PROCEEDINGS OF SCIENCE

Conformal Prediction for Reliable Uncertainty Quantification in Scientific AI Models: A HiggsML Signal-Strength Case Study

Luca Clissa,[a,b,*] **Leonardo Plini,**[c,d] **Daniele Bonacorsi**[a,b] **and Iacopo Vivarelli**[a,b]

[a] *Department of Physics and Astronomy "Augusto Righi", University of Bologna*

[b] *Istituto Nazionale di Fisica Nucleare, Sezione di Bologna*

[c] *Department of Computer, Control and Management Engineering "Antonio Ruberti" Sapienza University of Rome*

[d] *Istituto Nazionale di Fisica Nucleare, Laboratori Nazionali di Frascati*

E-mail: luca.clissa2@unibo.it

Reliable uncertainty quantification is essential in scientific applications, where predictive results must be accompanied by a transparent assessment of confidence rather than by point estimates alone. Among the large family of uncertainty-aware methods, Conformal Prediction (CP) is especially attractive because it is model-agnostic, can be applied post hoc on top of an already trained predictor, and offers finite-sample marginal coverage guarantees under mild assumptions. In this work we investigate the use of CP for signal-strength estimation in a high-energy physics setting based on the FAIR Universe HiggsML Uncertainty benchmark.

We propose a split-conformal workflow to distinguish Higgs signal from background events, combining predictions to obtain 1-σ (68.27%) confidence intervals for the signal strength. We compare several estimators and interval construction alternatives, and explore coverage and width properties of the obtained conformal intervals. Results indicate that CP attains empirical coverage close to the nominal level in these real-world-like settings, reaching competitive with a likelihood-based interval baseline. Notably, even when the underlying estimator is substantially biased, CP is able to shift the interval so as to recover the desired coverage. These findings candidate CP as a promising building block for trustworthy and reproducible scientific AI workflows.

International Symposium on Grids and Clouds (ISGC2026)
15–20 March, 2026
Academia Sinica, Taipei, Taiwan

*Speaker

© Copyright owned by the author(s) under the terms of the Creative Commons Attribution-NonCommercial-NoDerivatives 4.0 International License (CC BY-NC-ND 4.0) All rights for text and data mining, AI training, and similar technologies for commercial purposes, are reserved. ISSN 1824-8039 . Published by SISSA Medialab.

PoS(ISGC2026)024

1. Introduction

Standard Machine Learning (ML) systems typically output a point prediction that is interpreted as the model's best guess for a quantity of interest. However that is rarely enough, especially in scientific applications. Experimental science requires not only a central estimate, but also a rigorous characterization of its uncertainty, since downstream interpretation, comparison with theory, and decision making depend critically on the credibility attached to the reported result. This is especially true in High-Energy Physics (HEP), where interval estimates, nuisance parameters, and coverage properties are part of the standard language of inference [1, 2]. Consequently, uncertainty estimation is a central issue in HEP, so that there have been cross-experiment initiative to join efforts by the whole community to tackle this problem. The most recent example is the FAIR Universe HiggsML Uncertainty Challenge, an official NeurIPS 2024 competition designed to benchmark uncertainty-aware AI techniques for High-Energy Physics (HEP) [2, 19]

The landscape of techniques developed for uncertainty quantification is very broad, including likelihood-based methods, Bayesian inference, resampling techniques, Monte Carlo dropout, and more. Although powerful, these approaches also come with nontrivial assumptions and practical drawbacks, which hamper their practical usefulness. *Conformal Prediction* (CP) offers an appealing alternative thanks to its desirable properties like no distributional assumptions, model-agnostic nature and finite-sample coverage guarantees.

In this work we investigate Conformal Prediction for uncertainty estimation in a HiggsML-style signal-strength measurement. The full challenge requires a model to estimate the Higgs signal strength μ and provide a 1-σ (68.27%) confidence interval while accounting for systematic variations induced by nuisance parameters. Here we focus on a controlled first step involving nominal HiggsML data with $\mu = 1$ and no nuisance parameter variation. This setting allows us to isolate and test whether conformal calibration delivers the expected empirical coverage for μ.

Our contributions are as follows. First, we formulate signal-strength interval estimation as a split-conformal calibration problem over exchangeable pseudo-experiment blocks. Second, we compare two score constructions based on simple and absolute residuals, thus leading to free-form or symmetric intervals, respectively. Third, we test two alternative point estimators for μ, and explore the impact of bias on interval interpretability and empirical coverage. Finally, we compare the resulting interval widths and empirical coverage against a simple likelihood-based baseline.

2. Background

2.1 Uncertainty quantification in scientific domains

Scientific domains differ from many predictive applications because the downstream goal is often parameter estimation rather than only instance-level classification. In HEP, event classifiers are typically used as intermediate summary statistics. The final object of interest may be a signal strength, a cross section, or a constraint on a model parameter. For such quantities, reporting a point estimate without a confidence interval is insufficient, as intervals are needed to assess compatibility with theoretical expectations and to combine measurements across experiments.

Likelihood-based inference provides the conventional HEP solution. Let D denote anexperiment and let θ denote nuisance parameters. A profile-likelihood construction compares hypotheses

PoS(ISGC2026)024

for the parameter of interest μ using the likelihood ratio

$$\lambda(\mu) = \frac{L(D \mid \mu, \widehat{\widehat{\theta}}(\mu))}{L(D \mid \widehat{\mu}, \widehat{\theta})}, \tag{1}$$

where $\widehat{\widehat{\theta}}(\mu)$ maximizes the likelihood at fixed μ and $(\widehat{\mu}, \widehat{\theta})$ are the global maximum-likelihood estimates. Under regularity and large-sample conditions, Wilks' theorem motivates intervals based on the asymptotic distribution of $-2\log\lambda(\mu)$ [4]. This approach is statistically principled, but it requires an explicit knowledge of the likelihood and can become expensive or fragile when high-dimensional ML summaries and many nuisance parameters are involved.

Bayesian Neural Networks and related methods attempt to propagate parameter uncertainty through predictive distributions [6, 7]. They are conceptually attractive because they provide posterior uncertainty, but posterior quality depends on prior and likelihood specification, and scalable approximate inference can be difficult to validate. Monte Carlo dropout [8] and deep ensembles provide practical approximations to predictive variability, but empirical studies show that such methods can be miscalibrated under distribution shift [11, 14]. Quantile regression is another practical alternative, directly estimating conditional quantiles [10]. The common limitation is that a heuristic uncertainty estimate is not itself a coverage guarantee. Conformal Prediction addresses precisely this gap by wrapping a heuristic predictor and calibrating its output using held-out data.

2.2 Split Conformal Prediction

Although the approaches above are very powerful and widely adopted, they typically rely on one or more of limiting factors as rigid assumptions, large samples approximation, high computational costs, and limited theoretical guarantees, which are often difficult to check in real practice. Thus, in scientific applications where uncertainty statements are expected to be both operationally useful and statistically interpretable, these limitations motivate complementary tools that can provide coverage guarantees while making minimal modelling assumptions. *Conformal Prediction* (CP) offers such a route. Rather than deriving uncertainty from a fully specified probabilistic model, CP calibrates prediction sets or intervals using a held-out calibration sample. This intuitive but powerful approach thus provides mathematical coverage guarantees even for finite sample, with only mild assumptions. Also, it can be queued on top of any pre-trained model, making it very versatile and widely applicable.

Let $(X_i, Y_i)_{i=1}^n$ be calibration data and $(X_{\text{test}}, Y_{\text{test}})$ be a future test point. For a trained predictor $\widehat{f}$, split Conformal Prediction is based on the definition of *nonconformity scores*

$$s_i = s(X_i, Y_i; \widehat{f}) \in \mathbb{R}, \qquad i = 1, \ldots, n, \tag{2}$$

that measure the agreement between the target value y and the model output at x, where larger scores indicate worse performance.

For a given miscoverage probability, α, the procedure then relies on the empirical distribution of these scores. Namely, we define the finite-sample conformal quantile

$$\widehat{q}_{1-\alpha} = s_{(\lceil (n+1)(1-\alpha) \rceil)}, \tag{3}$$

PoS(ISGC2026)024

where $s_{(k)}$ is the k-th order statistic, with the conventional clipping at the largest observed score if needed. The prediction set is then formed including all target values that are within $\widehat{q}_{1-\alpha}$ distance from model's prediction:

$$C(X_{\text{test}}) = \left\{ y : s(X_{\text{test}}, y; \widehat{f}) \leq \widehat{q}_{1-\alpha} \right\}. \tag{4}$$

Provided that the calibration and test examples are exchangeable, i.e. the joint distribution $P(X, Y)$ does not change under data permutations, then it is possible to show that

$$\mathbb{P}\left[Y_{\text{test}} \in C(X_{\text{test}})\right] \geq 1 - \alpha. \tag{5}$$

Notably, this marginal coverage guarantee is model-agnostic, distribution-free and holds also for finite samples [13]. It does not require the model to be correct, nor does it require a parametric assumption for the data distribution. The informativeness of the set, however, depends strongly on the quality of the score function: a poor score can still give coverage, but only through wide or poorly adapted sets.

2.3 HiggsML uncertainty challenge

The FAIR Universe HiggsML Uncertainty Challenge extends the original HiggsML benchmark [18] from event classification to uncertainty-aware parameter estimation. The benchmark consists of simulated proton–proton collision events representative of Large Hadron Collider (LHC) analyses, with features describing visible decay products, missing transverse energy, jets, and derived kinematic quantities. The public dataset contains approximately 280 million events and 28 features, with events labelled as Higgs signal or as one of three background processes. The challenge

Process	Number Generated	LHC Events	Label
Higgs	52 040 227	1 015	signal
Z Boson	160 383 358	1 002 395	background
Di-Boson	605 118	3 783	background
tt	7 070 398	44 192	background

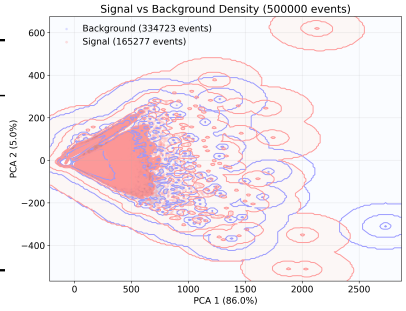


Figure 1: Left: dataset composition. Right: PCA projection with signal vs background density.

is evaluated not on isolated events, but on *pseudo-experiments*, i.e. samples of simulated collision events generated under a specified signal strength and nuisance parameter settings. Each pseudo-experiment is therefore treated as one hypothetical realization of the data that an LHC analysis would observe. In practice, the total yield of a pseudo-experiment is treated as a Poisson counting process characterized as an aleatoric superposition of signal and background contributions. More formally:

$$N_{\text{ev}} = N_{\text{sig}} + N_{\text{bkg}} \sim \text{Pois}(\nu(\mu)), \qquad \text{with} \qquad \nu(\mu) = \mu\,\gamma + \beta, \tag{6}$$

where γ and β denote the expected signal and background yields, respectively, and μ is the signal strength. Operationally, this corresponds to drawing a total number of events from the mixture

PoS(ISGC2026)024

model and then sampling event-level features according to the underlying signal or background processes.

The full challenge evaluates pseudo-experiments generated at varying values of the signal strength and potentially affected by six nuisance parameters that model detector and normalization systematic effects. The final goal is to return a 1-σ (68.27%) confidence interval for μ [2, 19].

3. Experimental setup

In this work, we restrict the analysis to the nominal configuration with $\mu_{\text{true}} = 1$ and no nuisance parameter variations. This simplified setting is deliberate to allows us studying CP behaviour under controlled exchangeability assumptions before considering more realistic scenarios where systematic shifts may induce distributional differences between calibration and evaluation samples. The full challenge settings are out of the scope of this work.

In summary, our experimental design is structured as follows. The ML task is first posed as event-level classification of signal versus background. Once the training is done, event-level predictions are combined to produce the event count in each pseudo-experiment, which is converted into an estimate $\widehat{\mu}$ of the signal strength. Conformal Prediction is then used to construct Confidence Intervals (C.I.) for the signal strength using calibration data. Finally, separate test data are used to assess model classification performance and CP coverage properties.

All the code to reproduce this study is available in the official project repository [22].

3.1 Data splits

The experiments use HiggsML benchmark data in the nominal setting $\mu_{\text{true}} = 1$, with no nuisance parameter variations. The event-level classifier is trained and validated on disjoint labelled splits. A separate reference split is used to estimate the selected signal and background yields (γ^*, β^*) needed for the corrected estimator (see Eq. (9)). Calibration and evaluation are performed on disjoint pseudo-experiment blocks of approximately 10^4 events each. The current setup uses approximately 10 million training events, 5 million validation events, about 2 million reference events, and about 10 million events for calibration and testing, corresponding to roughly 10^3 pseudo-experiment blocks.

3.2 Signal-strength estimators

Let $\widehat{f}(x)$ be a trained classifier returning a signal score and let $t = 0.5$ be the selection threshold. For a pseudo-experiment block $B_k = \{x_i\}_{i=1}^{n_k}$, we define the selected count through *hard-counting*:

$$n_{\text{sel}}(B_k) = \sum_{i=1}^{n_k} \mathbf{1}\{\widehat{f}(x_i) > t\}. \tag{7}$$

We study two estimators. The *raw estimator* ignores selection efficiencies and background contamination:

$$\widehat{\mu}_{\text{raw}}(B_k) = \frac{n_{\text{sel}}(B_k)}{\gamma}. \tag{8}$$

PoS(ISGC2026)024

This estimator is intuitively based on the definition of signal strength as the ratio between observed and expected signal counts, but it can be biased when the classifier has non-unit signal efficiency or fails to reject background events.

The *corrected estimator*, instead, uses reference estimates of the selected signal and background yields, denoted by γ^* and β^*, to correct possible errors in classifying signal versus background events. These estimates are computed on a disjoint reference sample:

$$\widehat{\mu}_{\text{MLE}}(B_k) = \frac{n_{\text{sel}}(B_k) - \beta^*}{\gamma^*}. \tag{9}$$

This corresponds to the maximum likelihood estimate (MLE) of μ for a Poisson counting process with known selected signal and background yields. Hence, this estimator inherits all MLE properties, in particular asymptotic unbiasedness.

3.3 Conformal intervals construction

The key design choice is applying Conformal Prediction at the pseudo-experiment level instead of individual-event level. Each calibration unit is a pseudo-experiment block B_c, with a known μ_{true}. Once training is completed, we compute $\widehat{\mu}_c$ and a nonconformity score comparing $\widehat{\mu}_c$ to μ_{true} for each calibration block. We compare two alternatives also for this step: *symmetric* and *free-form* interval constructions.

Symmetric intervals. The first construction uses an absolute residual difference:

$$s_c^{\text{absolute}} = |\mu_{\text{true}} - \widehat{\mu}_c|\,, \qquad c = 1, \ldots, C. \tag{10}$$

We then compute $\widehat{q}_{68}$ as the $1 - \alpha = 0.6827$ quantile of the nonconformity scores distribution, and define the one-sided interval for a new pseudo-experiment B_{new} as:

$$C\mathcal{I}^{\text{sym}}(B_{\text{new}}) = [\widehat{\mu}(B_{\text{new}}) - \hat{q}_{68}, \widehat{\mu}(B_{\text{new}}) + \widehat{q}_{68}]\,. \tag{11}$$

This interval is symmetric by construction and always contains the point estimate. However, when $\widehat{\mu}$ is biased, the interval must widen enough to cover the true value from a biased center.

Free-form residual intervals. The second construction uses signed residual difference:

$$s_c^{\text{signed}} = \mu_{\text{true}} - \widehat{\mu}_c, \qquad c = 1, \ldots, C. \tag{12}$$

In this case, we compute a two-sided central interval on the nonconformity scores distribution. More formally, we define $\widehat{q}_{\alpha/2}$ and $\widehat{q}_{1-\alpha/2}$ as the lower and upper rank-corrected empirical quantiles of the residual distribution (e.g. approximately the 16-th and 84-th percentiles for $1 - \alpha = 0.6827$). The resulting interval is then computed as:

$$C\mathcal{I}^{\text{free}}(B_{\text{new}}) = \left[\widehat{\mu}(B_{\text{new}}) + \widehat{q}_{\alpha/2}, \widehat{\mu}(B_{\text{new}}) + \widehat{q}_{1-\alpha/2}\right]. \tag{13}$$

Unlike Eq. (11), the free-form interval is not necessarily symmetric around $\widehat{\mu}$.

3.4 Coverage target and empirical evaluation

The target confidence level is $1 - \alpha = 0.6827$, corresponding to the 1-σ central probability of a Gaussian distribution. Empirical coverage is computed over test pseudo-experiments as:

$$\text{coverage} = \frac{1}{E}\sum_{e=1}^{E} \mathbf{1}\left\{\mu_{\text{true}} \in C\mathcal{I}(B_e)\right\}. \tag{14}$$

PoS(ISGC2026)024

Table 1: Validation performance of the event-level classifiers and reference selected-yield factors.

Model	Accuracy	Precision	Recall	F1	β^*_{ref}	γ^*_{ref}
GLM	0.7367	0.6869	0.3761	0.4860	0.0848	0.03765
RF	0.7952	0.7298	0.6057	0.6619	0.1111	0.6057
MLP	0.8334	0.7917	0.6745	0.7284	0.0880	0.6740

3.5 Classifiers

We evaluate three standard classifiers implemented with simple configurations: *Generalised Linear Model with logit link function*, *Random Forest*, and *Multi Layer Perceptron*. All these models are trained with mostly default 'sklearn' library config values for simplicity, as the goal of this work is to explore Conformal Prediction properties rather than optimizing model performance.

4. Results

We evaluate results from multiple angles to assess classification performance, empirical coverage and average interval width. For comparison, we also report a simple likelihood-based interval where:

$$n_{\text{sel}} \sim \text{Pois}(\lambda(\mu)), \qquad \lambda(\mu) = \mu\gamma^* + \beta^*. \tag{15}$$

The corresponding maximum-likelihood estimate is Eq. (9), and a 1-σ interval can be obtained from the profile-likelihood condition:

$$-2\,[\log L(\mu) - \log L(\widehat{\mu})] \leq 1. \tag{16}$$

This baseline is not intended to represent the full challenge profile-likelihood machinery with nuisance parameters, but it provides a local comparison for the simpler settings considered here.

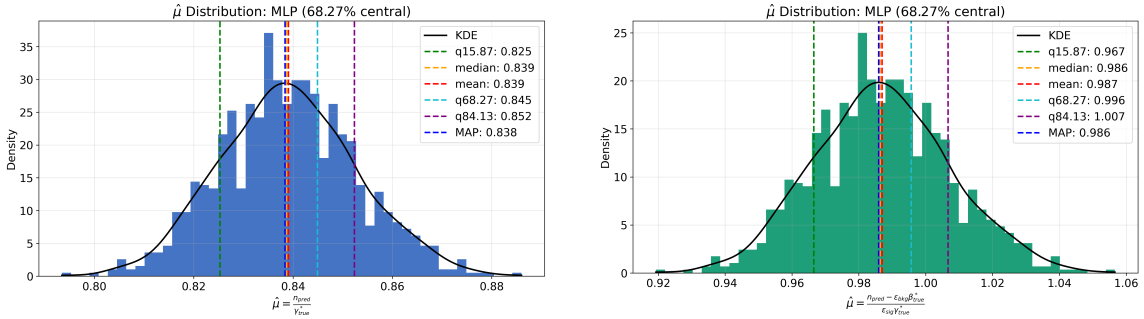
4.1 Classification performance

Although proper model selection is left out of the scope of this work, we report classification performance (Table 1) since it influences Conformal Prediction results and provides an insightful perspective on the properties of CP intervals. The models achieve discrete results, although not extraordinarily high performance. This is due the inherent complexity of the task, but also a consequence of the lack of hyperparameter optimization. The interval results below mainly use the MLP classifier unless otherwise stated, since it gives the best validation classification performance among the tested models. However, similar considerations apply to all other models.

4.2 Signal strength estimation

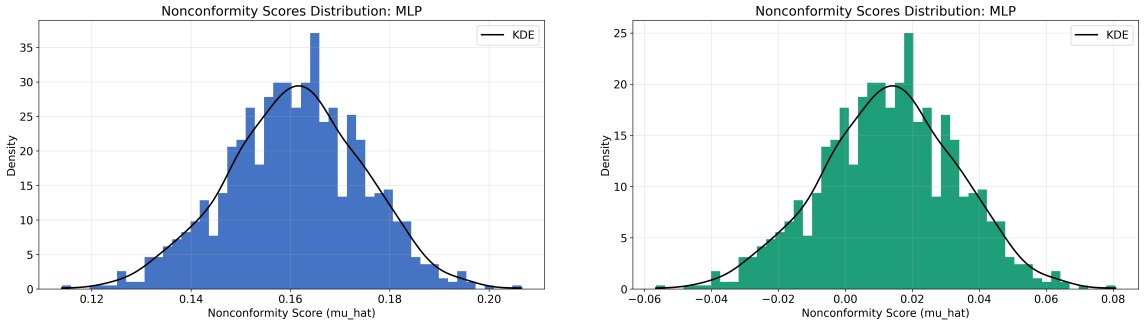
The distribution of signal strength estimates in the calibration sample is reported in Fig. 2. The *raw estimator* in Eq. (8) is strongly biased in the current configuration, with $\widehat{\mu}_{\text{raw}}$ concentrated around 0.839, substantially below the true value $\mu_{\text{true}} = 1$. This produces a positive shift in the signed nonconformity score residuals, s_c^{signed} (see Fig. 3a). On the contrary, the *corrected estimator* hugely mitigates this issue, providing an almost unbiased distribution centered on 0.986 (see Fig. 2b), with scores translated about 0 (see Fig. 3b).

PoS(ISGC2026)024



(a) Raw estimator **(b)** Efficiency-corrected estimator

Figure 2: Signal-strength estimates before and after efficiency correction. a) raw estimator, biased toward low μ values. b) efficiency-corrected estimator, recentered near $\mu_{\text{true}} = 1$.



(a) Raw estimator **(b)** Efficiency-corrected estimator

Figure 3: Signed nonconformity scores distributions. a) raw esitmator scores are skewed towards positive values due to underfitting. b) MLE estimator corrects the distortion; as a result the distribution is recentered around roughly 0.

Table 2: Empirical coverage and average interval width for μ at the target 68.27% confidence level.

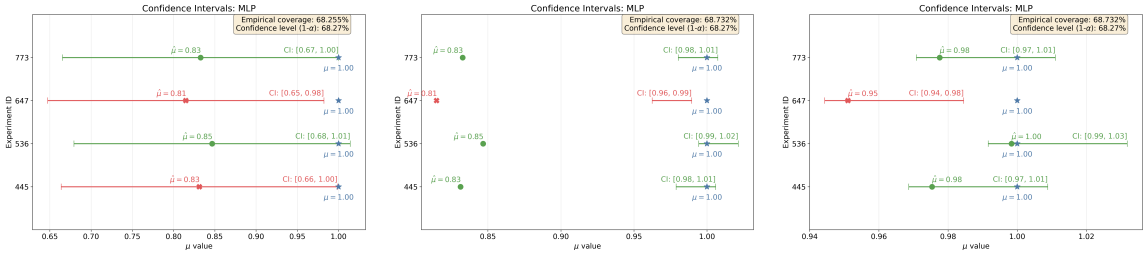
Method	Empirical coverage	Average width
Symmetric	69.11%	~ 0.05
Free-form	68.73%	~ 0.04
Likelihood	67.016%	~ 0.05

4.3 Coverage and width

Table 2 reports empirical coverage and average interval width for the two conformal constructions and the likelihood baseline. All methods are close to the target 68.27% coverage in the nominal setting. The free-form conformal interval is the narrowest among the reported methods, with empirical coverage very close to the target. The symmetric conformal interval is slightly wider, as expected, because it must absorb residual asymmetry through a single absolute-error quantile.

Also, conformal calibration produces intervals with coverage comparable to the likelihood-based construction, with the free-form residual approach using efficiency correction providing

PoS(ISGC2026)024



(a) $\widehat{\mu}_{\text{raw}}, CI^{\text{sym}}$ **(b)** $\widehat{\mu}_{\text{raw}}, CI^{\text{free}}$ **(c)** $\widehat{\mu}_{\text{MLE}}, CI^{\text{free}}$

Figure 4: Effect of conformal calibration under biased and corrected estimators. a) Raw estimator, symmetric intervals widen substantially to compensate bias and recover coverage. b) Raw estimator, free-form intervals restore coverage by shifting away from the biased $\widehat{\mu}$, potentially excluding it. c) MLE, free-form intervals recenter $\widehat{\mu}$ near μ_{true}, thus regaining the standard interpretation as uncertainty bands around the estimate.

slightly better coverage and a modest reduction in width.

Remarkably, conformal calibration corrects interval coverage without requiring the base point estimator to be unbiased (see Fig. 4). This however comes at the cost of either largely widening the predicted interval (see Fig. 4a), or moving the interval far from $\widehat{\mu}$ point-estimate (see Fig. 4b). This means that, for a biased estimator, calibrated free-form intervals may not contain $\widehat{\mu}$, which is statistically valid but counterintuitive. Nonetheless, the efficiency-corrected estimator in Eq. (9) resolves this tension by recentring $\widehat{\mu}$ near μ_{true}. Thus, conformal intervals based on MLE-like estimator remain calibrated while regaining the usual interpretation of an uncertainty band around the estimate (see Fig. 4c).

5. Discussion

Our work explores theoretical properties of Conformal Prediction in a real-science application. The results show how CP represents an appealing candidate approach to provide uncertainty estimation in HEP studies. The observed empirical coverage and interval width support this claim, highlighting performance on-par with standard likelihood-based methods in this simple setting. Beyond this competitiveness, conformal prediction offers a key theoretical advantage: the distribution of its coverage can be characterized analytically [13]. In principle, this enables prospective experimental design, where the number of calibration pseudo-experiments is increased to control the variability of coverage and progressively shrink confidence intervals toward a target width.

Notably, conformal calibration can recover the desired coverage even when the underlying estimator is visibly biased. This is a direct consequence of calibrating the residual distribution on exchangeable pseudo-experiments. If the evaluation pseudo-experiments share the same residual behavior as the calibration pseudo-experiments, the conformal quantiles shift the reported intervals accordingly. This property is valuable for scientific ML because model bias can arise depending on many factors. Conformal calibration provides a principled way to use held-out pseudo-experiments to correct the resulting uncertainty intervals.

The comparison between symmetric and free-form intervals illustrates a second trade-off. Symmetric intervals are more intuitive to interpret and always include $\widehat{\mu}$, but they are less efficient when residuals are skewed or shifted. Instead, free-form intervals adapt to bias and asymmetry.

PoS(ISGC2026)024

This suggests that adopting free-form interval construction in combination with unbiased estimators is a powerful yet flexible strategy for uncertainty estimation.

6. Limitations and Outlook

This study is intended as a preliminary exploration of Conformal Prediction applied in a HEP-like use-case, as we restrict to the nominal HiggsML setting with $\mu_{\text{true}} = 1$ and no nuisance-parameter variation. Extending the calibration protocol to this more realistic setting is an obvious next step.

In that respect, it is worth mentioning that the conformal guarantee holds at the level of exchangeable pseudo-experiment blocks. If calibration and evaluation data are generated under different nuisance-parameter distributions, exchangeability may break, and standard split conformal prediction loses its finite-sample validity without adaptation. Moreover, the reported coverage is marginal, i,e, it does not ensure conditional validity across μ_{true} values, nuisance configurations, or feature-space regions. Methods such as group-balanced, weighted, or nuisance-stratified conformal prediction are natural extensions to address this limitation [13, 16, 17].

Looking forward, we forsee three main immediate directions. First, extend calibration to varying μ and nuisance regimes, for instance via stratified or locally weighted conformal procedures, to test coverage beyond the marginal level. Second, improve the base estimator by moving beyond thresholded classifiers toward approaches such as conformalized quantile regression with the goal of reducing interval width while maintaining validity. Third, compare conformal calibration results with HiggsML challenge winners for a more thorough assessment of CP potential for robust uncertainty quantification in scientific applications.

7. Conclusions

We presented a conformal-prediction approach to uncertainty quantification for HiggsML signal-strength estimation. Using event-level classifiers and pseudo-experiment-level calibration, we constructed 68.27% confidence intervals for μ in the nominal $\mu = 1$, no-nuisance setting. The empirical coverage of the conformal intervals is close to the target level. The free-form residual-quantile construction gives slightly narrower intervals than the symmetric absolute-residual construction and is competitive with a simple likelihood-based interval baseline.

The results show that conformal calibration can correct interval coverage even when the point estimator is biased. After applying an efficiency correction, the estimator becomes centered and the conformal intervals retain both coverage and interpretability. These findings support further investigation of Conformal Prediction as a promising candidate for model-agnostic uncertainty estimation in scientific ML pipelines.

Acknowledgements

Funding This research was partly funded by PNRR - M4C2 - Investimento 1.3, Partenariato Esteso PE00000013 - "FAIR - Future Artificial Intelligence Research" - Spoke 8 "Pervasive AI" and the European Commission under the NextGeneration EU programme. The collection of original

PoS(ISGC2026)024

images was supported by funding from the University of Bologna and the European Space Agency (Research agreement collaboration 4000123556).

References

[1] G. Cowan, K. Cranmer, E. Gross, and O. Vitells, *Asymptotic formulae for likelihood-based tests of new physics*, Eur. Phys. J. C **71**, 1554 (2011). doi:10.1140/epjc/s10052-011-1554-0

[2] W. Bhimji et al., *FAIR Universe 2024: Higgs ML Uncertainty Challenge*, EPJ Web Conf. **337**, 01200 (2025). doi:10.1051/epjconf/202533701200

[3] Y. Pawitan, *In All Likelihood: Statistical Modelling and Inference Using Likelihood*, Oxford University Press (2001).

[4] S. S. Wilks, *The large-sample distribution of the likelihood ratio for testing composite hypotheses*, Ann. Math. Statist. **9**(1), 60–62 (1938). doi:10.1214/aoms/1177732360

[5] B. Efron and R. J. Tibshirani, *An Introduction to the Bootstrap*, Chapman and Hall (1993).

[6] D. J. C. MacKay, *A practical Bayesian framework for backpropagation networks*, Neural Computation **4**(3), 448–472 (1992). doi:10.1162/neco.1992.4.3.448

[7] C. Blundell et al., *Weight uncertainty in neural networks*, Proc. ICML, PMLR **37**, 1613–1622 (2015). link

[8] Y. Gal and Z. Ghahramani, *Dropout as a Bayesian approximation: Representing model uncertainty in deep learning*, Proc. ICML, PMLR **48**, 1050–1059 (2016). link

[9] B. Lakshminarayanan, A. Pritzel, and C. Blundell, *Simple and scalable predictive uncertainty estimation using deep ensembles*, Adv. Neural Inf. Process. Syst. **30** (2017). link

[10] R. Koenker and G. Bassett, *Regression quantiles*, Econometrica **46**(1), 33–50 (1978). doi:10.2307/1913643

[11] Y. Ovadia et al., *Can you trust your model's uncertainty? Evaluating predictive uncertainty under dataset shift*, Adv. Neural Inf. Process. Syst. **32** (2019). arXiv:1906.02530

[12] G. Shafer and V. Vovk, *A tutorial on conformal prediction*, J. Mach. Learn. Res. **9**, 371–421 (2008). link

[13] A. N. Angelopoulos and S. Bates, *Conformal prediction: A gentle introduction*, Found. Trends Mach. Learn. **16**(4), 494–591 (2023). doi:10.1561/2200000101

[14] A. Ashukha et al., *Pitfalls of in-domain uncertainty estimation and ensembling in deep learning*, ICLR (2020). arXiv:2002.06470

[15] V. Vovk, A. Gammerman, and G. Shafer, *Algorithmic Learning in a Random World*, Springer (2005). doi:10.1007/b106715

PoS(ISGC2026)024

[16] Y. Romano, E. Patterson, and E. J. Candès, *Conformalized quantile regression*, Adv. Neural Inf. Process. Syst. **32** (2019). arXiv:1905.03222

[17] R. J. Tibshirani, R. F. Barber, E. J. Candès, and A. Ramdas, *Conformal prediction under covariate shift*, Adv. Neural Inf. Process. Syst. **32** (2019). arXiv:1904.06019

[18] C. Adam-Bourdarios et al., *The Higgs boson machine learning challenge*, PMLR **42**, 19–55 (2015). link

[19] L. Benato et al., *FAIR Universe HiggsML Uncertainty Dataset and Competition*, arXiv:2410.02867 (2025). arXiv:2410.02867

[20] K. Cranmer, J. Pavez, and G. Louppe, *Approximating likelihood ratios with calibrated discriminative classifiers*, arXiv:1506.02169 (2015). arXiv:1506.02169

[21] A. Ghosh, B. Nachman, and D. Whiteson, *Uncertainty-aware machine learning for high energy physics*, Phys. Rev. D **104**, 056026 (2021). doi:10.1103/PhysRevD.104.056026

[22] L. Clissa, *Conformal Predictions for Signal Strength Estimation*, GitHub repository (2026). link

PoS(ISGC2026)024