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

FAIR Universe

Website: https://fair-universe.lbl.gov/

HiggsML Uncertainty Challenge

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

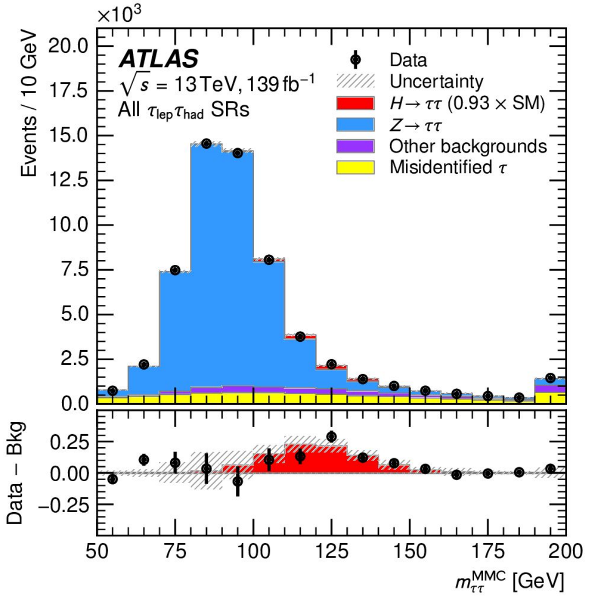
How this is different from HiggsML challenge?

- 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

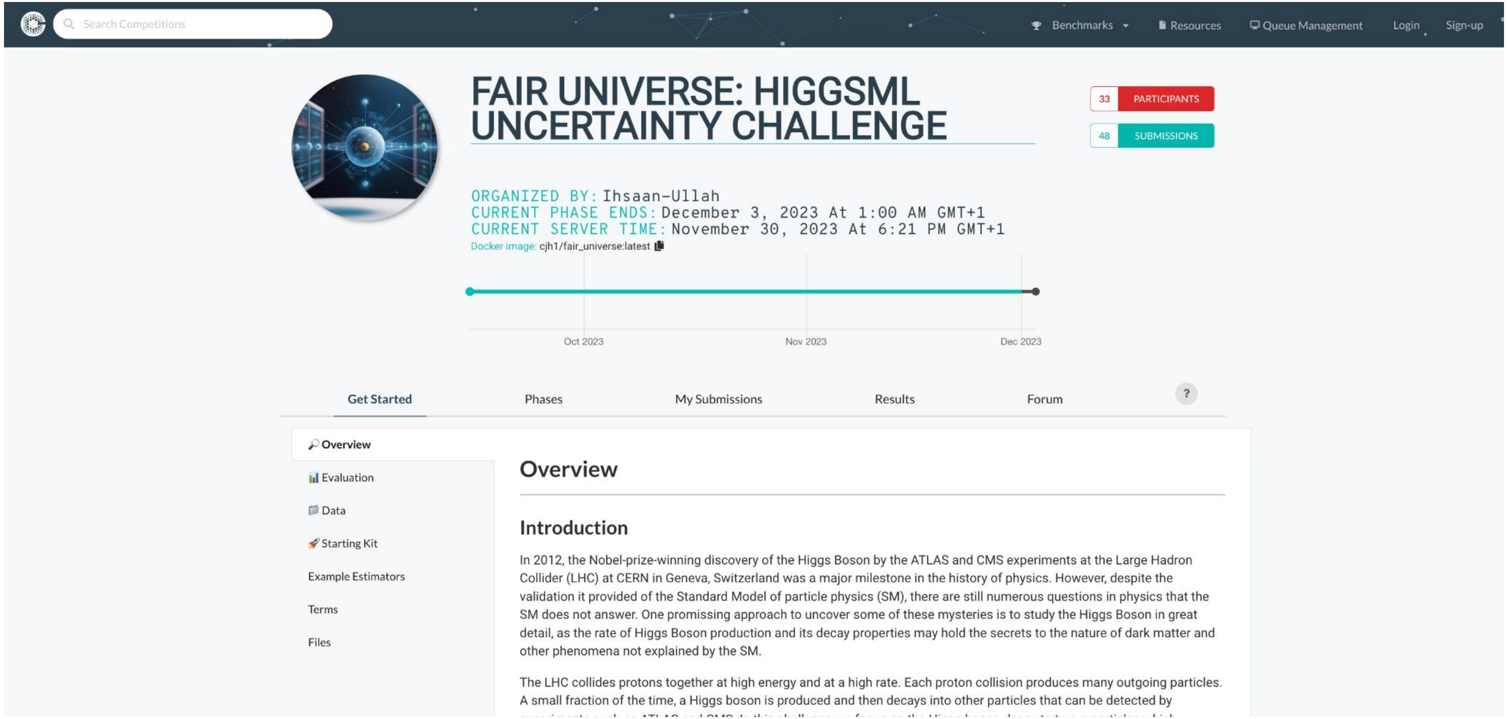
- Signal strength (μ)
- Uncertainty on signal strength ($\Delta\mu$)
- 16% and 84% quantiles



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

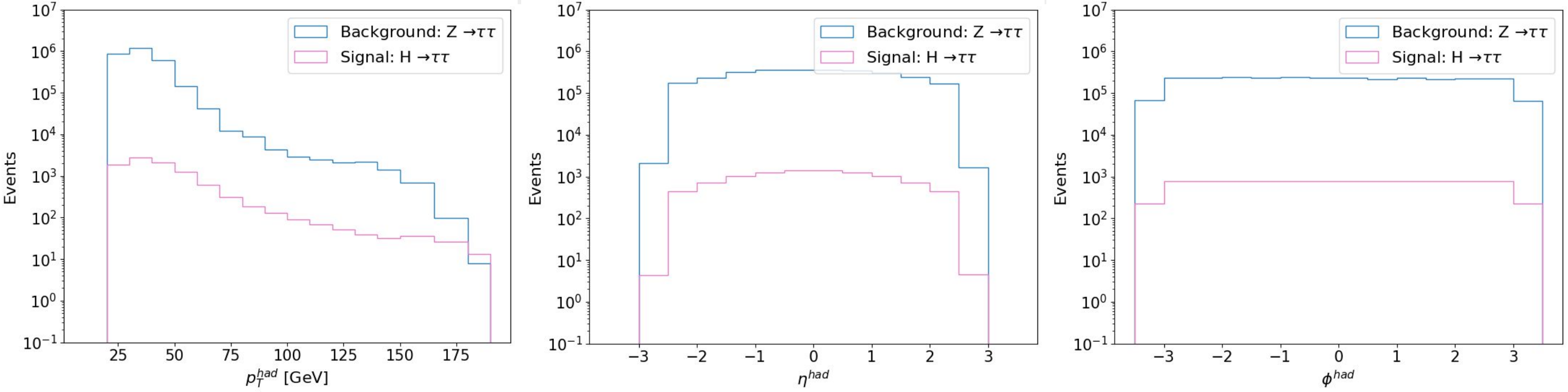


Problem Dataset

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 systematics

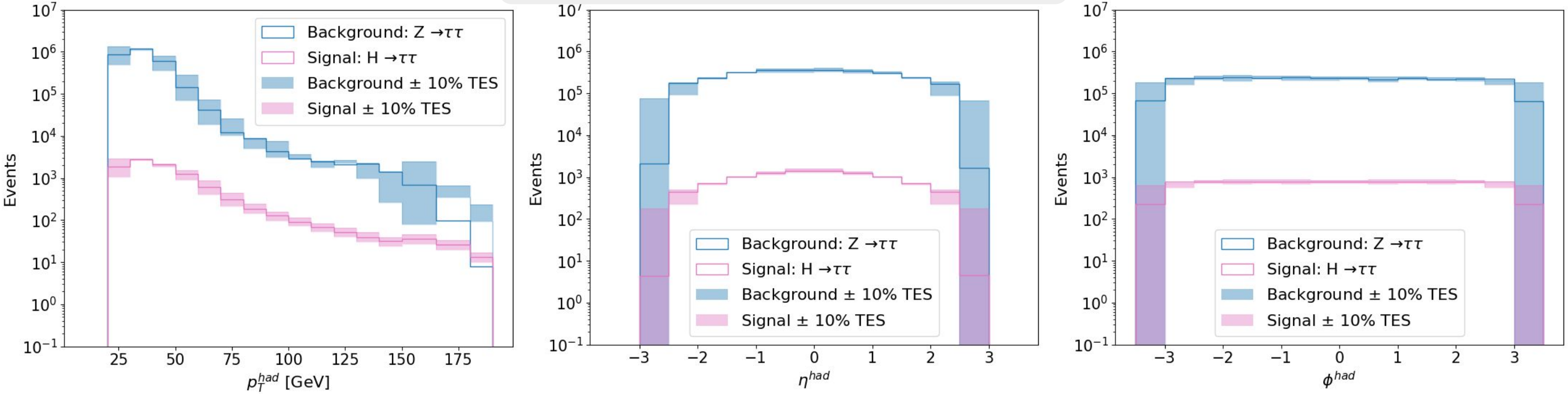
We made the problem harder by adding a systematic uncertainty

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	laurensslu	20	2023-12-01	cheat4	0.16	0.852	**0.71**	
2	laurensslu	20	2023-12-01	cheat7	0.61	0.544	**0.68**	
3	laurensslu	20	2023-12-01	cheat7	0.68	0.504	**0.63**	
4	laurensslu	20	2023-12-01	Cheat2	-0.44	1.55	**0.62**	
5	laurensslu	20	2023-12-01	cheat4	0.31	0.732	**0.61**	
6	laurensslu	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	laurensslu	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**	

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 μ on dataset without systematics
2. Predict μ and $\Delta\mu$ on dataset without systematics
3. Predict μ 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 μ value(s)

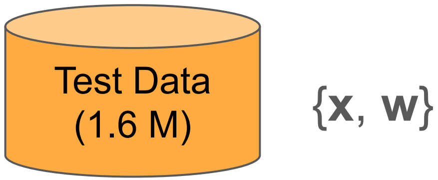
Observation: It might not be clear to the participants that the model has to work for different μ values

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

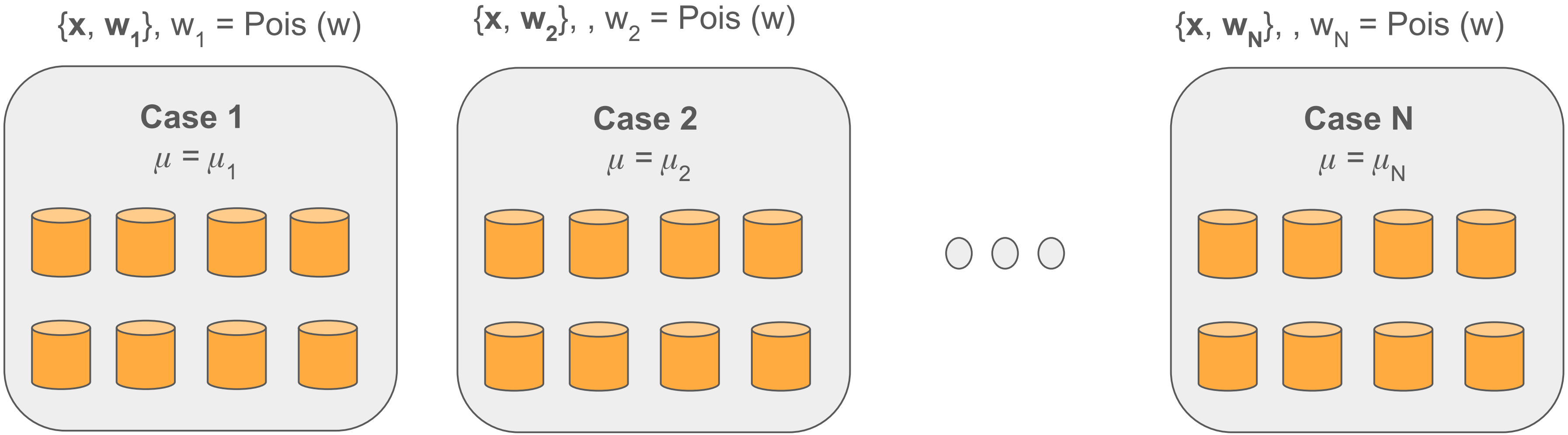
Describe and/or provide example

- 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



Bootstrap issues due to large event weight

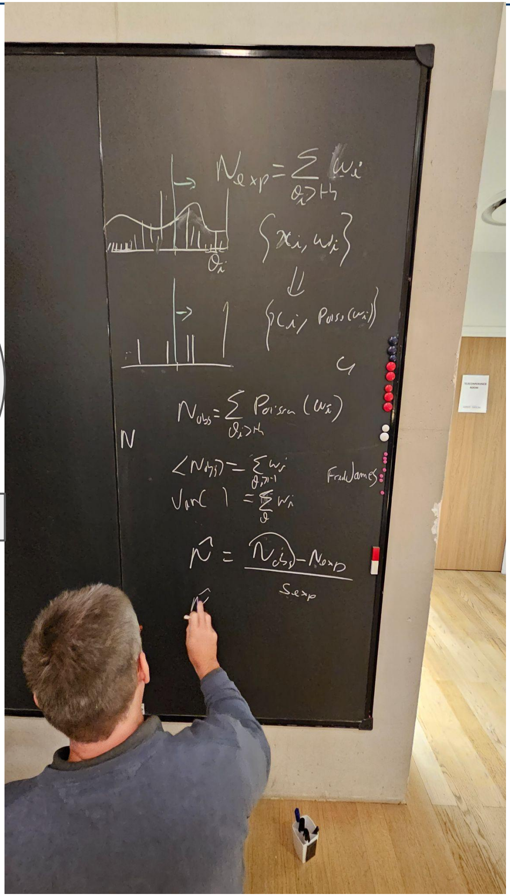
The challenge is considering a scenario of analyzing $139\ \mathrm{fb}^{-1}$ of proton-proton collision data
→ Collected by the ATLAS experiment during the Run-II phase (2015-2018) of the LHC.

Events weights:

- ~0.015, for signals
- ~10 for background

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

David trying to convince that large weight should not matter



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 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 μ values

Observation: Calculating score by taking average coverage across multiple μ values

- Averaging the coverage over multiple μ values might obscure performance variations at different μ values

We will use a different strategy for scoring

- Combine the scores from different test sets corresponding to different μ 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 μ_{true} to get coverage c:

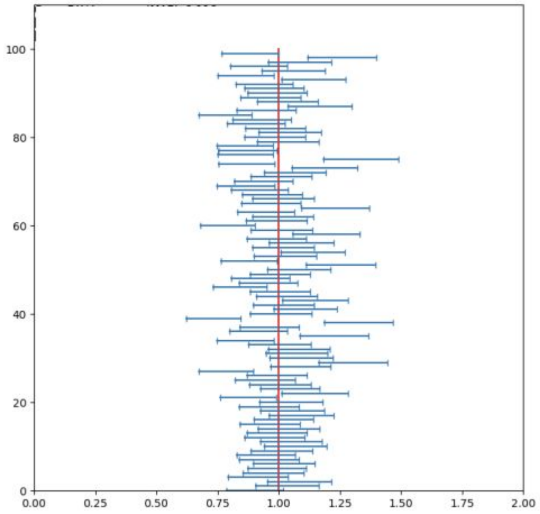
    $$c = \frac{1}{N}\sum_{i=0}^{N} 1 \text{ if}(\mu_{\text{true},i} \in [\mu_{84} - \mu_{16}]_i)$$

  - 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 μ interval was not used for the width calculation (w)

- Allows negative values
- It can almost cancel the argument of log → 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. μ vs μ^2)
$\rightarrow$ Investigate impact and ways to mitigate

Only 68% coverage is included
$\rightarrow$ Investigate inclusion of 95% and 99% intervals as well

Overcoverage is already discouraged by inclusion of width
$\rightarrow$ 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 μ, it is easier to cheat

It will be much more difficult when we will test it over multiple μ values

Results

Task:				Fact Sheet Answers	Higgs Uncertainty Challenge			
#	Participant	Entries	Date of last entry	Method Name	Quantile Score	Interval	Coverage	Detailed Results
1	laurenssIu	20	2023-12-01	cheat4	0.16	0.852	**0.71**	
2	laurenssIu	20	2023-12-01	cheat7	0.61	0.544	**0.68**	
3	laurenssIu	20	2023-12-01	cheat7	0.68	0.504	**0.63**	
4	laurenssIu	20	2023-12-01	Cheat2	-0.44	1.55	**0.62**	
5	laurenssIu	20	2023-12-01	cheat4	0.31	0.732	**0.61**	
6	laurenssIu	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	laurenssIu	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

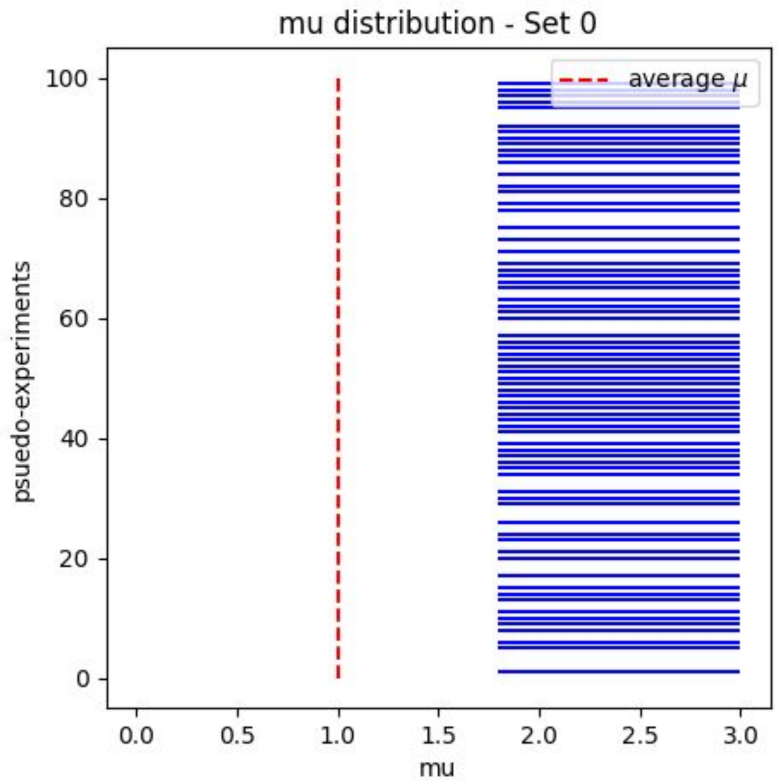
Interval:

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

It predicts a constant μ value every time

For a single μ is is easy to get a good estimation of the interval by multiple submissions

→ the situation will change when we have multiple μ 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)

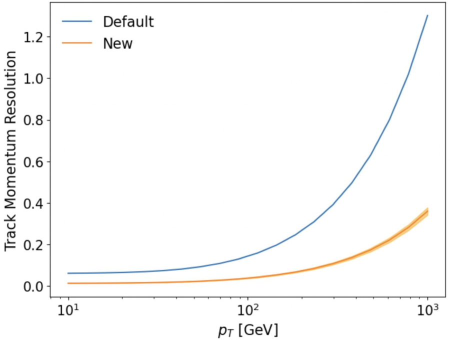
We value your feedback!

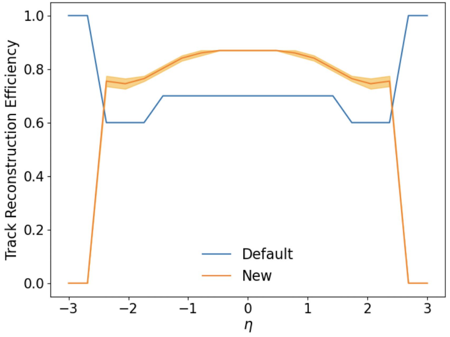
Please let us know how we should modify the challenge such that you can participate with your uncertainty-aware method (you might currently have)

Systematics with Delphes

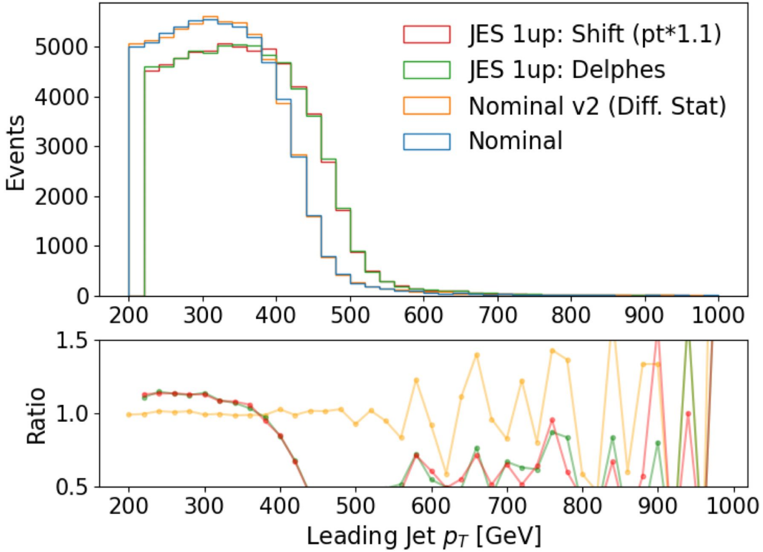
We have updated the ATLAS Delphes Card

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





Systematics added via Delphes and post-hoc shifting



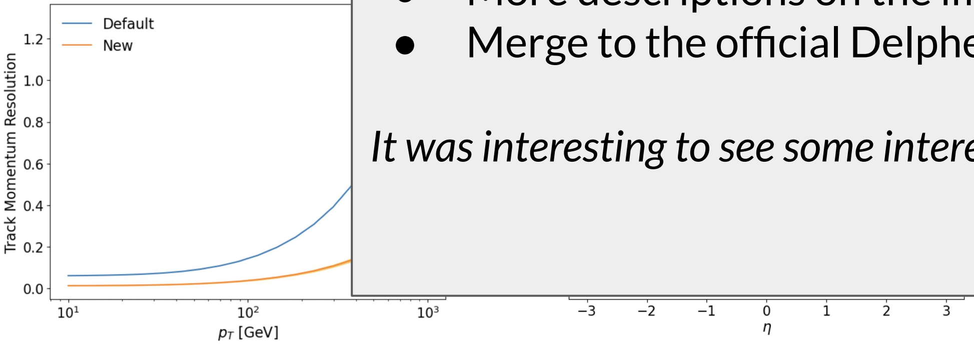
Systematics with Delphes

We have updated the ATLAS Delphes Card

- Include lat
- Define alte
  systemati

Systematics added via Delphes

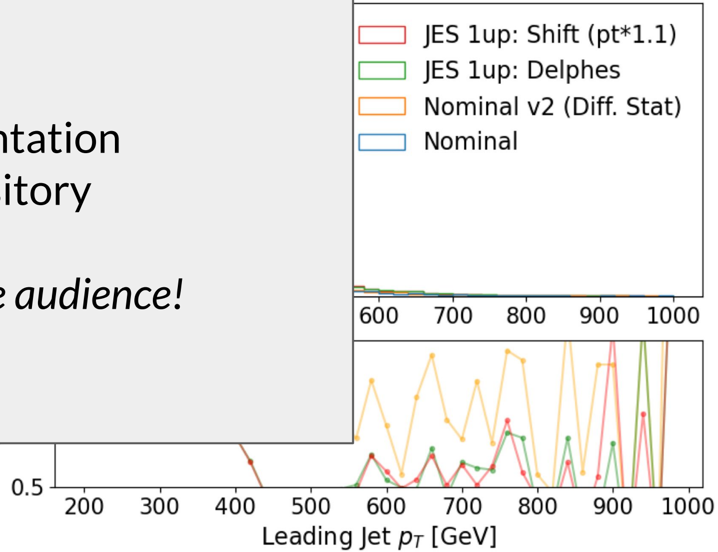
ting



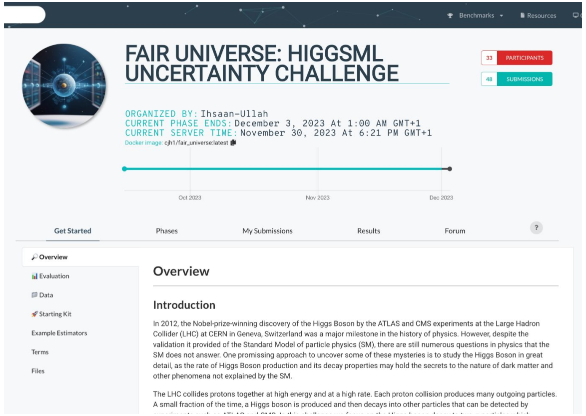
Coming Soon!

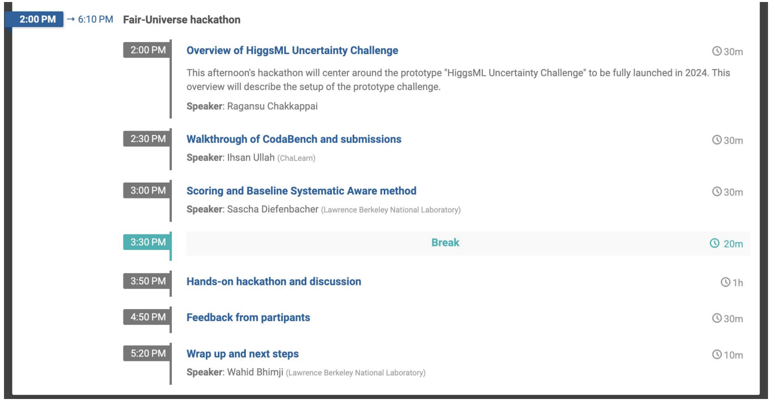
- More descriptions on the implementation
- Merge to the official Delphes repository

It was interesting to see some interest in the audience!



Continue working and send us Feedback!





Join the Google Group: Fair-Universe-Announcements

#fair-universe-hackathon channel on AIUPHYS2023 slack workspace

Collaborations, questions, comments:
Wahid Bhimji wbhimji@lbl.gov

Thank You!!

Other Possible Metrics

- Someone suggested looking into the CRPS metric, which is appearently used a lot in environmental science
- Someone suggested looking into the 'Coverage width based criteria' metric, which is used in math I think