in this paper shows that competing approaches each address only part of this objective. Bayesian methods provide efficient intervals when correctly specified but lack robustness under misspecification. Bootstrap resampling is flexible but computationally demanding and unstable in small samples. Quantile regression captures conditional structure but does not inherently guarantee empirical coverage. None of these methods simultaneously deliver model-agnostic deployment, finite-sample validity, and freedom from distributional assumptions.

In a hypothetical landscape without conformal prediction, practitioners would have to rely more heavily on methods tied to model-specific assumptions and case-by-case diagnostics, without access to a general framework offering distribution-free validity guarantees. The lack of such a unifying principle could make it harder to develop and analyze reliable machine-learning procedures.

From this perspective, conformal prediction can be viewed not only as a practically useful tool, but also as a methodological framework that has provided a coherent way to articulate and extend ideas about validity and uncertainty beyond its original formulation.

References

[1] Yuzhou Qian, Keng L. Siau, and Fiona F. Nah. Societal impacts of artificial intelligence: Ethical, legal, and governance issues. *Societal Impacts*, 3, 2024. doi: 10.1016/j.socimp.2024.100040.

[2] Vineeth Balasubramanian, Shen-Shyang Ho, and Vladimir Vovk. *Conformal prediction for reliable machine learning: theory, adaptations and applications*. Newnes, 2014.

[3] Ralph C Smith. *Uncertainty quantification: theory, implementation, and applications*. SIAM, 2024.

[4] C. Saunders, A. Gammerman, and V. Vovk. Transduction with confidence and credibility. In *Proceedings of the Sixteenth International Joint Conference on Artificial Intelligence (IJCAI '99)*, pages 722–726, 1999.

[5] Harris Papadopoulos, Kostas Proedrou, Volodya Vovk, and Alex Gammerman. Inductive confidence machines for regression. In *Proceedings of the European Conference on Machine Learning (ECML 2002)*, pages 345–356, 2002.

[6] Vladimir Vovk, Alex Gammerman, and Glenn Shafer. *Algorithmic Learning in a Random World*. Springer, 2005.

[7] Alexander Gammerman and Vladimir Vovk. Kolmogorov complexity: Sources, theory and applications. *Computer Journal*, 42(4), 1999.

[8] Matteo Fontana, Gianluca Zeni, and Simone Vantini. Conformal prediction: a unified review of theory and new challenges. *Bernoulli*, 29(1):1–23, 2023.

[9] Jing Lei, Max G'Sell, Alessandro Rinaldo, Ryan J Tibshirani, and Larry Wasserman. Distribution-free predictive inference for regression. *Journal of the American Statistical Association*, 113(523):1094–1111, 2018.

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

[11] Rina Foygel Barber, Emmanuel J. Candes, Aaditya Ramdas, and Ryan J. Tibshirani. Predictive inference with the jackknife+, 2020.

[12] Ryan J. Tibshirani, Rina Foygel Barber, Emmanuel J. Candès, and Aaditya Ramdas. Conformal prediction under covariate shift. In *Proceedings of the 33rd International Conference on Neural Information Processing Systems*, pages 2530–2540, Red Hook, NY, USA, 2019. Curran Associates Inc.

[13] Victor Chernozhukov, Kaspar Wüthrich, and Yinchu Zhu. Distributional conformal prediction. *Proceedings of the National Academy of Sciences*, 118(48):e2107794118, 2021. doi: 10.1073/pnas.2107794118.

[14] Robin Dunn, Larry Wasserman, and Aaditya Ramdas. Distribution-free prediction sets for two-layer hierarchical models. *Journal of the American Statistical Association*, 118(544):2491–2502, 2023.

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

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

[17] Yaniv Romano, Evan Patterson, and Emmanuel Candes. Conformalized quantile regression. *Advances in Neural Information Processing Systems*, 32, 2019.

[18] George Casella and Roger Berger. *Statistical inference*. Chapman and Hall/CRC, 2024.

[19] Roger Koenker and Kevin F Hallock. Quantile regression. *Journal of Economic Perspectives*, 15(4): 143–156, 2001.

[20] Laurent Valentin Jospin, Hamid Laga, Farid Boussaid, Wray Buntine, and Mohammed Bennamoun. Hands-on bayesian neural networks—a tutorial for deep learning users. *IEEE Computational Intelligence Magazine*, 17(2):29–48, 2022.

[21] Bradley Efron. *Large-Scale Inference: Empirical Bayes Methods for Estimation, Testing, and Prediction*. Cambridge University Press, Cambridge, 2010.

[22] Vladimir N Vapnik. An overview of statistical learning theory. *IEEE Transactions on Neural Networks*, 10 (5):988–999, 1999.

[23] N Cristianini and J Shawe-Taylor. An introduction to support vector machines and other kernel-based methods, 2000.

[24] Sylvain Arlot and Alain Celisse. A survey of cross-validation procedures for model selection. *Statistics Surveys*, 4, 2010. doi: 10.1214/09-ss054.

[25] Bradley Efron and Robert J Tibshirani. *An introduction to the bootstrap*. Chapman and Hall/CRC, 1994.

[26] Athanasios Tsanas and Angeliki Xifara. Energy efficiency. UCI Machine Learning Repository, 2012. doi:10.24432/C51307.

[27] Thomas Melluish, Craig Saunders, Ilia Nouretdinov, and Volodya Vovk. Comparing the bayes and typicalness frameworks. In *European Conference on Machine Learning*, pages 360–371, 2001.

[28] Charles Blundell, Julien Cornebise, Koray Kavukcuoglu, and Daan Wierstra. Weight uncertainty in neural network. In *International Conference on Machine Learning*, pages 1613–1622, 2015.

[29] Matteo Sesia and Emmanuel J Candès. A comparison of some conformal quantile regression methods. *Stat*, 9(1):e261, 2020.

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