FAIR Universe

BERKELEY LAB

ATLAS EXPERIMENT

Overview of HiggsML Uncertainty Challenge

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

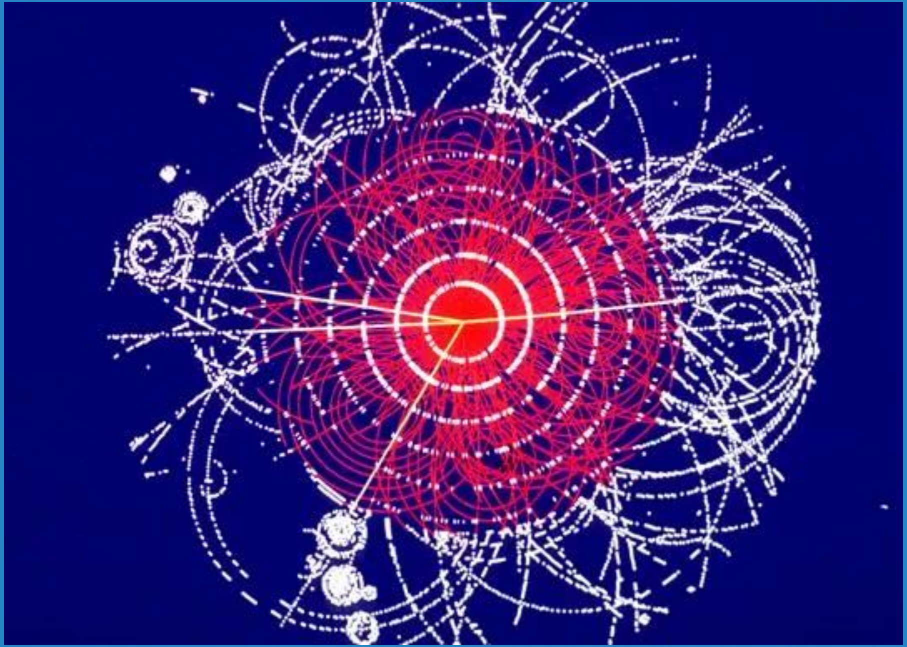
université PARIS-SACLAY

IJCLab Irène Joliot-Curie

Contents

- Introduction
- Data Generation and Processing
- Uncertainty and Evaluation
- Conclusion

Introduction



The Higgs Boson is one of the most important discoveries of the 21th century

Validates and completes the standard model

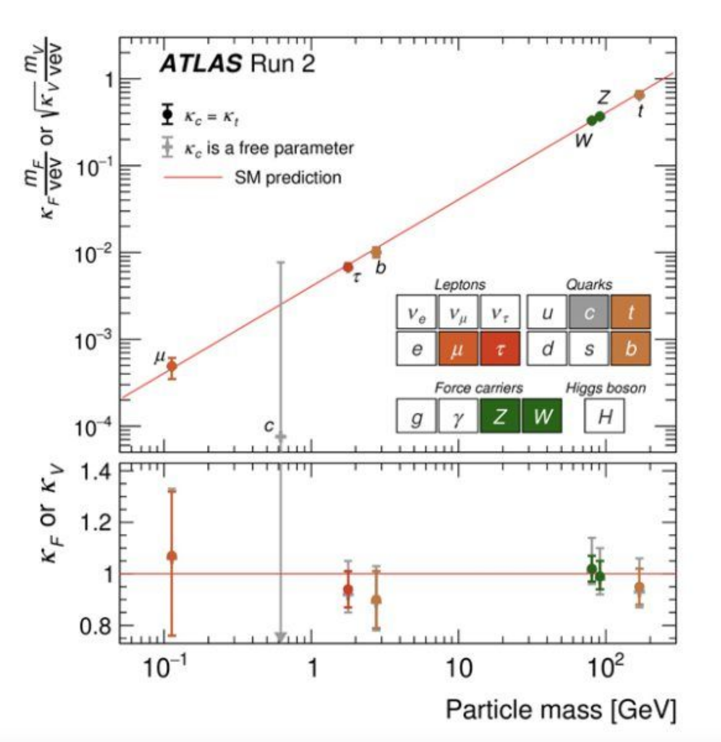
The Higgs boson gives mass to fermions by the Yukawa coupling and to Weak Bosons by the spontaneous breaking of symmetry.

Introduction

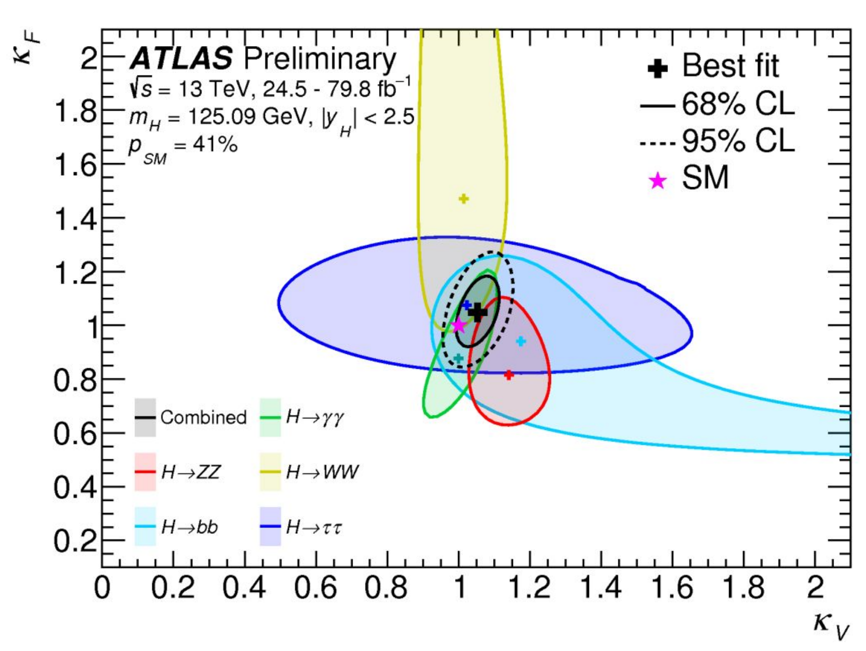
But the story is not complete :

Out of the 9 massive fermions the Higgs coupling is only precisely measured in 3 (top bottom and tau) fermions.

The rest of the fermions have very low cross section hence making discovery difficult.

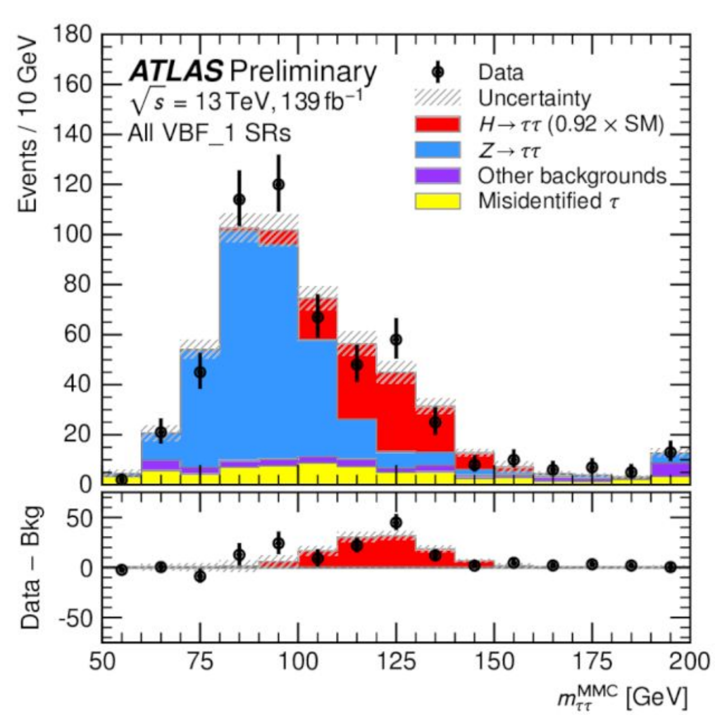


Introduction



The Standard Model is not the end of the story there are many unanswered question in HEP. Precise measurement of higgs cross section can open doors to Beyond Standard Model physics

Machine learning Approaches



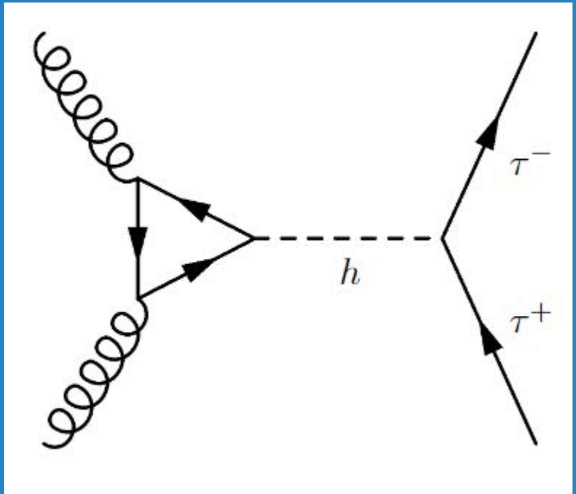
One of the way to increase signal significance is to construct a signal rich region. This was demonstrated by the 2014 HIggsML challenge.

But ML methods are sensitive to mis-modeling in the presence of systematic uncertainty, Hence bringing a need for Uncertainty Aware Artificial Intelligence.

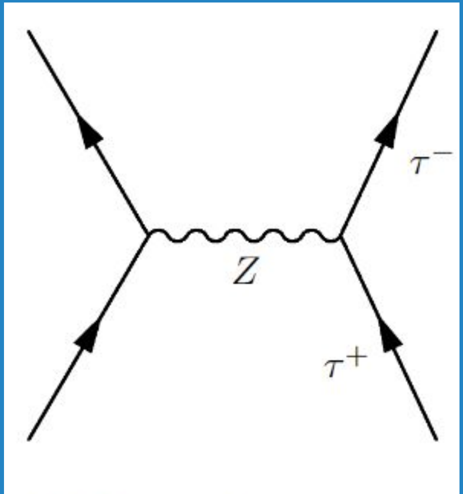
Higgs production and Background

The Objective of the challenge is to determine the signal strength μ with high confidence.

The problem at hand is the Higgs boson cross section in the Higgs to tau tau channel. Specifically in the 1 hadron 1 lepton case



(a) ggF Production



(a) $Z \to ll$ Background

Data generation

The data is generated with combination of Pythia for the event generation and Delphes for the fast detector simulation.

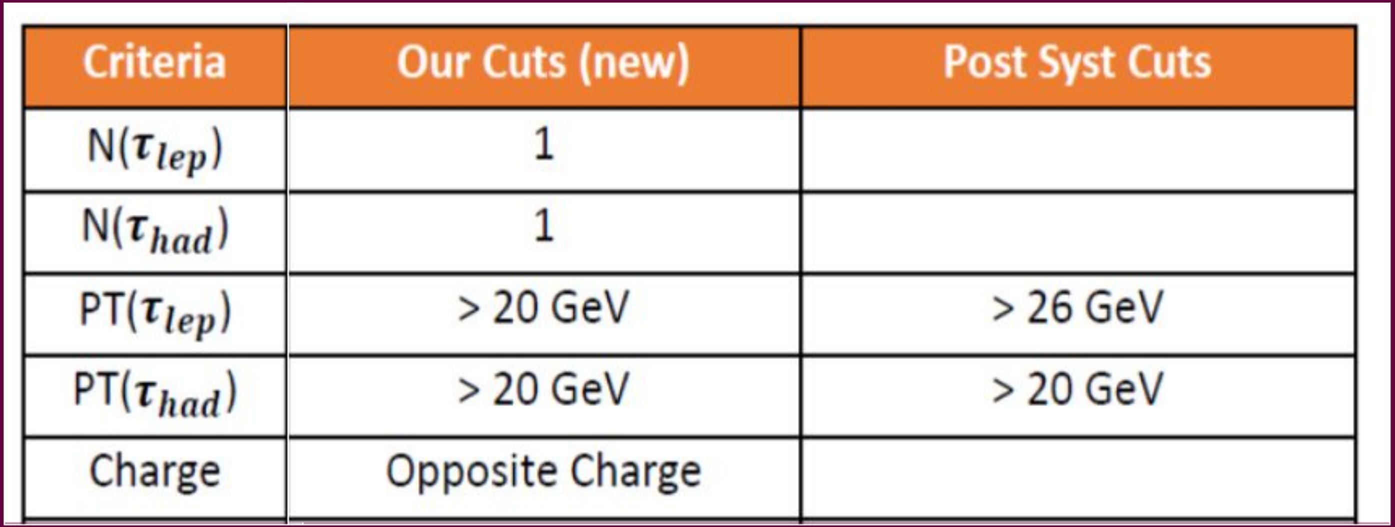
Though they are not at as accurate as Madgrap or Sherpa with GEANT4 they are sufficient for the purpose at hand. By making this sacrifice we gain tremendous advantage in speed as we can now generate around 1000 events in a minute in single cpu. Ultimately allowing for larger datasets in the competition or even generation within the competition.

DELPHES
fast simulation

Data generation - Preprocessing

The data set undergo a preselections and get split into the **train set** and the test set.

The **test set** undergo more selections (Post syst cuts) after the systematics is applied.



Criteria	Our Cuts (new)	Post Syst Cuts
N(τ_{lep})	1	
N(τ_{had})	1	
PT(τ_{lep})	> 20 GeV	> 26 GeV
PT(τ_{had})	> 20 GeV	> 20 GeV
Charge	Opposite Charge	

Cross-section and Weighs

In simulated data more often than not the number of events (data points) we have of signal and background are in amounts which are different from what is experimentally observed. This happens sometimes due to the properties of the generator but many a time this is intentionally done to make the number of signal and background data points equal to help the learning process in ML. To compensate for this effect we add weights to this data set.

$$N_{=} \sum_{i \in \mathcal{G}} w_i$$

The total number of actual signal (γ) and background (β) events will then be given by

$$\gamma = \sum_{i \in \mathcal{S}} w_i \text{ and } \beta = \sum_{i \in \mathcal{B}} w_i$$

Systematic Uncertainty

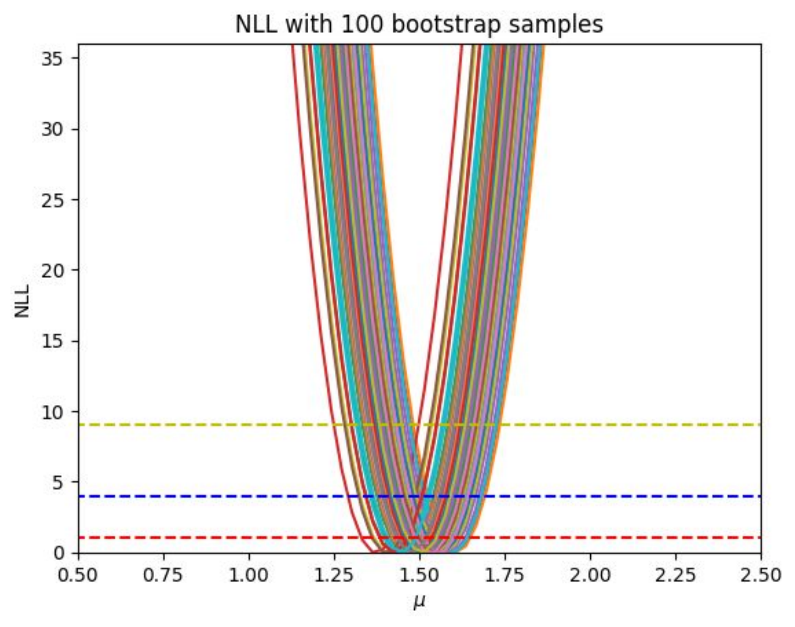
“Currently” only systematic uncertainty for this challenge is the **Tau-hadron Energy Scale (TES)**

This is the systematic uncertainty in the energy measurement of the to tau hadrons.

This is implemented by recomputing the 4-Vector of the tau-hadron and then scaling the energy of the particle. This effects the 4-vector of the missing energy thus missing energy terms are recomputed.

Since Primary variables are shifted it is necessary to recompute the Derived variables.

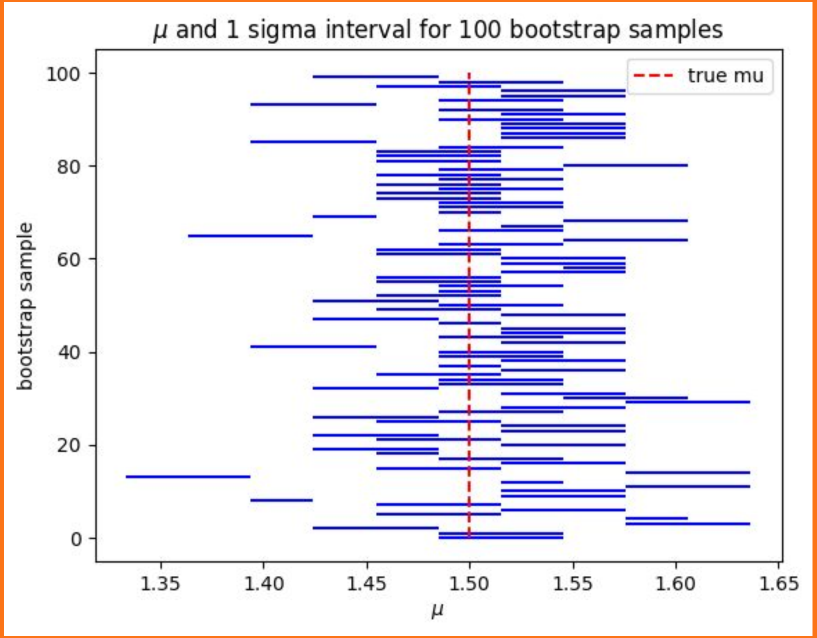
Bootstrapping and Pseudo-Experiment



One way to emulate counting experiment from prior probability is by bootstrapping. bootstrapped data set is created by reweighting the data sets with the Poisson of the weights.

$$\mathcal{W}_i = poisson(w_i) \forall i \in \mathcal{G}$$

Signal Strength and Coverage

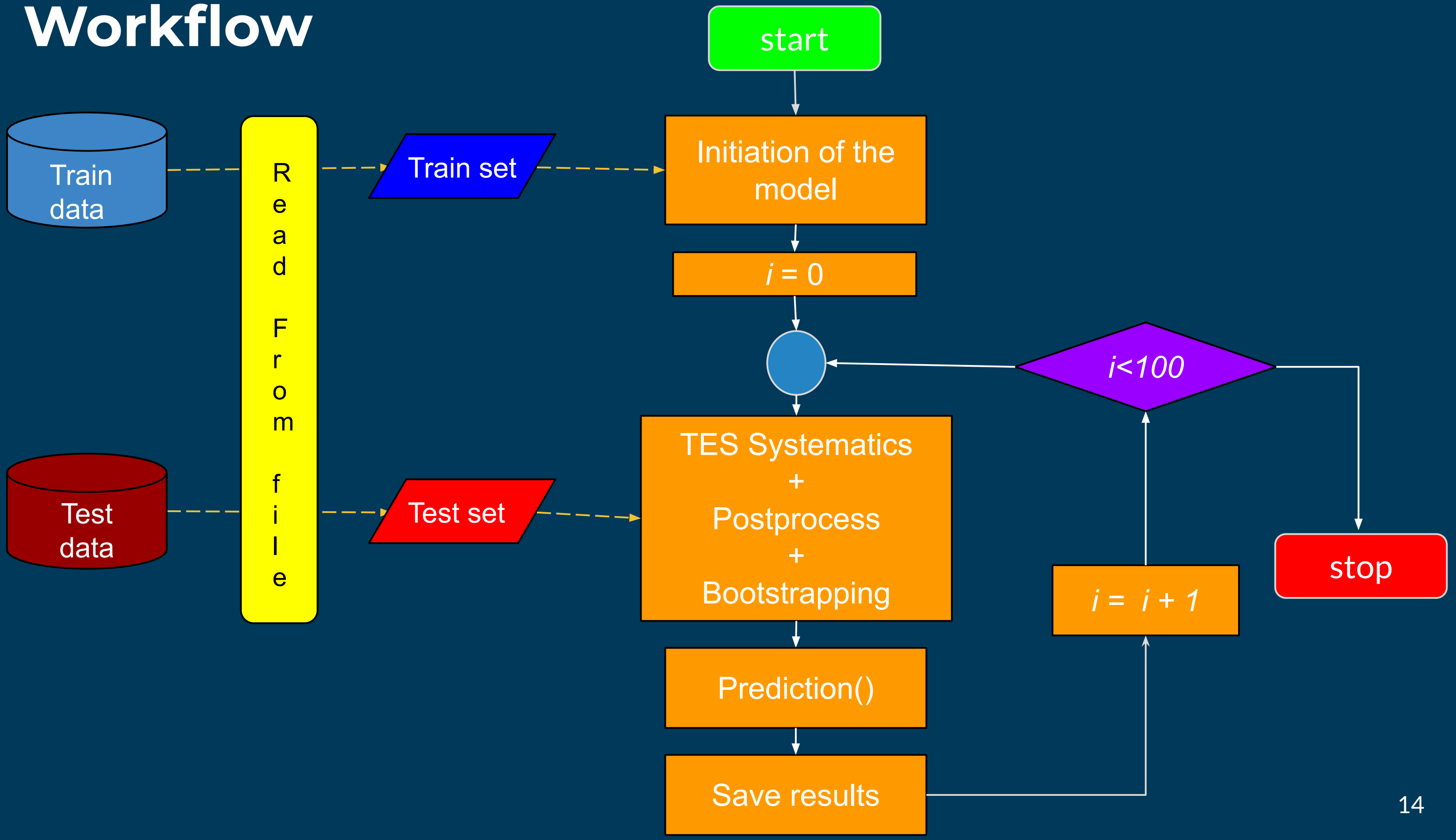


Since the aim of the competition is to design model which predicts μ and its 1σ CL (uncertainty)

They way we test this is by counting the number of times the true μ falls within the proposed interval or it’s coverage.

The Scoring will be hence based on coverage and interval more on this from Sascha.

Workflow


start
Train data
Test data
Read From file
Train set
Test set
Initiation of the model
i = 0
TES Systematics + Postprocess + Bootstrapping
Prediction()
Save results
i = *i* + *1*
i<*100*
stop

Remarks and Conclusion

The competition is designed to test if its possible to develop a ML algorithm which can predict μ with good CL in the presence of systematics

This is done with the help of 100 bootstrap datasets each with different Tau-hadron Energy scale systematics.

Thank you for your attention!