A Platform, Dataset, and Challenge for Uncertainty-Aware Machine Learning

Wahid Bhimji Lawrence Berkeley National Laboratory (LBNL)
Paolo Calafiura LBNL
Ragansu Chakkappai Université Paris-Saclay, CNRS/IN2P3, IJCLab (IJClab) and ChaLearn
Po-Wen Chang LBNL Yuan-Tang Chou University of Washington, Seattle
Sascha Diefenbacher LBNL Jordan Dudley University of California, Berkeley and LBNL
Steven Farrell LBNL Aishik Ghosh University of California, Irvine and LBNL
Isabelle Guyon ChaLearn Chris Harris LBNL Shih-Chieh Hsu U Washington
Elham E Khoda U Washington Rémy Lyscar IJCLab Alexandre Michon IJClab
Benjamin Nachman LBNL Peter Nugent LBNL
David Rousseau IJCLab and ChaLearn Benjamin Sluijter Universiteit Leiden and LBNL
Benjamin Thorne LBNL Ihsan Ullah ChaLearn Yulei Zhang U Washington

fair-universe@lbl.gov

Abstract

A new challenge has been set up that focuses on measuring the physics properties of elementary particles with imperfect simulators due to differences in modelling systematic errors. Additionally, the challenge has leveraged a large-compute-scale AI platform for sharing datasets, training models, and hosting machine learning competitions. Our challenge has brought together the physics and machine learning communities to advance our understanding and methodologies in handling systematic (epistemic) uncertainties within AI techniques. The challenge is ongoing at the time of submission of this paper, it will be completed in March 2025, with intermediate results available for NeurIPS 2024. A comparison of the various methods used by participants and the first lessons will be presented at the workshop.

Introduction

For several decades, the discovery space in almost all branches of science has been accelerated dramatically due to increased data collection brought on by the development of larger, faster instruments. More recently, progress has been further accelerated by the emergence of powerful AI approaches, including deep learning, to exploit this data. However, an unsolved challenge that remains, and *must* be tackled for future discovery, is how to effectively quantify and tackle uncertainties, including understanding and controlling *systematic* uncertainties (also named *epistemic* uncertainties in other fields). This is widely true across scientific and industrial applications involving measurement instruments (medicine, biology, climate science, chemistry, and physics, to name a few). A compelling example is found in analyses to further our fundamental understanding of the universe through analysis of the vast volumes of particle physics data produced at CERN, in the Large Hadron Collider (LHC).

High energy physics (HEP) relies on statistical analysis of aggregated observations. Therefore, the interest in uncertainty-aware ML methods in HEP is nearly as old as the application of ML in the field. Advanced efforts began with initial investigations in the use of Bayesian networks for uncertainty quantification [1], as well as with the development of uncertainty-minimising inference methods [2]. There have been several recent developments in this area, with the introduction of

Machine Learning and the Physical Sciences Workshop, NeurIPS 2024.

multiple uncertainty-aware methods capable of dealing with systematic uncertainties in a given dataset [3, 4, 5, 6], as well as in the application of previous methods to actual measurement data [7].

We aim to address the issue of systematic errors within a specific domain. Yet, the techniques developed by the challenge participants will apply to identifying, quantifying, and correcting systematic errors in other domains. This effort will also intersect with critical topics in machine learning, including data bias and fairness. We plan to keep our submission platform accessible even after the challenge concludes, establishing it as a lasting benchmark. This initiative should significantly influence research in uncertainty-aware ML/AI techniques, which currently suffer from a critical shortage of datasets and benchmarks dedicated to their research and development.

A more detailed description of the challenge shall be found in the FAIR Universe HiggsML Uncertainty Challenge Competition whitepaper [8].

1 Novelty

Previous challenges concerning High Energy Physics (HiggsML data challenge [9], the TrackML Challenges (NeurIPS 2018 competition) [10, 11], the LHC Olympics [12]) have not addressed the issue of systematic biases. This challenge introduces a significant change w.r.t previous challenges using simulated data that includes biases (or *systematics*) in the test dataset. In addition, participants are not asked to provide a measurement but to provide a confidence interval on a measurement. We have developed an innovative metric to assess their performance.

While there have been previous challenges focusing on meta-learning and transfer-learning, such as the NeurIPS 2021 and 2022 meta-learning challenges [13, 14], Unsupervised and Transfer Learning[15], challenges related to bias e.g. Crowd bias challenge [16], and those addressing distribution shifts, like the Shifts challenge series, and CCAI@UNICT 2023 [17], to the best of our knowledge, this is the first challenge that requires participants to handle systematic uncertainty.

2 Dataset

We are using a simulated particle physics dataset for this competition to produce data representative of high energy proton collision data collected by the ATLAS experiment [18] at the Large Hadron Collider (LHC) [19]. The dataset is created using two widely used simulation tools, Pythia 8.2 [20] and Delphes 3.5.0 [21]. We have organised the dataset into a tabular format where each row corresponds to a collision event, the measurements recorded from a single proton bunch crossing of interest. Each row has 28 features that describe the particle properties of the event. The events are divided into two categories. The signal category includes collision events with a Higgs boson decaying into tau pairs, while the background category includes other processes (subcategories) leading to a similar final state.

Due to its complexity, the process of generating events is computationally intensive; use of a supercomputer allowed to create a vast amount of data, about 80 million events, which is two orders of magnitude larger than for the HiggsML competition. It will be made publicly available after the competition under the Creative Commons Attribution license to serve as a benchmark after the competition.

In addition, we have developed a biasing script capable of manipulating a dataset by introducing six parameterised distortions (the systematics). For example, a detector miscalibration can cause a bias in other features in a cascade way, or a new source of noise is added to a feature, or in another case, the magnitude of a particular background (e.g. the $t\bar{t}$) contribution can change so that the composition of the background (thus the feature distributions) can be different. In both cases, the inference would be done on a dataset not i.i.d to the training dataset. The biasing script is provided to the participants so that they know what biases they should expect; the biases are "known unknowns". The case of "unknown unknowns" is beyond this challenge.

3 Tasks and application scenarios

The participant aims to develop an estimator for the Higgs boson count in a dataset analogous to results from Large Hadron Collider experiments. Such measurement is typical to those carried over

at the Large Hadron Collider, which allows us to strengthen (or invalidate!) our understanding of the fundamental laws of nature.

The primary metric, signal strength (μ), defaults to one, aligning with Standard Model predictions for Higgs boson occurrences. The challenge involves estimating μ's true value, μ_{true}, which may vary from one and is inherently unknown. For challenge purposes, pseudo-experiments simulate data across μ_{true} values from 0.5 to 3, evaluating participant estimators.

Participants are tasked with generating a 68% Confidence Interval (CI) for μ, incorporating both aleatoric (random) and epistemic (systematic) uncertainties rather than a single-point estimate.

The primary simulation dataset assumes a μ of one. Participants receive a training subset, labelled for particle identification, and unlabeled test sets, each with a different value of μ_{true} and biased differently. For each test set, they must predict a CI for μ. The organizers provide the script to generate test data from the primary simulation dataset.

In a machine learning context, the task resembles a transduction problem with distribution shift: it requires constructing a μ interval estimator from labelled training data and biased unlabelled test data. A potential approach involves training a classifier to distinguish Higgs boson from the background, with robustness against bias achieved possibly through data augmentation via the provided script.

This challenge shifts focus from the qualitative discovery of individual Higgs boson events to the quantitative estimation of overall Higgs boson counts in test sets, akin to assessing disease impact on populations rather than diagnosing individual cases.

4 Metrics

Participants must submit a model to the challenge platform that can analyze a dataset to determine (μ_{16}, μ_{84}), which represents the bounds of the 68% Confidence Interval (CI) for μ.

The model's performance will be assessed based on two criteria:

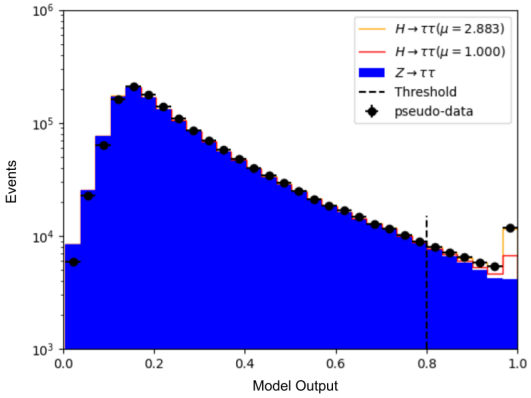
- **Precision:** The narrowness of the CI (narrower is preferable).
- **Coverage:** The accuracy of the CI in reflecting the measurement's uncertainty, meaning there should be a 68%

To evaluate the model performance, we are using the very large dataset created to generate *pseudo-experiments*, which are small datasets with a mixture of events representative of what could be obtained in a real experiment. The number of events is Poisson-fluctuated, and events are drawn from the large dataset so that features are randomised. In addition, the six biases are Gaussian randomised, and different values of μ are tested. Nevertheless the model should calculate a (μ_{16}, μ_{84}) interval which should include the true μ value 68% of the time. The score of the model is obtained from the narrowness of the CI, penalised by how much the measured coverage departs from the nominal 68% (see Fig.1b). Thanks to the large dataset and significant computing resources available, the participant model is evaluated on 1.000 independent pseudo-experiments for each submission, increased to 10.000 independent pseudo-experiments for the final ranking of participants.

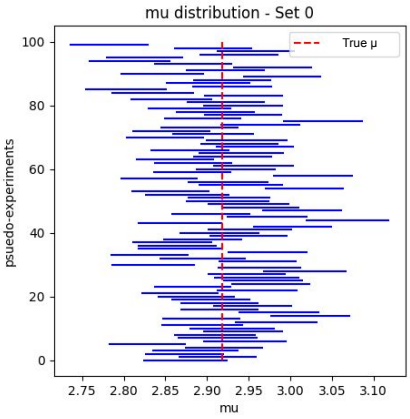
5 Baselines

A **Starting Kit** is available on the challenge website. This kit includes code for installing necessary packages, loading and visualising data, training and evaluating a model and preparing a submission for the competition.

The Baseline method estimates μ by merging standard techniques without addressing systematics for simplicity. Initially, it utilizes a binary classifier (XGBoost Boosted Decision Tree or a simple PyTorch neural network) trained on a subset of training data to filter events, enhancing signal event density and reducing μ estimator variance. Although adjustable for variance optimization, the classifier's decision threshold is fixed heuristically. μ is then estimated from these filtered events, assuming Poisson distribution for Large Hadron Collider events, enabling point-wise and interval maximum likelihood estimation. Further refinement involves binning events as per classifier selection and estimating μ per bin, akin to a voting ensemble. A reserved training dataset segment assesses



(a) *Histogram of events for Model Output*: Unlabeled test pseudo-data (black), hold-out data for (1) background events $Z \to \tau\tau$ (blue), (2) signal events $H \to \tau\tau$ for $\mu = 1$ (orange), and (3) signal events fitted histogram to test pseudo-data, leading to estimated $\mu = 2.883$.



(b) *Coverage plot*: All predicted Confidence Intervals (blue lines) for each pseudo experiment generated for a given μ_{true} (vertical dotted line).

Figure 1: Baseline method results

signal-background ratio per bin for μ=1. This calibration step then permits using unlabeled test data (pseudo-data) for μ estimation. The alignment of maximum likelihood estimation (orange line) with empirical data (black line), in particular in the right-most bins, which are the most signal pure, indicates method success (Fig.1a). The maximum likelihood also yields the Confidence Interval.

To address the problem of systematic errors, participants are encouraged to enhance the Baseline model, for instance, by adopting a Domain Adversarial Neural Network to improve resilience against biases, attempting to directly model the biases, or refining the estimator through a bias-aware model.

Conclusions and Outlook

The challenge is ongoing at the time of submission of this paper, it will be completed in March 2025, with intermediate results available for NeurIPS 2024. Participants will have to submit detailed documentation of their model in addition to the code they will have submitted during the challenge. Hence we will be able to compare the different approaches and evaluate them in detail, beyond the one metric used to rank them. What techniques are novel ? Which technique works best for which bias? Which technique can maintain good results when training dataset size is reduced? Or when fewer training computing resources is used ? Which technique is likely to be adopted by the field? With the release of the large dataset associated with the biasing script allowing the simulation of the impact of six distinct systematic sources, we believe the challenge will trigger new developments in statistics and Machine Learning dealing with uncertainties.

References

[1] S. Bollweg, M. Haußmann, G. Kasieczka, M. Luchmann, T. Plehn, and J. Thompson, *Deep-Learning Jets with Uncertainties and More*, SciPost Phys. **8** (2020) 006, arXiv:1904.10004 [hep-ph].

[2] P. De Castro and T. Dorigo, *INFERNO: Inference-Aware Neural Optimisation*, Comput. Phys. Commun. **244** (2019) 170, arXiv:1806.04743 [stat.ML].

[3] S. Wunsch, S. Jörger, R. Wolf, and G. Quast, *Optimal Statistical Inference in the Presence of Systematic Uncertainties Using Neural Network Optimization Based on Binned Poisson Likelihoods with Nuisance Parameters*, Comput. Softw. Big Sci. **5** (2021) 4, arXiv:2003.07186 [physics.data-an].

[4] A. Ghosh, B. Nachman, and D. Whiteson, *Uncertainty-aware machine learning for high energy physics*, Phys. Rev. D **104** (2021) 056026, arXiv:2105.08742 [physics.data-an].

[5] P. Feichtinger *et al.*, *Punzi-loss: a non-differentiable metric approximation for sensitivity optimisation in the search for new particles*, Eur. Phys. J. C **82** (2022) 121, arXiv:2110.00810 [hep-ex].

[6] N. Simpson and L. Heinrich, *neos: End-to-End-Optimised Summary Statistics for High Energy Physics*, J. Phys. Conf. Ser. **2438** (2023) 012105, arXiv:2203.05570 [physics.data-an].

[7] L. Layer, T. Dorigo, and G. Strong, *Application of Inferno to a Top Pair Cross Section Measurement with CMS Open Data*, arXiv:2301.10358 [hep-ex].

[8] W. Bhimji, P. Calafiura, R. Chakkappai, Y.-T. Chou, S. Diefenbacher, J. Dudley, S. Farrell, A. Ghosh, I. Guyon, C. Harris, S.-C. Hsu, E. E. Khoda, R. Lyscar, A. Michon, B. Nachman, P. Nugent, M. Reymond, D. Rousseau, B. Sluijter, B. Thorne, I. Ullah, and Y. Zhang, *Fair universe higgsml uncertainty challenge competition*, 2024. `https://arxiv.org/abs/2410.02867`.

[9] C. Adam-Bourdarios, G. Cowan, C. Germain, I. Guyon, B. Kégl, and D. Rousseau, *The Higgs boson machine learning challenge*, in *Proceedings of the NIPS 2014 Workshop on High-energy Physics and Machine Learning*, G. Cowan, C. Germain, I. Guyon, B. Kégl, and D. Rousseau, eds. PMLR, Montreal, Canada, 13 Dec, 2015. `https://proceedings.mlr.press/v42/cowa14.html`.

[10] S. Amrouche, L. Basara, P. Calafiura, V. Estrade, S. Farrell, D. R. Ferreira, L. Finnie, N. Finnie, C. Germain, V. V. Gligorov, T. Golling, S. Gorbunov, H. Gray, I. Guyon, M. Hushchyn, V. Innocente, M. Kiehn, E. Moyse, J.-F. Puget, Y. Reina, D. Rousseau, A. Salzburger, A. Ustyuzhanin, J.-R. Vlimant, J. S. Wind, T. Xylouris, and Y. Yilmaz, *The tracking machine learning challenge: Accuracy phase*, in *The NeurIPS 2018 Competition*, pp. 231–264. Springer International Publishing, Nov., 2019. arXiv:1904.06778 [hep-ex].

[11] S. Amrouche, L. Basara, P. Calafiura, D. Emeliyanov, V. Estrade, S. Farrell, C. Germain, V. V. Gligorov, T. Golling, S. Gorbunov, H. Gray, I. Guyon, M. Hushchyn, V. Innocente, M. Kiehn, M. Kunze, E. Moyse, D. Rousseau, A. Salzburger, A. Ustyuzhanin, and J.-R. Vlimant, *The Tracking Machine Learning Challenge: Throughput Phase*, Comput. Softw. Big Sci. **7** (2023) 1, arXiv:2105.01160 [cs.LG].

[12] G. Kasieczka, B. Nachman, D. Shih, O. Amram, A. Andreassen, K. Benkendorfer, B. Bortolato, G. Brooijmans, F. Canelli, J. H. Collins, B. Dai, F. F. De Freitas, B. M. Dillon, I.-M. Dinu, Z. Dong, J. Donini, J. Duarte, D. A. Faroughy, J. Gonski, P. Harris, A. Kahn, J. F. Kamenik, C. K. Khosa, P. Komiske, L. Le Pottier, P. Martín-Ramiro, A. Matevc, E. Metodiev, V. Mikuni, C. W. Murphy, I. Ochoa, S. E. Park, M. Pierini, D. Rankin, V. Sanz, N. Sarda, U. Seljak, A. Smolkovic, G. Stein, C. M. Suarez, M. Szewc, J. Thaler, S. Tsan, S.-M. Udrescu, L. Vaslin, J.-R. Vlimant, D. Williams, and M. Yunus, *The lhc olympics 2020 a community challenge for anomaly detection in high energy physics*, Reports on Progress in Physics **84** (Dec., 2021) 124201. `http://dx.doi.org/10.1088/1361-6633/ac36b9`.

[13] A. E. Baz, I. Ullah, and etal, *Lessons learned from the neurips 2021 metadl challenge: Backbone fine-tuning without episodic meta-learning dominates for few-shot learning image classification*, PMLR (2022, to appear) .

[14] D. Carrión-Ojeda, H. Chen, A. E. Baz, S. Escalera, C. Guan, I. Guyon, I. Ullah, X. Wang, and W. Zhu, *Neurips'22 cross-domain metadl competition: Design and baseline results*, 2022.

[15] I. Guyon, G. Dror, V. Lemaire, D. L. Silver, G. Taylor, and D. W. Aha, *Analysis of the ijcnn 2011 utl challenge*, Neural Networks **32** (2012) 174.

[16] M. L. Danula Hettiachchi, *Crowd bias challenge*, 2021. `https://kaggle.com/competitions/crowd-bias-challenge`.

[17] S. P. Federica Proietto, Giovanni Bellitto, *Ccai@unict 2023*, 2023. `https://kaggle.com/competitions/ccaiunict-2023`.

[18] ATLAS Collaboration, *The atlas experiment at the cern large hadron collider*, JINST **3** (2008) S08003.

[19] L. Evans and P. Bryant, *LHC machine*, JINST **3** (aug, 2008) S08001.

[20] T. Sjöstrand, S. Ask, J. R. Christiansen, R. Corke, N. Desai, P. Ilten, S. Mrenna, S. Prestel, C. O. Rasmussen, and P. Z. Skands, *An introduction to PYTHIA 8.2*, Comput. Phys. Commun. **191** (2015) 159, arXiv:1410.3012 [hep-ph].

[21] DELPHES 3, J. de Favereau, C. Delaere, P. Demin, A. Giammanco, V. Lemaître, A. Mertens, and M. Selvaggi, *DELPHES 3, A modular framework for fast simulation of a generic collider experiment*, JHEP **02** (2014) 057, arXiv:1307.6346 [hep-ex].