Conformal Prediction for Reliable Uncertainty Quantification in Scientific AI Models

Luca Clissa, Leonardo Plini, Daniele Bonacorsi, Iacopo Vivarelli

luca.clissa2@unibo.it

Outline

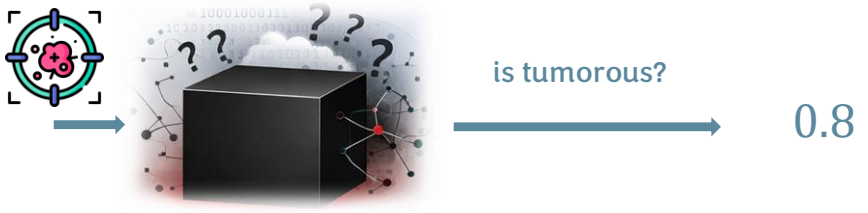
- Uncertainty Quantification
- Conformal Preditctions
- FAIR HiggsML Uncertainty Challenge
- Experimental setup
- Preliminary results
- Conclusions & outlook

Background & motivation

- Machine Learning (ML) models typically provide point-estimates
- No measure of uncertainty or confidence
- **Assumption:** past performance is good indicator of future performance

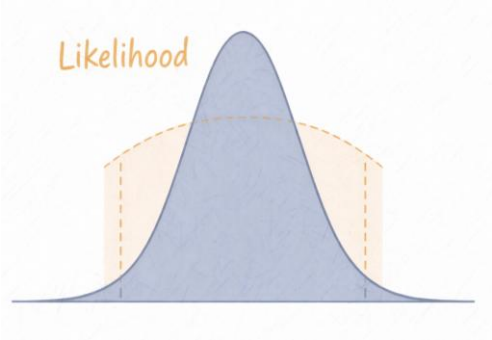
- Highly precise point estimate may not be enough
  - E.g. high-stakes domains like medicine or science
- Ideally, we want:
  - Estimate of the variability
  - Measure of confidence about the prediction

$$\hat{f}(x_i) \rightarrow \hat{y}_i$$



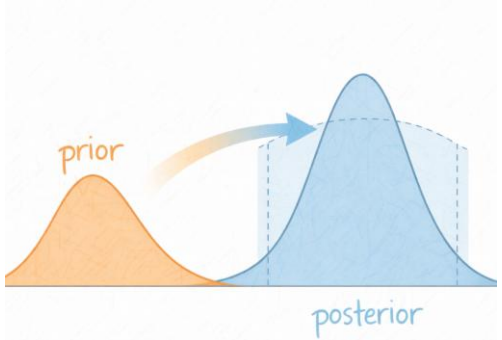
How to estimate uncertainty?

Methodological landscape of Uncertainty Quantification



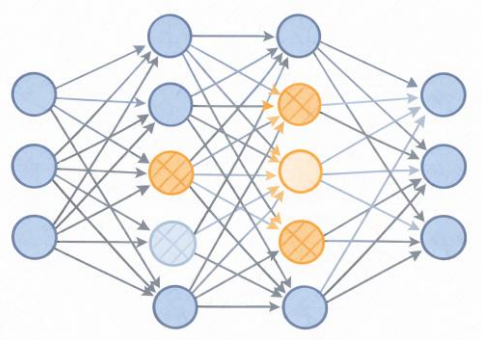
Likelihood-based inference

Uncertainty from parametrized statistical models



Bayesian modelling

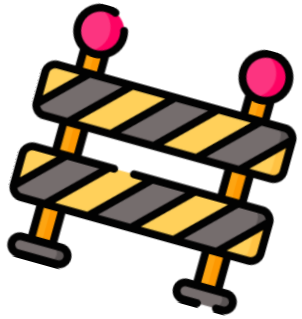
Naturally provides posterior distributions



Monte Carlo Dropout [1]

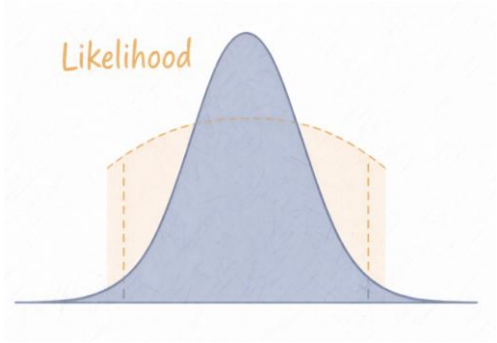
Sampling at inference to simulate model variance

Limitations



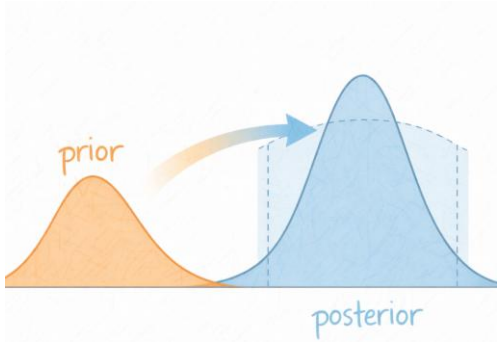
Methodological landscape of Uncertainty Quantification

Likelihood-based inference



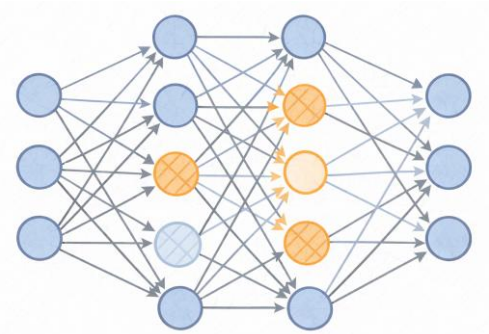
- Strict distributional assumptions
- Rarely met in complex, high-dimensional real-world data

Bayesian modelling



- Computationally heavy (posterior sampling)
- Need retraining

Monte Carlo Dropout [1]



- No mathematical guarantees

Conformal Predictions (CP)

Goal: find prediction set $\mathcal{C}(X_{test}) \subset Y$ ***such that:*** $\mathbb{P}\left(Y_{test} \in \mathcal{C}(X_{test})\right) \geq 1 - \alpha$

Strategy:

1. Split into train, calibration and test data
2. Train model on train data, $\hat{f}(x), f: X \to Y$
3. Define heuristic measure of uncertainty/error, nonconformity scores: $s\left(Y_i, \hat{f}(X_i)\right) \to \mathbb{R}$
4. Compute $\hat{q}$ as the $(1-\alpha)$-th quantile of calibration scores $\left\{s_i = s\left(Y_i, \hat{f}(X_i)\right)\right\}_{i=1}^{n_{calib}}$
5. Compute prediction set $\mathcal{C}(X_{test}) = \left\{y : \ s\left(Y_{test}, \hat{f}(X_{test})\right) \leq \hat{q}\right\}$

Assumptions: calibration and test data are exchangeable (i.e. permutations do not alter distribution)

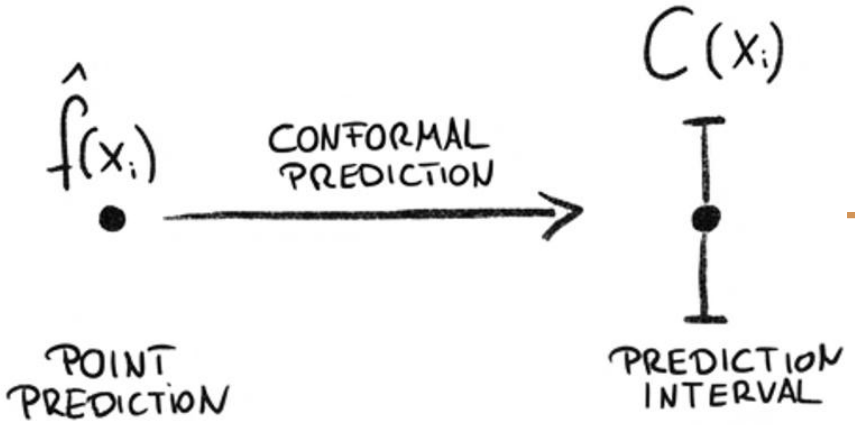
Properties

Model-agnostic

Work with any pre-trained model, no need for retraining

Distribution-free

No assumptions about underlying data distribution P(X,Y)



Marginal coverage

Mathematical coverage guarantees (true value in predicted set with a given confidence level $(1-\alpha)$

Finite-sample guarantees

Coverage properties hold even for small sample

Goal

- Stress-test CP theoretical guarantees in a complex scientific application
- Use case: FAIR Universe Higgs Uncertainty Challenge [12]
- Task:
  - Discriminate signal (Higgs) VS background events
  - Estimate Higgs signal strength, μ_{true}, from simulated collisions at CERN's LHC
  - Provide 68.27% (1σ) Confidence Interval (CI) for μ_{true}
- Challenges:
  - Class imbalance
  - MC samples have distribution shifts:
    - Varying nominal signal strength
    - 6 nuisance parameters

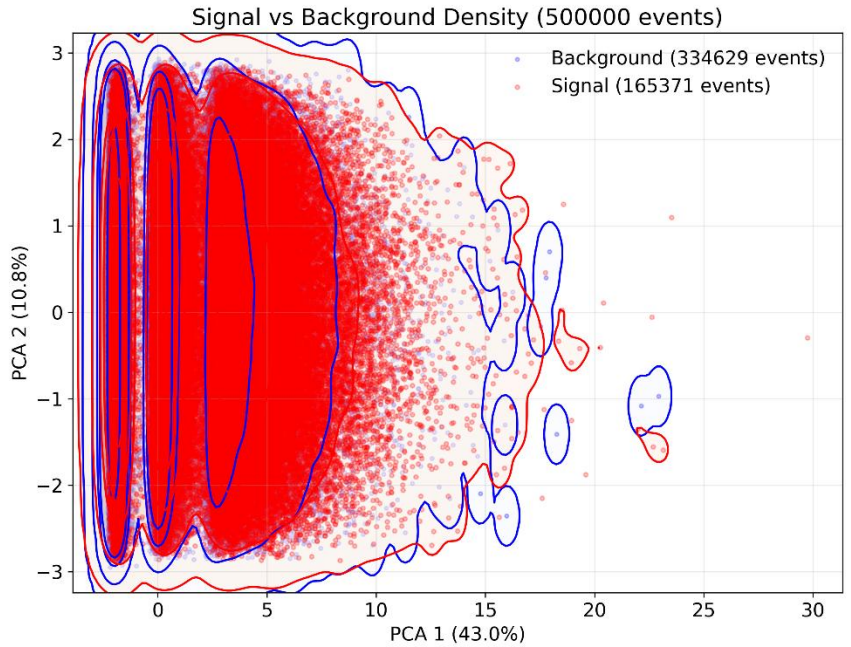
Data generation process

- MC samples contain events of 4 processes

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**

- Events are generated as a counting process:
  - $N_{ev} \sim Pois(\nu)$, where ν is the expected total yield
  - $\nu = \mu \cdot \gamma + \beta$, where:
    - γ is the expected signal yield
    - β is the expected background yield
    - μ is the signal strength (default to 1)

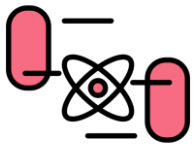
- In practice:
  - Draw $N_{sig} \sim Pois(\mu \cdot \gamma)$ events, then sample signal features
  - Draw $N_{bkg} \sim Pois(\beta)$ events, then sample background features



Our proposal

1. **Training:**
    1. signal VS background classification
2. **Calibration:**
    1. Collect B pseudo-experiments (calib blocks)
    2. For each block, b:
        1. Compute $\mathrm{n_{sig}} = \sum_{i=1}^{n_b} \mathbb{1}\left(\hat{f}(x_i) > 0.5\right)$
        2. Estimate $\hat{\mu}$ from $\mathrm{n_{sig}}$ and expected yields
        3. Compute $s(\mu_{true}, \hat{\mu})$
    3. Compute $\hat{q}$ from B scores (one per block)
3. **Testing**
    1. Collect E pseudo-experiments (eval blocks)
    2. For each block, e:
        1. Estimate $\hat{\mu}$ as in 2.2.2
        2. Use $\hat{q}$ to get $\mathcal{C}(X_{test})$
    3. Check empirical coverage across pseudo-experiments

Experimental setup



Data

Simulated data	HiggsML data
• Two scenarios: easy, hard	• at $\mu = 1$

Estimation

Raw	Corrected
• $\hat{\mu} = \frac{n_{sig}}{\gamma}$	• $\hat{\mu} = \frac{n_{sig} - \beta^*}{\gamma^*}$

Interval

	Symmetric	Free-form
Scoring	• $s(\mu_{true}, \hat{\mu}) = \lvert \mu_{true} - \hat{\mu} \rvert$	• $s(\mu_{true}, \hat{\mu}) = \mu_{true} - \hat{\mu}$
Quantile	• q_{68}	• q_{16}, q_{84}
CI	• $\mathcal{C}(X_{test}) = \hat{\mu} \pm \hat{q}_{68}$	• $\mathcal{C}(X_{test}) = (\hat{\mu} + \hat{q}_{16}, \hat{\mu} + \hat{q}_{84})$

Preliminary results: classification

- Data: HiggsML
  - Train/validation: 10M events
  - **Reference: 10M events (for β^*, γ^*)**
  - Calib/test: 10M events (**1049 block of 10k events each**)
- Models:
  - GLM: LogisticRegression(penalty="l2", solver="lbfgs", max_iter=1000)
  - RF: RandomForestClassifier(n_estimators=50, criterion="gini")
  - MLP: MLPClassifier(_layer_sizes=(32, 16), activation="relu", max_iter=1000)
- Results:
  - Simple models and configs
  - Performance: not bad but biased
  - MLP > RF > GLM

(val)	Accuracy	Precision	Recall	f1	β^*_{ref}	γ^*_{ref}
GLM	0.7367	0.6867	0.3762	0.4861	0.0848	0.03761
RF	0.7954	0.7299	0.6065	0.6625	0.1111	0.6067
MLP	0.8343	**0.8046**	**0.6595**	**0.7249**	0.0794	0.6601

Preliminary results: signal strength

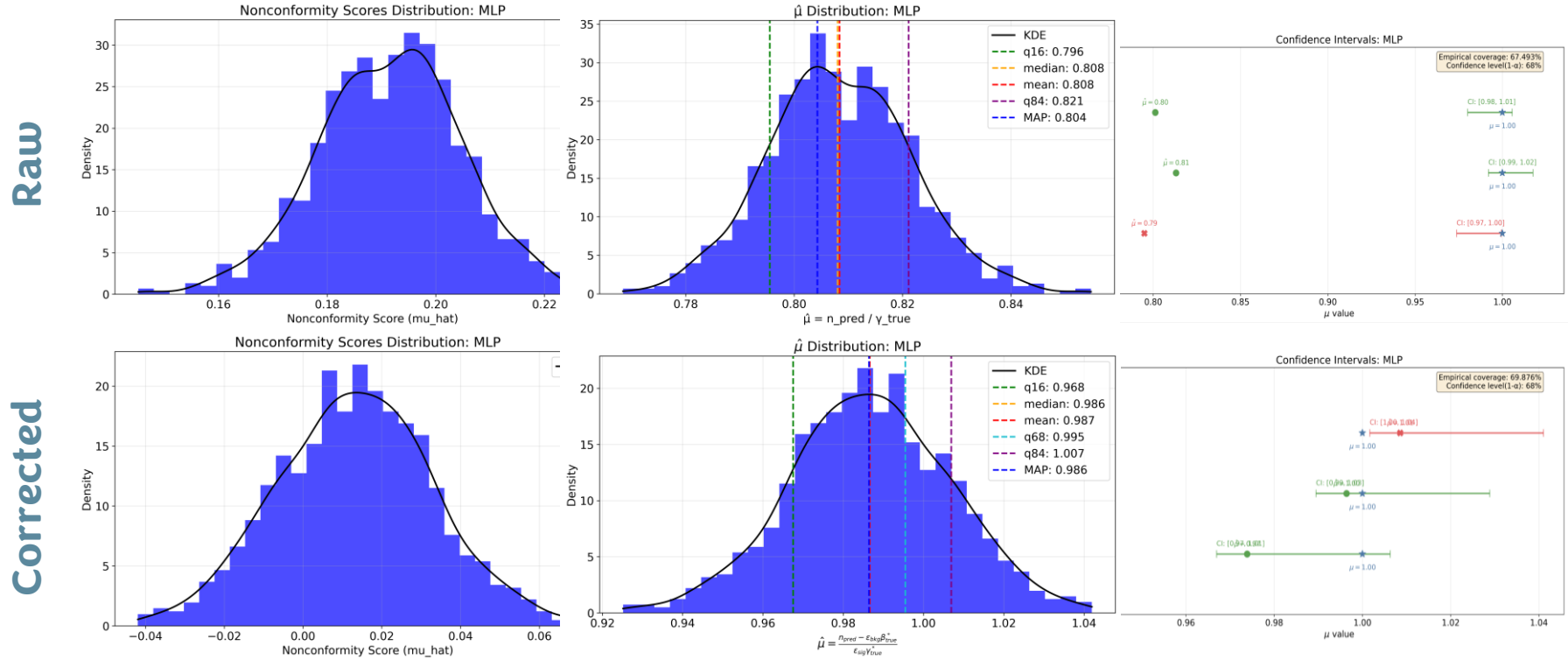
Estimation strategy:

- Raw estimates are clearly biased (underestimation)

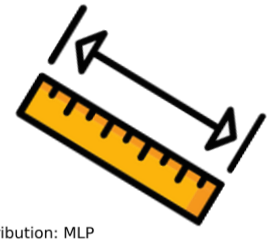
  → positive shift in scores distribution

- Efficiency correction fixes $\hat{\mu}$ estimates

Nontheless CP fixes CI!



Preliminary results: coverage

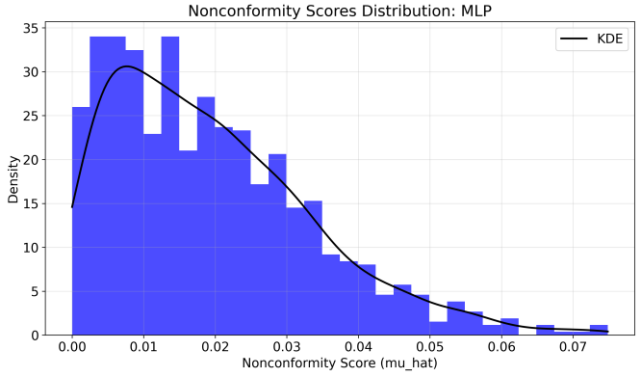


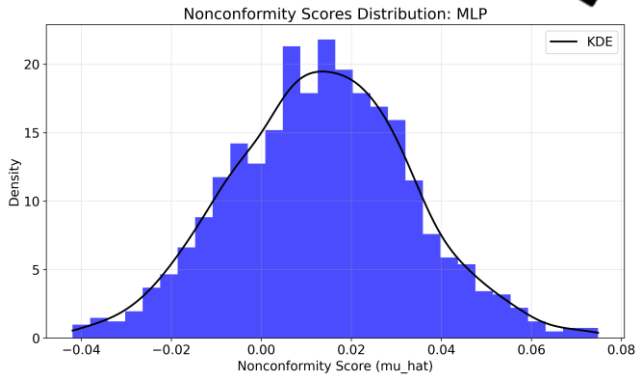
Interval construction:

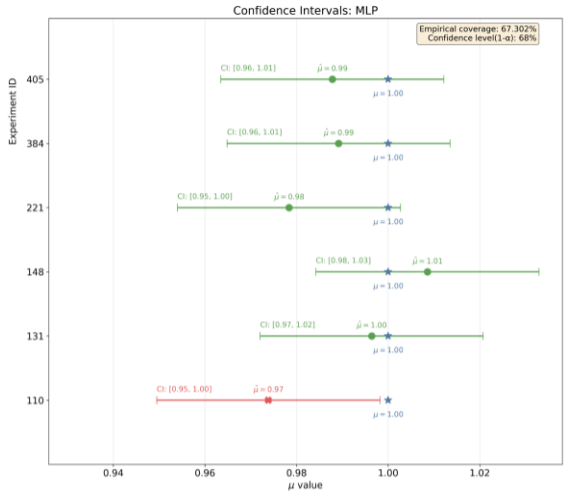
- $\mathcal{C}(X_{test}) = \hat{\mu} \pm \hat{q}_{68}$:
  - Intervals are symmetric by construction
  - Intuitive interpretation

  $\rightarrow \hat{\mu} \in \mathcal{C}(X_{test})$ always!
- $\mathcal{C}(X_{test}) = (\hat{\mu} + \hat{q}_{16}, \hat{\mu} + \hat{q}_{84})$
  - symmetry depends on bias
  - Biased model → asymmetric interval

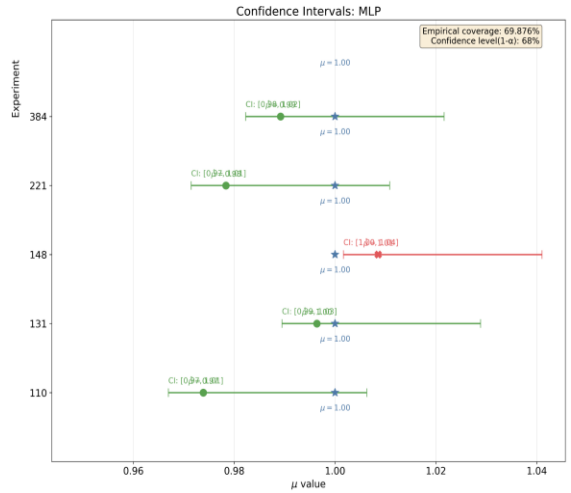
	Empirical coverage	Width
Symmetric	67.30%	~0.05
Free-form	69.88%	~0.04







Symmetric



Free-form

Conclusions & outlook

CP provides a powerful tool for uncertainty estimation

- Model-free
- Little assumptions
- Finite-sample coverage guarantees

Preliminary results look promising

- **Efficiency correction + free-form construction seems better**
- **Coverage checked despite biased models!**
- **Reasonable width**

Future work: full HiggsML challenge

- Variations of signal strength + nuisance
- Explore extensions for non-exchangeable data (**group-balanced CP**)
- More estimation strategies: (**direct quantile regression**)

Questions?

Contacts: luca.clissa2@unibo.it

References

[1] Angelopoulos AN, Bates S (2023), "Conformal Prediction: A Gentle Introduction". *Foundations and Trends in Machine Learning*, Vol. 16 No. 4 pp. 494–591, doi: https://doi.org/10.1561/2200000101

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

[3] Wilks SS (1938), "The Large-Sample Distribution of the Likelihood Ratio for Testing Composite Hypotheses". Annals of Mathematical Statistics, Vol. 9 No. 1 pp. 60–62, doi: https://doi.org/10.1214/aoms/1177732360

[4] MacKay DJC (1992), "A Practical Bayesian Framework for Backpropagation Networks". Neural Computation, Vol. 4 No. 3 pp. 448–472, doi: https://doi.org/10.1162/neco.1992.4.3.448

[5] Blundell C, Cornebise J, Kavukcuoglu K, Wierstra D (2015), "Weight Uncertainty in Neural Networks". Proceedings of the 32nd International Conference on Machine Learning (ICML), pp. 1613–1622.

[6] Gal Y, Ghahramani Z (2016), "Dropout as a Bayesian Approximation: Representing Model Uncertainty in Deep Learning". Proceedings of the 33rd International Conference on Machine Learning (ICML), pp. 1050–1059.

[7] Ovadia Y, Fertig E, Ren J, Nado Z, Sculley D, Nowozin S, Dillon JV, Lakshminarayanan B, Snoek J (2019), "Can You Trust Your Model's Uncertainty? Evaluating Predictive Uncertainty Under Dataset Shift". Advances in Neural Information Processing Systems (NeurIPS), Vol. 32.

[8] Ashukha A, Lyzhov A, Molchanov D, Vetrov D (2020), "Pitfalls of In-Domain Uncertainty Estimation and Ensembling in Deep Learning". International Conference on Learning Representations (ICLR).

[9] Koenker R, Bassett G (1978), "Regression Quantiles". Econometrica, Vol. 46 No. 1 pp. 33–50, doi: https://doi.org/10.2307/1913643

[10] Romano Y, Patterson E, Candès EJ (2019), "Conformalized Quantile Regression". Advances in Neural Information Processing Systems (NeurIPS), Vol. 32.

[11] Melki G, Lee N, Musacchio J, Schuster T (2023), "Group-Conditional Conformal Prediction via Quantile Regression Calibration". Proceedings of the IEEE/CVF International Conference on Computer Vision Workshops (ICCVW).

[12] W. Bhimji, P. Calafiura, R. Chakkappai, P. Chang, Y. Chou, S. Diefenbacher, J. Dudley, S. Farrell, A. Ghosh, I. Guyon, C. Harris, S. Hsu, E. Khoda, R. Lyscar, A. Michon, B. Nachman, P. Nugent, M. Reymond, D. Rousseau, B. Sluijter, B. Thorne, I. Ullah, Y. Zhangì (2025), "FAIR Universe 2024: Higgs ML Uncertainty Challenge", *EPJ Web Conf. 337 01200*, DOI: 10.1051/epjconf/202533701200

Backup

Experimental setup

Data

Simulated data	HiggsML data
• Two scenarios: easy, hard	• at $\mu = 1$

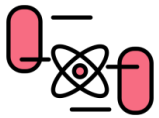
Estimation

Raw	Corrected
• $\hat{\mu} = \frac{n_{sig}}{\gamma}$	• $\hat{\mu} = \frac{n_{sig} - \beta^*}{\gamma^*}$

Uncertainty

	One-sided	Two-sided
Scoring	• $s(\mu_{true}, \hat{\mu}) = \lvert \mu_{true} - \hat{\mu} \rvert$	• $s(\mu_{true}, \hat{\mu}) = \mu_{true} - \hat{\mu}$
Quantile	• q_{68}	• $q_{16},\ q_{84}$
CI	• $\mathcal{C}(X_{test}) = \hat{\mu} \pm \hat{q}_{68}$	• $\mathcal{C}(X_{test}) = (\hat{\mu} + \hat{q}_{16}, \hat{\mu} + \hat{q}_{84})$

Experimental setup



Data

Simulated data	HiggsML data
• Two scenarios: easy, hard	• at $\mu = 1$

Uncertainty

	One-sided	Two-sided
Scoring	• $s(\mu_{true}, \hat{\mu}) = \mid \mu_{true} - \hat{\mu} \mid$	• $s(\mu_{true}, \hat{\mu}) = \mu_{true} - \hat{\mu}$
Quantile	• q_{68}	• q_{16}, q_{84}
CI	• $\mathcal{C}(X_{test}) = \hat{\mu} \pm \hat{q}_{68}$	• $\mathcal{C}(X_{test}) = (\hat{\mu} + \hat{q}_{16}, \hat{\mu} + \hat{q}_{84})$

Estimation

Raw	Corrected
• $\hat{\mu} = \frac{n_{sig}}{\gamma}$	• $\hat{\mu} = \frac{n_{sig} - \beta}{\gamma}$