FAIR Universe – the challenge of handling uncertainties in fundamental science NeurIPS 2024 Competition Proposal

David Rousseau * Wahid Bhimji † Paolo Calafiura Ragansu Chakkappai
Yuan-Tang Chou Sascha Diefenbacher Steven Farrell Aishik Ghosh
Isabelle Guyon Chris Harris Elham E Khoda Benjamin Nachman
Yulei Zhang Ihsan Ullah
fair-universe@lbl.gov
https://github.com/FAIR-Universe/HEP-Challenge
https://fair-universe.lbl.gov/

Abstract

We propose a challenge organised in conjunction with the Fair Universe project, a collaborative effort funded by the US Department of Energy and involving the Lawrence Berkeley National Laboratory, Université Paris-Saclay, University of Washington, and ChaLearn. This initiative aims to forge an open AI ecosystem for scientific discovery. The challenge will focus on measuring the physics properties of elementary particles with imperfect simulators due to differences in modelling systematic errors. Additionally, the challenge will leverage a large-compute-scale AI platform for sharing datasets, training models, and hosting machine learning competitions. Our challenge will bring together the physics and machine learning communities to advance our understanding and methodologies in handling systematic (otherwise known as epistemic) uncertainties within AI techniques.

Keywords

Particle Physics, High Energy Physics, Uncertainties, Uncertainty-aware AI

1 Competition description

Fair Universe is a scientific competition where participants have to deal with a biased dataset and estimate a confidence interval on a parameter of interest in fundamental physics.

1.1 Background and impact

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

*Lead Organizer
†Backup Organizer

Preprint. Under review.

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).

Ten years ago, part of our team organised the Higgs Boson Machine Learning Challenge (HiggsML) [1], the most popular Kaggle challenge at the time attracting 1785 teams. This challenge has significantly heightened interest in applying Machine Learning (ML) techniques within High-Energy Physics and, conversely, has exposed physics issues to the ML community. Whereas previously, the most effective methods predominantly relied on boosted decision trees, Deep Learning has since gained prominence (see, e.g., HEP ML living review). While the LHC has not discovered new physics beyond the Higgs boson, it has accumulated vast data. Consequently, the most excellent discovery prospect is precision analyses of this Higgs boson data.

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 [2], as well as with the development of uncertainty-minimising inference methods [3]. There have been a number of recent developments in this area, with the introduction of multiple uncertainty-aware methods capable of dealing with systematic uncertainties in a given dataset [4, 5, 6, 7], as well as in the application of previous methods to actual measurement data [8].

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, which inspired the name of our project: Fair Universe. If we judge by the popularity of our previous challenge and the importance of such topics to ML, we anticipate that our challenge will attract hundreds of participants. 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.

1.2 Novelty

This entirely new public competition will build on our experience running several competitions in particle physics and broader. These include the original HiggsML data challenge, the TrackML Challenge (NeurIPS 2018 competition), the LHC Olympics, AutoML/AutoDL, and other competitions. Building on the foundation of the original HiggsML challenge, this competition introduces a significant change by 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 [9, 10], Unsupervised and Transfer Learning[11], challenges related to bias e.g. Crowd bias challenge [12], and those addressing distribution shifts, like the Shifts challenge series, and CCAI@UNICT 2023 [13], to the best of our knowledge, this is the first challenge that requires participants to handle systematic uncertainty.

Moreover, this project has connected the Perlmutter system at NERSC, a large-scale supercomputing resource featuring over 7000 NVIDIA A100 GPUs, with Codabench, a new version of the renowned open-source benchmark platform Codalab competitions. Our challenge is set to be among the first to utilise this platform for a public competition, marking a significant milestone in accessible, high-performance computing for AI research.

1.3 Data

We will use a simulated particle physics dataset for this competition, to produce data representative of high energy proton collision data collected by the ATLAS experiment [14] at the Large Hadron Collider (LHC) [15]. The dataset is created using two widely used simulation tools, Pythia 8.2 [16] and Delphes 3.5.0 [17]. We have organised the dataset into a tabular format where each row corresponds to a collision event. Each row has 32 features that describe the particles and properties of the event. The events are divided into two categories (see Table 1): signal and background. The signal

Table 1: Summary of the dataset for each category and subcategory. "Number Generated" is the number of events available, while "LHC events" is the average number in this category in a pseudo-experiment.

Process	Number Generated	LHC Events	Label
Higgs	16214520	9220	**signal**
Z Boson	14135841	2569787	**background**
W Boson	287514800	2964267	**background**
$t\bar{t}$	31921500	320318	**background**

category includes collision events with a Higgs boson decaying into tau pairs, while the background category includes other processes (sub-categories) leading to a similar final state.

Due to its complexity, the process of generating events is computationally intensive; use of the Perlmutter supercomputer allowed to create a vast amount of data, about 50 million events, which is two orders of magnitude larger than for the HiggsML competition. It will be made publicly available 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 a total of five parameterised distortions (the systematics). For example, a detector miscalibration can cause a bias in other features in a cascade way, or in another case, the magnitude of the W boson 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 dataset, with all labels, has been handled on the private disks of just two people from the team before being uploaded to Codabench where it is secured. The random generator seeds which could be used to reproduce the dataset are also secured.

1.4 Tasks and application scenarios

The participant's objective is to develop an estimator for the Higgs boson count in a dataset, analogous to results attainable from Large Hadron Collider experiments. This estimation facilitates the determination of key physical parameters.

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 used 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 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 (which was the focus of our first challenge) to the quantitative estimation of overall Higgs boson counts in test sets, akin to assessing disease impact on populations rather than diagnosing individual cases.

1.5 Metrics

Participants must submit a model to the Codabench 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% probability that μ_{truth} falls within the CI.

The model that predicts (μ_{16}, μ_{84}) serves as an estimator. In statistics, the effectiveness of an estimator is not evaluated based on a single case, but rather over numerous instances. Therefore, the model undergoes evaluation across a series of N_{test} independent pseudo-experiments, each characterized by randomized biases and μ_{truth}, to assess it against the specified criteria. We introduce a novel uncertainty metric, which we call **Coverage Score**. This metric is divided into two components:

The first component is the **Average Interval Width** w, calculated as follows:

$$w = \frac{1}{N_{test}} \sum_{i=0}^{N} |\mu_{84,i} - \mu_{16,i}| \ , \tag{1}$$

where $\mu_{16,i}$ and $\mu_{84,i}$ are the bounds of the 68% CI for each experiment i within the range $[0, N_{\text{test}}]$.

The second component c quantifies the frequency with which the true value of μ_{true} falls within the 68% Confidence Interval (CI), as illustrated in Figure 2a:

$$c = \frac{1}{N_{test}} \sum_{i=0}^{N} 1 \text{ if } \mu_{\text{true},i} \in [\mu_{84,i} - \mu_{16,i}] \ . \tag{2}$$

If the confidence interval $[\mu_{16,i}, \mu_{84,i}]$ accurately represents the 68% quantile, the true value of μ should lie within this interval in 68% of the pseudo-experiments. Consequently, we employ a function f that penalises models that deviate from this 68% reference.

$$c \in [0.68 - 2\sigma_{68}, 0.68 + 2\sigma_{68}] : f(c) = 1 \tag{3}$$

$$c < 0.68 - 2\sigma_{68} : f(c) = 1 + |\frac{c - (0.68 - 2\sigma_{68})}{\sigma_{68}}|^4 \tag{4}$$

$$c > 0.68 + 2\sigma_{68} : f(c) = 1 + |\frac{c - (0.68 + 2\sigma_{68})}{\sigma_{68}}|^3 \tag{5}$$

Here, σ_{68} is the standard deviation for a binomial distribution with $p = 0.68$, given by

$$\sigma_{68} = \sqrt{\frac{(1 - 0.68)0.68}{N_{\text{test}}}} \ , \tag{6}$$

this parameter establishes a tolerance zone around the anticipated value of 0.68, accommodating statistical variances inherent in the evaluation process. A graphical representation of $f(c)$ is depicted in Fig. 1. We opted for an asymmetric penalty function because, within the field of High Energy Physics (HEP), overestimating uncertainty is deemed more acceptable than underestimating it. Hence, coverage exceeding 68% incurs a lesser penalty compared to coverage falling below 68%.

The final **Coverage Score** used to rank participants is calculated as follows:

$$\text{score} = -\ln((w + \epsilon) f(c)) \tag{7}$$

where, w represents the average width of the Confidence Interval, c is the coverage parameter, and $\epsilon = 10^{-2}$ serves as a regularization term to guard against submissions that report unrealistically

narrow CIs. The use of the $-\ln()$ function ensures that higher scores are awarded to superior submissions and that the score variations remain within a modest range.

To ensure efficient use of resources and timely updates to the leaderboard, each participant's model inference is executed across 100 pseudo-experiments times five distinct values of μ_{truth}. In the Final phase of the competition, the best submission from each participant will be evaluated over 10,000 pseudo-experiments, times ten different μ_{truth} values, to minimize the luck factor.

1.6 Baselines, code, and material provided

A **Starting Kit** will be made available on the challenge website. This kit will include code for installing necessary packages, loading and visualising data, training and evaluating a model and preparing a submission for the competition. This should facilitate local testing and adjustments by participants before their official submissions, and allow the participants to train and evaluate models. Moreover, the competition adheres to a specific interface, details of which are documented in the Starting Kit and demonstrated through the provided sample code.

The Baseline method estimates μ by merging standard techniques without addressing systematics for simplicity. Initially, it utilizes a binary classifier (a simple PyTorch neural network) trained on a BaselineTrain subset of training data to filter particles, enhancing signal event density and reducing μ estimator variance. The classifier's decision threshold, although adjustable for variance optimization, 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 particles as per classifier selection and estimating μ per bin, akin to a voting ensemble. BaselineHoldout, a reserved training dataset segment, assesses 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) indicates method success (Fig.2b).

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. To lower the barrier to entry, a parallel competition will be hosted on the Codabench platform, mirroring the main competition except for the exclusion of all systematic errors. This auxiliary competition, lacking any prizes or publicity, serves solely as an experimental platform for participants to acquaint themselves with the problem.

1.7 Website, tutorial and documentation

We have set up a dedicated GitHub repository for this competition, which will be linked to the competition's Codabench website. The GitHub repository covers the following points:

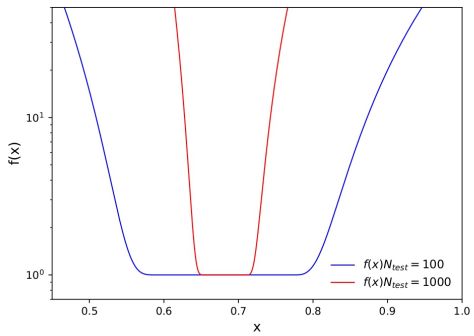
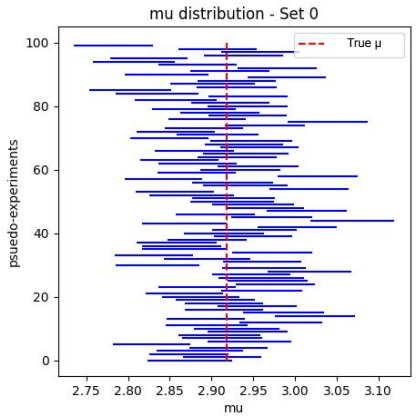
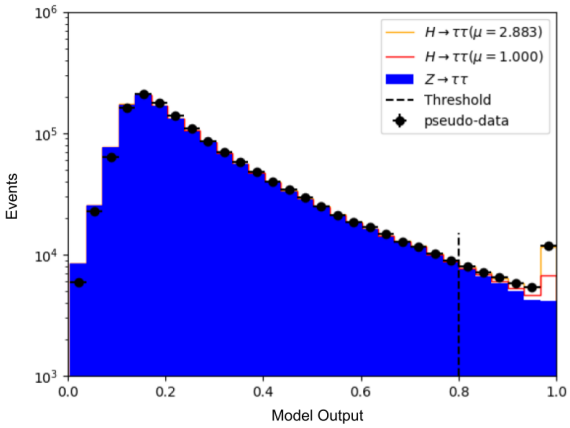


Figure 1: Evaluation function output as a function of the coverage value c. The shape of the function is determined by uncertainty on the coverage evaluation, governed by N_{test}, the number of pseudo-experiments. A larger N_{test} leads to narrower function.



(a) *Coverage plot*: all the predicted intervals (blue lines) for each pseudo experiment generated for a given μ_{true} (vertical dotted line).



(b) *Histogram of events for Model Output*: Unlabeled test pseudo-data (black), and HoldOut 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 $\mu = 2.883$.

Figure 2: Baseline method results

- Competition introduction and instructions for setting up the complete environment.
- Complete details of the evaluation process.
- Information about how to submit.
- Troubleshooting instructions for possible issues and contact details for reporting issues.
- Link to a dedicated forum on the Codabench platform for efficient communication.

In addition, a code tutorial is provided for: (i) loading and discovering properties of data. (ii) explaining the coding structure and expected functions to be implemented in submissions. (iii) providing instructions and examples for running the baseline methods on the public datasets.

2 Organizational aspects

2.1 Protocol

The competition will be hosted on Codabench platform with a dedicated webpage on Codabench. Participants are required to create accounts on the platform and register for the competition. Participants can see and access all the public details without registering for the competition. Both signing up to Codabench and registering for the competition have no fee.

Participants are provided with a detailed overview of the competition, evaluation procedure, data description and starting kit.The starting kit is a tutorial notebook designed to familiarise participants with the competition protocol, competition data, scoring, unit tests to test their submissions and code to prepare their submissions for the competition website. Participants are also given some ready-to-submit submissions in the GitHub repository to get them started with libraries such as Pytorch [18], Tensorflow [19], scikit-learn [20] etc.

Once participants are registered, they can submit their solutions (code submissions, possibly including training), which will be executed on resources at NERSC (see Section 3.1). Participants will get their submission feedback once per day, as displayed on the competition leaderboard.

The competition consists of two phases: **Phase 1: Public Phase** – Participants submit their model **Phase 2: Final Phase** – Participants now submit only one best submission from the previous phase for final evaluation (by default it would be one with the highest score).

Every submission is executed in the same environment with the same maximum execution time and identical resources. Each participant is allowed to make only five submissions per day and a

maximum of 100 submissions during the challenge. To enable the participants to perform other experiments on their hardware, they can use the public data provided in Phase 1.

The proposed protocol was tested in two limited simplified competitions during the Nov. 2023 AI Uncertainty Conference in Paris [21] during the March 2024 ACAT conference in Stony Brook[22].

Furthermore, since we have team members of the Codabench platform as co-organisers (Isabelle Guyon, Ihsan Ullah) as well as staff at NERSC (Wahid Bhimji, Steven Farrell, Chris Harris) in the competition, we will be able to address possible Codabench bugs and issues efficiently.

2.2 Rules and Engagement

- **General Terms**: This challenge is governed by the General ChaLearn Contest Rule Terms, the Codabench Terms and Conditions, and the specific rules set forth.
- **Announcements**: To receive announcements and be informed of any change in rules, the participants must provide a valid email. We will use the Competition forum to communicate with the participants about any announcements, changes, new releases, phase changes, etc.
- **Conditions of participation**: Participation requires complying with the challenge's rules. Prize eligibility is restricted by US government export regulations; see the General ChaLearn Contest Rule Terms. The organisers, sponsors, their students, close family members, as well as any person having had access to the truth values or to any information about the data or the challenge design, giving them an unfair advantage, are excluded from participation. ChaLearn and the organisers reserve the right to evaluate any entry made in the challenge for scientific purposes and whether or not it qualifies for prizes.
- **Dissemination**: The challenge is part of the official selection for NeurIPS 2024. There will be publication opportunities for reports co-authored by organisers and participants.
- **Registration**: The participants must register to Codabench with a valid email address. Teams must register and provide a group email forwarded to all team members. Teams or solo participants registering multiple times may be disqualified.
- **Anonymity**: The participants who do not present their results at the conference can elect to remain anonymous by using a pseudonym.
- **Submission method**: The results must be submitted through this Codabench competition site. The number of submissions per day and maximum total computational time are restrained and subject to change according to the number of participants. In case of a problem, send email to `fair-universe@lbl.gov`. The entries must be formatted as specified on the Instructions page.
- **Reproducibility**: The participant should try to guarantee their method's reproducibility (for example, by fixing all random seeds involved).
- **Eligibility to Prizes**: The submission submitted to the Final Phase is used for final evaluation. Possible prize recipients must fill out a fact sheet and they are required to make their code (both for training and inference) publicly available under an OSI-approved license, within a week after the final submission deadline. Entrants who decline their prize eligibility retain all their rights to their entries and are not obliged to release their code publicly.
- **Monetary Prizes**: The three top-ranking eligible participants in the Final phase of blind testing will receive the monetary prizes, as well as funded invitations to NeurIPS
- **Jury Prizes**: Special jury prizes will be attributed by a jury to participants with a reasonable score, and the solution appears to be novel, practical, and frugal. The special jury prizes will be funded invitations to NeurIPS 2024 or a dedicated workshop at CERN, Geneva, Switzerland. It should be remembered that the jury prize for the 2014 HiggsML challenge (an invitation to CERN) was given to Tianqi Chen et al, for their very first release of XGBoost[23] announced on the challenge's forum[24] and the rest is history.

Discussion: The rules have been designed with the criteria of *inclusiveness for all participants* and *openness of results* in mind. We aim to achieve inclusiveness by providing them with computation cycles (for the Public phase and Final phase) on our compute resources for evaluation on multiple test sets at submission time. This way, participants who do not have ample computing resources will still

have a fair chance to win the challenge. We aim to achieve openness of results by asking all willing participants to upload their code and, afterwards, fill in a fact sheet about the methods used.

Cheating prevention: The testing dataset will remain hidden in the Codabench platform. We will also monitor submissions and contact participants with suspicious submission patterns. Finally, the candidate prize winners must open-source code to claim their prize (training and inference code). All other participants will individually scrutinise their code before they earn an award. The μ_{truth} values are randomised for each submission during the public phase to avoid participants trying to guess them through multiple attempts. However, in the final phase, all participants are evaluated on the same new set of μ_{truth} values to reduce the luck factor.

2.3 Schedule and readiness

The competition preparations started in November 2023. The running duration of the competition will be four months, from June 2024 to September 2024 (Table 2). We are finishing the baseline preparation and setting up of the Codabench platform competition. The competition data and protocol preparation have already been completed. We're writing a white paper detailing the protocol and state of the art. We're also improving the documentation of the protocol and starting kit.

Table 2: Envisioned competition schedule.

Date	Phase	Description
November 2023 - December 2023	Preparation	Data preparation
January 2024 - February 2024	Preparation	Protocol preparation
March 2024 - May 2024	Preparation	Baselines preparation and Finalising up challenge environment. White-paper writing
June 2024-mid Oct 2024	Public Phase	Start of the public phase, publicity
mid Oct 2024	Final Phase	Evaluating performance on hidden dataset
end October 2024	Results	Notification of winners

2.4 Competition promotion and incentives

To promote the competition, we will use the following channels: i) mailing list from hundreds of participants from past challenges we organised, ii) advertisement on Codabench, MLNews, comp.ai.neural-nets groups, fair-universe announcement groups, iii) n-network advertisement, e.g. topical/local/regional mailing lists, personal Twitter accounts and personal emails.

ChaLearn `http://chalearn.org` will donate a prize pool of 4000 USD, which will be distributed as: (1^{st} rank) 2000 USD, (2^{nd} rank) 1500 USD and (3^{rd} rank) 500 USD. The jury prizes will include funded invitations to NeurIPS 2024 and to a post-competition workshop at CERN, Geneva. Furthermore, we will invite the most relevant participants to work on a post-challenge collaborative paper [25]. We already have experience working on such collaborative papers thanks to the analysis of the NeurIPS 2019 TrackML competition [26], the NeurIPS 2019 AutoDL challenge [27], NeurIPS 2021 MetaDL challenge [9], and NeurIPS 2022 Cross-Domain MetaDL Challenge [10]

3 Resources

3.1 Resources provided by organisers

We are relying on the following resources:

- **Competition infrastructure:** We will use the public instance of Codabench hosted by Université Paris-Saclay and LISN as competition platform. We will use *Perlmutter*, an HPE (Hewlett Packard Enterprise) Cray EX supercomputer, to process participant submissions at NERSC. Compute workers dedicated to the competition will be launched on Perlmutter to scale the number of submissions. Each compute worker node will be equipped with 4 NVIDIA A100, 2x AMD EPYC 7763 and 512 GB of DDR4 memory.
- **Support Staff:** The Codabench platform is administered by dedicated engineering staff at Université Paris-Saclay. During the competition, the organisers will be available to support the participants through the forum of the challenge.

3.2 Support requested

We will take the main responsibility for the publicity of our competition, ensuring many participants from the NeurIPS community. To support us in this publicity, we count on the NeurIPS organisation for the following matters: i) display of the competition on the NeurIPS 2024 website, ii) referring participants to the various competitions that are organised iii) time slot during the program, among which we can announce the winners and discuss the competition.

3.3 Organizing team

The following does not count for the 8 page limit, as per NeurIPS 2024 template.

- **Wahid Bhimji**
  Lawrence Berkeley National Laboratory, USA - wbhimji@lbl.gov
  He leads the Data and AI Services Group at NERSC, Berkeley Lab. His group tackles all aspects of AI on NERSC supercomputers, including software platforms, joint projects, support and education for scientists running large-scale AI for science. He also has a PhD and several years of experience in particle physics. He coordinates all aspects of the project, including the development of the datasets and competition, as well as the integration of the Codabench platform with the HPC resources at NERSC.
- **Paolo Calafiura**
  Lawrence Berkeley National Laboratory, USA - pcalafiura@lbl.gov
  He leads the Physics and X-Ray Science Computing group in the Scientific Data Division, Berkeley Lab. His group researchers design, develop, and deploy data analysis and exploration methods at scale to satisfy the computing needs of mission-critical experimental collaborations. Paolo has a Ph.D. in Particle Physics. His research interests include pattern recognition methods for noisy experimental data; data analysis methods, tools, and environments; and the design, development, deployment, and production operation of robust, secure scientific applications and cyberinfrastructure. He contributes to the competition's organisation and evaluation.
- **Ragansu Chakkappai**
  Université Paris-Saclay, CNRS/IN2P3, IJCLab, 91405 Orsay, France
  ragansu.chakkappai@ijclab.in2p3.fr
  He is a PhD student from Université Paris-Saclay working at IJCLab Orsay in collaboration with the ATLAS Experiment in CERN. Ragansu has completed his Masters in Particle and Nuclear Physics from Université Paris-Saclay and works on Machine Learning methods for Higgs boson analysis under the supervision of Dr. David Rousseau. He contributes to the data generation, design of baseline methods, and implementing of the competition protocol. He will also do the post-challenge analysis.
- **Yuan-Tang Chou**
  University of Washington, Seattle, USA - ytchou@uw.edu
  Yuan-Tang is a postdoc in an experimental particle group at the University of Washington working on the ATLAS Experiment at CERN. He is also a member of the A3D3 Institute, focusing on deploying Machine learning methods with heterogeneous computing systems at scale. He obtained his Ph.D. in Physics from the University of Massachusetts, USA. He contributes to the validation of competition and evaluation metrics.
- **Sascha Diefenbacher**
  Lawrence Berkeley National Laboratory, USA - sdiefenbacher@lbl.gov
  Sascha is a postdoctoral scholar at Lawrence Berkeley National Laboratory. She obtained her doctorate at the University of Hamburg, Germany, on generative modelling in high energy physics, covering both generative applications to high-dimensional data sets and the precision of generative models in general. She contributes to the competition design baseline methods and the establishment, design and development of evaluation metrics, as well as the post-challenge analysis.
- **Steven Farrell**
  (https://www.nersc.gov/about/nersc-staff/data-analytics-services/steven-farrell/)
  NERSC, Lawrence Berkeley National Laboratory, USA - SFarrell@lbl.gov
  Steven is a machine learning engineer and the lead for AI services at NERSC. He supports AI workflows on the NERSC supercomputers and collaborates with scientists who are applying AI methods to scientific problems. His background is in high-energy physics, and he has co-organised the TrackML competitions. He also co-founded the HPC working group in MLCommons, led the development of the MLPerf HPC benchmark suite, and organised the MLPerf HPC submission rounds from 2020-2022. He contributes to integrating the Codabench platform with the HPC resources at NERSC and to general software development and testing for the competition.

- **Aishik Ghosh**
  University of California, Irvine, USA - aishikg@uci.edu
  Lawrence Berkeley National Laboratory, USA
  Aishik is a UC Irvine postdoctoral scholar and a Berkeley Lab affiliate. He develops machine learning methods for particle and astrophysics, and one of his main focuses is uncertainty quantification and propagation. He contributes to the competition design, baseline methods, and design of performance metrics.
- **Isabelle Guyon** (https://guyon.chalearn.org/)
  ChaLearn, USA and Google, USA - guyon@chalearn.org
  She is president of ChaLearn (a non-profit organisation dedicated to organising challenges in Machine Learning), a professor at Université Paris-Saclay, France, and a Research Director at Google. Her research is on machine learning, with an interest in bias in data, privacy and fairness, and applications of large language models to human-AI co-creation. She is a long-time challenge organiser, having organised dozens of challenges (including 7 NeurIPS competitions), and creator of the NeurIPS competition track. She is the community lead of the open-source projects Codalab and Codabench of competition platforms. She contributes to the competition design and the coordination of its implementation.
- **Chris Harris**
  Lawrence Berkeley National Laboratory, USA - cjh@lbl.gov
  He is a Scientific Data Architect in the Data and AI services Group at NERSC, Berkeley Lab. He focuses on all aspects of scientific software development. He develops the software to integrate the Codabench platform with NERSCs HPC resources.
- **Elham E Khoda**
  University of California, San Diego, USA - ekhoda@ucsd.edu
  Elham is a computational data science researcher at the San Diego Supercomputing Centre (SDSC) at the UC San Diego. He was a postdoc in the experimental particle physics group of the University of Washington when he joined the project. Elham contributes to the data generation and competition design.
- **Benjamin Nachman** (http://go.lbl.gov/nachmangroup)
  Lawrence Berkeley National Laboratory, USA - bpnachman@lbl.gov
  He is a staff scientist and leader of the Machine Learning for Fundamental Physics Group in the Physics Division at Berkeley Lab. His group develops, adapts, and deploys deep learning solutions to particle, nuclear, and astrophysics. He contributes to the development and curation of the challenge dataset as well as the establishment of benchmarks and evaluation metrics.
- **David Rousseau** (https://users.ijclab.in2p3.fr/david-rousseau/en/home/)
  Université Paris-Saclay, CNRS/IN2P3, IJCLab, 91405 Orsay, France
  rousseau@ijclab.in2p3.fr
  He is a particle physicist who participated in the ATLAS experiment at the Large Hadron Collider. He has been working on various uses of AI in High Energy Physics. He was one of the organisers of the 2014 HiggsML challenge on Kaggle and the 2019 TrackML challenge of Kaggle and Codalab (the first version of Codabench). He supervises the competition protocol and implementation.
- **Ihsan Ullah** (https://ihsaan-ullah.github.io/)
  ChaLearn, USA - ihsan2131@gmail.com
  He is a Research Software Engineer at ChaLearn, USA. He is working on challenge organisation, data preparation, machine learning and software development. He contributes to implementing the competition protocol and setting up competition on the Codabench platform.

References

[1] 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.

[2] 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].

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

[4] 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].

[5] 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].

[6] 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].

[7] 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].

[8] 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].

[9] 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) .

[10] 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.

[11] 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.

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

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

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

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

[16] 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].

[17] 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].

[18] A. Paszke, S. Gross, F. Massa, A. Lerer, J. Bradbury, G. Chanan, T. Killeen, Z. Lin, N. Gimelshein, L. Antiga, A. Desmaison, A. Köpf, E. Yang, Z. DeVito, M. Raison, A. Tejani, S. Chilamkurthy, B. Steiner, L. Fang, J. Bai, and S. Chintala, *Pytorch: An imperative style, high-performance deep learning library*, 2019.

[19] M. Abadi, A. Agarwal, P. Barham, E. Brevdo, Z. Chen, C. Citro, G. S. Corrado, A. Davis, J. Dean, M. Devin, S. Ghemawat, I. Goodfellow, A. Harp, G. Irving, M. Isard, Y. Jia, R. Jozefowicz, L. Kaiser, M. Kudlur, J. Levenberg, D. Mané, R. Monga, S. Moore, D. Murray, C. Olah, M. Schuster, J. Shlens, B. Steiner, I. Sutskever, K. Talwar, P. Tucker, V. Vanhoucke, V. Vasudevan, F. Viégas, O. Vinyals, P. Warden, M. Wattenberg, M. Wicke, Y. Yu, and X. Zheng, *TensorFlow: Large-scale machine learning on heterogeneous systems*, 2015. https://www.tensorflow.org/. Software available from tensorflow.org.

[20] F. Pedregosa, G. Varoquaux, A. Gramfort, V. Michel, B. Thirion, O. Grisel, M. Blondel, P. Prettenhofer, R. Weiss, V. Dubourg, J. Vanderplas, A. Passos, D. Cournapeau, M. Brucher, M. Perrot, and E. Duchesnay, *Scikit-learn: Machine learning in Python*, Journal of Machine Learning Research **12** (2011) 2825.

[21] *Artificial Intelligence and the Uncertainty Challenge in Fundamental Physics*, 2023. https://indico.cern.ch/e/aiuphys2023.

[22] *Advanced Computing and Analysis Techniques in Physics Research*, 2024. https://indico.cern.ch/e/acat2024.

[23] T. Chen and C. Guestrin, *XGBoost: A scalable tree boosting system*, in Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, KDD '16. ACM, New York, NY, USA, 2016. http://doi.acm.org/10.1145/2939672.2939785.

[24] B. Xu, *First public annoucement of XGBoost*, 2014. https://www.kaggle.com/competitions/higgs-boson/discussion/8184.

[25] A. Marot, D. Rousseau, and Z. Xu, *Ai competitions and benchmarks: towards impactful challenges with post-challenge papers, benchmarks and other dissemination actions*, in AI Competitions and Benchmarks: the science behind the contests. 2024. arXiv:2312.06036 [cs.LG].

[26] 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].

[27] Z. Liu, A. Pavao, Z. Xu, S. Escalera, F. Ferreira, I. Guyon, S. Hong, F. Hutter, R. Ji, J. C. S. J. Junior, G. Li, M. Lindauer, Z. Luo, M. Madadi, T. Nierhoff, K. Niu, C. Pan, D. Stoll, S. Treguer, J. Wang, P. Wang, C. Wu, Y. Xiong, A. Zela, and Y. Zhang, *Winning solutions and post-challenge analyses of the chalearn autodl challenge 2019*, IEEE Transactions on Pattern Analysis and Machine Intelligence **43** (2021) 3108.