

# Fair Universe HiggsML Uncertainty Challenge

## Lessons Learned and Plans

Elham E Khoda  
University of Washington

1st December, 2023

**Artificial Intelligence and the Uncertainty challenge in Fundamental Physics**

---

### FAIR Universe Team

Wahid Bhimji, Paolo Calafiura, Ragansu Chakkappai, Yuan-Tang Chou, Sascha Diefenbacher, Steven Farrell, Aishik Ghosh, Isabelle Guyon, Chris Harris, Shih-Chieh Hsu, Elham E Khoda, Benjamin Nachman, Benjamin Thorne, Peter Nugent, Mathis Reymond, David Rousseau, Ihsan Ullah, Daniel Whiteson

# Project Goals

---

## US Dept. of Energy, AI for HEP project

- **Large-compute-scale AI ecosystem** for sharing datasets, training large models, fine-tuning those models, and **hosting challenges and benchmarks**
  - Participants were able to run on NERSC Perlmutter ( one of the DOE supercomputers at the Berkeley Lab) → *started testing this week*
- Create **public datasets**
  - The dataset is public
- Measuring and minimizing the **effects of systematic** uncertainties
  - This was the first hackathon and demo challenge



**Website:** <https://fair-universe.lbl.gov/>

# HiggsML Uncertainty Challenge

Improve Higgs boson ( $H \rightarrow \tau\tau$  decay mode) signal strength ( $\mu$ ) in the presence of the background  $Z \rightarrow \tau\tau$  process

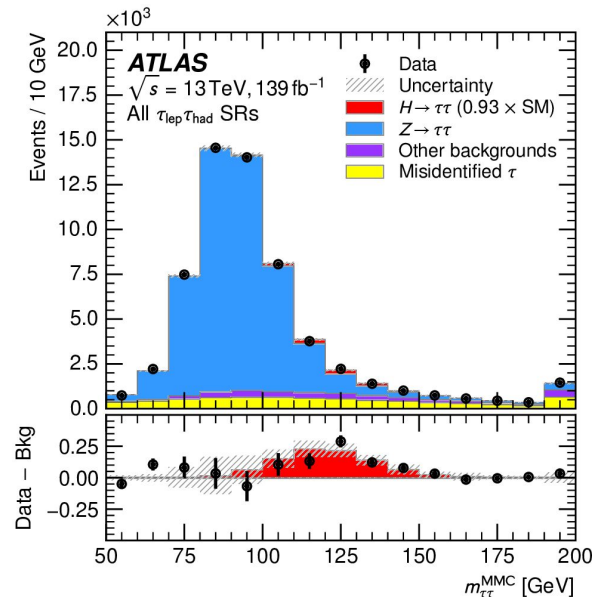
How this is different from HiggsML challenge?

- The effect of systematic uncertainty is included in the problem
- One uncertainty corresponding to the Tau Energy Scale (TES)
- Also the dataset will be much larger

## Objective

Your algorithm should predict

- Signal strength ( $\mu$ )
- Uncertainty on signal strength ( $\Delta\mu$ )
- 16% and 84% quantiles



[ATLAS HIGG-2019-09](#), [JHEP 08 \(2022\) 175](#)

# HiggsML Uncertainty Challenge: Paris Version

[https://www.codabench.org/competitions/1299/?secret\\_key=28d9c0fc-fe66-44c8-be89-0f2c712b4514](https://www.codabench.org/competitions/1299/?secret_key=28d9c0fc-fe66-44c8-be89-0f2c712b4514)

The screenshot shows the competition page for 'FAIR UNIVERSE: HIGGSML UNCERTAINTY CHALLENGE'. The page features a dark header with navigation links: 'Search Competitions', 'Benchmarks', 'Resources', 'Queue Management', 'Login', and 'Sign-up'. The main content area includes a circular image of a globe with data points, the challenge title, and statistics: 33 PARTICIPANTS and 48 SUBMISSIONS. It also lists the organizer (Insaan-Ullah), current phase end date (December 3, 2023), and current server time (November 30, 2023). A progress bar shows the timeline from October to December 2023. Below the main content is a navigation menu with 'Get Started', 'Phases', 'My Submissions', 'Results', and 'Forum'. On the left, there is a sidebar menu with 'Overview', 'Evaluation', 'Data', 'Starting Kit', 'Example Estimators', 'Terms', and 'Files'. The main content area has an 'Overview' section with an 'Introduction' paragraph.

**FAIR UNIVERSE: HIGGSML UNCERTAINTY CHALLENGE**

33 PARTICIPANTS  
48 SUBMISSIONS

ORGANIZED BY: Insaan-Ullah  
CURRENT PHASE ENDS: December 3, 2023 At 1:00 AM GMT+1  
CURRENT SERVER TIME: November 30, 2023 At 6:21 PM GMT+1  
Docker image: cjh1/fair\_universe:latest

Oct 2023 Nov 2023 Dec 2023

Get Started Phases My Submissions Results Forum

Overview

Evaluation  
Data  
Starting Kit  
Example Estimators  
Terms  
Files

### Overview

#### Introduction

In 2012, the Nobel-prize-winning discovery of the Higgs Boson by the ATLAS and CMS experiments at the Large Hadron Collider (LHC) at CERN in Geneva, Switzerland was a major milestone in the history of physics. However, despite the validation it provided of the Standard Model of particle physics (SM), there are still numerous questions in physics that the SM does not answer. One promising approach to uncover some of these mysteries is to study the Higgs Boson in great detail, as the rate of Higgs Boson production and its decay properties may hold the secrets to the nature of dark matter and other phenomena not explained by the SM.

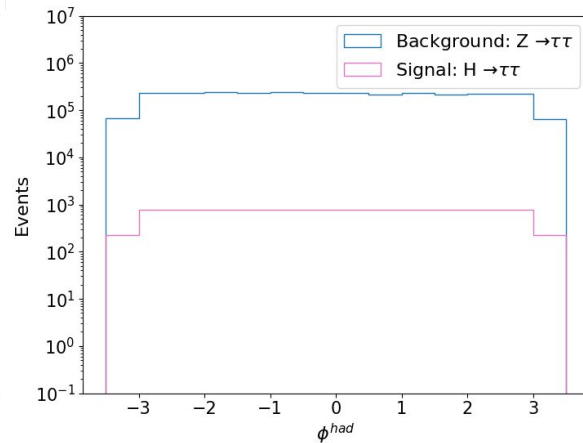
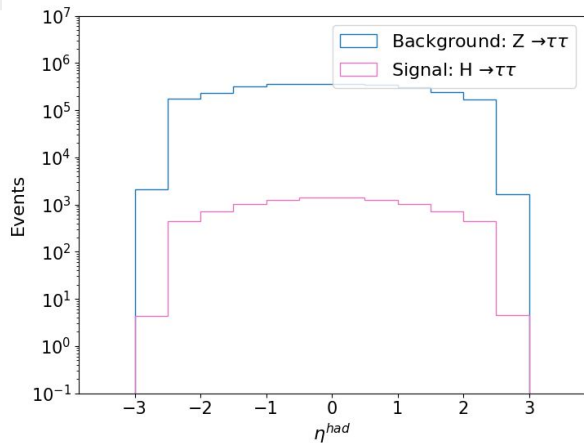
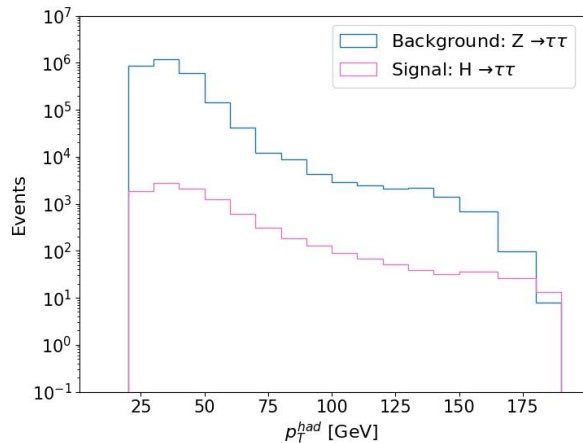
The LHC collides protons together at high energy and at a high rate. Each proton collision produces many outgoing particles. A small fraction of the time, a Higgs boson is produced and then decays into other particles that can be detected by

# Problem Dataset

Signal (label = 1) and Background (label = 0) events are mixed

*We have over 30 feature variables in the dataset*

Some example features



# Problem Dataset: 1 systematic

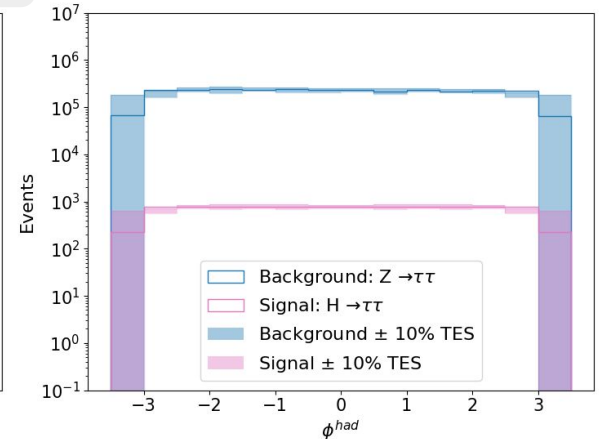
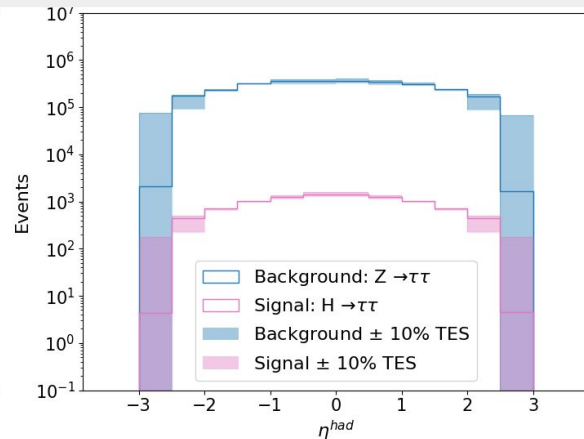
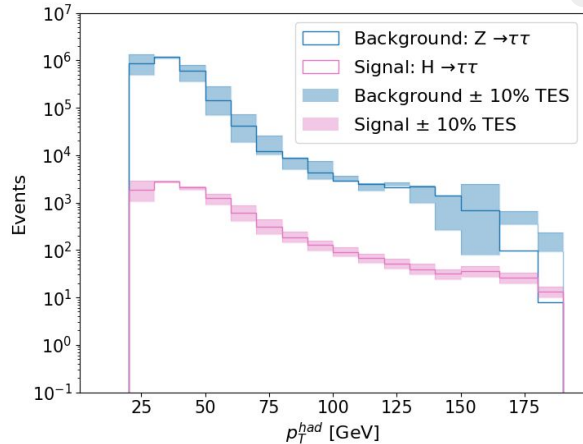
We made the problem harder by adding a systematic uncertainty

In this challenge:

- Only 1 systematic uncertainty: Tau Energy Scale (TES)

*We have over 30 feature variables in the dataset*

Some example features



# Current Results

Thanks to all of you who where here on Wednesday afternoon!

- There are a few few initial submissions
- Many of you already already working

Results								
Task:				Fact Sheet Answers	Higgs Uncertainty Challenge			
#	Participant	Entries	Date of last entry	Method Name	Quantile Score	Interval	Coverage	Detailed Results
1	laurens slu	20	2023-12-01	cheat4	0.16	0.852	0.71	
2	laurens slu	20	2023-12-01	cheat7	0.61	0.544	0.68	
3	laurens slu	20	2023-12-01	cheat7	0.68	0.504	0.63	
4	laurens slu	20	2023-12-01	Cheat2	-0.44	1.55	0.62	
5	laurens slu	20	2023-12-01	cheat4	0.31	0.732	0.61	
6	laurens slu	20	2023-12-01	Cheat2	-0.74	1.375	0.55	
7	ragansu	11	2023-12-01	tes_finder	-0.95	1.124	0.54	
8	laurens slu	20	2023-12-01	Cheat2	-1.59	1.325	0.53	
9	ravalin	10	2023-12-01	1binNLL	-2.9	1.233	0.5	
10	ihsanchalearn	15	2023-12-01	1 bin NLL	-2.9	1.233	0.5	
11	ravalin	10	2023-12-01	1binNLL	-2.9	1.233	0.5	
12	ihsanchalearn	15	2023-12-01	test - starting kit submission	-7.16	0.324	0.22	
13	ihsanchalearn	15	2023-12-01	XGB 2	-7.86	0.324	0.15	
14	ihsanchalearn	15	2023-12-01	1 bin NLL	-8.5	0.34	0.08	
15	ravalin	10	2023-12-01	1binNLL	-8.5	0.34	0.08	
16	ihsanchalearn	15	2023-12-01	test - starting kit submission	-7.19	0.084	0.07	
17	ihsanchalearn	15	2023-12-01	XGB 2	-7.3	0.081	0.05	
18	ihsanchalearn	15	2023-12-01	XGB 2	-7.44	0.08	0.03	
19	ihsanchalearn	15	2023-12-01	XGB NLL	-8.67	0.279	0.03	
20	ihsanchalearn	15	2023-12-01	XGB 1	-10.52	1.652	0.02	7

# Build up the complexity in multiple steps

---

**Observation:** Folks have difficulty understanding the challenge problem and try solutions

- Debugging is not easy

## Set up a hierarchy of tasks

1. Predict  $\mu$  on dataset without systematics
2. Predict  $\mu$  and  $\Delta\mu$  on dataset without systematics
3. Predict  $\mu$  and  $\Delta\mu$  on dataset with systematics



# Adding systematics to the training data

---

**Observation:** It was not very clear to the participants that systematics is not included in the training data

## Improve description and provide example

- We will make it more clear in the description
- The starting-kit has an example how to use the [systematics class](#)
- Also we will provide a cleaner stand-alone example such that it becomes clear to the participants how to use it

# Starting-kit: complicated directory structure

---

**Observation:** The github repo has too many directories

- It is confusing for the first-time users to find necessary information

## Simply the GitHub repository

- The repository contains the other examples we studied
- We will move to a new GitHub
  - It will have simpler directory structure

# Model will be tested on different $\mu$ value(s)

---

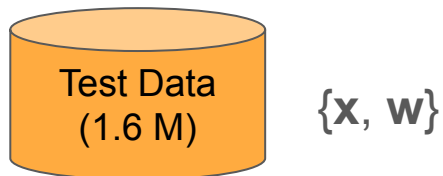
**Observation:** It might not be clear to the participants that the model has to work for different  $\mu$  values

- Default training corresponds  $\mu = 1$
- This effect should be included in the training process

## Describe and/or provide example

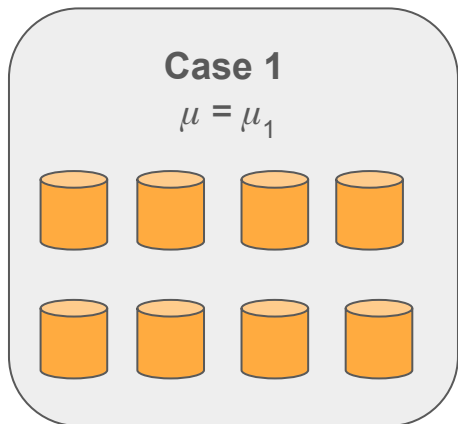
- How to simulate different  $\mu$  values in the training data
  - Mix different amount of signal and background

# Test Sets

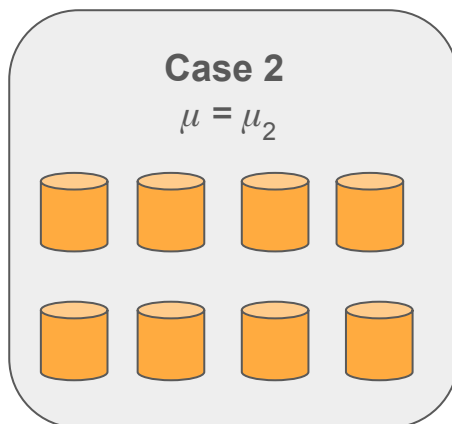


Create different test cases, Bootstrap to get 100 sets for each case

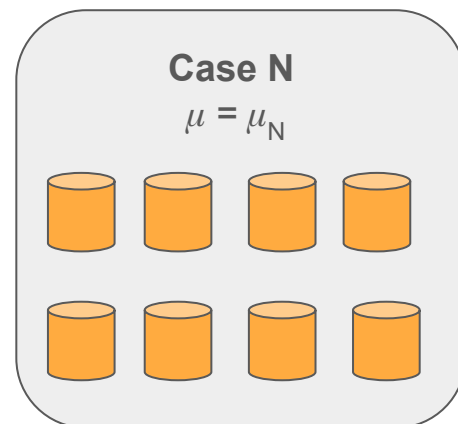
$\{x, w_1\}, w_1 = \text{Pois}(w)$



$\{x, w_2\}, w_2 = \text{Pois}(w)$



$\{x, w_N\}, w_N = \text{Pois}(w)$



# Bootstrap issues due to large event weight

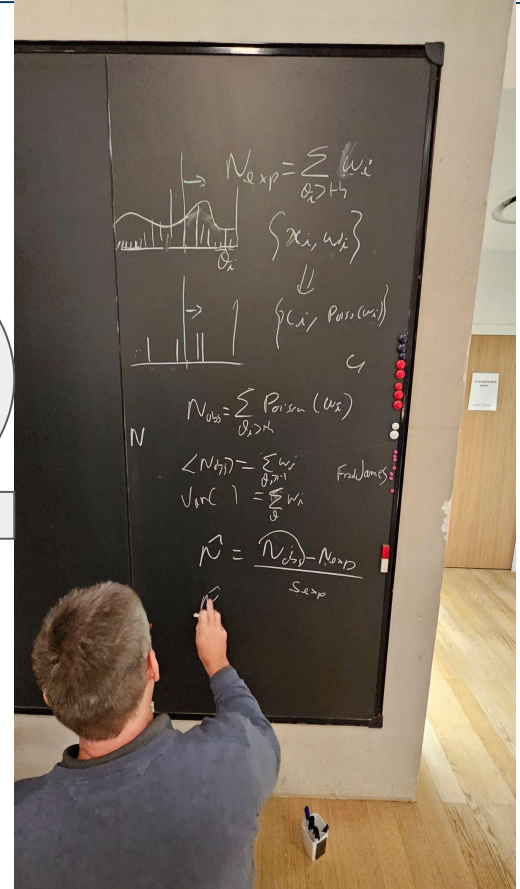
The challenge is considering a scenario of analyzing  $139 \text{ fb}^{-1}$  of proton-proton collision data  
→ Collected by the ATLAS experiment during the Run-II phase (2015-2018) of the LHC.

## Events weights:

- $\sim 0.015$ , for signals
- $\sim 10$  for background

David trying to convince that large weight should not matter

Bootstrapping based on Poisson pseudo-experiments had issues due to the large event weights



# Current Strategy

---

Event Weight = Cross-Section x Luminosity / total number of events generated

**Reason for having large event weights:**

- Not enough event to match the target luminosity of  $139 \text{ fb}^{-1}$

Through sampling and bootstrapping we were effectively counting a single event multiple times

**Solution:**

We are generating many more events such that we do not have event weights  $> 1$

# How to calculate score for multiple $\mu$ values

---

**Observation:** Calculating score by taking average coverage across multiple  $\mu$  values

- Averaging the coverage over multiple  $\mu$  values might obscure performance variations at different  $\mu$  values

**We will use a different strategy for scoring**

- Combine the scores from different test sets corresponding to different  $\mu$  values

# Uncertainty Quantification Metric

- For  $N$  test sets and predicted  $[\mu_{16}, \mu_{84}]_i, i \in [0, N]$ 
  - Calculate fraction of times interval contains  $\mu_{\text{true}}$  to get coverage  $c$ :

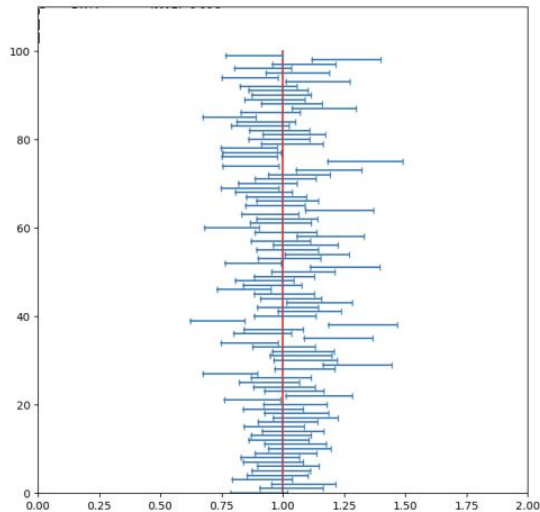
$$c = \frac{1}{N} \sum_{i=0}^N 1 \text{ if } (\mu_{\text{true},i} \in [\mu_{84} - \mu_{16}]_i)$$

- Calculate average interval width  $w$ :

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

- Combine both values for score  $s$ :

$$s = w f(c)$$



Ref: [Sasch's slides from Monday](#)



# Absolute value of the interval

---

**Observation:** Absolute value of  $\mu$  interval was not used for the width calculation ( $w$ )

- Allows negative values
- It can almost cancel the argument of log  $\rightarrow$  getting a high score value

$$s = -\ln[(w + \epsilon) f(c)]$$

Use absolute width ( $w$ ) values for the score calculation

- It is already fixed

# Other Comments related to scoring

---

Width (and therefore score) is sensitive to parameter scaling (e.g.  $\mu$  vs  $\mu^2$ )

→ Investigate impact and ways to mitigate

Only 68% coverage is included

→ Investigate inclusion of 95% and 99% intervals as well

Overcoverage is already discouraged by inclusion of width

→ Investigate if overcoverage penalty just through width is sufficient

**Alternative Metrics to look at for insight/inspiration:**

- CRPS metric
- Coverage width based criteria

# Current Winner!

Guess which method is winning at the moment!

As we are testing on one  $\mu$ , it is easier to cheat

It will be much more difficult when we will test it over multiple  $\mu$  values

Results								
Task:				Fact Sheet Answers	Higgs Uncertainty Challenge			
#	Participant	Entries	Date of last entry	Method Name	Quantile Score	Interval	Coverage	Detailed Results
1	laurens slu	20	2023-12-01	cheat4	0.16	0.852	0.71	
2	laurens slu	20	2023-12-01	cheat7	0.61	0.544	0.68	
3	laurens slu	20	2023-12-01	cheat7	0.68	0.504	0.63	
4	laurens slu	20	2023-12-01	Cheat2	-0.44	1.55	0.62	
5	laurens slu	20	2023-12-01	cheat4	0.31	0.732	0.61	
6	laurens slu	20	2023-12-01	Cheat2	-0.74	1.375	0.55	
7	ragansu	11	2023-12-01	tes_finder	-0.95	1.124	0.54	
8	laurens slu	20	2023-12-01	Cheat2	-1.59	1.325	0.53	
9	ravalin	10	2023-12-01	1binNLL	-2.9	1.233	0.5	
10	ihsanchalearn	15	2023-12-01	1 bin NLL	-2.9	1.233	0.5	
11	ravalin	10	2023-12-01	1binNLL	-2.9	1.233	0.5	
12	ihsanchalearn	15	2023-12-01	test - starting kit submission	-7.16	0.324	0.22	
13	ihsanchalearn	15	2023-12-01	XGB 2	-7.86	0.324	0.15	
14	ihsanchalearn	15	2023-12-01	1 bin NLL	-8.5	0.34	0.08	
15	ravalin	10	2023-12-01	1binNLL	-8.5	0.34	0.08	
16	ihsanchalearn	15	2023-12-01	test - starting kit submission	-7.19	0.084	0.07	
17	ihsanchalearn	15	2023-12-01	XGB 2	-7.3	0.081	0.05	
18	ihsanchalearn	15	2023-12-01	XGB 2	-7.44	0.08	0.03	
19	ihsanchalearn	15	2023-12-01	XGB NLL	-8.67	0.279	0.03	
20	ihsanchalearn	15	2023-12-01	XGB 1	-10.52	1.652	0.02	

# One of winning cheat solutions

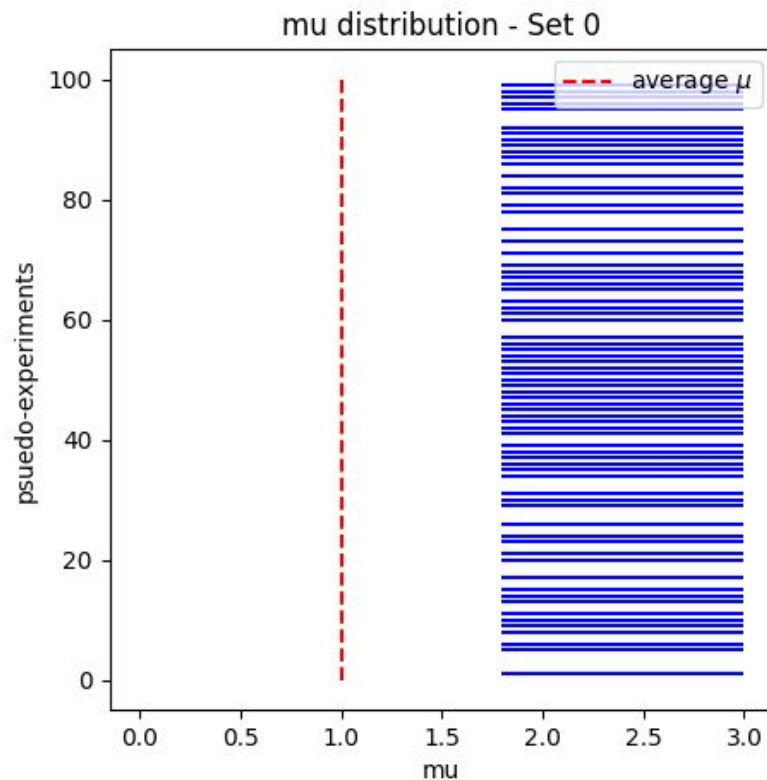
## Interval:

- It predicts a fixed interval ~60 % of time
- Otherwise it predicts interval = 0

It predicts a constant  $\mu$  value every time

For a single  $\mu$  is is easy to get a good estimation of the interval by multiple submissions

→ the situation will change when we have multiple  $\mu$  value



# Next Steps: Short Term

---

**The competition will remain open for next 2-3 months**

- Please continue working and send us feedback
- **We appreciate your patience and support!**

**Few expected upcoming changes:**

- Updated dataset once we have more simulated events
- Re-think about the scoring criteria
- Add multiple tasks with increasing complexity

# Next Steps: Longer Term

---

**Update the competition to make it closer to the real scenario** → **make it a public challenge hopefully as a NeurIPS 2024 competition**

**Add more background processes:**

- Currently we only had one background process (Z boson)
- We will add 3 other processes

**Add more systematics:**

- 3-4 other experimental systematics will be added (like MET, JES, bkg comp)

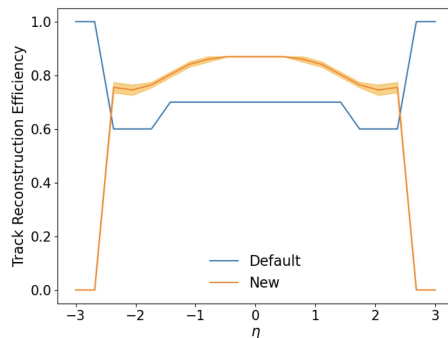
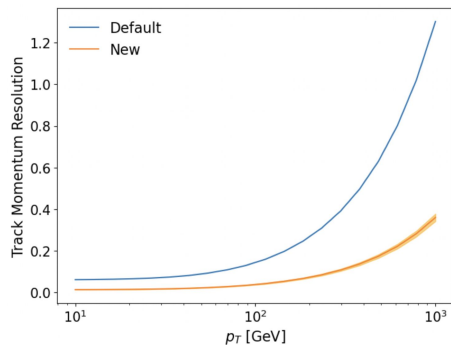
**We value your feedback!**

**Please let us know how we should modify the challenge such that you can participate with your uncertainty-aware method (you might currently have)**

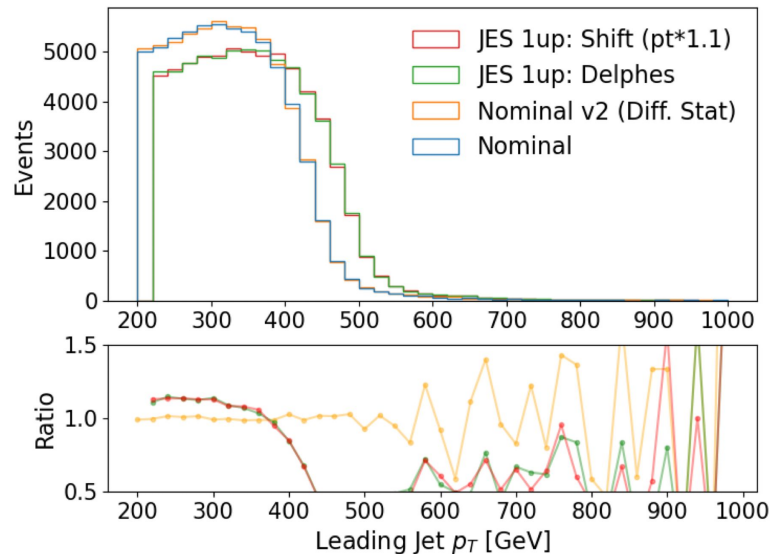
# Systematics with Delphes

## We have updated the ATLAS Delphes Card

- Include latest ATLAS results
- Define alternative functions to create systematics variations



## Systematics added via Delphes and post-hoc shifting



# Systematics with Delphes

## We have updated the ATLAS Delphes Card

- Include late
- Define alternative systematic

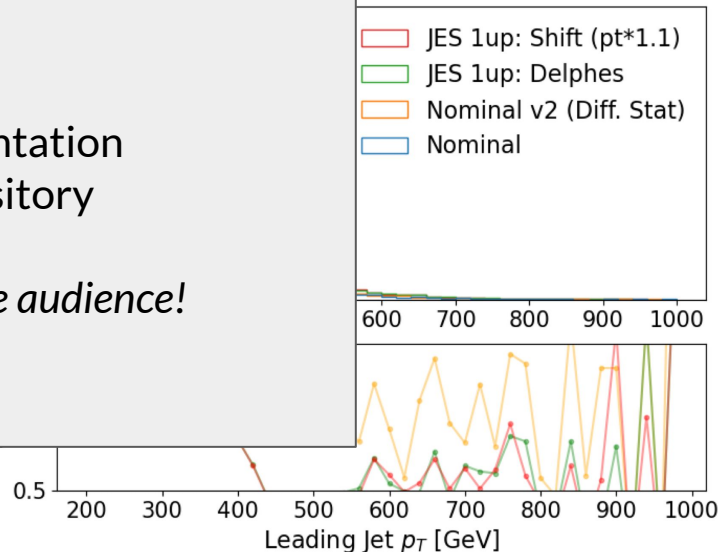
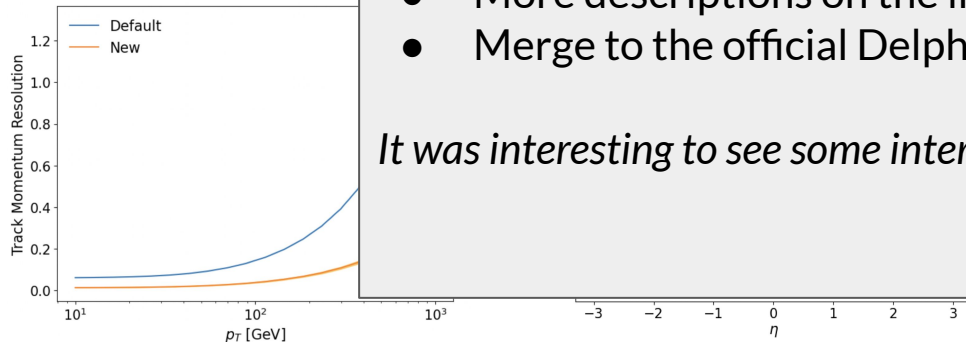
## Systematics added via Delphes

Plotting

**Coming Soon!**

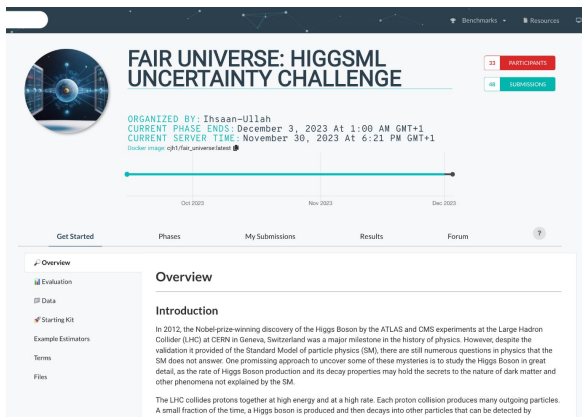
- More descriptions on the implementation
- Merge to the official Delphes repository

*It was interesting to see some interest in the audience!*

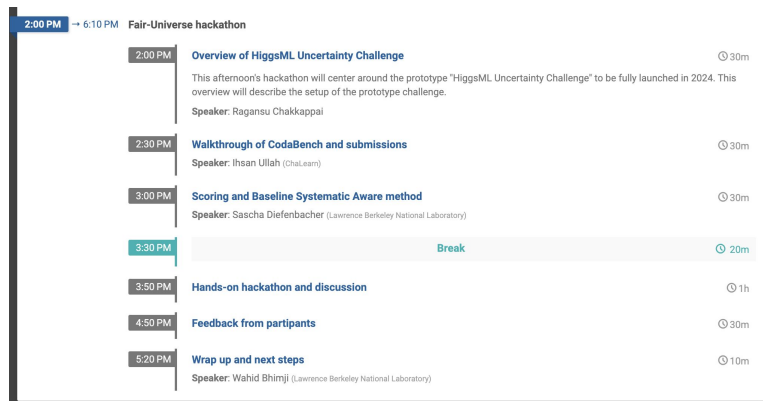




# Continue working and send us Feedback!



The screenshot shows the 'FAIR UNIVERSE: HIGGSML UNCERTAINTY CHALLENGE' website. It features a navigation bar with 'Benchmarks' and 'Resources'. The main content area includes a circular logo, the challenge title, and organizational details: 'ORGANIZED BY: Ihsaan-Ullah', 'CURRENT PHASE ENDS: December 9, 2023 At 1:00 AM GMT+1', and 'CURRENT SERVER TIME: November 30, 2023 At 6:21 PM GMT+1'. A progress bar shows the timeline from October 2023 to December 2023. Below the progress bar are tabs for 'Get Started', 'Phases', 'My Submissions', 'Results', and 'Forum'. A sidebar on the left contains a menu with 'Overview', 'Evaluation', 'Data', 'Starting Kit', 'Example Estimators', 'Terms', and 'Files'. The main content area is titled 'Overview' and contains an 'Introduction' section with text about the Higgs Boson discovery and the challenge's goal.



The screenshot shows the 'Fair-Verse hackathon' schedule. The title bar indicates the time range from 2:00 PM to 6:10 PM. The schedule is as follows:

- 2:00 PM** Overview of HiggsML Uncertainty Challenge (30m)  
This afternoon's hackathon will center around the prototype 'HiggsML Uncertainty Challenge' to be fully launched in 2024. This overview will describe the setup of the prototype challenge.  
Speaker: Ragansu Chakkappai
- 2:30 PM** Walkthrough of CodaBench and submissions (30m)  
Speaker: Ihsan Ullah (Chat:eam)
- 3:00 PM** Scoring and Baseline Systematic Aware method (30m)  
Speaker: Sascha Diefenbacher (Lawrence Berkeley National Laboratory)
- 3:30 PM** Break (20m)
- 3:50 PM** Hands-on hackathon and discussion (1h)
- 4:50 PM** Feedback from participants (30m)
- 5:20 PM** Wrap up and next steps (10m)  
Speaker: Wahid Bhimji (Lawrence Berkeley National Laboratory)

Join the Google Group: [Fair-Universe-Announcements](#)

[#fair-universe-hackathon](#) channel on [AIUPHYS2023 slack workspace](#)

**Collaborations, questions, comments:**  
Wahid Bhimji [wbhimji@lbl.gov](mailto:wbhimji@lbl.gov)

# Thank You!!





# Other Possible Metrics

---

Someone suggested looking into the CRPS metric, which is apparently used a lot in environmental science

Someone suggested looking into the 'Coverage width based criteria' metric, which is used in math I think





