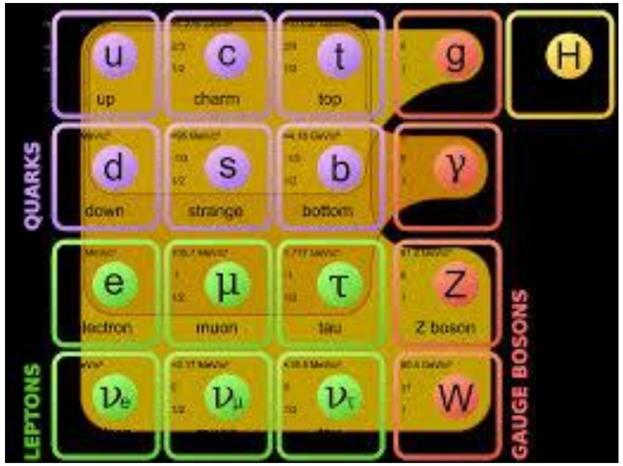
Black Swans
Higgs particle detection

Enseignement d'Intégration - Collaboration B

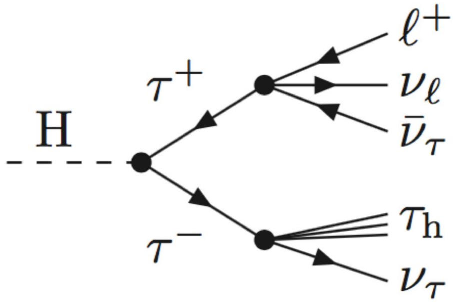
0 Context : Higgs mechanism

- Higgs boson discovery : 2012
  - Explains mass of W and Z bosons



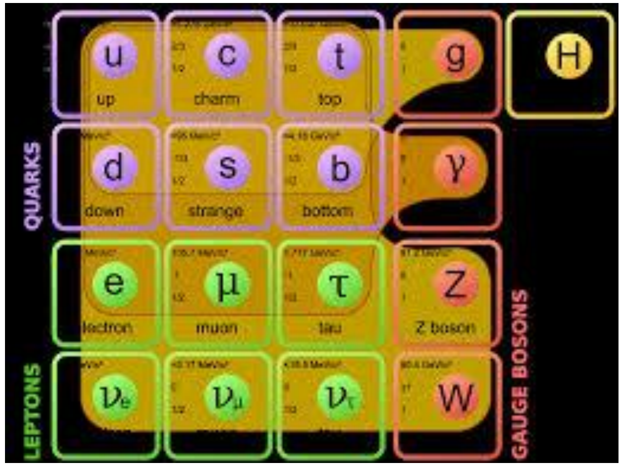
- Can Higgs also explain fermion particles' mass ?
  - Other mechanisms could be at work…

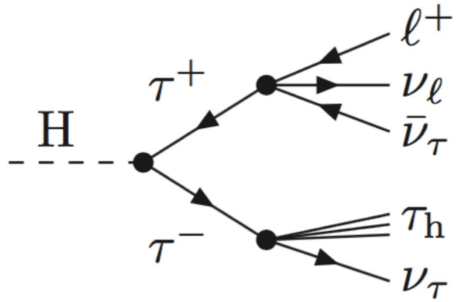
- How to prove Higgs coupling to fermions ?
  - Search for H $\rightarrow \tau^+ + \tau^-$ decay to validate Higgs mechanism
    - Tough because



0 Context : search for H $\rightarrow \tau^+ + \tau^-$

- Why is the search for H $\rightarrow \tau^+ + \tau^-$ decay tough ?
  - τ leptons decay quickly into neutrinos (invisible) + hadrons or leptons
  - τ cannot be spotted directly – just a cloud of noise around it
  - Z $\rightarrow \tau^+ + \tau^-$ mimics the signal : background
- Objective
  - Separate signal from background with given dataset
  - Determine which decay comes from Higgs Boson





Outline

1. Feature Analysis
2. Boosted Decision Tree
3. Neural Network
4. Statistics
5. Systematics

Outline

1. **Feature Analysis**
2. Boosted Decision Tree
3. Neural Network
4. Statistics
5. Systematics

Pipeline objectives

FA : Select appropriate features

- ➢ Avoid correlation
- ➢ Reduce systematic bias
- ➢ Avoid error amplification

Outline

1. Feature Analysis
2. **Boosted Decision Tree**
3. **Neural Network**
4. Statistics
5. Systematics

Pipeline objectives

FA : Select appropriate features

- ➢ Avoid correlation
- ➢ Reduce systematic bias
- ➢ Avoid error amplification

BDT / NN : implement trained models

- ➢ choose evaluation criteria
- ➢ optimise hyper-parameters

Outline

1. Feature Analysis
2. Boosted Decision Tree
3. Neural Network
4. **Statistics**
5. Systematics

Pipeline objectives

FA : Select appropriate features

- ➢ Avoid correlation
- ➢ Reduce systematic bias
- ➢ Avoid error amplification

BDT / NN : implement trained models

- ➢ choose evaluation criteria
- ➢ optimise hyper-parameters

STAT : estimate $\mu S + B$

- ➢ Assess signal sensitivity μ
- ➢ introduce ΔS and ΔB

Outline

1. Feature Analysis
2. Boosted Decision Tree
3. Neural Network
4. Statistics
5. **Systematics**

Pipeline objectives

FA : Select appropriate features

- ➢ Avoid correlation
- ➢ Reduce systematic bias
- ➢ Avoid error amplification

BDT / NN : implement trained models

- ➢ choose evaluation criteria
- ➢ optimise hyper-parameters

STAT : estimate $\mu S + B$

- ➢ Assess signal sensitivity μ
- ➢ introduce ΔS and ΔB

SYST : provide systematic uncertainties

- ➢ limit impact on signal sensitivity μ
- ➢ use dataset to evaluate models

Feature Analysis

Lucas DESGRANGES

Arthur FILLON

Omar HAMOU TAHRA

Madeleine POTIER

Raphaël ROY

Martin SCHNEIDER

Working group 1

1 Feature analysis - Intro

- ★ Signal VS Background
- ★ Correlation matrix
- ★ Systematic bias
- ★ Shift and feature importance

1 Feature analysis - What are the features ?

Features

Primary features (PRI_)

What is measured by ATLAS

Exemple : PRI_had_pt, transverse momentum of the hadron

Derived features (DER_)

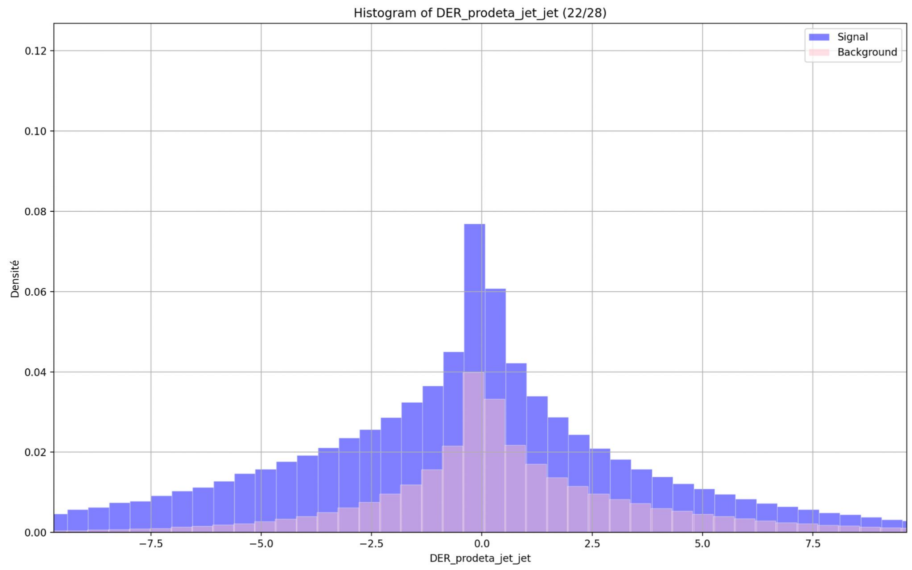
Derived from the primary features

Exemple : DER_pt_ratio_lep_tau, The ratio of the transverse momenta of the lepton and the hadronic tau

Has physical significance
→ Will help BDT or NN

1 Feature analysis - Signal VS Background

Same distribution for signal and background
→ a priori, the feature is non discriminating



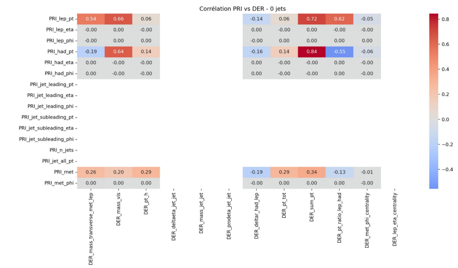
Signal distribution different from background distribution
→ a priori, the feature interesting to use

1 Feature analysis - Correlation matrix

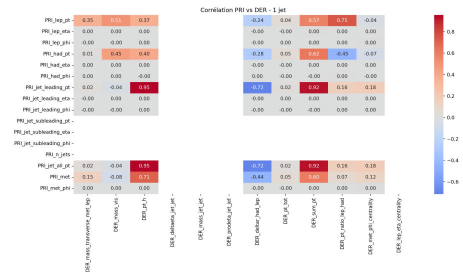
★ Analyse correlations between derived and primary features

★ Deduce which derived and primary features might be redondant

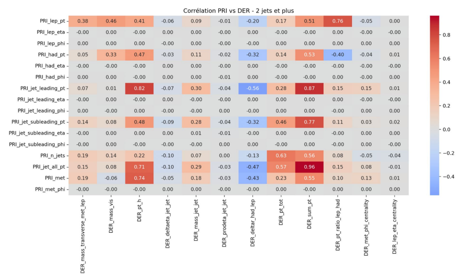
0 jets



1 jets



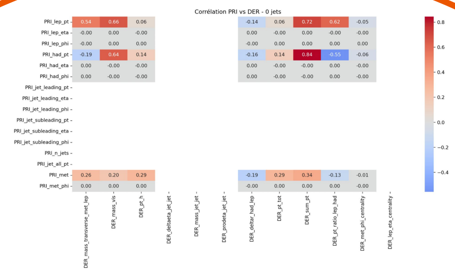
2 and + jets



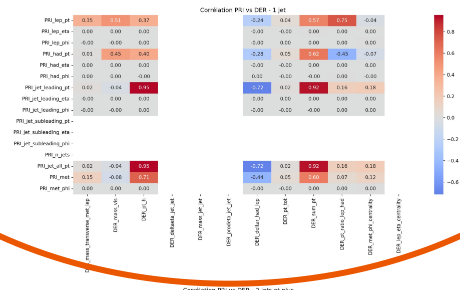
1 Feature analysis - Correlation matrix

Undefined values for 0 and 1 jets
→ Missing correlations
→ Focus on 2 and + jets

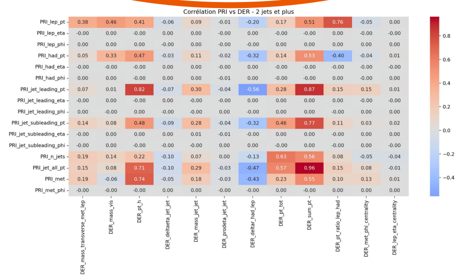
0 jets



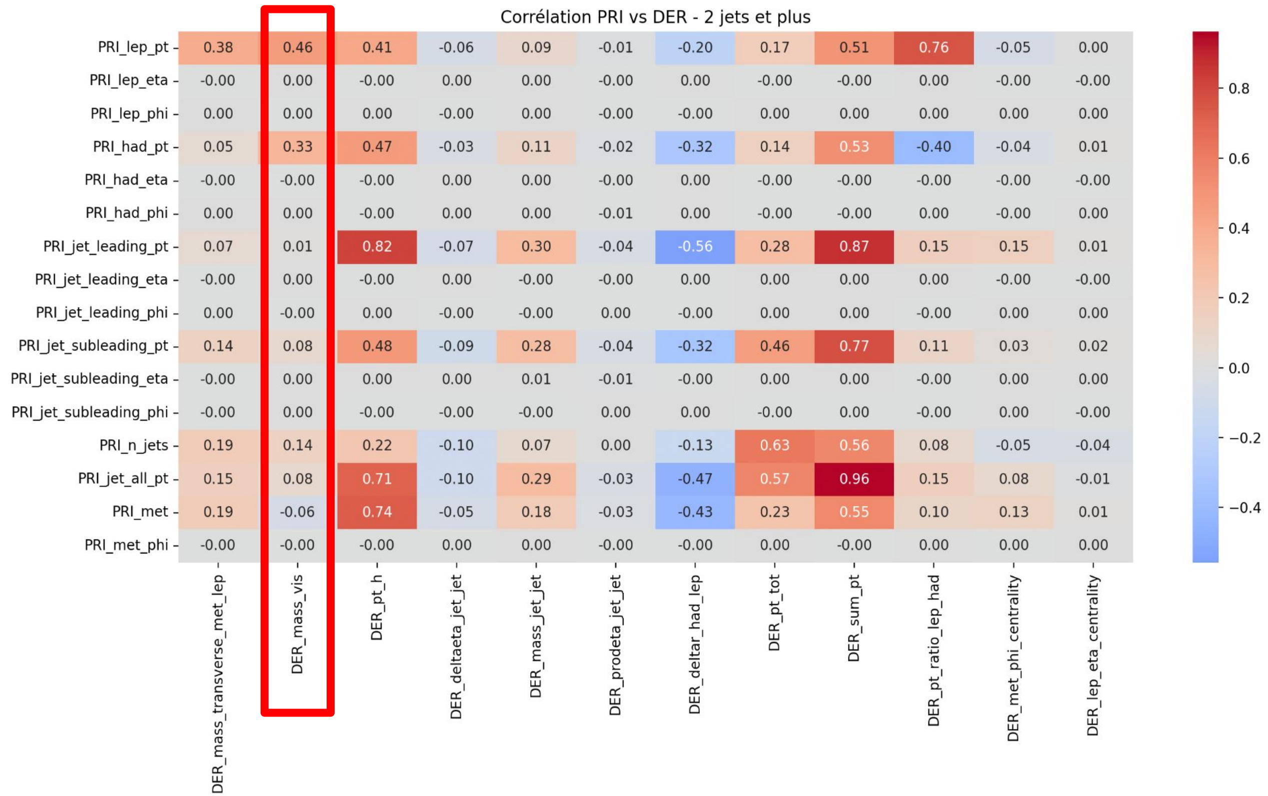
1 jets



2 and + jets



1 Feature analysis - Correlation matrix

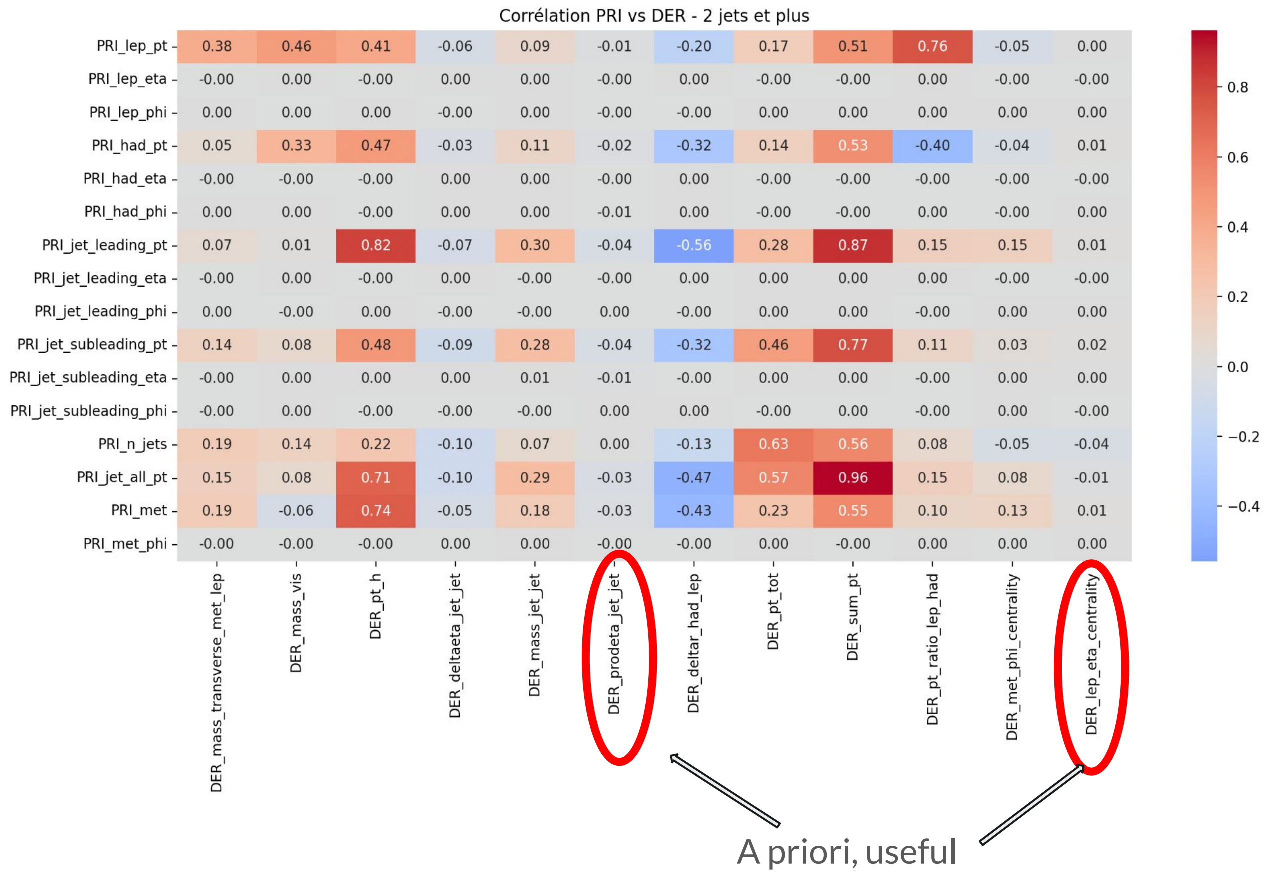


DER_mass_vis : The invariant mass of the hadronic tau and the lepton

→ A priori correlated to PRI_lep_pt and PRI_had_pt

→ Coherent with correlation matrix

1 Feature analysis - Correlation matrix



	DER_mass_transverse_met_lep	DER_mass_vis	DER_pt_h	DER_deltaeta_jet_jet	DER_mass_jet_jet	DER_prodeta_jet_jet	DER_deltar_had_lep	DER_pt_tot	DER_sum_pt	DER_pt_ratio_lep_had	DER_met_phi_centrality	DER_lep_eta_centrality
PRI_lep_pt	0.38	0.46	0.41	-0.06	0.09	-0.01	-0.20	0.17	0.51	0.76	-0.05	0.00
PRI_lep_eta	-0.00	0.00	-0.00	0.00	0.00	-0.00	0.00	-0.00	-0.00	0.00	-0.00	-0.00
PRI_lep_phi	0.00	0.00	0.00	-0.00	0.00	-0.00	-0.00	0.00	0.00	0.00	0.00	0.00
PRI_had_pt	0.05	0.33	0.47	-0.03	0.11	-0.02	-0.32	0.14	0.53	-0.40	-0.04	0.01
PRI_had_eta	-0.00	-0.00	-0.00	0.00	0.00	-0.00	0.00	-0.00	-0.00	-0.00	-0.00	-0.00
PRI_had_phi	0.00	0.00	-0.00	0.00	0.00	-0.01	0.00	-0.00	-0.00	0.00	-0.00	0.00
PRI_jet_leading_pt	0.07	0.01	0.82	-0.07	0.30	-0.04	-0.56	0.28	0.87	0.15	0.15	0.01
PRI_jet_leading_eta	-0.00	0.00	-0.00	0.00	-0.00	-0.00	0.00	0.00	0.00	0.00	-0.00	-0.00
PRI_jet_leading_phi	0.00	-0.00	0.00	-0.00	-0.00	0.00	-0.00	0.00	0.00	-0.00	0.00	0.00
PRI_jet_subleading_pt	0.14	0.08	0.48	-0.09	0.28	-0.04	-0.32	0.46	0.77	0.11	0.03	0.02
PRI_jet_subleading_eta	-0.00	0.00	0.00	0.00	0.01	-0.01	-0.00	0.00	0.00	-0.00	0.00	0.00
PRI_jet_subleading_phi	-0.00	0.00	-0.00	-0.00	-0.00	0.00	0.00	-0.00	0.00	-0.00	0.00	-0.00
PRI_n_jets	0.19	0.14	0.22	-0.10	0.07	0.00	-0.13	0.63	0.56	0.08	-0.05	-0.04
PRI_jet_all_pt	0.15	0.08	0.71	-0.10	0.29	-0.03	-0.47	0.57	0.96	0.15	0.08	-0.01
PRI_met	0.19	-0.06	0.74	-0.05	0.18	-0.03	-0.43	0.23	0.55	0.10	0.13	0.01
PRI_met_phi	-0.00	-0.00	-0.00	0.00	0.00	-0.00	-0.00	0.00	-0.00	0.00	0.00	0.00

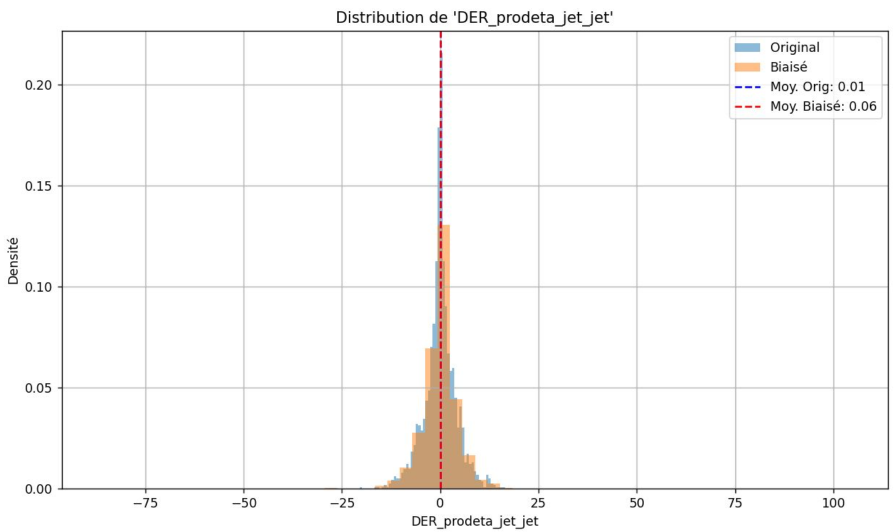
1 Feature analysis - Systematic bias

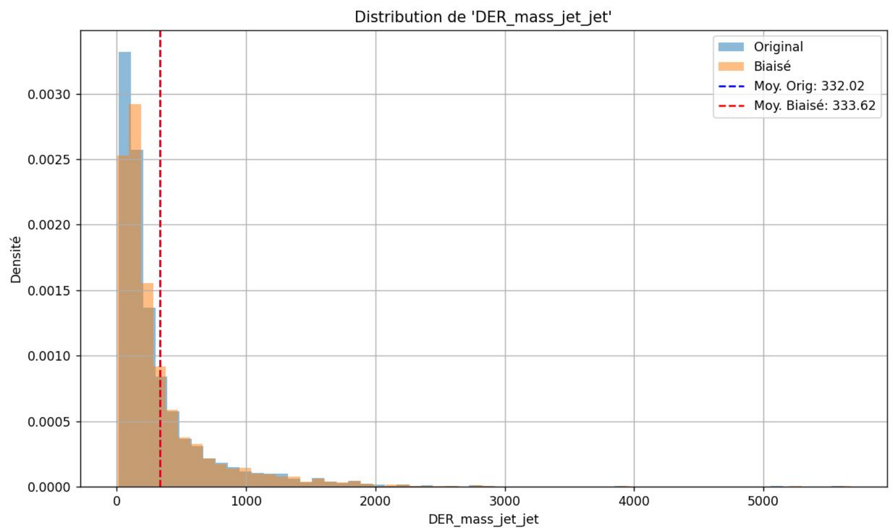
2 Nuisance parameters selected :

- ★ α_{tes} : Uncertainty related to the calibration of the resconstructed tau lepton energy
- ★ α_{jes} : Uncertainty related to jet energy calibration due to detector response

They do not apply to every PRI values
-> we use the systematics function to compute the biased features

1 Feature analysis - Systematic bias



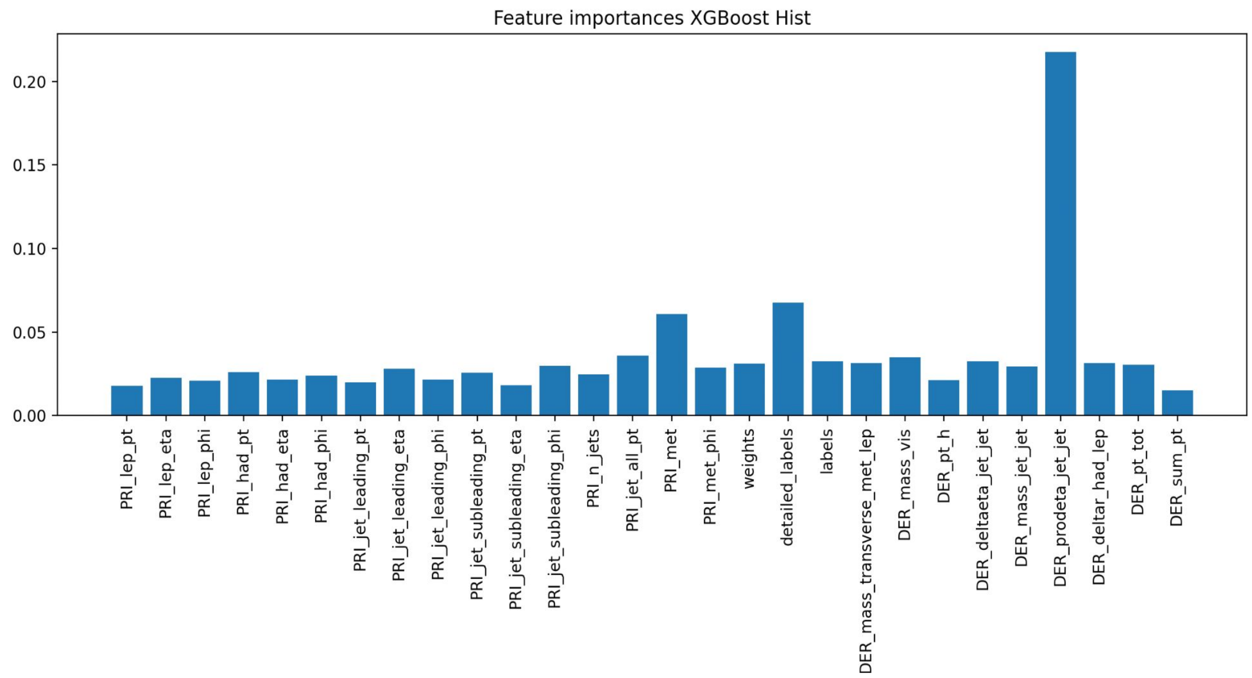


1 Feature analysis - Wasserstein Distance

- ★ Quantify the shift between distributions
- ★ Intuitive explanation : Minimal amount of work to turn a pile of soil into another

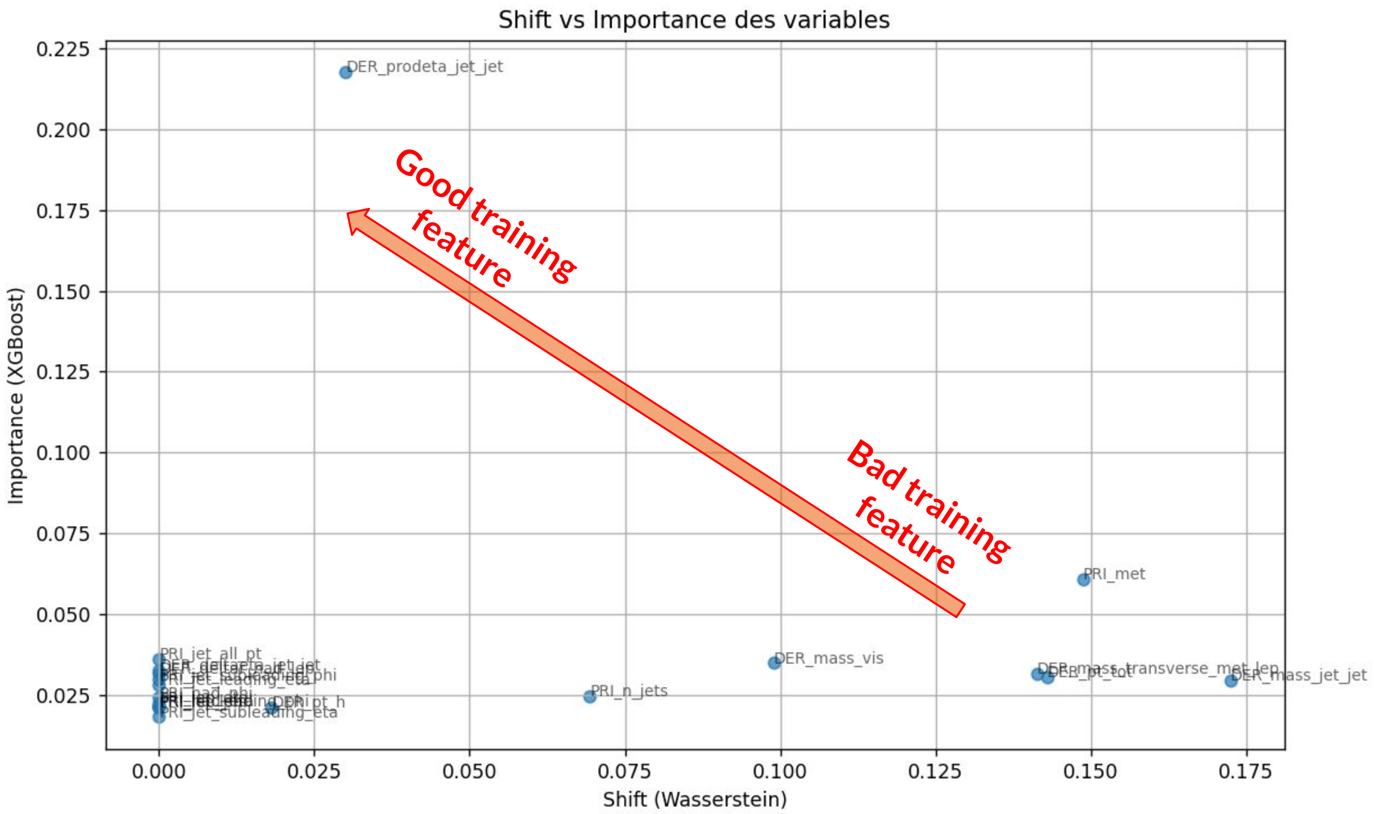
$$W_1(\mu_1, \mu_2) = \int_{\mathbb{R}} |F_1(x) - F_2(x)| \, \mathrm{d}x.$$

1 Feature analysis - Feature importance

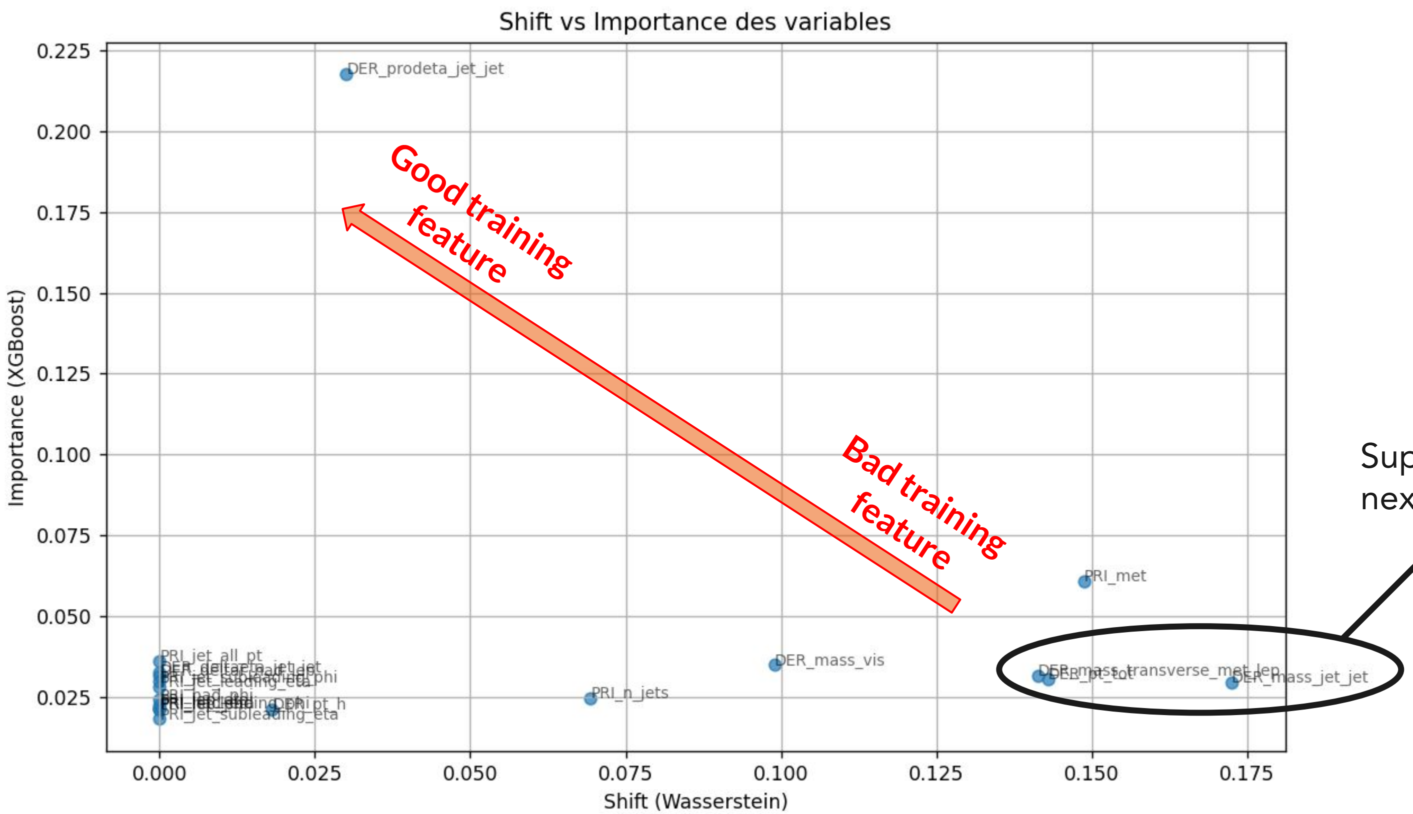


Calculated with the BDT model

1 Feature analysis - Relevant features



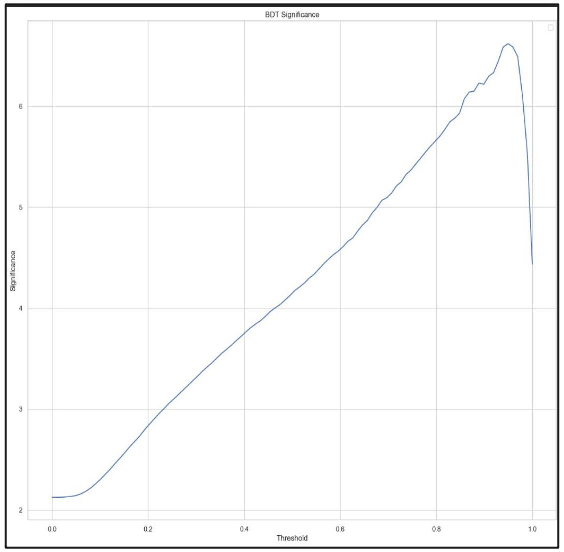
1 Feature analysis - Relevant features



Suppressed for next training

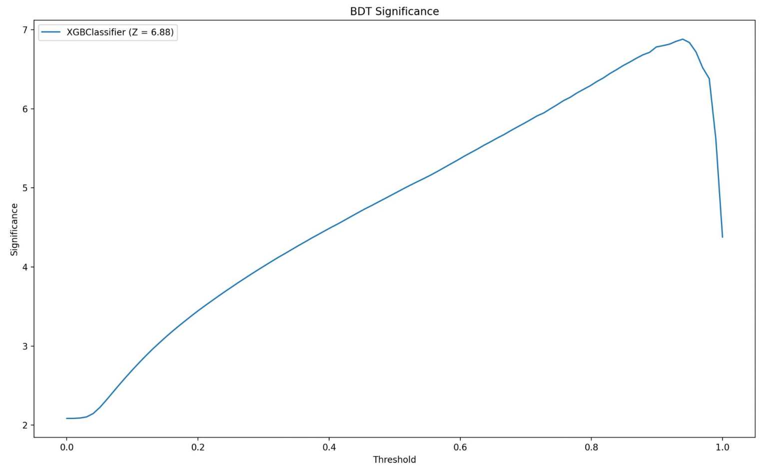
1 Feature analysis - Significance with BDT

For all features unbiased



Z = 6.61

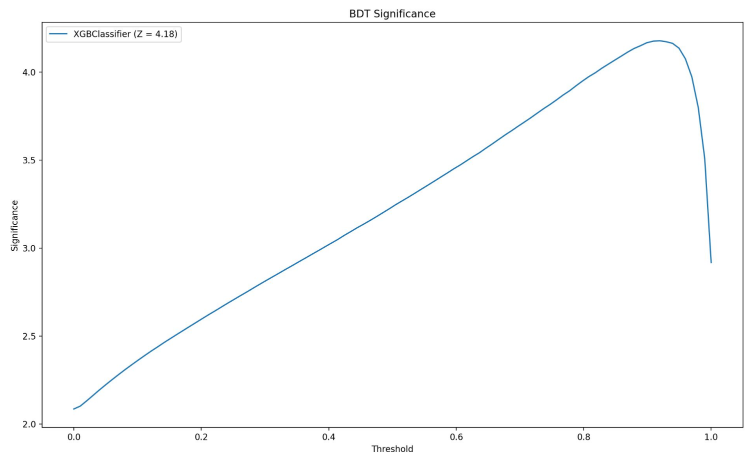
For selected features unbiased



Z = 6.88

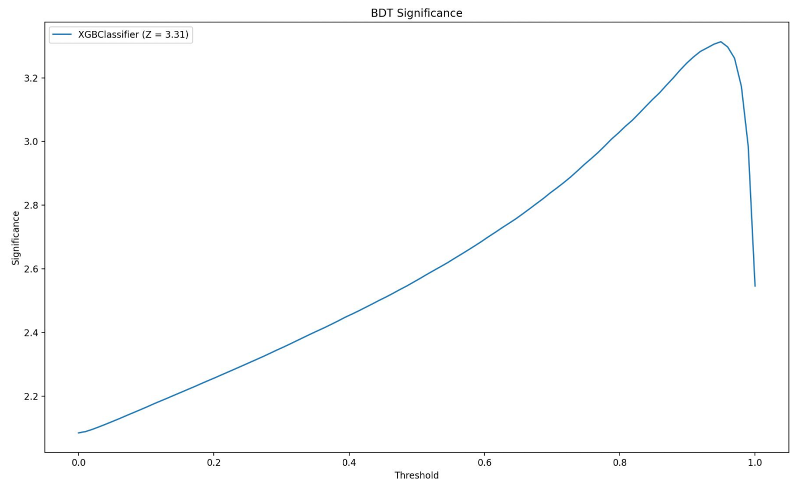
1 Feature analysis - Significance with BDT

For all features biased



$Z = 4.18$

For selected features biased



$Z = 3.31$

Boosted Decision Tree

Mahdi AYADI

Yasmine HAKMOUNI

Aymen AWAINIA

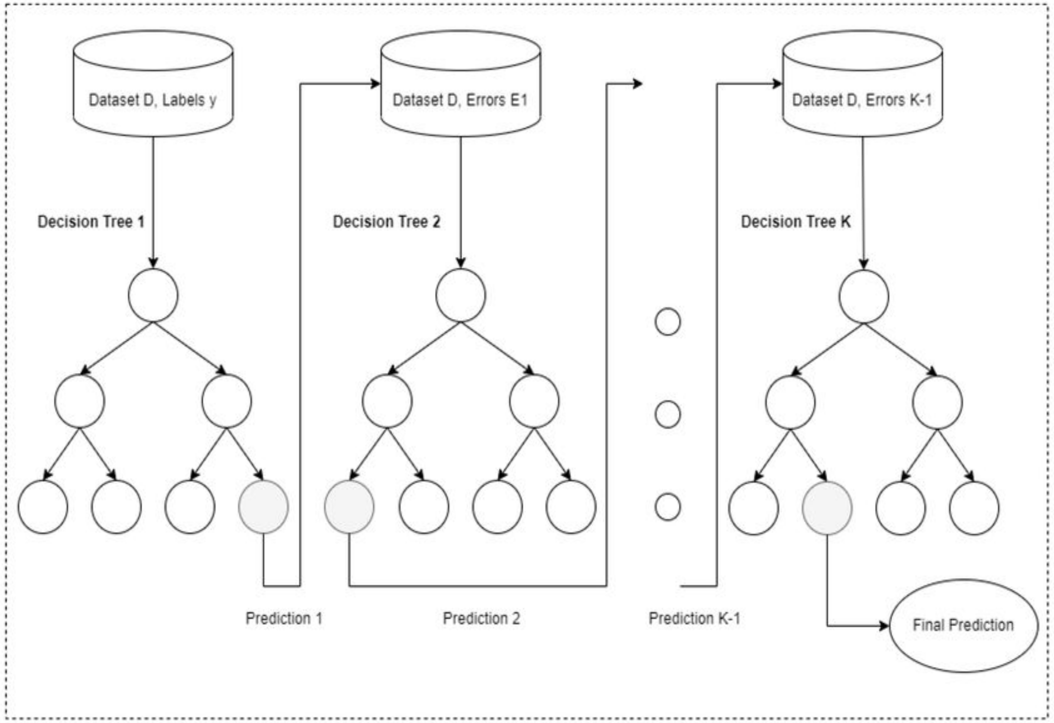
Paul MENARD

Lucie MOUGIN

Working group 2

2 Boosted Decision Trees - Intro

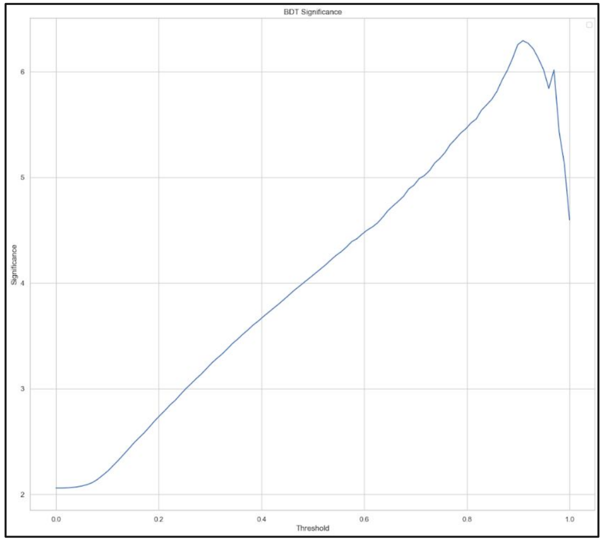
- ❖ BDTs?
  - Gradient boosting algorithm
  - Sequential model building
  - Error correction at passage
- ❖ Suitability
  - Complex data
  - Good performance
- ❖ Used Implementations
  - XGBoost
  - LightGBM
  - Scikit-learn GradientBoostingClassifier



2 Initial Setup - Model training

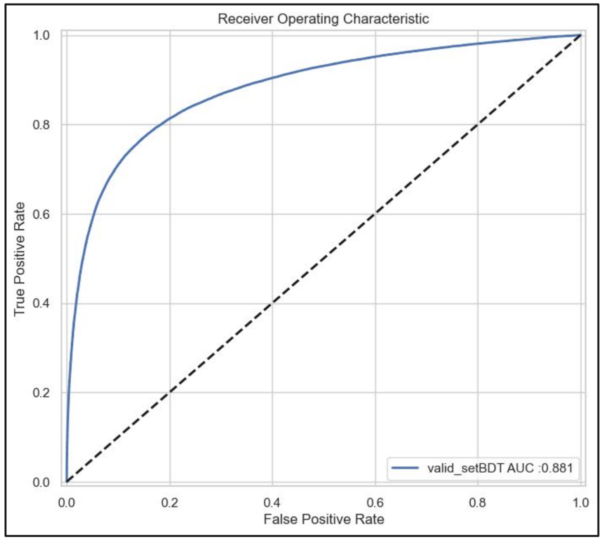
- Dataset events chunk size:
  - 500 000 Training set
  - 50 000 Holdout set
  - ~500 000 Test
- First XGBoost Model
- Default parameters
- Baseline evaluation Metrics:
  - Evaluate_AUC
  - Evaluate_significance

BDT Significance



Z=6.29

Receiver Operating Characteristic



AUC = 0.881

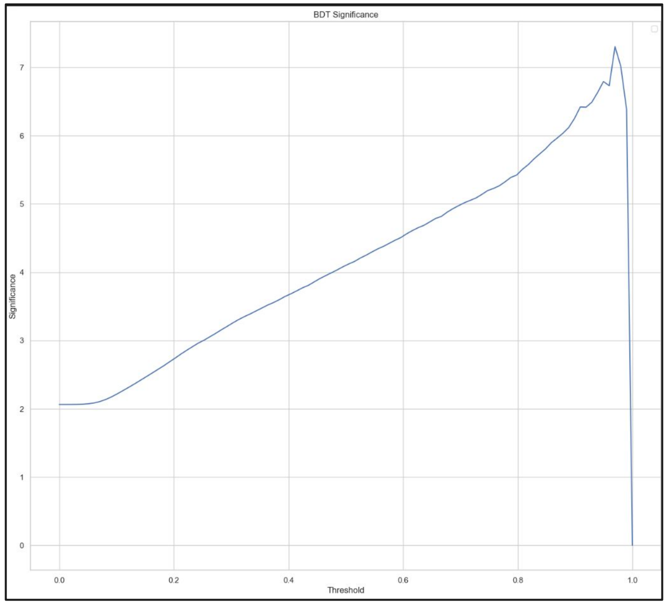
2 HyperParameter Optimisation

- Tuning is Crucial for performance
- Technique: RandomizedSearchCV
- integrated in fit( ) function
- Key parameters Tuned

```python
param_dist = {
    "max_depth": stats.randint(3, 10),
    "n_estimators": stats.randint(100, 300),
    "learning_rate": stats.uniform(0.05, 0.5),
}
gsearch = RandomizedSearchCV(
    estimator=self.model,
    param_distributions=param_dist,
    scoring="roc_auc",
    n_iter=120,
    cv=2,
    random_state=42,
    verbose=1,
    n_jobs=-1
)
```

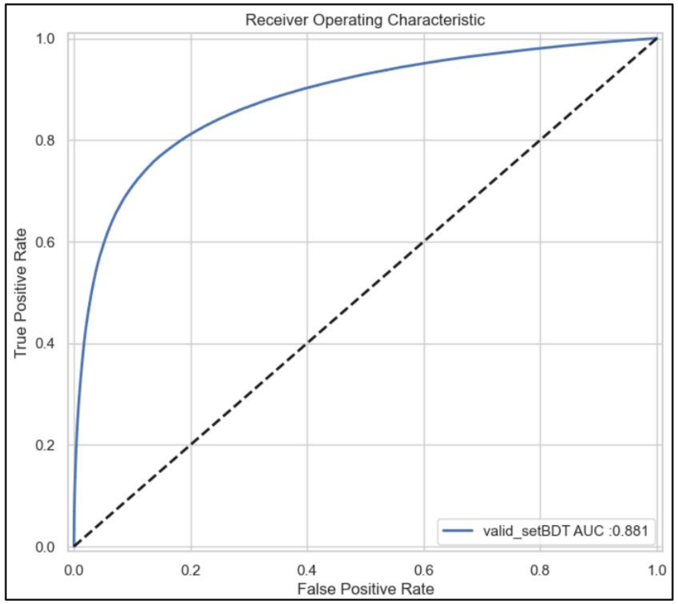
2 Training Xgboost with HPO

BDT Significance (120iter)



Z=7.3

Receiver Operating Characteristic



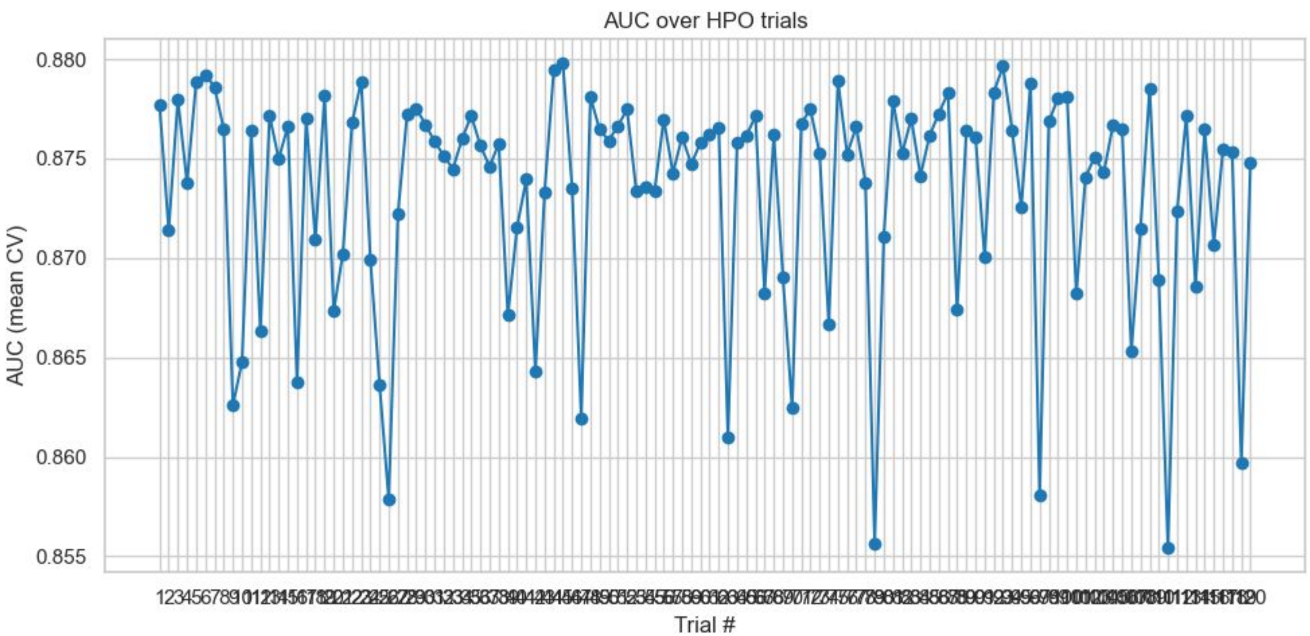
AUC = 0.882

Conclusion: AUC remains the same with and without HPO, but significance improves with HPO

2 AUC Throughout HPO

Tracing AUC curves as functions of iterations:

AUC over each iteration of Random Search CV



- Fluctuations
- No Influence

➡ Key variable: SIGNIFICANCE

2 In-depth evaluation - HPO params

Learning rate	0,005	0,3	0,5	0,995
Significance	4,66	6,18	6,10	5,07

Max Depth	1	6	10	20
Significance	4,36	4,36	6,04	5,75

n_estimators	10	150	500	5000
Significance	5.24	6.16	6.19	5.86

Significance based on parameters
(The others parameters are set by default)

Observing significance evolution with:

- Learning_rate
- Max_depth
- N_estimators

Conclusion:

➢ There is an optimal range of hyperparameter values where significance is maximized.

➢ Beyond this range, increasing the parameters brings no added benefit and may even reduce statistical performance.

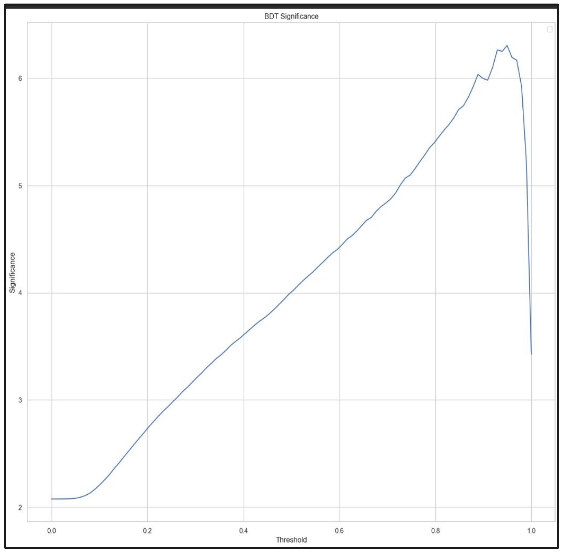
➢ Best parameters are :
Learning rate: 0.237
Max Depth: 7
n_estimators:114

2 Comparison of Models with HPO

20 iterations

LightGBM:

BDT Significance



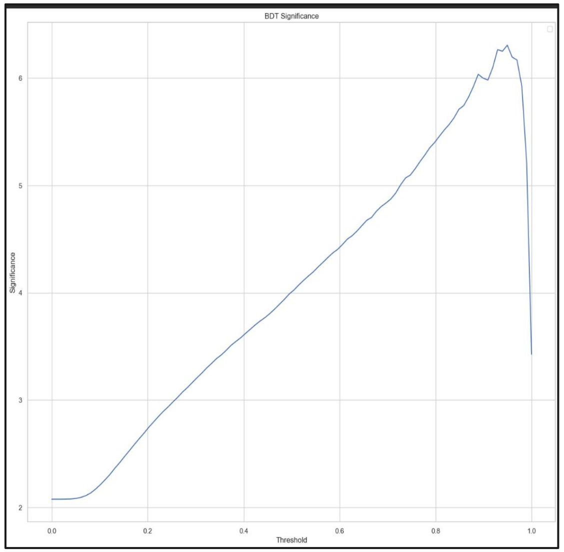
$Z = 6.30$

2 Comparison of Models with HPO

20 iterations

LightGBM:

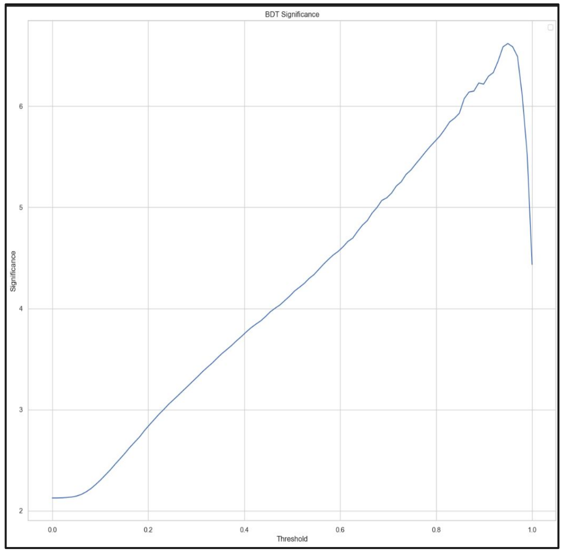
BDT Significance



Z = 6.30

XGBoost :

BDT Significance



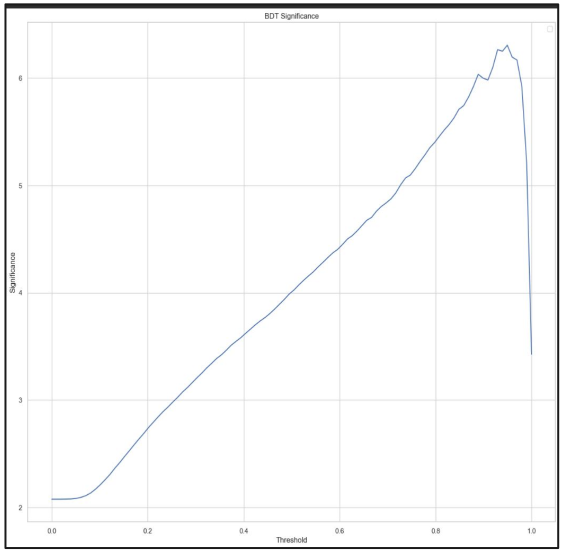
Z = 6.61

2 Comparison of Models with HPO

20 iterations

LightGBM:

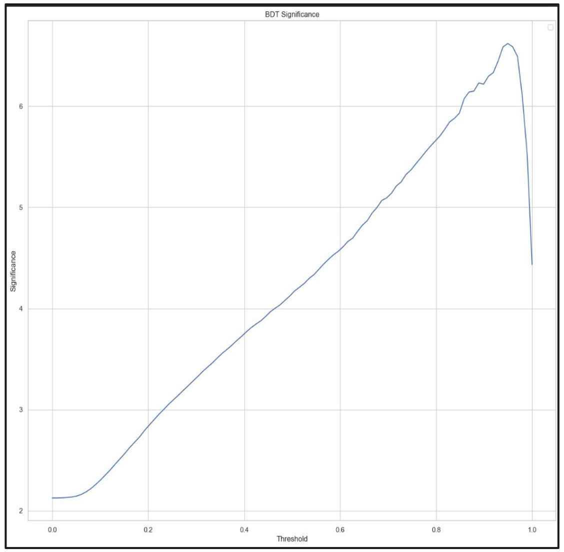
BDT Significance



Z = 6.30

XGBoost :

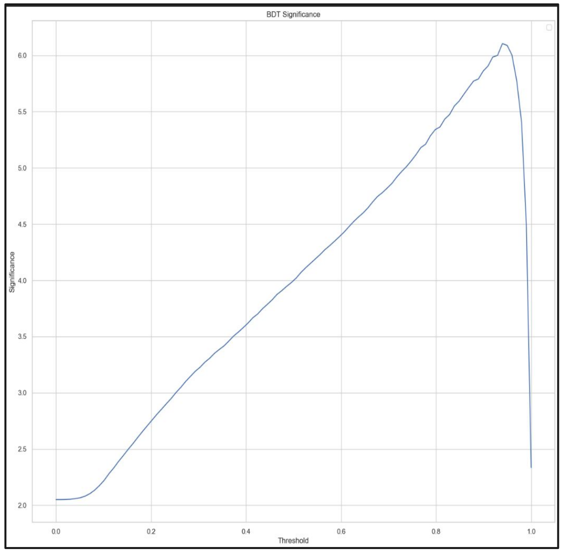
BDT Significance



Z = 6.61

SKLearn :

BDT Significance



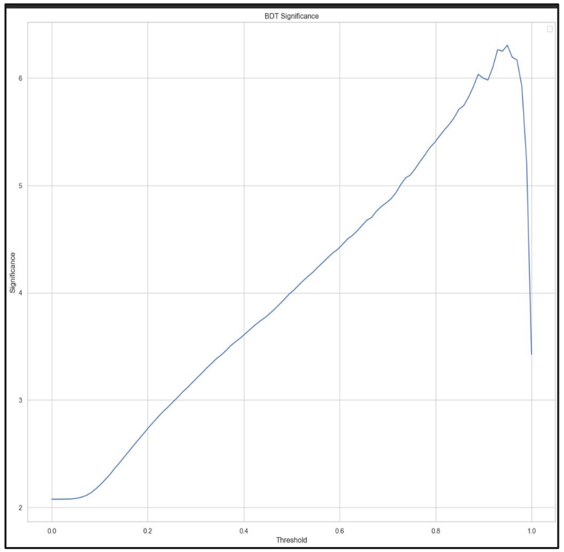
Z = 6.10

2 Comparison of Models with HPO

20 iterations

LightGBM:

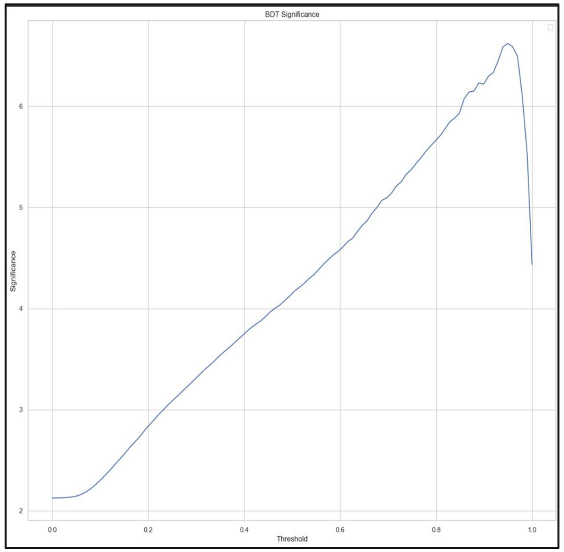
BDT Significance



Z = 6.30

XGBoost :

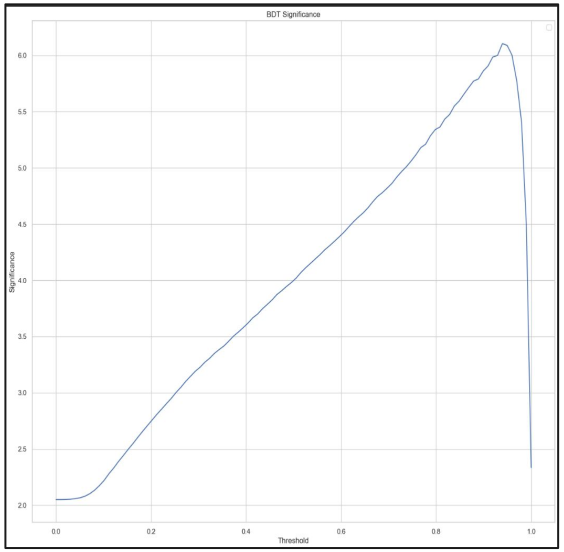
BDT Significance



Z = 6.61

SKLearn :

BDT Significance



Z = 6.10

2 Final Verdict - significance

- 3 model comparison
- Applied HPO (20 Iterations)
- Significance values:

- XGBoost: 6.61
- LightGBM: 6.30
- SKLearn: 6.10

- CONCLUSION:
  - Successfully identified the the best evaluation metric (significance)
  - Developed and optimised a powerful classifier:

    XGBoost with HPO (**n_iter : 120)** :

    Learning rate: 0.237

    Max Depth: 7

    N_estimators:114

    $\boxed{\Rightarrow Z = 7.3}$

Neural Network

Hugo GAUVRIT

Martin SÉGUINEAU DE PRÉVAL

Ana KOENIG

Grégoire EYMARD

Chloé-Ly GRIMONT

Working group 3

3 Neural Network - introduction

Goal: Neural Network to predict the right label thanks to a dataset around 2 million events.

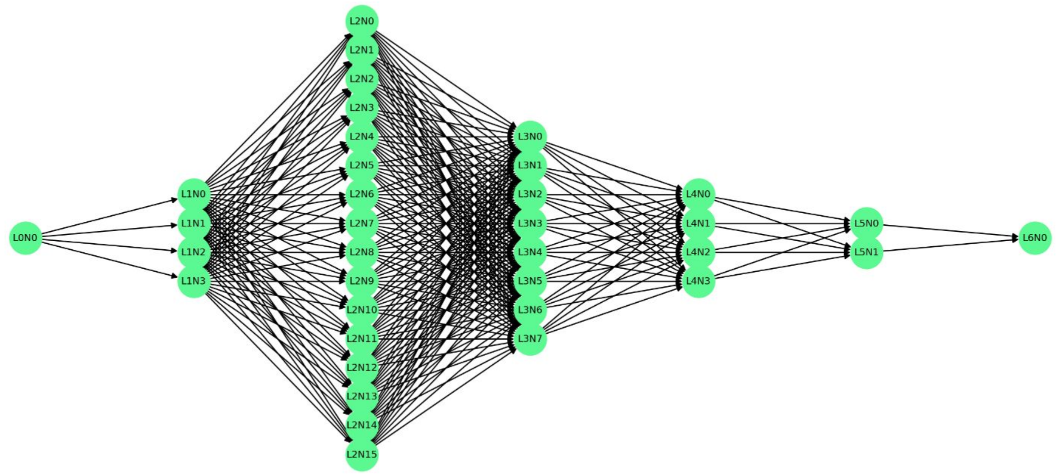
- Data with labels 0/1
- 0 → Background event
- 1 → Signal event ($H \rightarrow \tau\tau$)
- 30 physics-based features

3 Neural Network - the model

Parameters	Choice	Remark
Loss function	Binary cross-entropy	Log loss : usual for binary classification
Activation function	ReLu Sigmoid output	Keep only positive values Transformation to classify in 2 groups
Learning rate	10^{-4}	controls the speed at which the model updates its weights to minimize the loss
Optimiser	Adam	“Adaptive Moment Estimation”
Training dataset size	300 000 events	Used to train the model
Holdout dataset size	100 000 events	Estimate systematic effects
Validation dataset size	200 000 events	Used for model evaluation, significance calculation

3 Neural Network - HPO

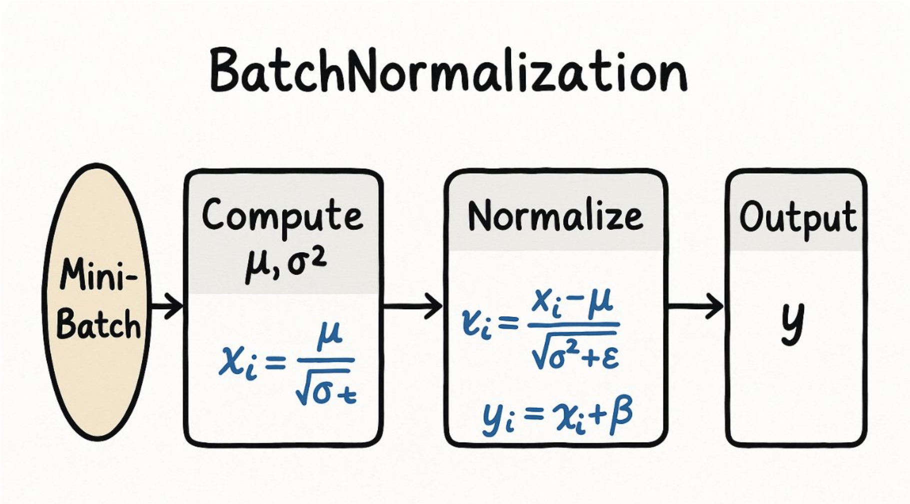
- Optimize the architecture
  - Bottleneck structure
    - To allow the network to compress informations (at the beginning, 30 informations, at the end only one)



Bottleneck-structured neural network. *Generated by ChatGPT.*

3 Neural Network - HPO

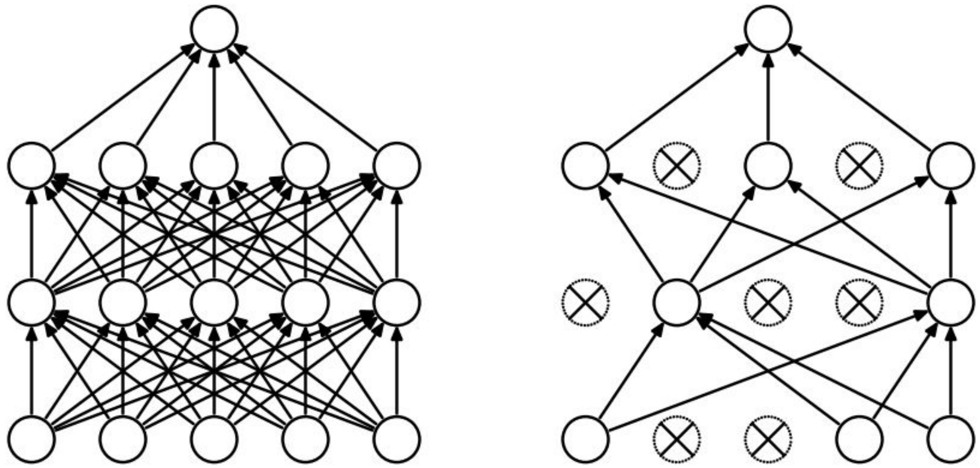
- Minimising the loss function
  - Batch normalisation
    - normalises weights between layers when training model
    - goal : obtain average = 0, variance = 1



How a batch is normalised. *Generated by ChatGPT.*

3 Neural Network - HPO

- Avoid overfitting
  - Dropout function
    - alternately deletes neurons when training
  - Early stopping
    - limit epochs once the model stops progressing



(a) Standard Neural Net (b) After applying dropout.

Dropout: A Simple Way to Prevent Neural Networks from Overfitting, Journal of Machine Learning Research 2014, Nitish Srivastava et al.

3 Neural Network - Final model

Hyper-parameters	Choice	Remark
Depth	9 layers (7 hidden layers)	With Bottleneck architecture
Width	2041 neurons	(1024, 512, 256, 128, 64, 32, 16, 8, 1)
Number of epochs	80	An excessive number however it is ok with early stopping
Early stopping	Stopped at 16	Impacts positively the training time
Drop out	0.2	On the 7 first layers
Batch normalisation	Yes	On each layer except the last one

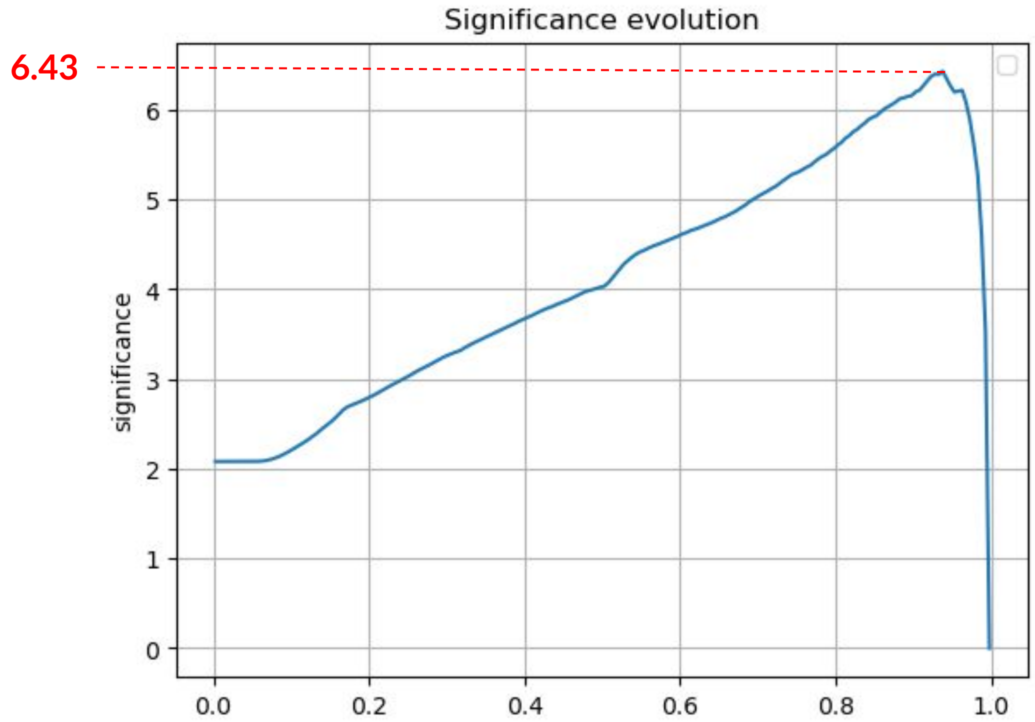
3 Neural Network - evaluation criteria

We focus the study on 5 evaluation criteria :

- Significance
- AUC - ROC
- Score
- Loss
- Training time

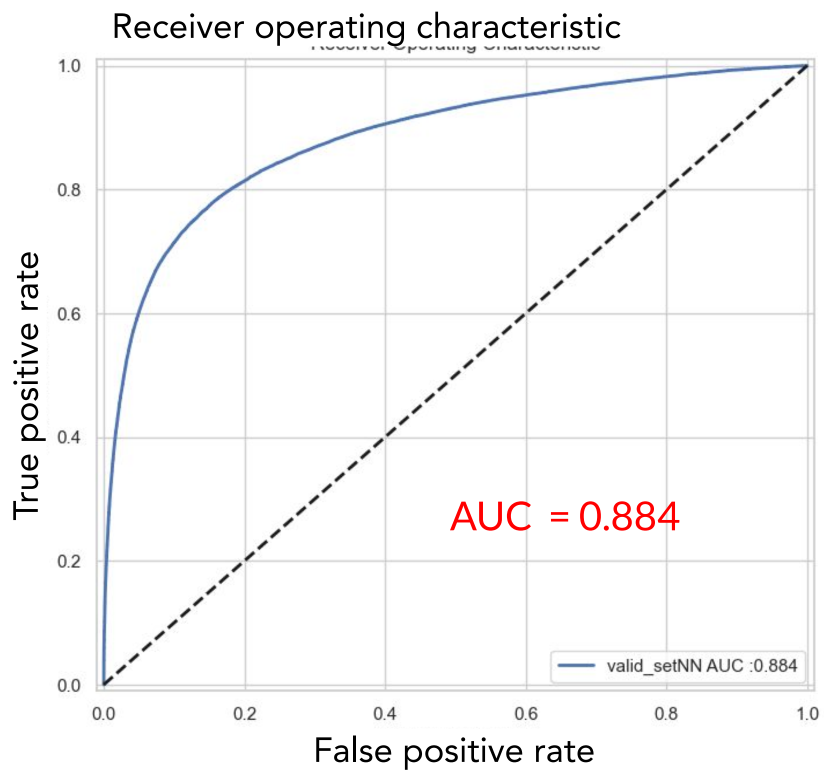
3 Neural Network - evaluation criteria

- Significance
- AUC - ROC
- Score
- Loss
- Training time



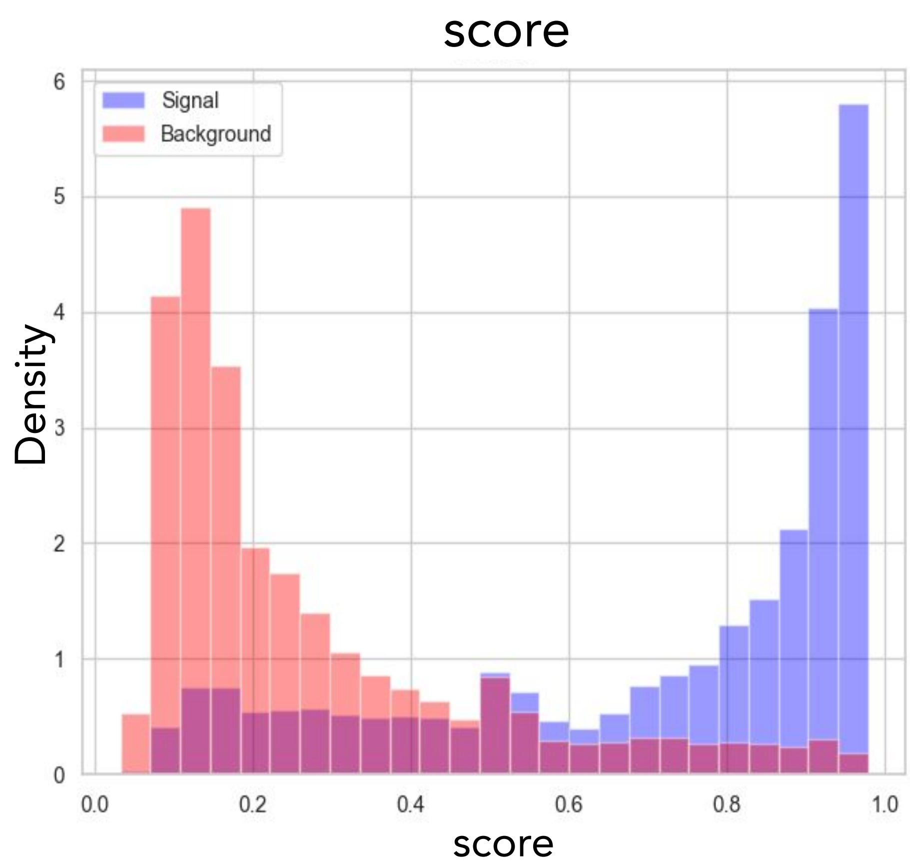
3 Neural Network - evaluation criteria

- Significance
- AUC - ROC
- Score
- Loss
- Training time



3 Neural Network - evaluation criteria

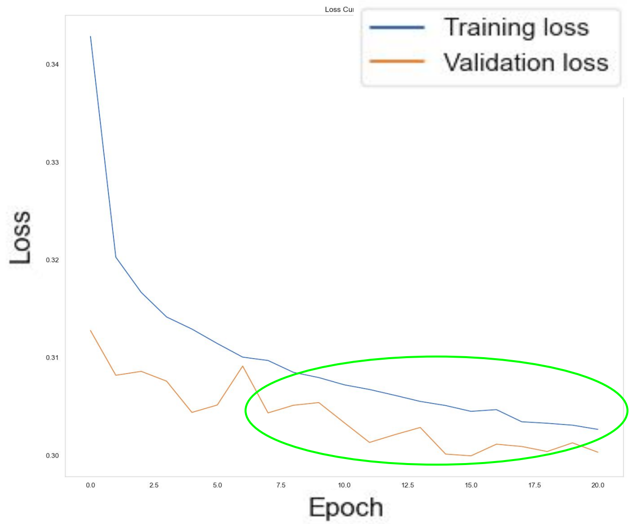
- Significance
- AUC - ROC
- Score
- Loss
- Training time

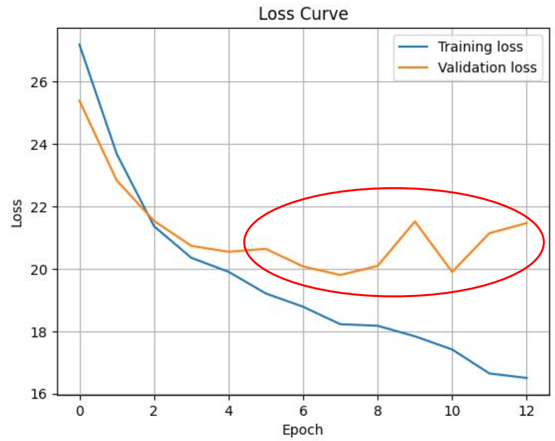


3 Neural Network - evaluation criteria

- Significance
- AUC - ROC
- Score
- Loss
- Training time

Loss curve

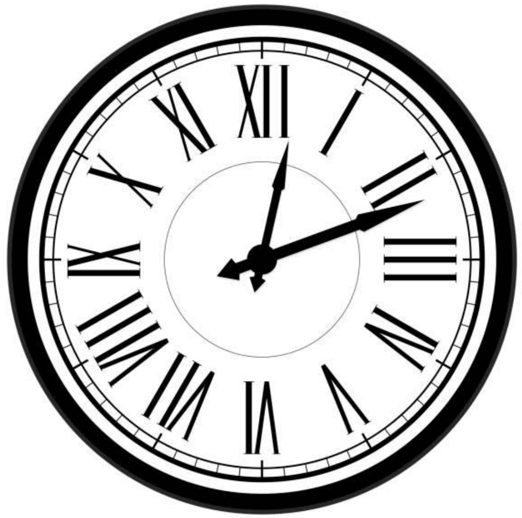




Overfitting

3 Neural Network - evaluation criteria

- Significance
- AUC - ROC
- Score
- Loss
- Training time



Approx. 25 min

3 Neural Network - Possible Improvements

What to explore & improve ?	Tools
Change activation functions	LeakyReLu, SiLu
HPO	With tools optuna
Prevent overfitting	More dropout, L2 regularisation
Custom loss function	Use Asimov significance loss
Schedule learning rate	Use ReduceLROnPlateau

Statistics

JAEGER,DELIGNON,BRETON,HENNART,LASCHON,VANDIER

Working group 4

4 Statistics : introduction

Goal : estimate μ and the statistical & systematic uncertainties ($\delta\mu_{stat}$, $\delta\mu_{sys}$)

Due to the randomness of data (arising from the finite size of the sample and random fluctuations)

Due to the lack of knowledge of the probability distribution of what we observe

4 Statistics : several approaches

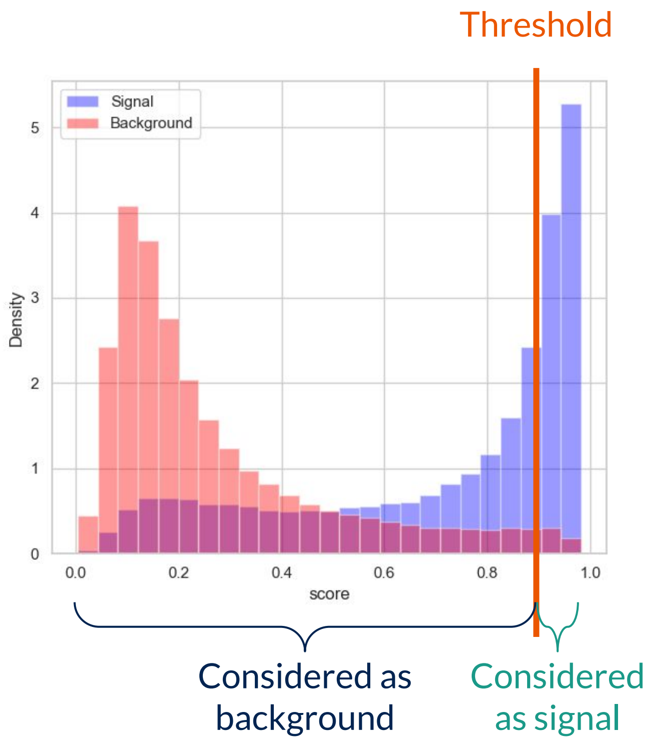
- Most basic approach : looking for a threshold to separate signal and background based on the score

- Improved approach : Using Minuit algorithm (gradient descent) on the negative log-likelihood (nll)

- nearly complete approach : Same as before but taking into account the systematics errors

4 Statistics : 1st approach

- Looking for a threshold to separate signal and background based on the score



The sum of the weights of the events above the threshold in denoted N

Estimation of μ and δμ:

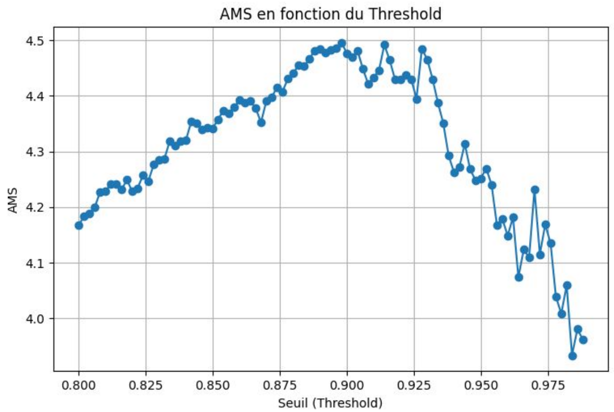
$$\mu_{counting\ method} = \frac{N - \beta}{\gamma}$$

$$\delta\mu_{stat,couting\ method} = \frac{\sqrt{\gamma + \beta}}{\gamma}$$

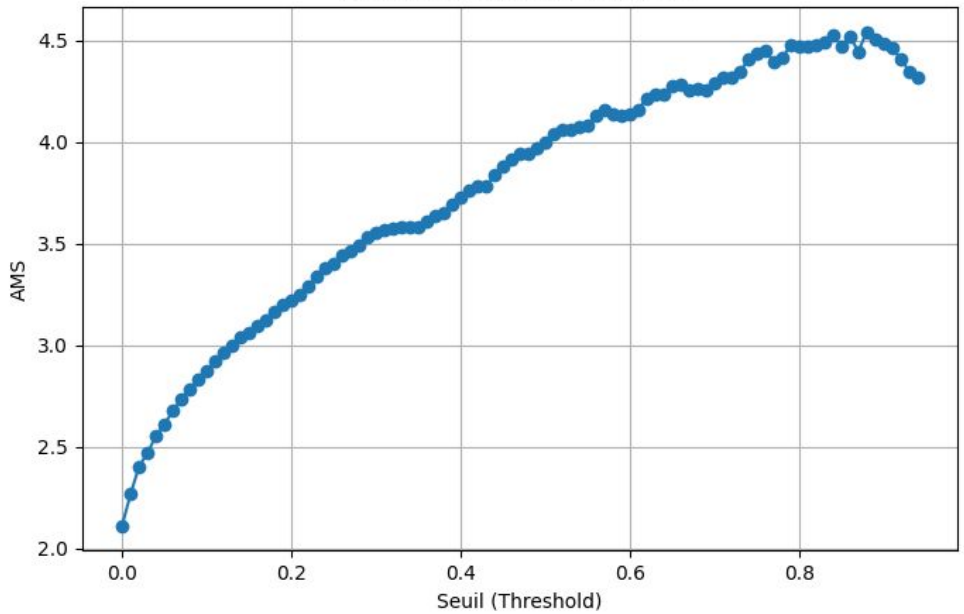
4 Statistics : 1st approach

AMS : Approximate Median Significance

$$AMS = \sqrt{2\left((s+b)\cdot\ln\left(1+\frac{s}{b}\right)-s\right)}$$



AMS as a function of Threshold (Zoom)



AMS as a function of Threshold

4 Statistics : 1st approach

```
Train Results:
        mu_hat  :  -1.1052031361101093
        del_mu_stat  :  0.17808067327043833
        del_mu_sys  :  0.0
        del_mu_tot  :  0.17808067327043833
Holdout Results:
        mu_hat  :  0.9999999999999987
        del_mu_stat  :  0.17808067327043833
        del_mu_sys  :  0.0
        del_mu_tot  :  0.17808067327043833
Valid Results:
        mu_hat  :  0.47725164812641335
        del_mu_stat  :  0.17808067327043833
        del_mu_sys  :  0.0
        del_mu_tot  :  0.17808067327043833
Significance (Asimov):
       Maximum Asimov significance: 6.1845
```

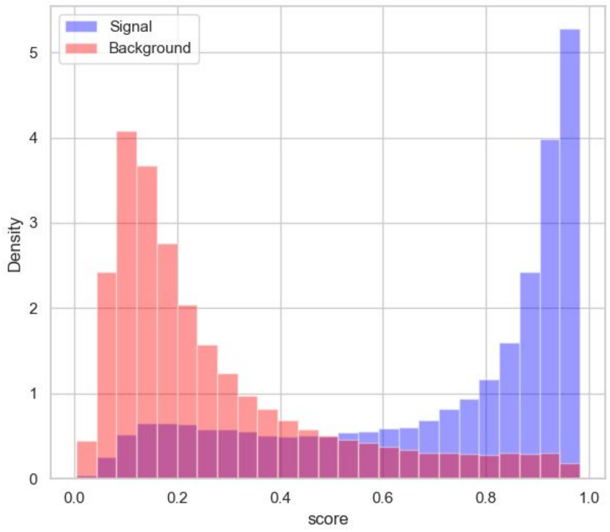
BDT results

```
Train Results:
        mu_hat  :  1.0686875333753154
        del_mu_stat  :  0.24292011036127267
        del_mu_sys  :  0.0
        del_mu_tot  :  0.24292011036127267
Holdout Results:
        mu_hat  :  0.9999999999999994
        del_mu_stat  :  0.24292011036127267
        del_mu_sys  :  0.0
        del_mu_tot  :  0.24292011036127267
Valid Results:
        mu_hat  :  1.2120187400591114
        del_mu_stat  :  0.24292011036127267
        del_mu_sys  :  0.0
        del_mu_tot  :  0.24292011036127267
Significance (Asimov):
       Maximum Asimov significance: 4.5195
```

NN results

4 Statistics : 2nd approach

- Refining the previous approach :



For each bin, we consider that the number of events follow a Poisson law with a parameter which is **specific to that bin** : $\mu\gamma_i + \beta_i$

(γ_i : weights of the Signal events in the i-th bin
□$_i$: weights of the Background events in the i-th bin)

We are going to estimate μ with the Maximum Likelihood Estimator :

Set : $y_i = \gamma_i +$ □$_i$ and $V_i \sim \text{Poisson}(\mu\gamma_i +$ □$_i)$

Log-likelihood for the i-th bin : $NLL_i(\mu) = -2\log(\boldsymbol{Pr}(V_i = y_i))$

Log-likelihood for the set of all bins : $NLL_{tot}(\mu) = \sum_{i=0}^{N_{bins}-1} NLL_i(\mu)$

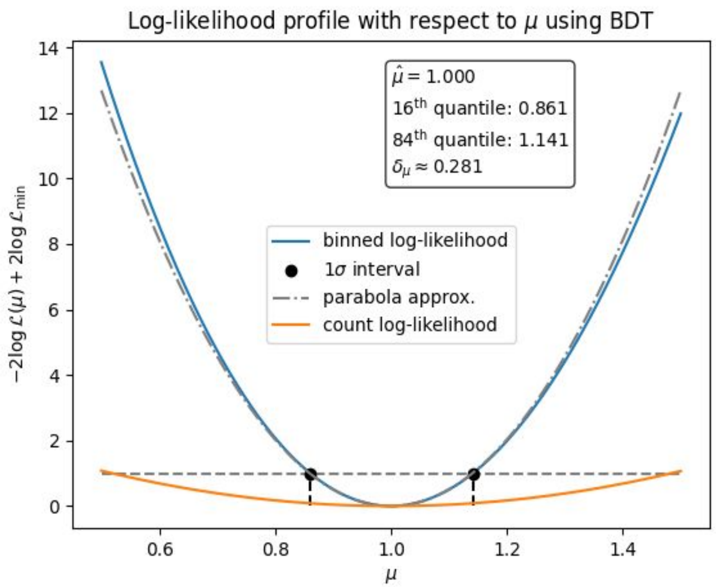
Estimated δμ ← Estimated μ ← iminuit

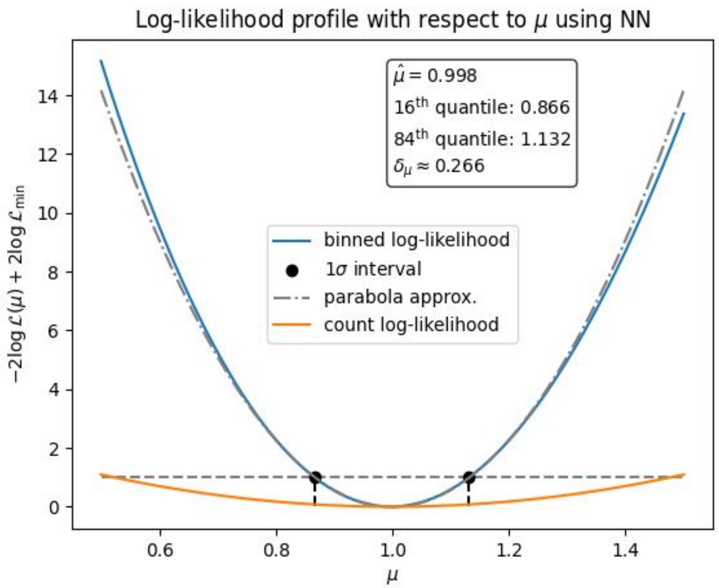
Minimization

4 Statistics : 2nd approach

Number of bins = 25

Here we plot the difference of the NLL with its minimum :

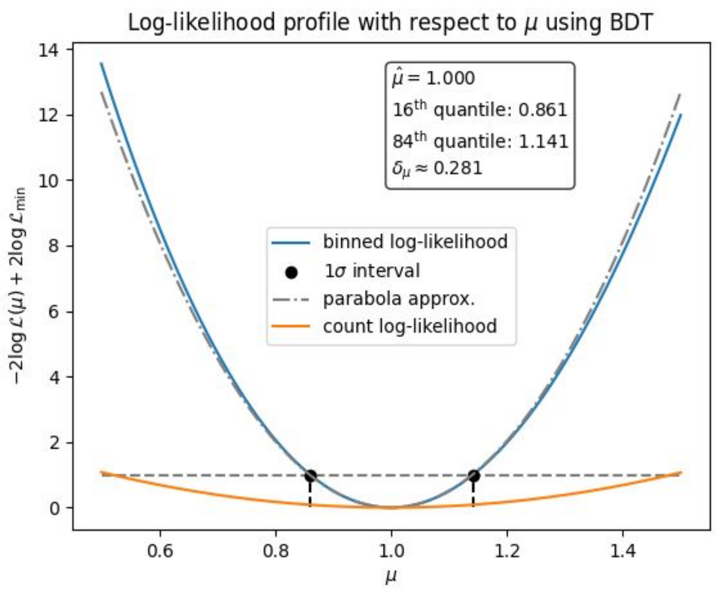


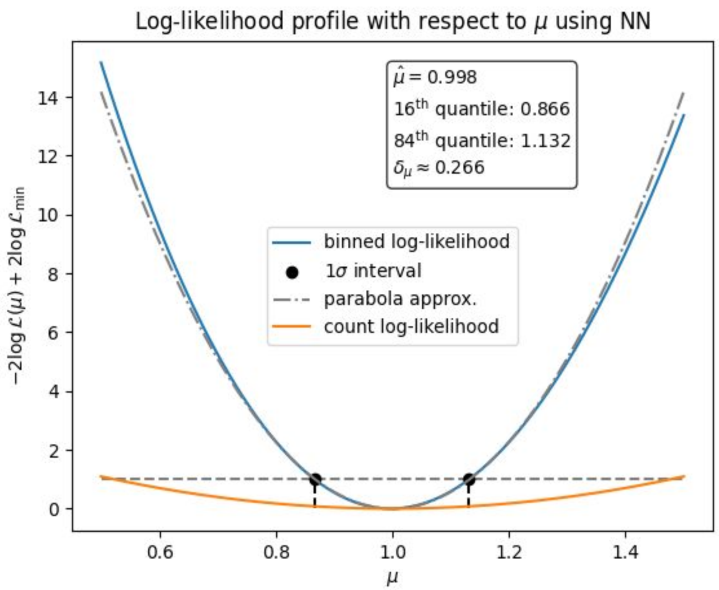


Intersections give the 1 sigma interval i.e. the 16th and 84th quantiles (gaussian approximation of the Poisson laws)

4 Statistics : 2nd approach

Number of bins = 25

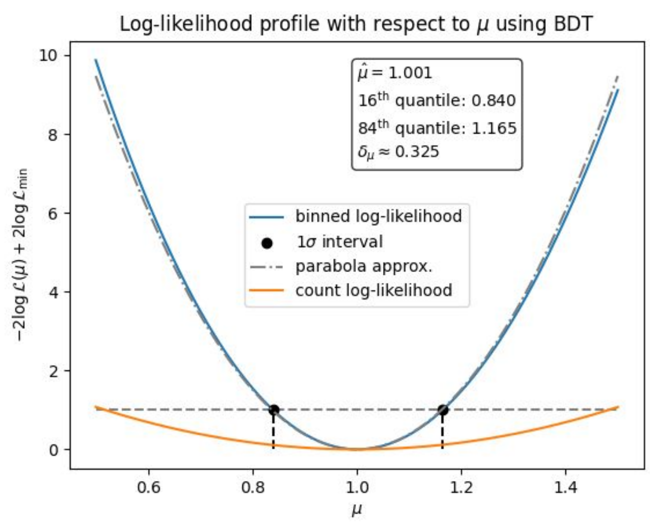


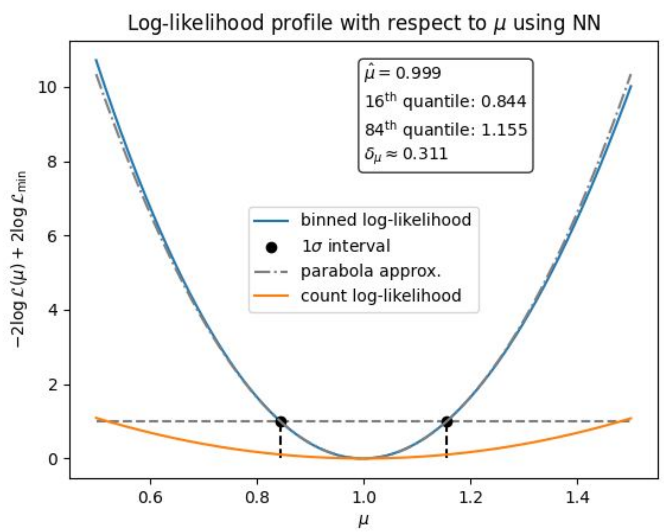


NN results are slightly better, but it is longer to train

4 Statistics : 2nd approach

Number of bins = **10**





When the number of Bins decreases, the δμ interval length increases

4 Statistics : TES and JES

- We have to propagate the uncertainty taking into account the systematic uncertainty

$$\mu\gamma_i + \beta_i \rightarrow \mu\gamma_i'(\alpha_{JES}, \alpha_{TES}) + \beta_i'(\alpha_{JES}, \alpha_{TES})$$

$$\gamma_i' = \gamma_i + \boxed{jesfitter_{signal}}(\alpha_{JES}) + \boxed{tesfitter_{signal}}(\alpha_{TES})$$

$$\beta_i' = \beta_i + \boxed{jesfitter_{background}}(\alpha_{JES}) + \boxed{tesfitter_{background}}(\alpha_{TES})$$

Functions provided by the Systematics team

4 Statistics : TES and JES

We do the same thing we did with binned likelihood taking our new functions, however we now need to minimize the NLL with respect to the 3 parameters, not just μ

→ We managed to call the jes_fitter and tes_fitter functions
→ However, we didn’t obtain the uncertainties on the 3 parameters, potentially due to problems with our utilization of Minuit (returning a lot of “NaN”), or to misunderstandings regarding the parts of the jes_fitter and tes_fitter outputs that we had to exploit

Systematics

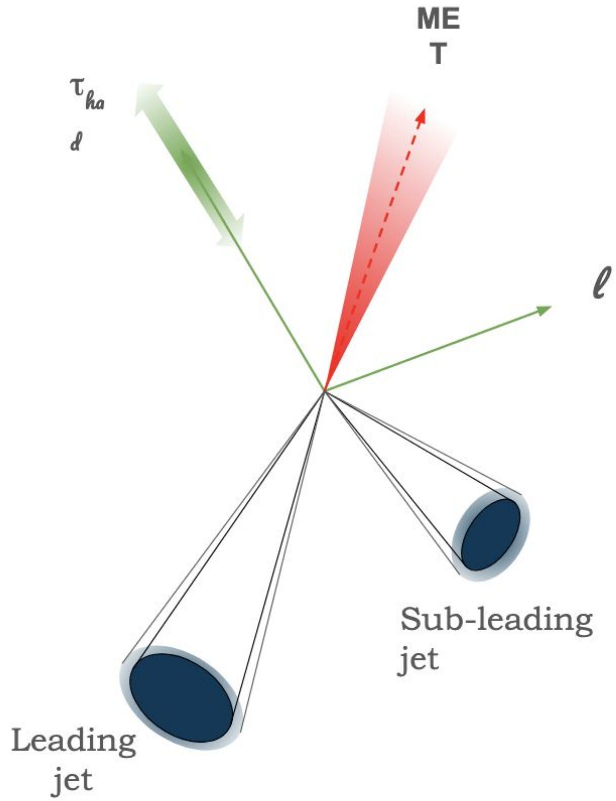
ALBESSARD, BOURRET, GHERBI, LECOINTRE, PLUMEJEAU, QUÉNARD

Working group 5

5 Systematics — Context

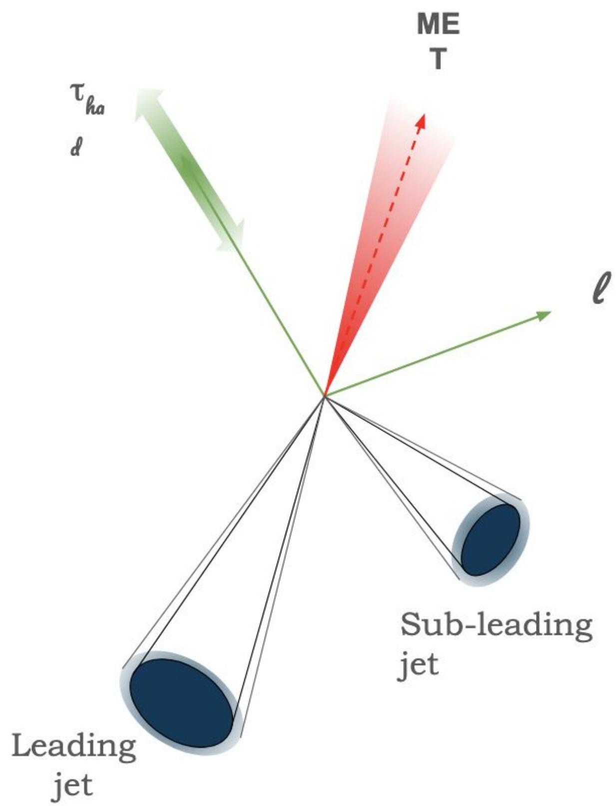
- Systematic error :
  - Error that is repeated the same way at each measurement
  - Acts as a bias
  - Predictable

- Causes :
  - Calibration of measure instruments
  - Experimental conditions
  - Small inaccuracy in the theoretical model



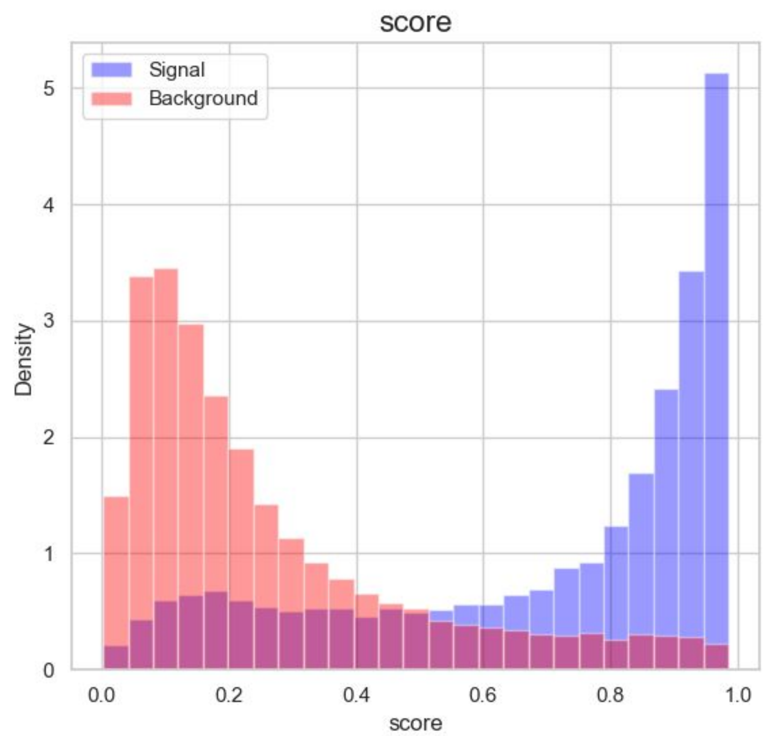
5 Systematics — Context

- ❖ Two parameters taken into account:
  - ➢ Hadron (tau) energy scale (TES)
  - ➢ Jet energy scale (JES)



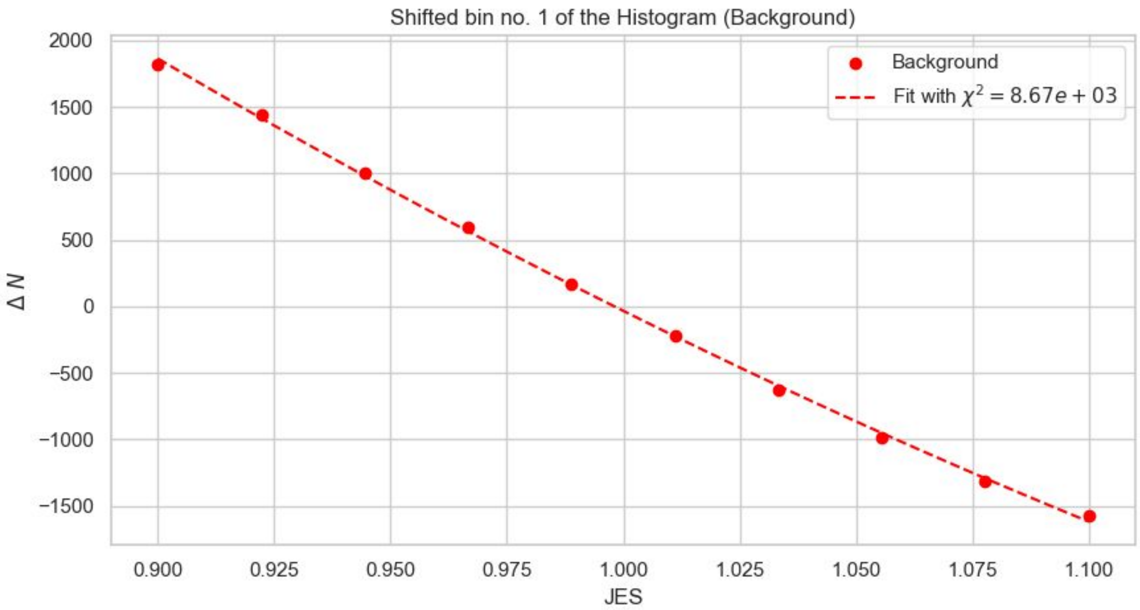
5 Systematics — Methods

- Results using BDT model
- Compute the influence of TES/JES on the score (bin by bin)

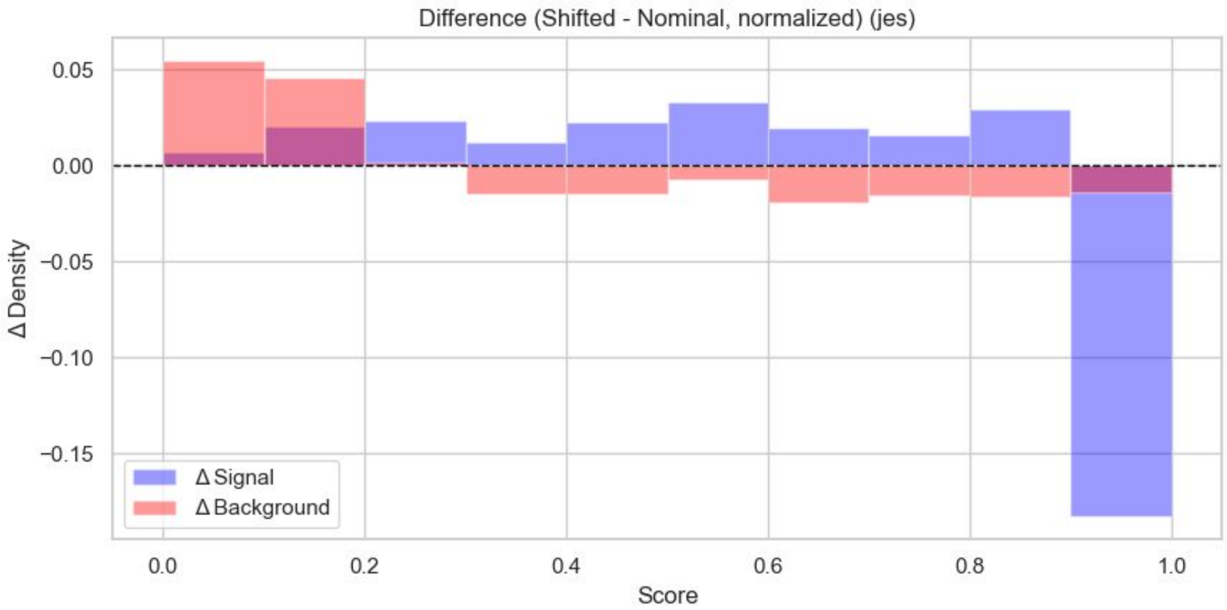


5 Systematics — Methods

- Isolate a bin to get its evolution of ΔN with respect to TES/JES
- Fit this function for all bins (polynomial of degree 1 or 2) (*For statistical team*)



5 Systematics — Methods



- ❖ Compute the influence of TES/JES on the score (bin by bin)
- ❖ Get the histogram of ΔN (variation of the number of events) with respect to the score for a chosen value for TES/JES

5 Systematics — Our functions

- Inputs
  - Model : BDT or NN
  - Train set with weights and labels
  - Number of bins (10 by default)

- Outputs
  - The fitting polynomial for each bin (for signal / for background)
  - Two functions (for signal / for background)
    - Inputs : best value of Tes/Jes (from Statistical team) and fitting polynomials
    - Output : histogram with score and shifted score

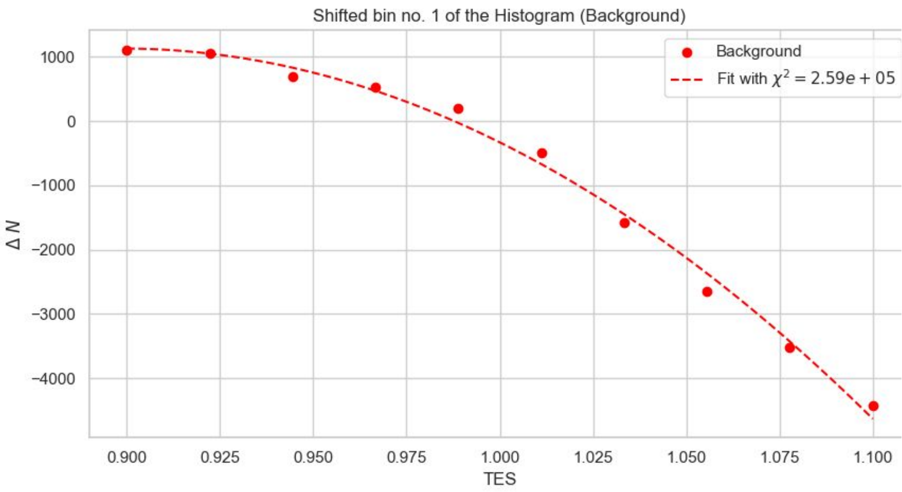
5 Systematics — Main issues

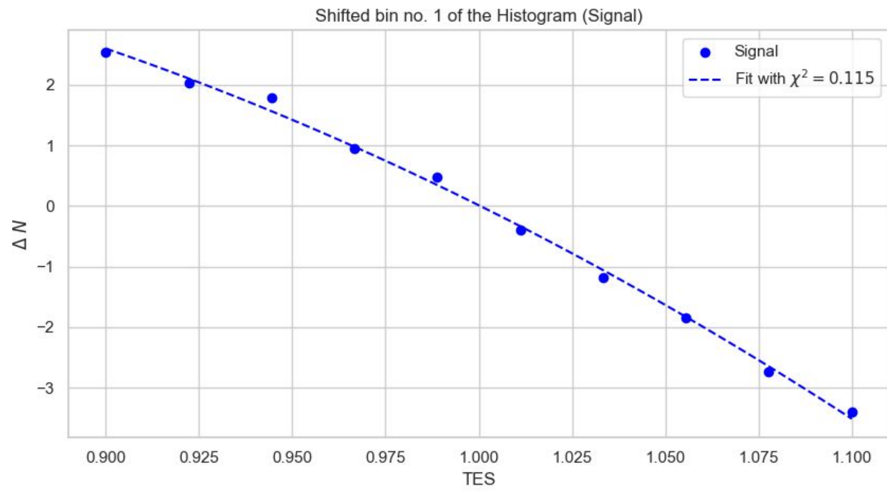
- Understanding our objectives and how the code worked (aka the fact that we could not run our code independently)
  - Solution : Systematic not to be run on its own, to be called from model ; run it in command line or from notebook

- Unidentified shift in the computation of ΔN with respect to TES
  - Solution : An error in our code put all weights to one and caused the scatter plots to be wrong

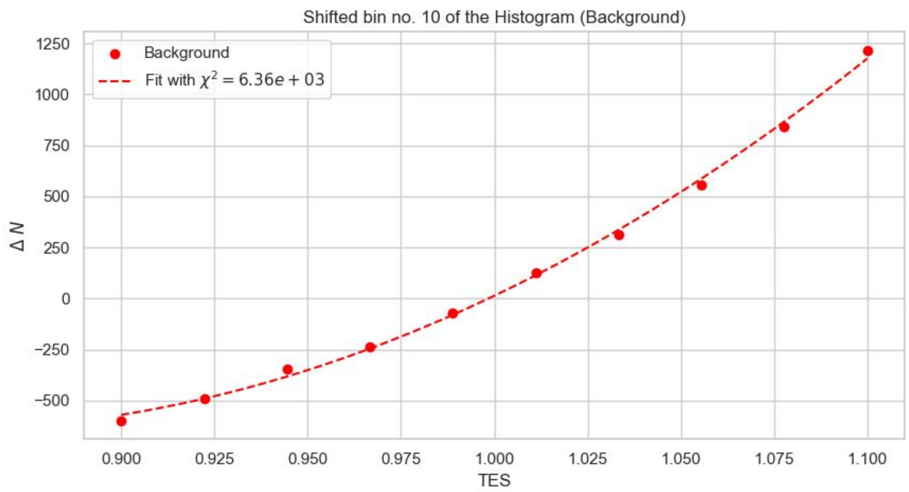
5 Systematics — Former Results

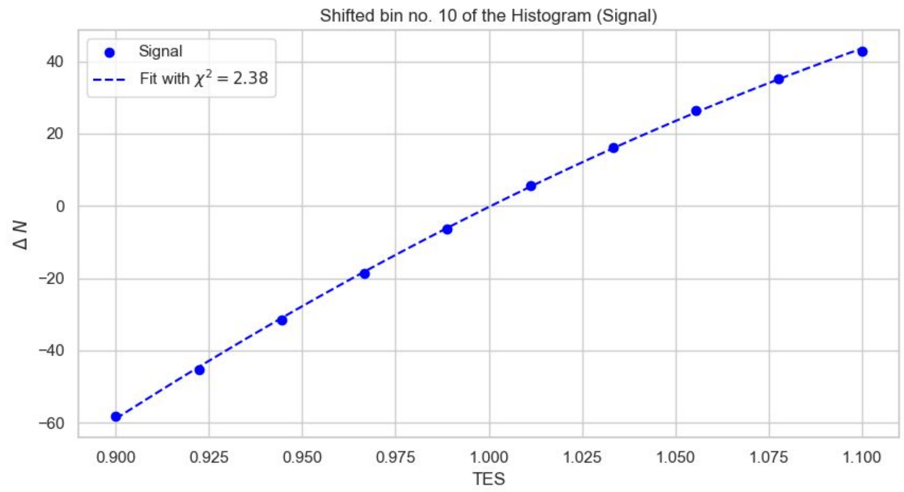
5 Systematics — Results (tes)



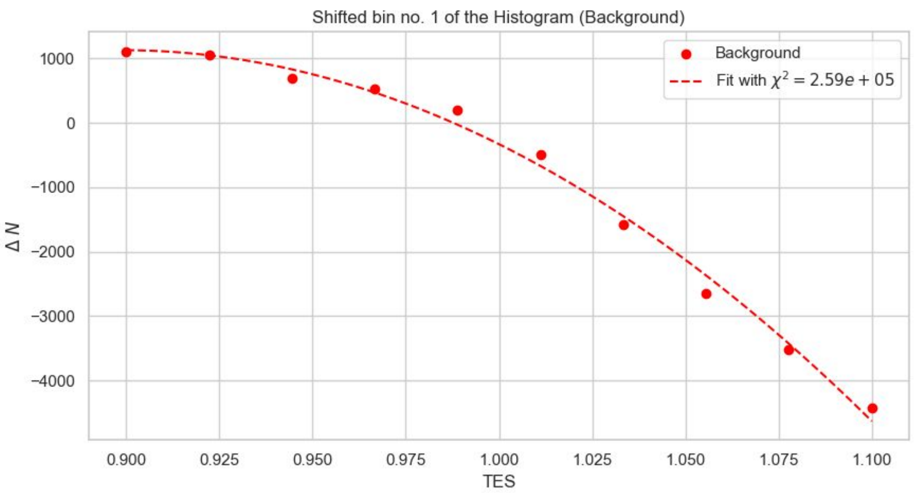


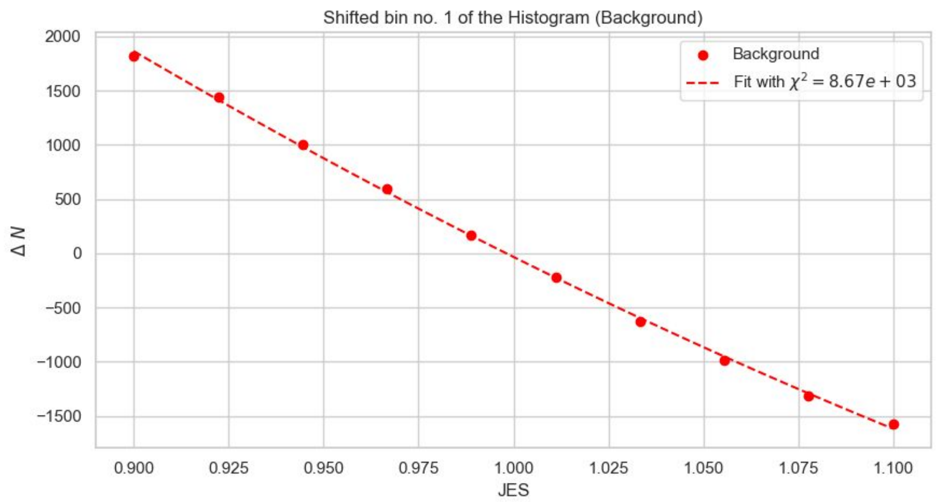
5 Systematics — Results (tes)



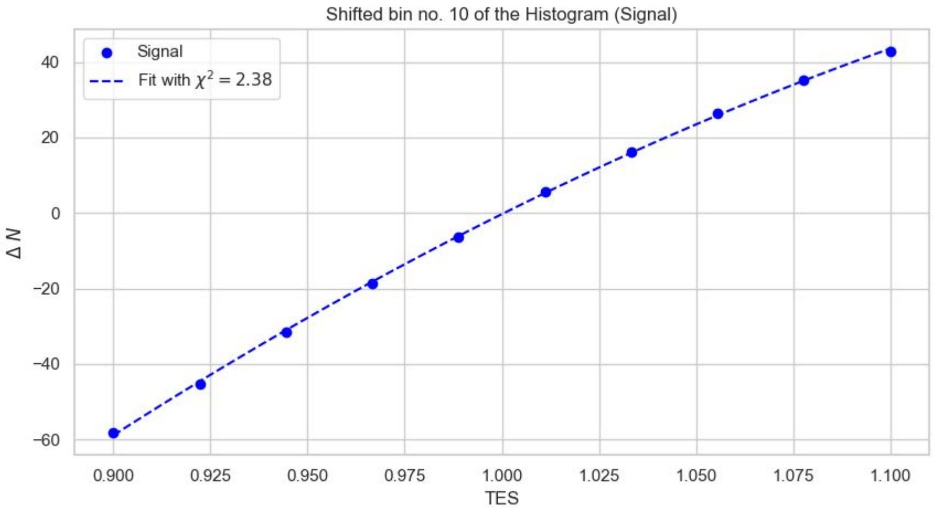


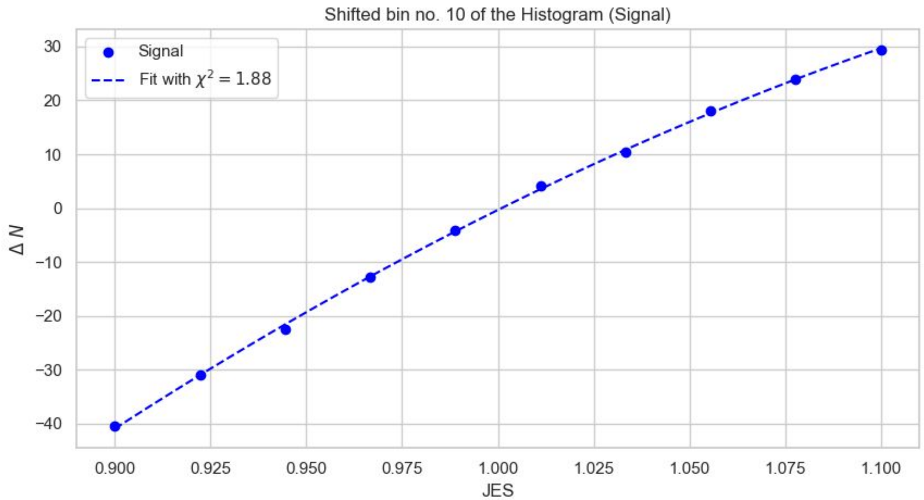
5 Systematics — Results (tes vs jes)



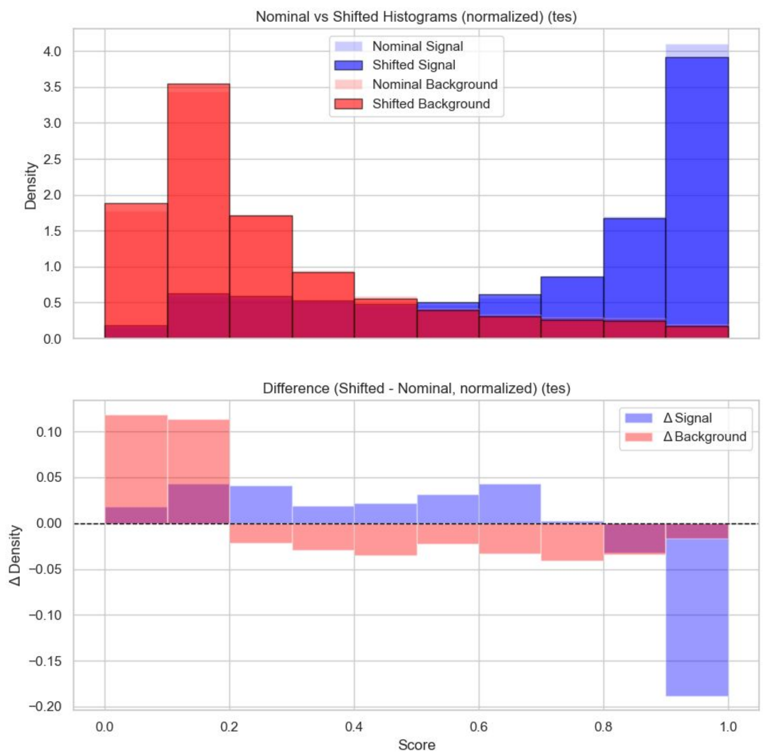


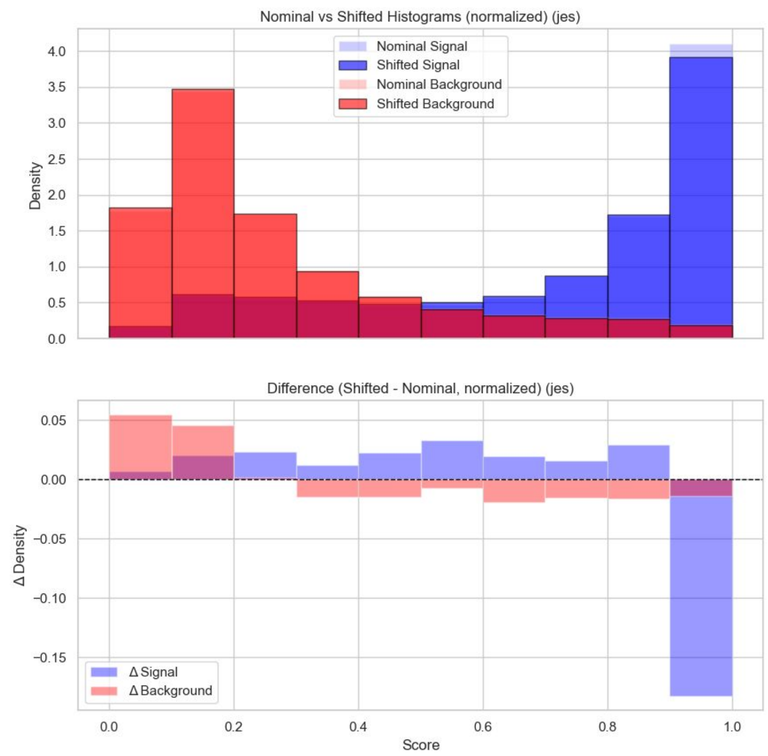
5 Systematics — Results (tes vs jes)



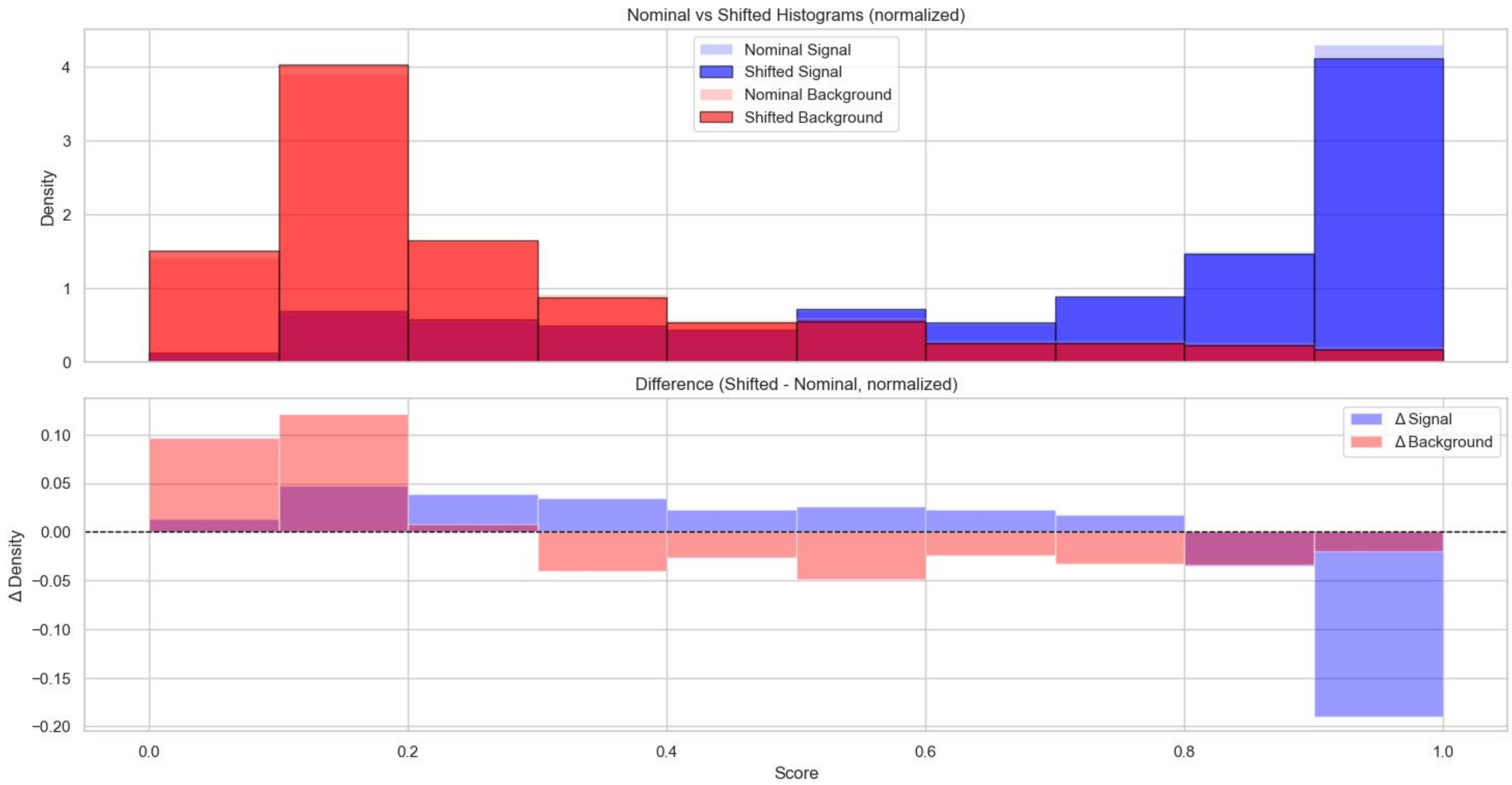


5 Systematics — Results (tes & jes)

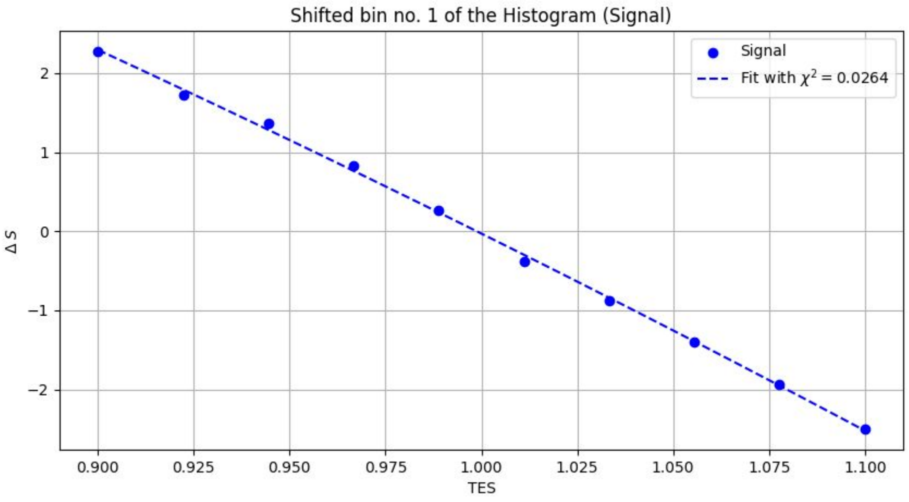




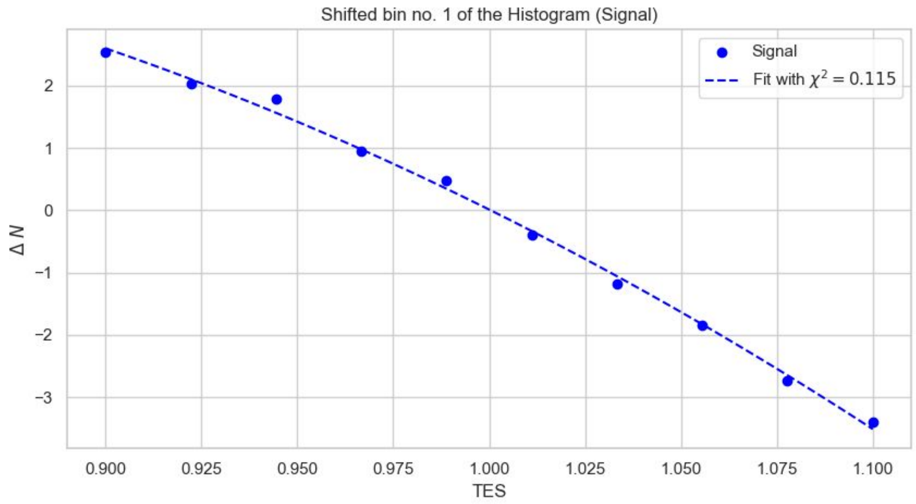
5 Systematics — Results NN (tes)



5 Systematics — Results (NN - BDT)



NN

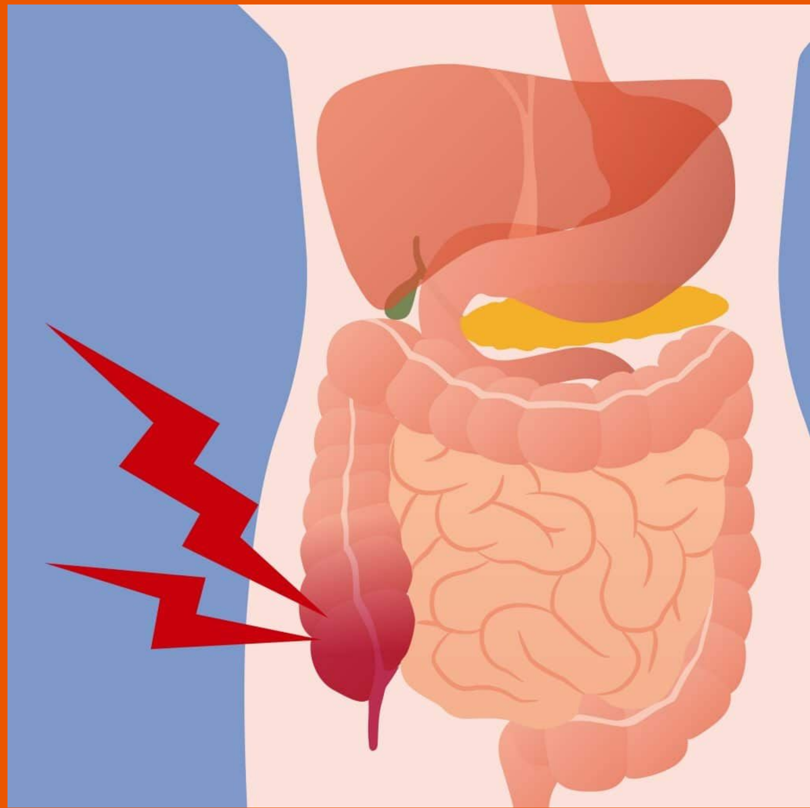


BDT

Conclusion

Thank you for your attention !

Appendix



0 [Group name]