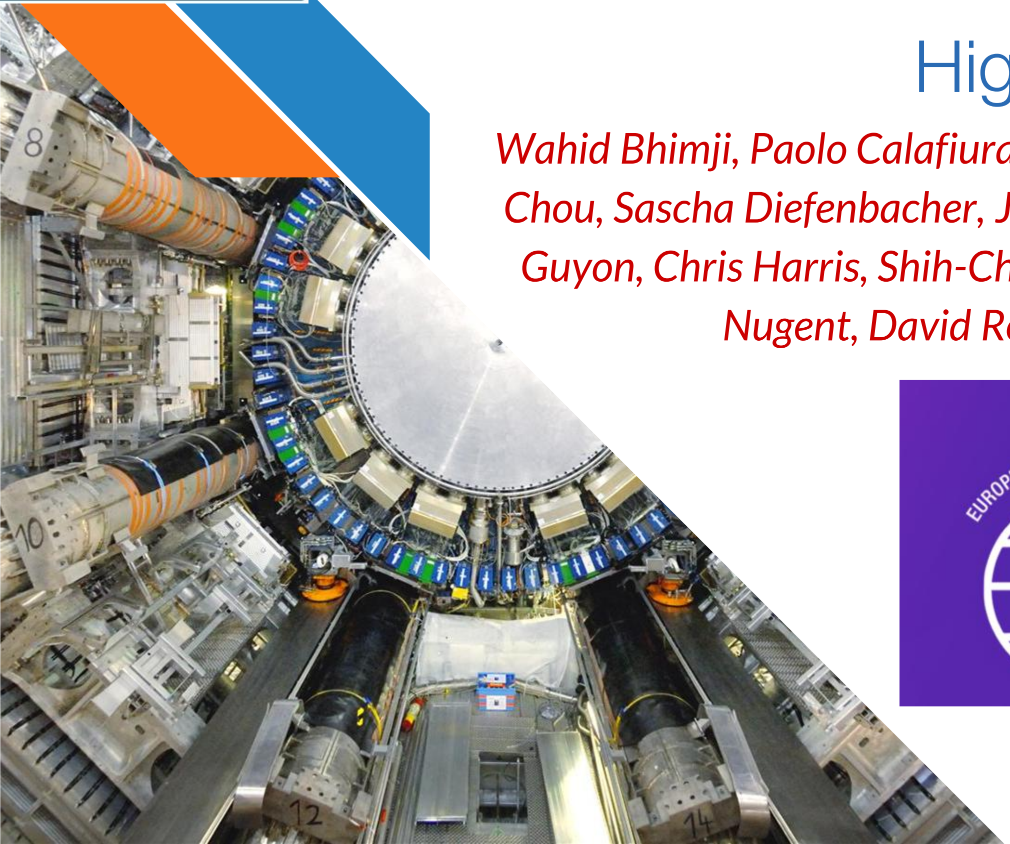
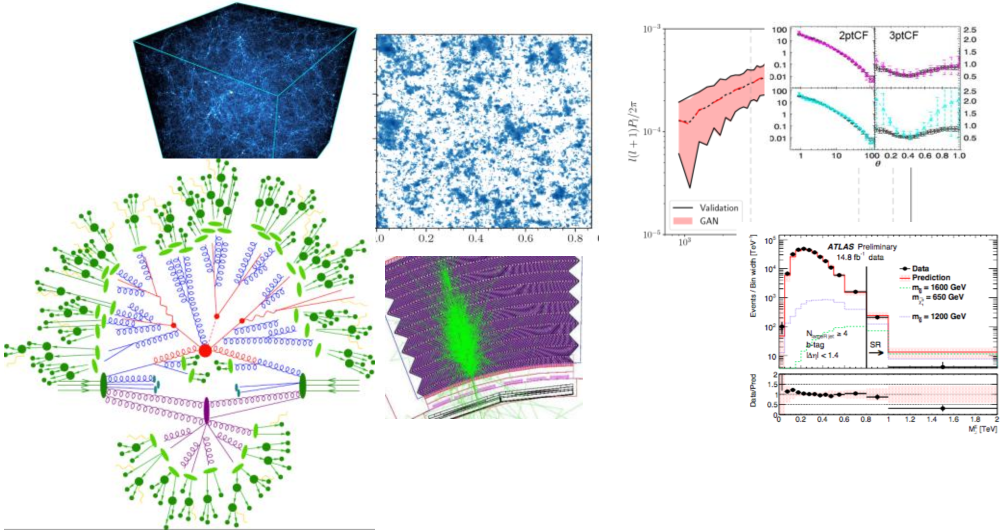
FAIR Universe
HiggsML Uncertainty Challenge

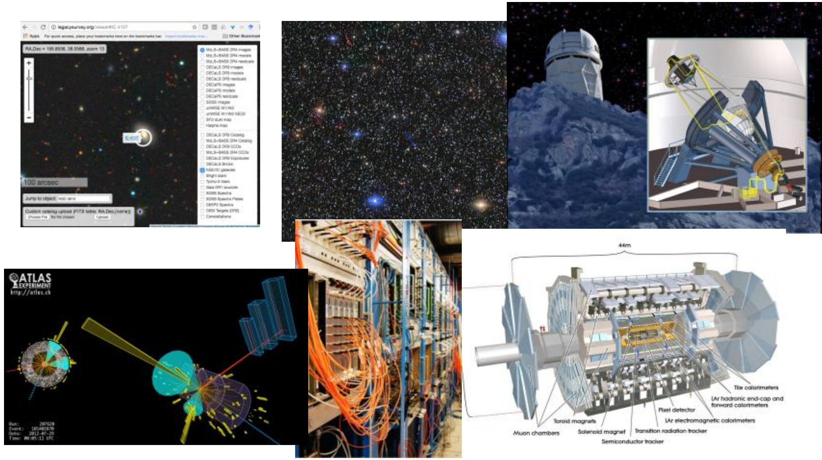
*Wahid Bhimji, Paolo Calafiura, **<u>Ragansu Chakkappai</u>**, Po-Wen Chang, Yuan-Tang Chou, Sascha Diefenbacher, Jordan Dudley, Steven Farrell, Aishik Ghosh, Isabelle Guyon, Chris Harris, Shih-Chieh Hsu, Elham E Khoda, Benjamin Nachman, Peter Nugent, David Rousseau, Benjamin Thorne, Ihsan Ullah, Yulei Zhang*



EPS HEP2025 Marseille

Bias and uncertainty in Physics



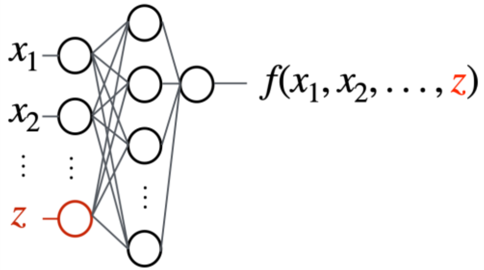


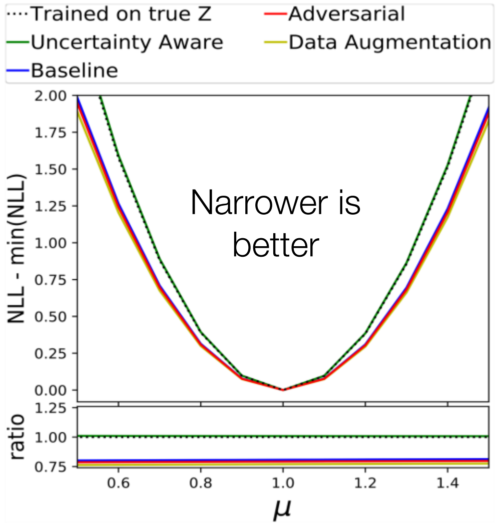
Exp/Obs reconstruction

Differences between simulation and data can bias measurements

Landscape of Uncertainty Aware Learning

- “pivot” Louppe, Kagan, Cranmer : arXiv:1611.01046
- “Uncertainty-aware” approach of Ghosh, Nachman, Whiteson PhysRevD.104.056026
  - Parameterize classifier using Z
  - Measured on “Toy” 2D Gaussian Dataset and dataset from *HiggsML Challenge* modified to include systematic on tau-energy scale
  - Performs as well as classifier trained on true Z
- Other novel approaches e.g. (not comprehensive)
  - Inferno: arxiv:1806.04743
  - Direct profile-likelihood: e.g. arxiv:2203.13079
  - (Neuro) Simulation Based Inference has to include Z: arXiv:1911.01429





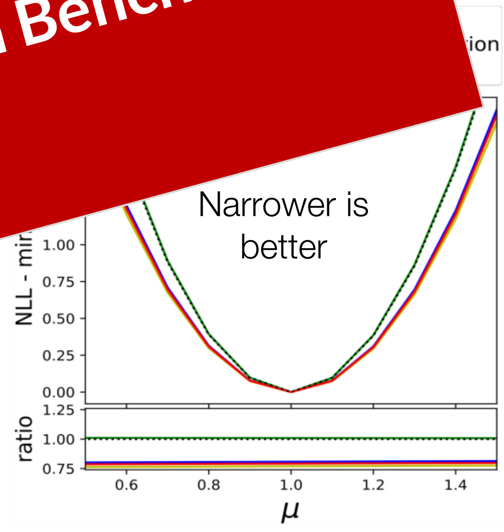
Landscape of Uncertainty Aware Learning

- “pivot” Louppe, Kagan, Cranmer : arXiv:1611.01046
- “Uncertainty-aware” approach of Ghosh, Nachman, Whi
  PhysRevD.104.056026
  - Parameterize classifier using Z
  - Measured on “To ”
    HiggsM

v.2203.13079

sed Inference has to include Z:

1.01429



(Signal Strength)

Many methods developed, but No Common Benchmark

Fair Universe: HiggsML Uncertainty Challenge

FAIR UNIVERSE - HIGGS UNCERTAINTY CHALLENGE

160 PARTICIPANTS

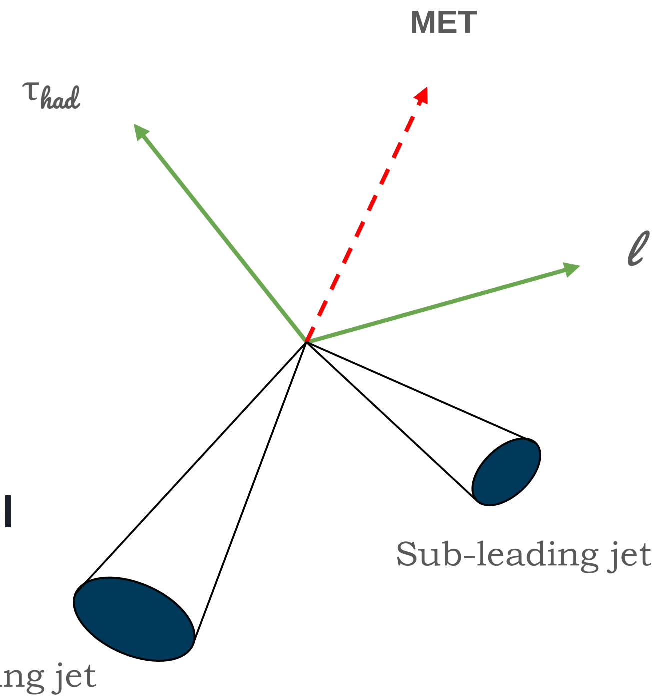
347 SUBMISSIONS

A pool of 4000 USD

- Full HiggsML Uncertainty Challenge Ran from September 12 to March 14th
- Accepted as NeurIPS competition 2024
- Dedicated workshop at NeurIPS (most important ML conference) - 2024 at December 14th
- Final results presented on May at CERN(satellite event of IML - 2025)

Physics Problem

- **Process** :
  - Signal : $H \rightarrow \tau^+ \tau^-$ (1 Lepton 1 Hardon)
  - Backgrounds : $Z \rightarrow \tau^+ \tau^-, VV, t\bar{t}$
- **Luminosity** : $10\ fb^{-1}$
- **Parameter of interest** : Signal Strength $\mu = \frac{N_{obs}}{N_{\exp(SM)}}$
- **Objective** : Measure $\boldsymbol{\mu}$ and its **1 $\boldsymbol{\sigma}$ Confidence interval** in the presence of **Systematics Uncertainty**.



Fair Universe: HiggsML Uncertainty Challenge

	2014 Higgs challenge	2024 FAIR Universe
Data Source	ATLAS Open Data (Sample of real data)	New simulated dataset
Data Quantity	800,000	400,000,000
Parameterized systematics	✗	6 Systematics
Inference	Classification accuracy	Signal strength (μ) CI (Confidence Interval) with pseudo-experiment
Metric	Z - Classification Significance	New metric aimed at CI Coverage

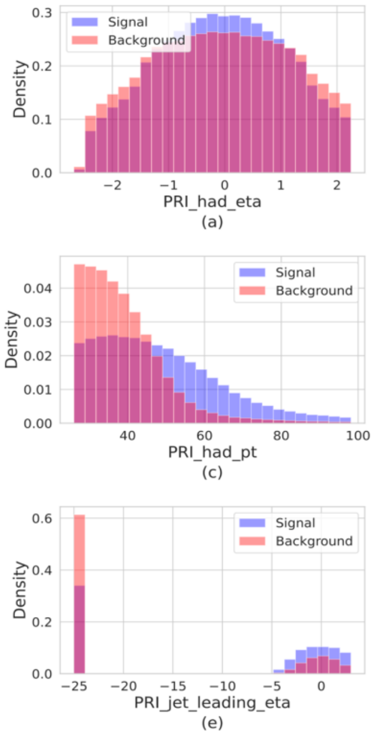
Challenge Dataset

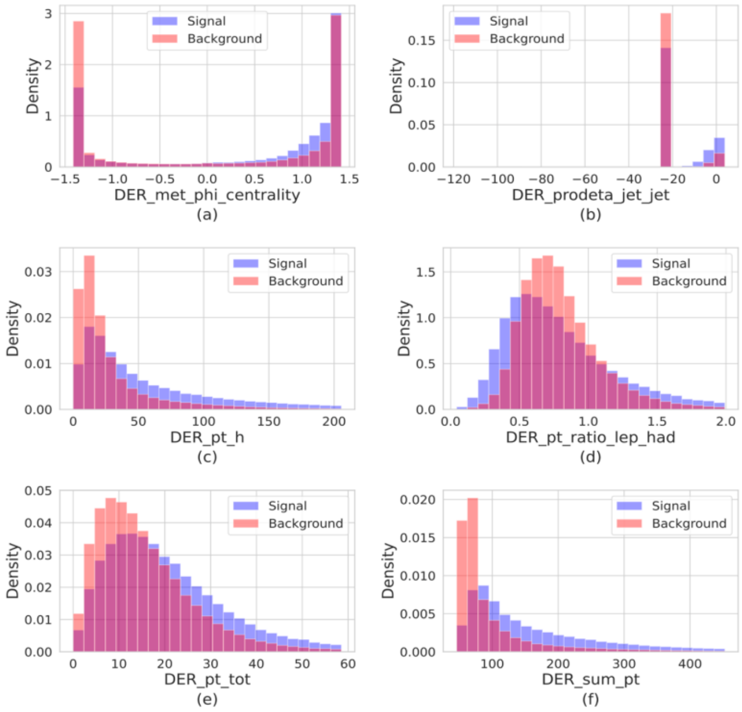
- Generated data with fast simulation with Pythia LO and Delphes detector simulation
- Using the updated Delphes ATLAS card
- Generated ~**280 Million** Events after initial cuts equivalent to **220** X **10**fb-1 + Higgs
- Data organised into **tabular** form with **28** features per event.

Process	Number Generated	LHC Events @10fb^{-1}	Label
Higgs	52101127	1015	**signal**
Z Boson	221724480	1002395	**background**
Di-Boson	2105415	3783	**background**
$t\bar{t}$	12073068	44190	**background**

Dataset permanently released on Zenodo in May 2025 for Future Benchmarks
https://zenodo.org/records/15131565

Dataset : some feature examples



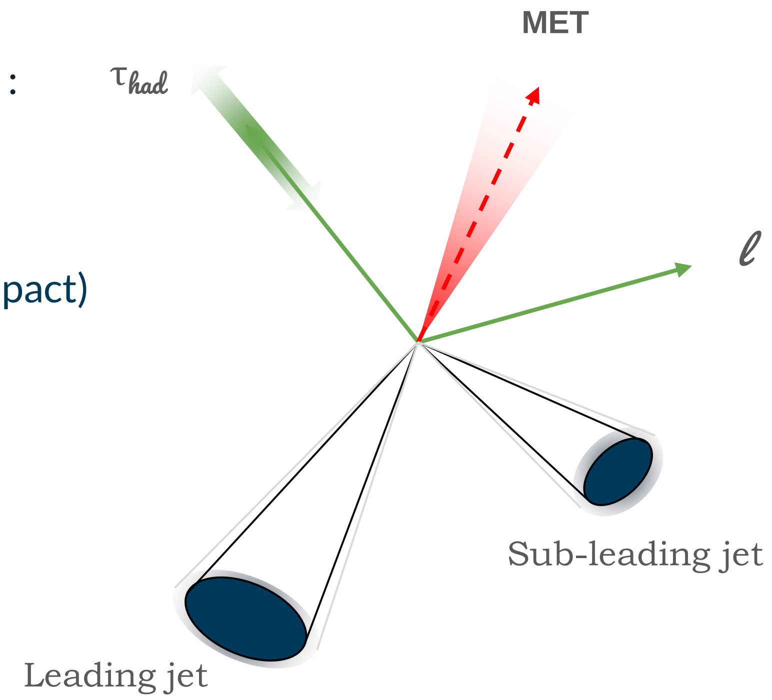


Challenge Datasets - Systematics

Apply parameterized systematics (Nuisance Parameters) :

- Feature distortions:
  - Tau Energy Scale (**TES**) (and correlated MET)
  - Jet Energy Scale (**JES**) (and correlated MET impact)
  - Additional randomised Soft MET

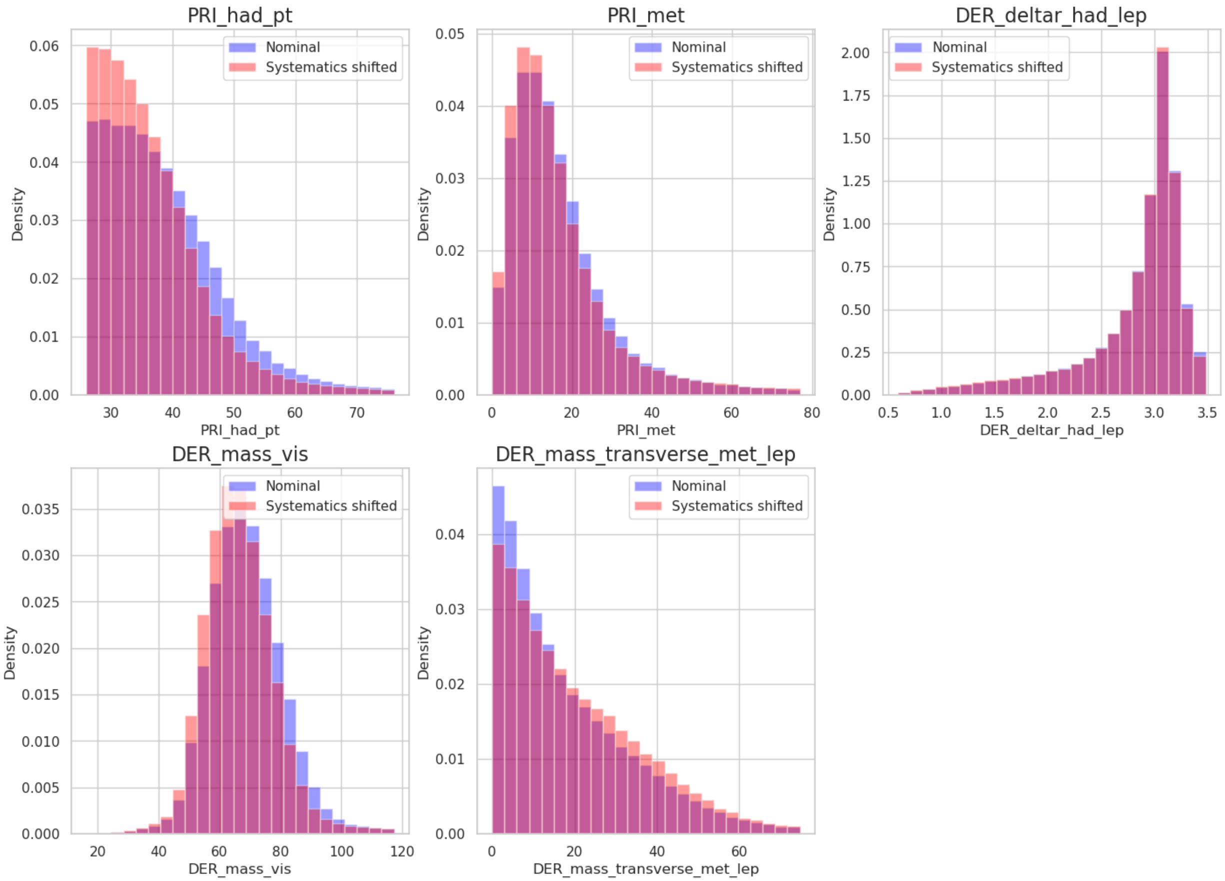
- Event category normalisation
  - Background overall normalisation
  - Di-boson background normalisation
  - ttbar background normalisation



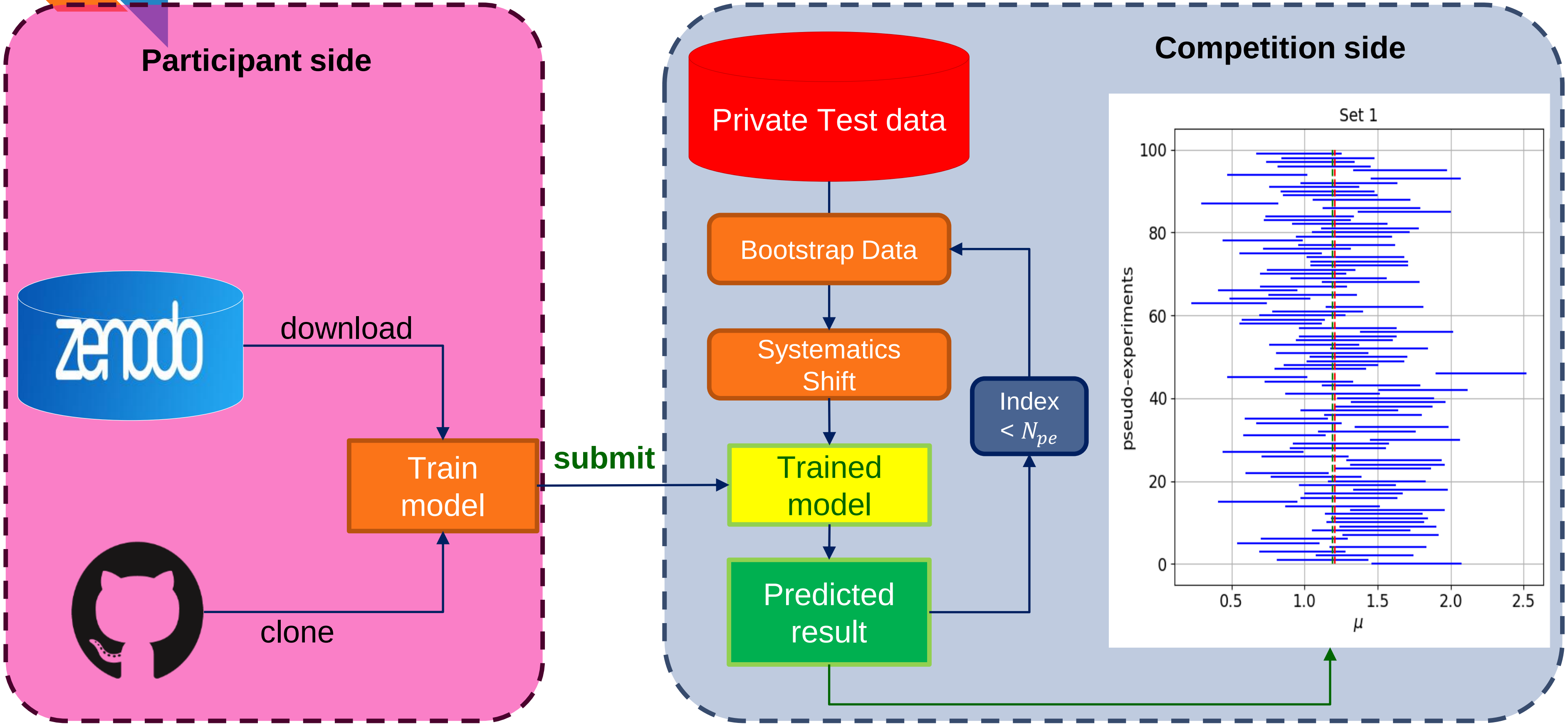
Tau Energy Scale Systematics Applied

Histogram between nominal (TES = 1) and shifted (TES = 0.9)

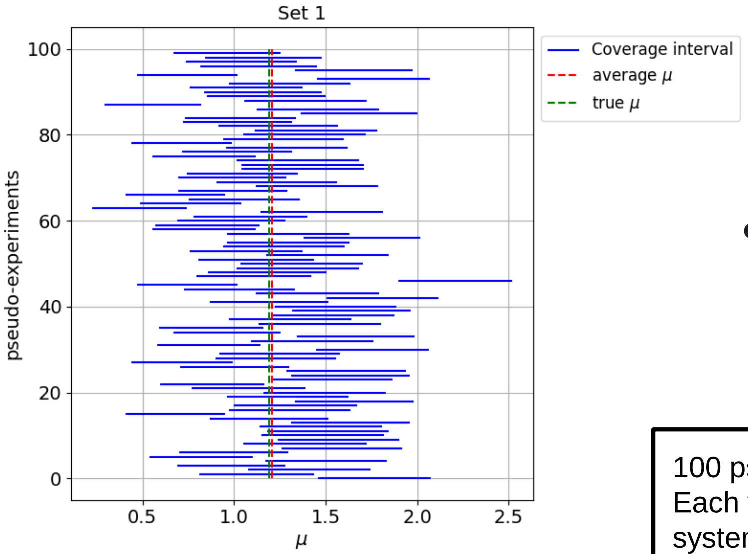
TES = 0.9, is an exaggeration, in practice it is sampled with a gaussian of 1 +/- 0.01



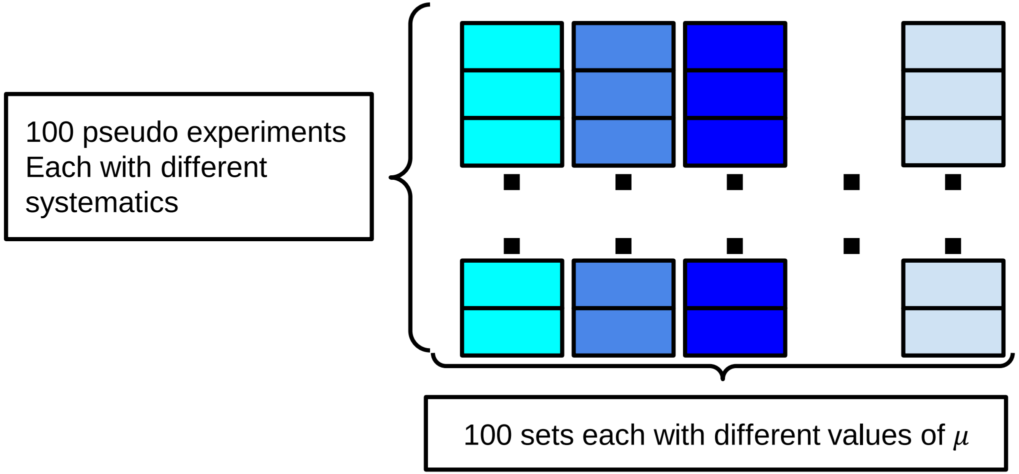
Competition Work Flow



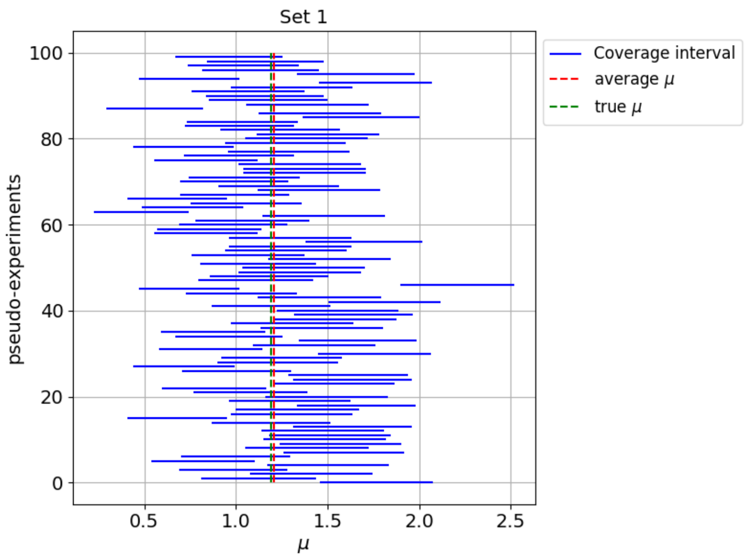
Confidence Interval Evaluation



- Form multiple pseudo-experiment test sets: different signal strengths ($\boldsymbol{\mu}$) and systematics
    - **10$\boldsymbol{\mu}$** times **100** pseudo-experiments (Online Phase)
    - **1000$\boldsymbol{\mu}$** times **100** pseudo-experiments (Offline Phase)
- Task: predict uncertainty interval $[\boldsymbol{\mu}_{16}, \boldsymbol{\mu}_{84}]$
    - E.g. 68% quantile of likelihood or assume 1σ



Confidence Interval Evaluation



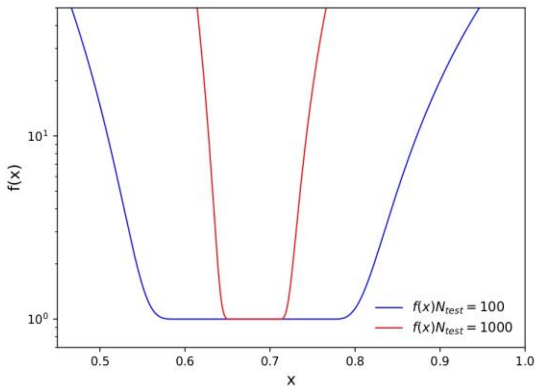
- Form multiple pseudo-experiment test sets: different signal strengths ($\boldsymbol{\mu}$) and systematics
  - **10$\boldsymbol{\mu}$** times **100** pseudo-experiments (Online Phase)
  - **1000$\boldsymbol{\mu}$** times **100** pseudo-experiments (Offline Phase)
- Task: predict uncertainty interval $[\boldsymbol{\mu}_{16}, \boldsymbol{\mu}_{84}]$
  - E.g. 68% quantile of likelihood or assume 1σ

Scoring formula

$$s = -\ln\left((w + \epsilon) f(c)\right)$$

Mean width

Coverage penalty



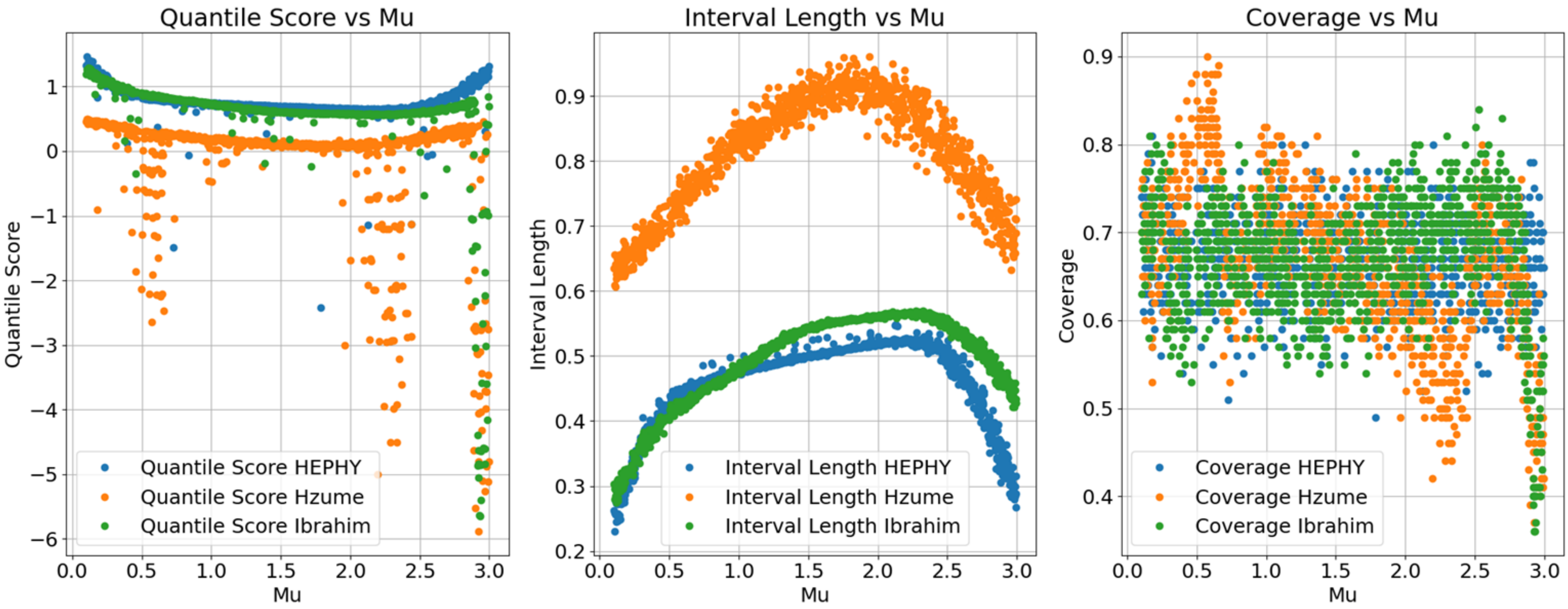
Results

Winners

Medal	Rank	Team	Avg Coverage	Avg Interval	Avg Quantile Score
	1 (Tie)	HEPHY	0.6683	0.4599	-0.5823
	1 (Tie)	IBRAHIME	0.6698	0.4974	-0.5761
	3	HZUME	0.6659	0.8134	-2.1650

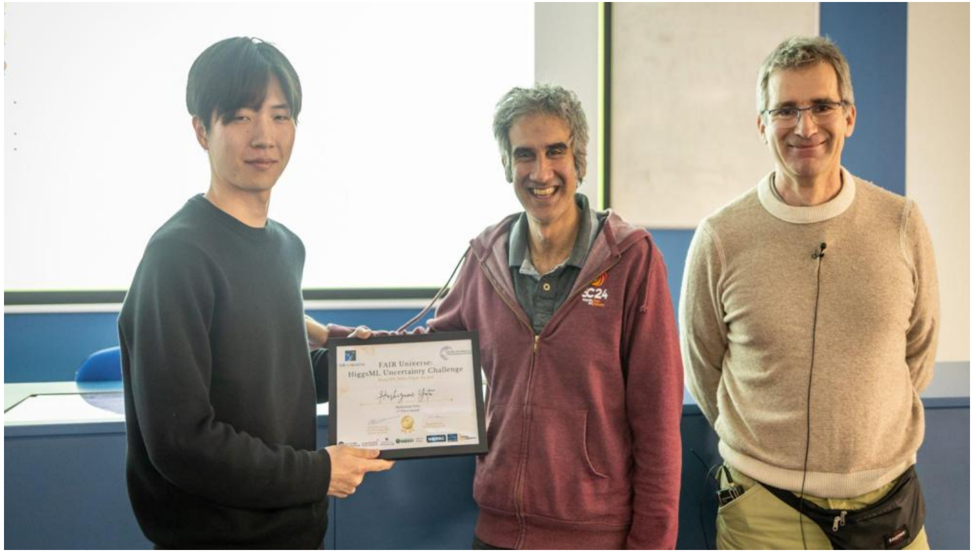
- HEPHY (Lisa Benato, Cristina Giordano, Claudius Krause, Ang Li, Robert Schöfbeck, Dennis Schwarz, Maryam Shooshtari, Daohan Wang) from Vienna's Institute of High Energy Physics (HEPHY) in Austria will win $2000.
- IBRAHIME (Ibrahim Elsharkawy) from University of Illinois at Urbana-Champaign will win $2000.
- HZUME (Hashizume Yota) from Kyoto University Japan will win $500.

Final evaluation on 1000 x 100 p-e



3rd - Hashizume Yota - Kyoto University Japan

Decision-Tree Aggregated Features and Hybrid Bin-Classifier/Quantile-Regressor



IML 2025 link :
https://indico.cern.ch/event/1523250/contributions/6456344/attachments/3070079/5431138/higgsml-3rd-place.pdf

1st (Tie) - Vienna's Institute of High Energy Physics (HEPHY)

UNBINNED MEASUREMENTS WITH REFINABLE SYSTEMATICS

Lisa Benato, Cristina Giordano, Claudius Krause, Ang Li, Robert Schöfbeck, Dennis Schwarz, Maryam Shooshtari, Daohan Wang

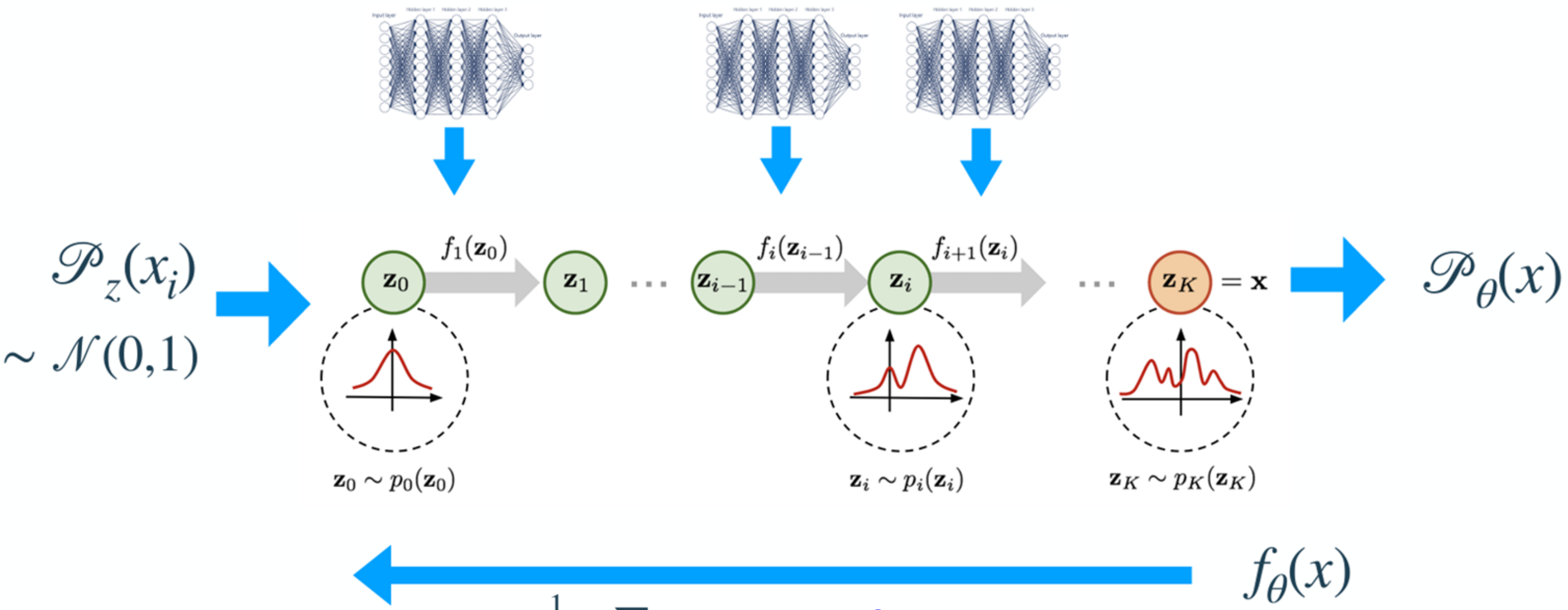
See more **in Robert Schöfbeck** (next) talk

1st (Tie) - Ibrahim Elsharkawy - University of Illinois at Urbana-Champaign

CONTRASTIVE NORMALIZING FLOWS FOR UNCERTAINTY-AWARE PARAMETER ESTIMATION

https://arxiv.org/abs/2505.08709

Contrastive Normalizing Flows



$$\mathscr{L} = \frac{1}{|\mathscr{D}|}\sum_{x,y\in\mathscr{D}} \{-\log p_\theta(x) + c \log p_\theta(y)\}$$

Hyper parameter
For tuning signal like
vs Background like

For Example: x = signal events, y = background events

DNN Training with CNF

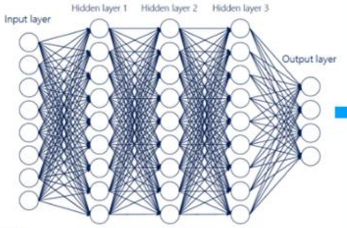
Train the same DNN with event data perturbed with a variety of nuisance parameters.

$$r(x_i, \nu_1, \nu_2) \approx \frac{P(x_i \,|\, Sig, \phi_s[x_i, \nu_1])}{P(x_i \,|\, Bg, \phi_{bg}[x_i, \nu_2])}$$

$\{NP_1\}$ → Event
$\{NP_2\}$ → Event
$\{NP_3\}$ → Event
$\{NP_4\}$ → Event

Sig CNF c=2
Sig CNF c=0.5
Bg CNF c=2
Bg CNF c=0.5

Event



$P_{sig}(x)$

Score Histogram and MLE

Inverse Neyman Construction

and 1-sigma Error Bars

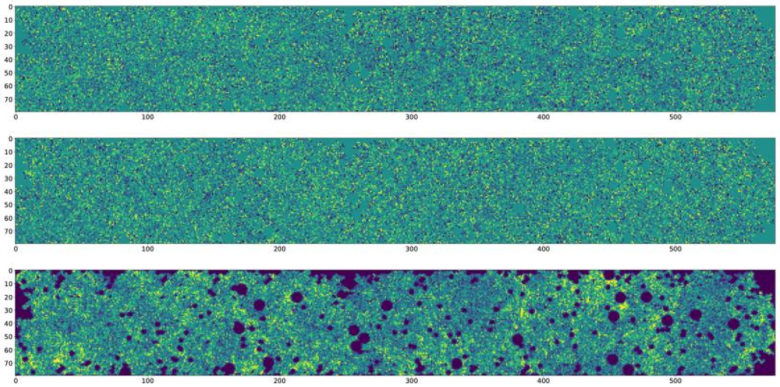
Loss: Binary Cross Entropy

$$\mathscr{L} = -\frac{1}{N}\sum_{i=1}^{N}\left[y_i \cdot \log(\hat{y}_i) + (1-y_i)\cdot \log(1-\hat{y}_i)\right]$$

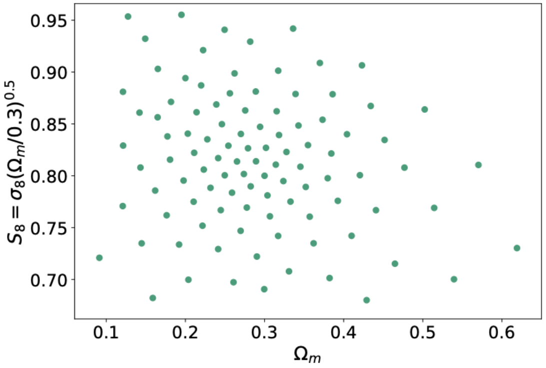
Learn more at: https://arxiv.org/abs/2505.08709

Fair Universe - Weak lensing Challenge

- NeurIPS 2025 Competition
- Goal to determine cosmological parameters (Ω_m and S_8), and uncertainties, from large simulated weak gravitational lensing dataset
- Various realistic systematic effects including baryonic effect, intrinsic alignment, photometric redshift uncertainty, shear measurement bias, point spread function, and source clustering effect
- Second phase test data with different (OoD) physical models



Shear maps γ_1 and γ_2 and weight map

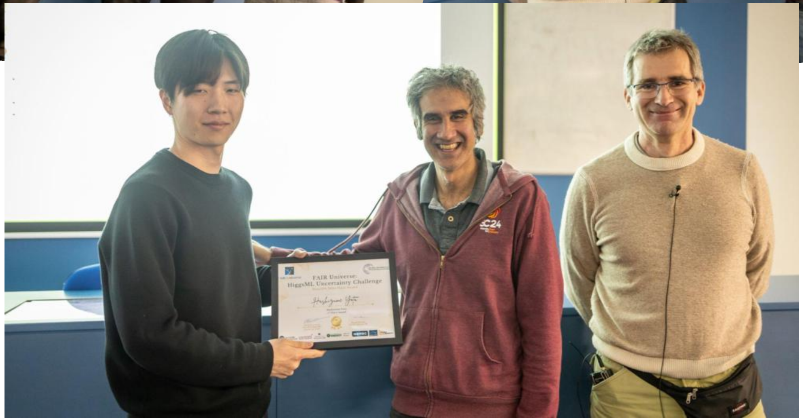


101 different cosmological models

Conclusion

- AI challenge which addresses Systematic Uncertainty in HEP problem.
- New Scoring to take **Coverage** and **Confidence interval** into account.
- Large Computing Infrastructure as backend
- Exciting submissions which could be applied to future HEP analysis
- Large **Public** Data Set with ~280M Events (signal + background) in **Zenodo**
- New **long term permanent** benchmark will be launched Soon.

Ongoing information Google Group: Fair-Universe-Announcements
Collaborations, questions, comments: fair-universe@lbl.gov

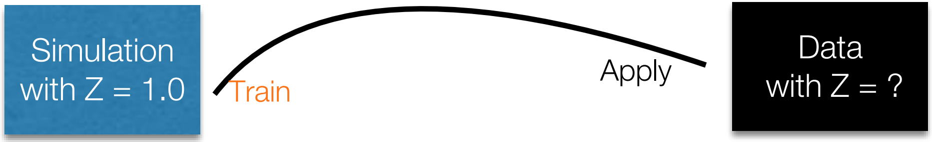


Thank you for your attention!

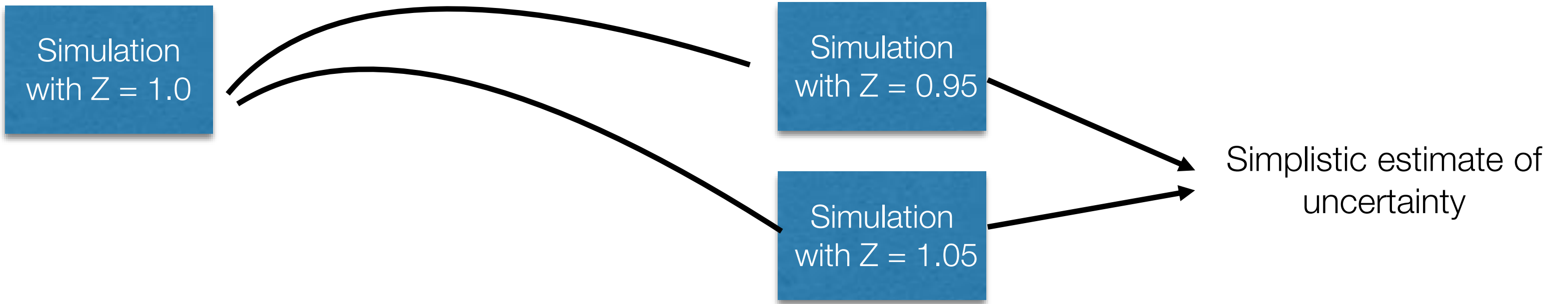
Back-up

Bias and uncertainty in ML in HEP

- ML methods in HEP are often trained based on simulation which has estimated systematic uncertainties ("Z")
- These are then applied in data with the different detector state Z=?



- Common baseline approach: Train classifier on nominal data (e.g. Z=1) and estimate uncertainties with alternate simulations. Shift Z and look at impact or perform full profile likelihood

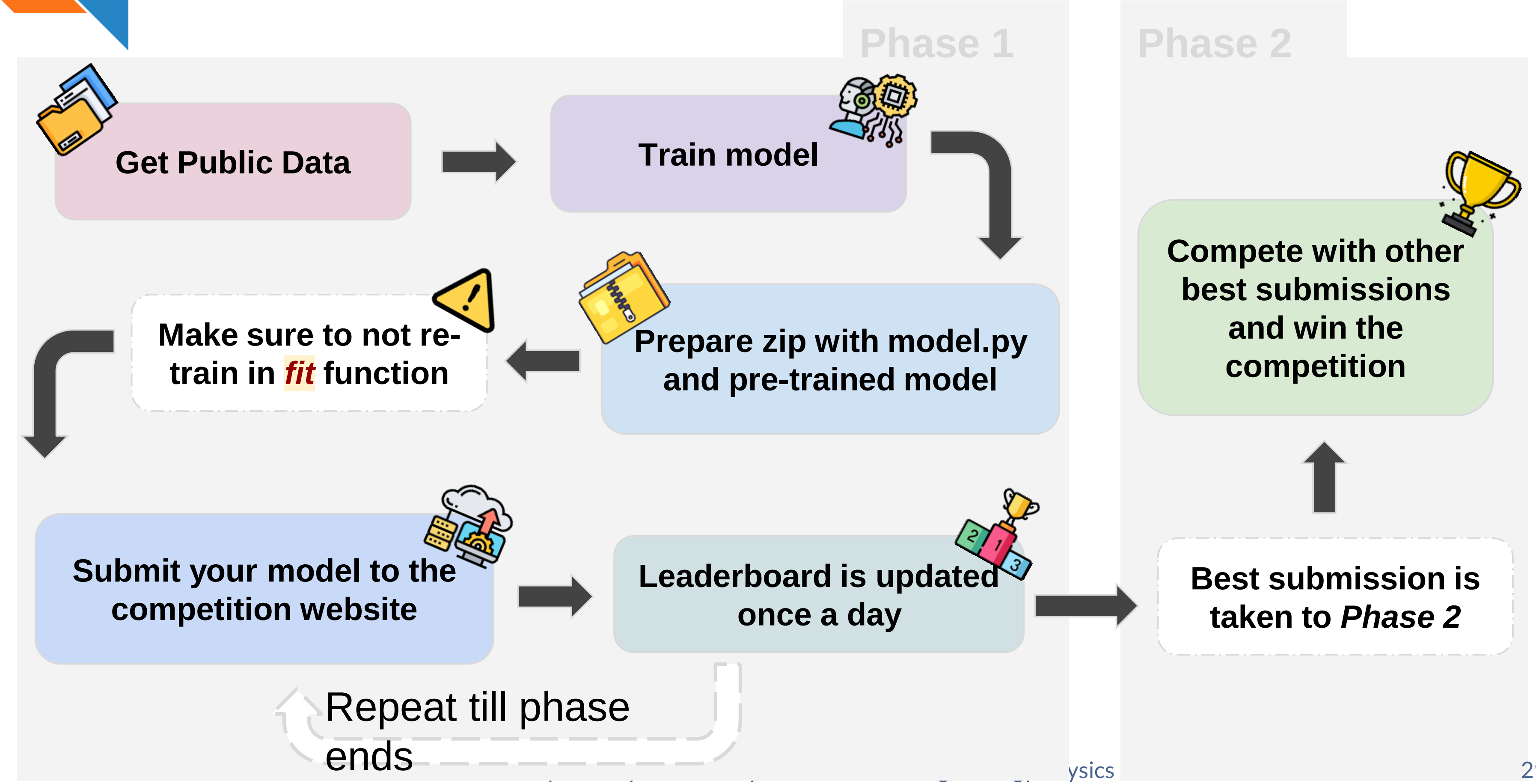


Background on Fair Universe Project

FAIR Universe

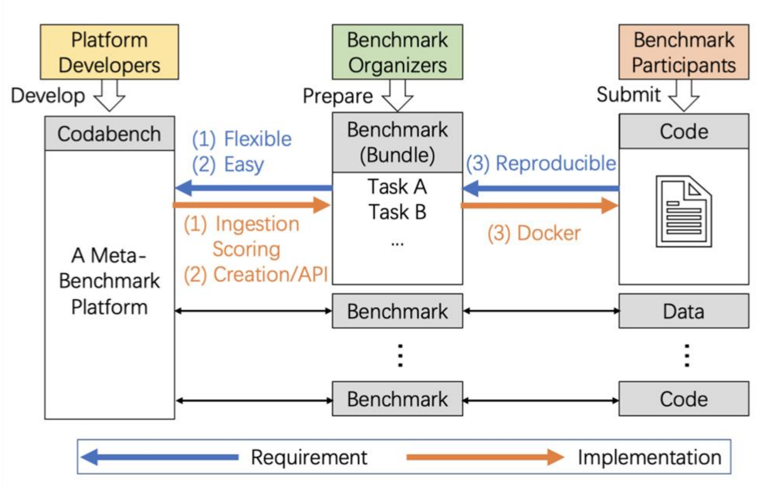
- 3 year US Dept. of Energy, AI for HEP project. Aims to:
  - Provide an open, **large-compute-scale AI ecosystem** for sharing datasets, training large models, fine-tuning those models, and **hosting challenges and benchmarks.**
  - **Organize a challenge series**, progressively rolling in tasks of increasing difficulty, based on novel datasets.
  - Tasks will focus on **measuring and minimizing the effects of systematic uncertainties** in HEP (particle physics and cosmology).
- This funding went to LBL, NERSC, U Washington, and Chalearn

Competition Flow



Large-compute-scale AI ecosystem … hosting challenges and benchmarks.

Codabench Platform

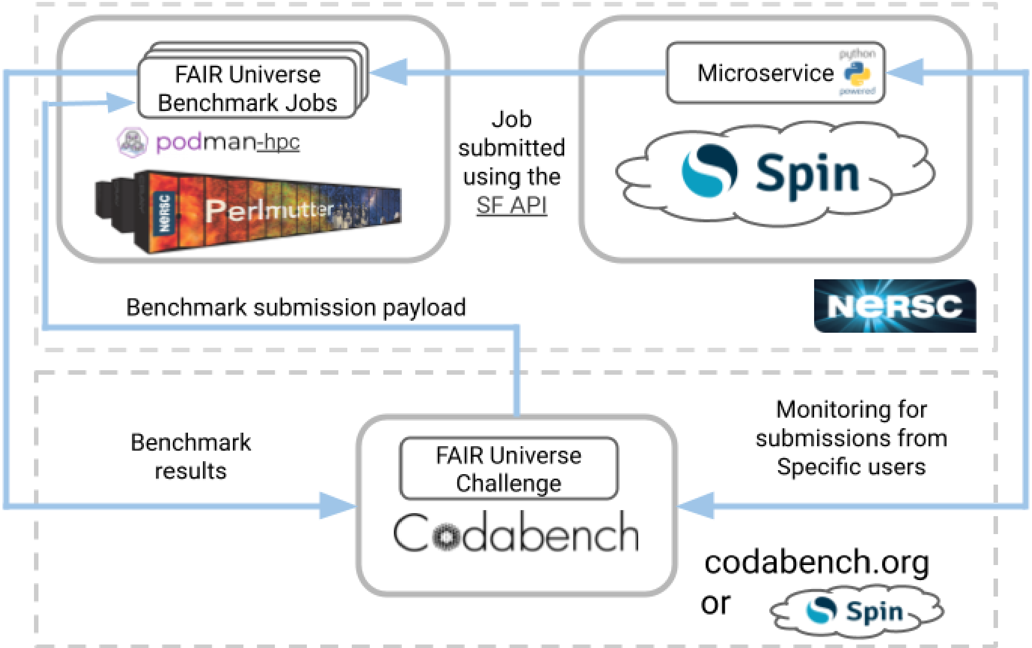


Codabench - open source platform for AI benchmarks and challenges

- Originally (CodaLab) Microsoft/Stanford now a Paris-Saclay/LISN led community
- > 600 challenges since 2013
- Completely open-ended competition design.
- Allows code submission as well as results e.g. for evaluation timing or reproducibility
- Also data-centric AI “inverted competitions”
- Queues for evaluation can run on diverse compute resources
- Platform itself can be deployed on different compute resources
- Ranked best challenge platform for ML by ML contests

Codabench

Fair Universe Platform: Codabench-NERSC integration



System Specifications

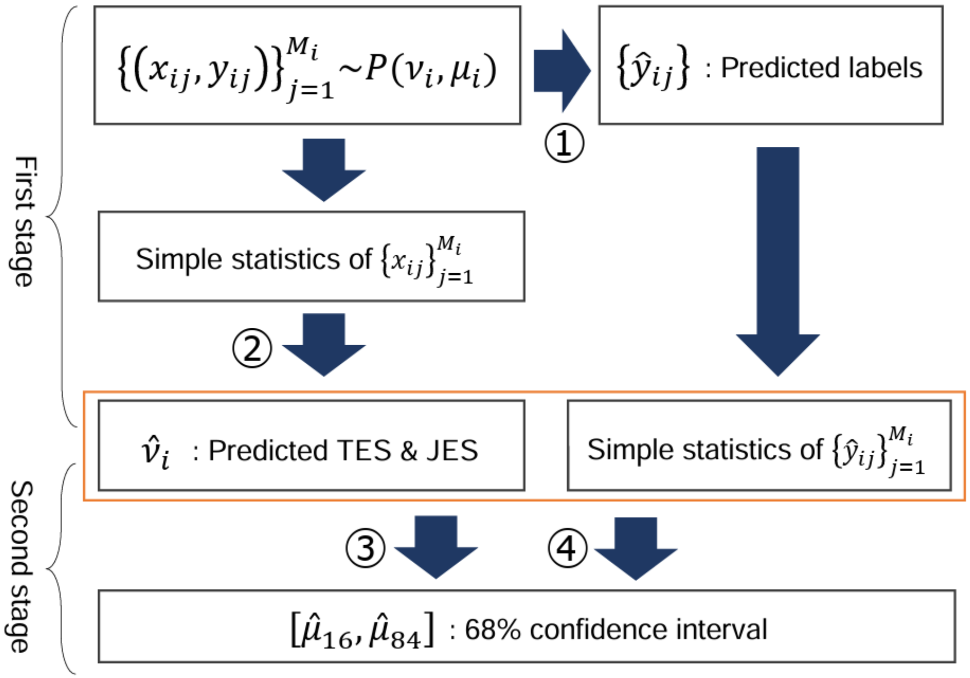
Partition	# of nodes	CPU	GPU
GPU	1536	1x AMD EPYC 7763	4x NVIDIA A100 (40GB)
	256	1x AMD EPYC 7763	4x NVIDIA A100 (80GB)

3rd Place Solution
Decision-Tree Aggregated Features
and Hybrid Bin-Classifier/Quantile-Regressor

Yota Hashizume (hzume)

Overview

- 2-stage, GPU-free GBDT-based model
- First stage
    - Aggregated features
    - Used 2 models (①,②)
- Second stage
    - Estimate 68% confidence interval by using aggregated features
    - Used 2 models (③,④) and merged their outputs



https://indico.cern.ch/event/1523250/contributions/6456344/attachments/3070079/5431138/higgsml-3rd-place.pdf