PhD days

Clément LOTTEAU - IP2I - ALICE group

Measurement of beauty production in proton-proton and Pb-Pb collisions with the ALICE experiment at the CERN LHC

Overview

PART I - My PhD subject

1. The ALICE experiment
2. Probing the quark and gluon plasma with heavy-flavour quarks
3. What is a jet?
4. PhD subject : a quick explanation
5. Jet Tagging - Track Counting algorithm vs. Boosted Decision Trees

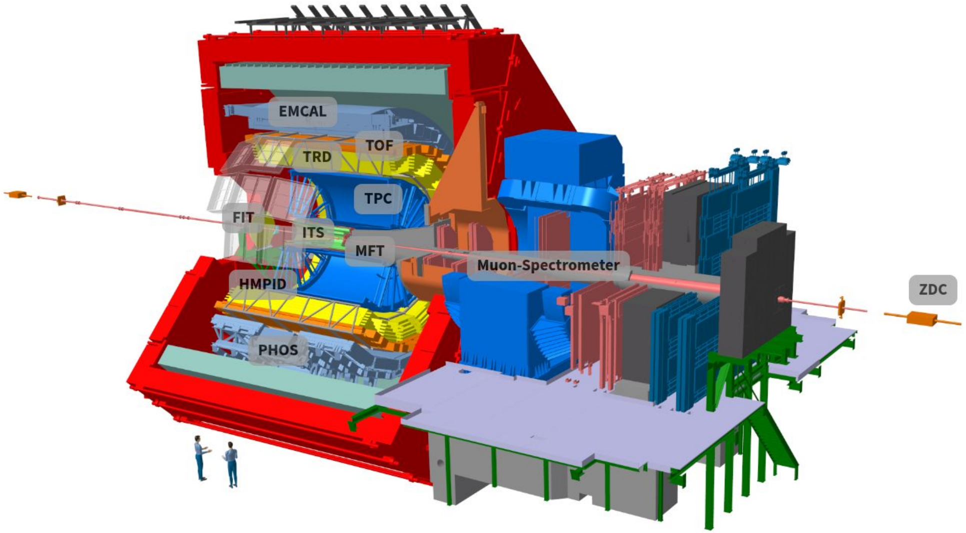
PART II - My service work for the ALICE collaboration

The ALICE experiment

Designed to study the physical properties of **strongly interacting matter** at extremely high temperature and energy densities reached in heavy-ion collisions at which a **Quark and Gluon Plasma (QGP)** is formed

Study **QGP** to better understand:

- Confinement
- Parton energy loss in the presence of free color charges
- Formation of hadronic bound states
- Restoration of chiral symmetry
- Primordial universe and compact objects like neutron stars



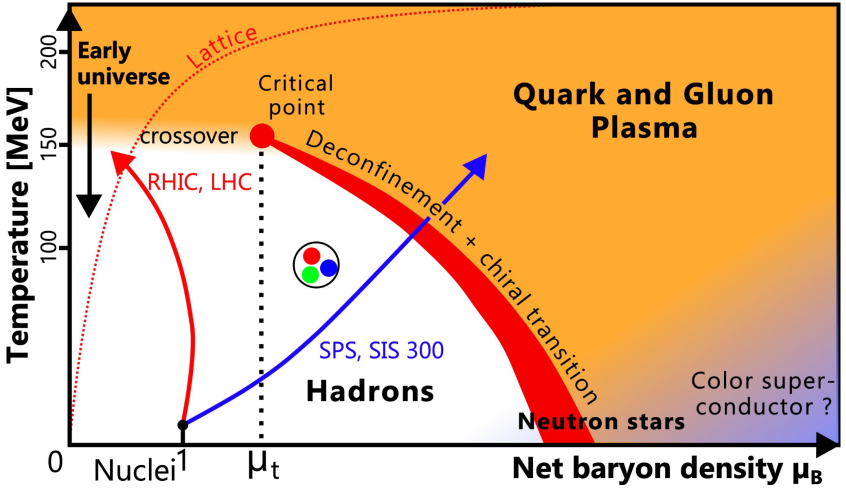
The ALICE detector

Probing QGP with heavy-flavor (HF) quarks

HF quarks (c,b) produced in hard scatterings at initial collision stages, before the formation of QGP
→ experience the entire QGP evolution

Energy-loss effects resulting from interaction with QGP constituents

b-quarks are more sensitive probes than c-quarks due to higher mass (less thermalized, radiative loss suppression…)



Phase diagram of QCD

What is a jet ?

High-energy partons (quarks of gluons) radiate gluons in the direction of their propagation, they become quark-antiquark pairs which radiate gluons…etc.

→ This process is called «**showering**» and is followed by **hadronization.**

The newly formed hadrons propagate approximately in a **cone** aligned with the direction of the primordial parton. This object is called a **hadronic jet.**

The **flavor** attributed to the jet is the flavor of the initial parton : light flavor, charm or beauty (lf,c,b)

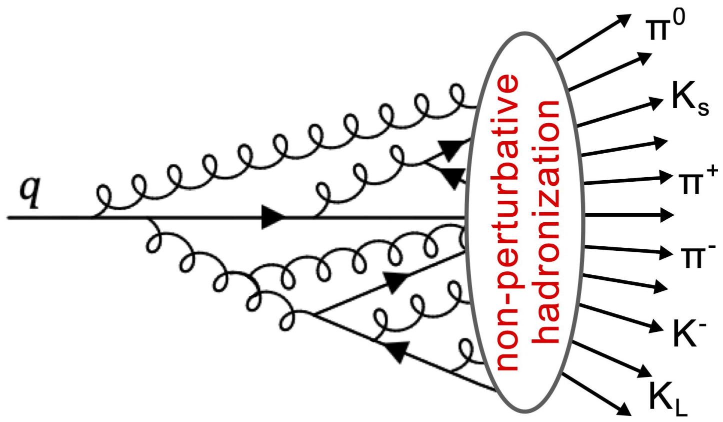


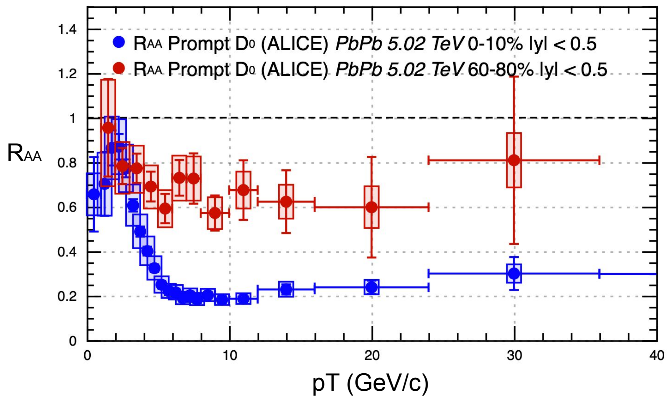
Illustration of jet production

PhD subject - quick explanation (1/2)

Measurement of b-jet production in pp and Pb-Pb collisions with **Run3** data → **Nuclear modification in Pb-Pb with respect to pp reference**

Advantages of studying **b-jet production** :

- Jet pT directly related to b quark pT
- Jet substructure provides unique insight into the energy loss mechanisms



Nuclear modification factor for prompt D0 production for central and peripheral collisions

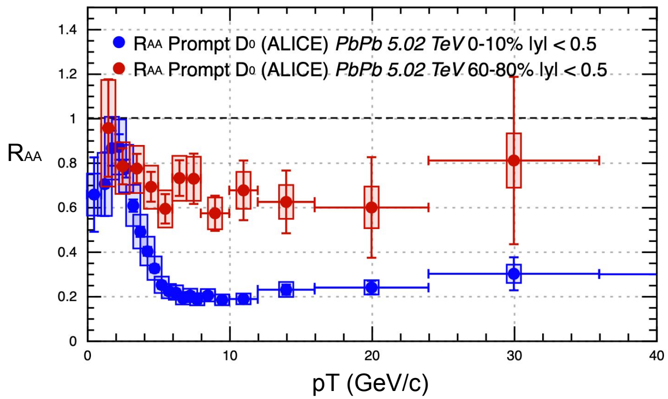
PhD subject - quick explanation (2/2)

Measurement of b-jets consists of :

1. Jet reconstruction
2. b-jet tagging/identification (separation from charm and light-flavour jets)
3. Proper treatment of huge background in Pb-Pb, especially at lower jet pT

Upgraded ALICE detector in **Run3** :

- Orders of magnitude higher statistics for b physics with respect to Run 1-2
- New Inner Tracking System → significant improvement of track impact parameter resolution → crucial for b-quark measurement



Nuclear modification factor for prompt D0 production for central and peripheral collisions

Jet Tagging - Tools for performance evaluation

Study in different jet pT intervals :

[5-10], [10-20], [20-40], [40-70], [70-120], [120-200] (GeV/c)

Two quantities to **evaluate the performances** of the tagging algorithms

$$\text{Efficiency} = \frac{\text{Number of selected b-jets}}{\text{Total number of b-jets}}$$

$$\text{Purity} = \frac{\text{Number of selected b-jets}}{\text{Total number of selected jets (b,c,lf)}}$$

Jet Tagging - Main parameters of the analysis : IP, IPs

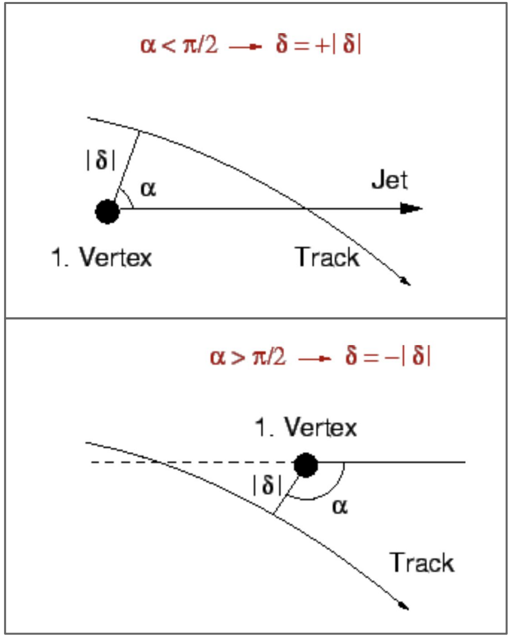
IP of track: distance of closest approach (DCA) of the track to the jet production vertex in the transverse plane of the detector (perpendicular to the beam)

IP significance (IPs): IP / σ
(σ = IP resolution)

We take into account the **sign of the IP**

In this presentation, the tagging was done **with IP**. Future analyses will be made with the IPs

Impact Parameter and its sign



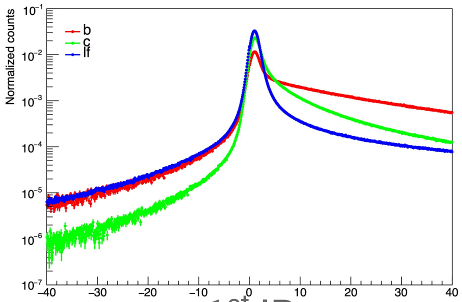
Jet Tagging - **Track counting** and BDTs

Impact Parameter significance with Track Counting

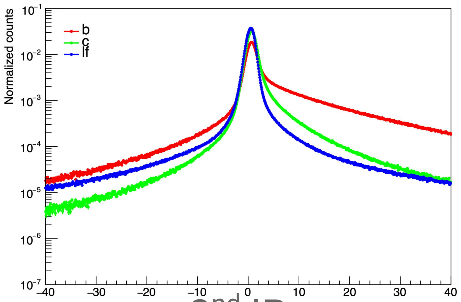
IPs of jets : **b** > **c** > **lf** → strong discriminating power

Track Counting algorithm:

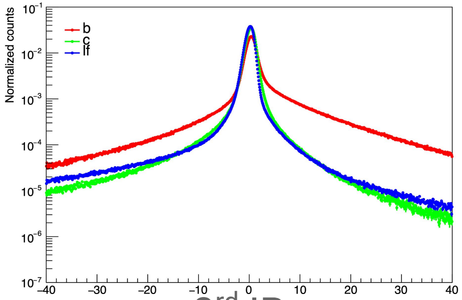
- Arranges IPs of tracks in jets in **descending order** (1st, 2nd, 3rd largest IPs)
- Jet tagged as **b** if Nth largest IPs > chosen threshold
- Nth largest IPs and threshold give different **tagging efficiency** and **background rejection**
- 3rd largest IPs has the highest **purity**



1st IPs



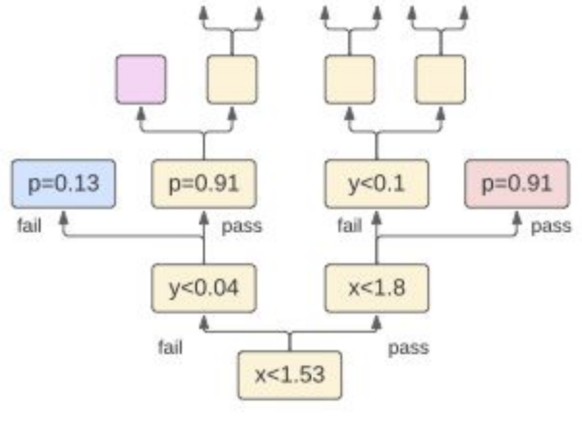
2nd IPs



3rd IPs

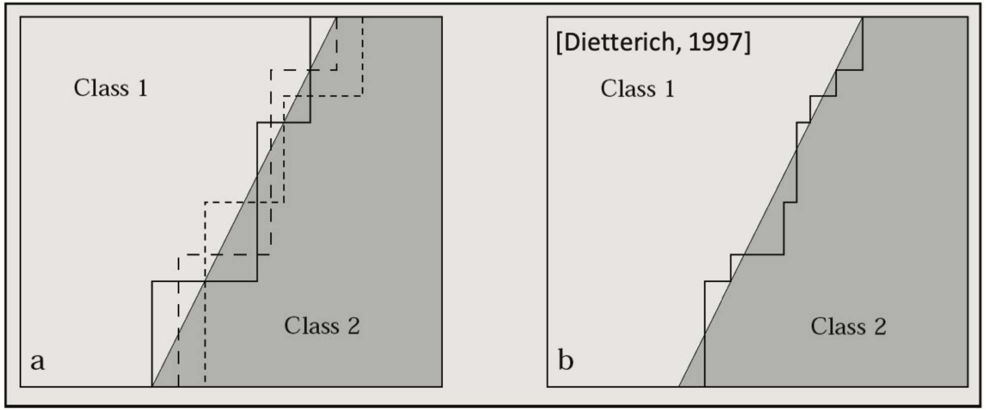
Jet Tagging - Track counting and **BDTs**

Classification with Boosted Decision Trees (BDTs)



Decision Tree

BDTs : many DTs and combination of results



Multiple trees (left) - Combination (right)

Jet Tagging - Track counting and **BDTs**

Classification with Boosted Decision Trees (BDTs)

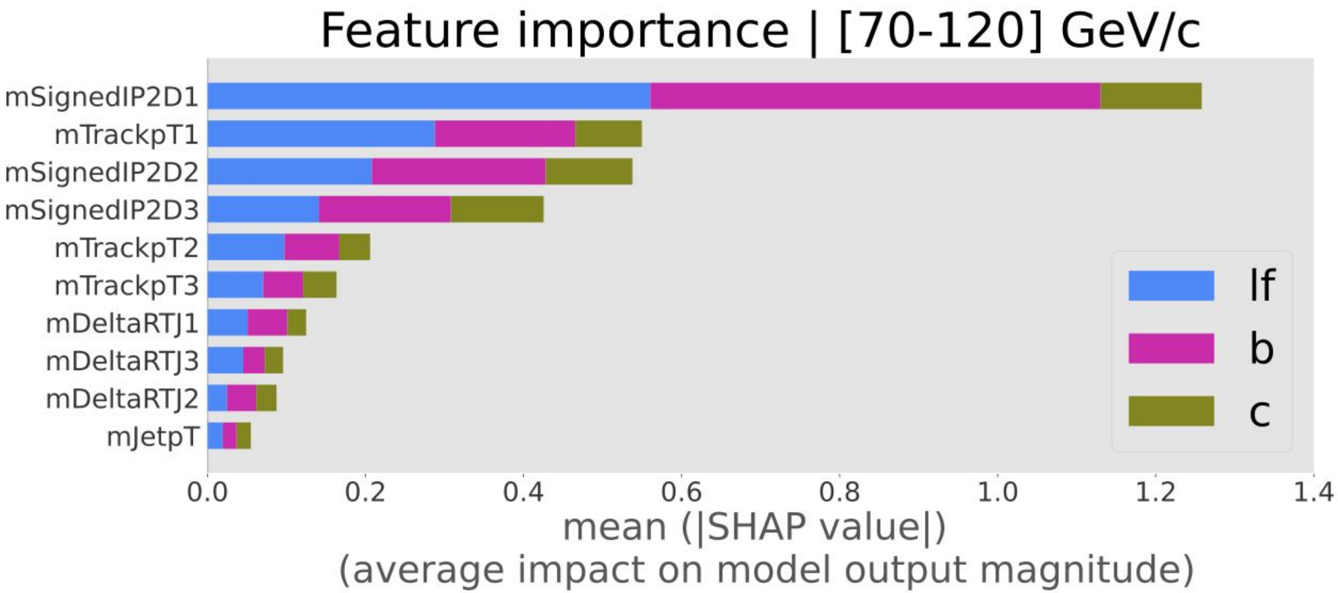
1. **Finding the right input** (importance and correlations)

2. Training and testing on Monte Carlo
   a. Separating the Monte Carlo dataset in two
   b. Training on one set
   c. Testing on the other set

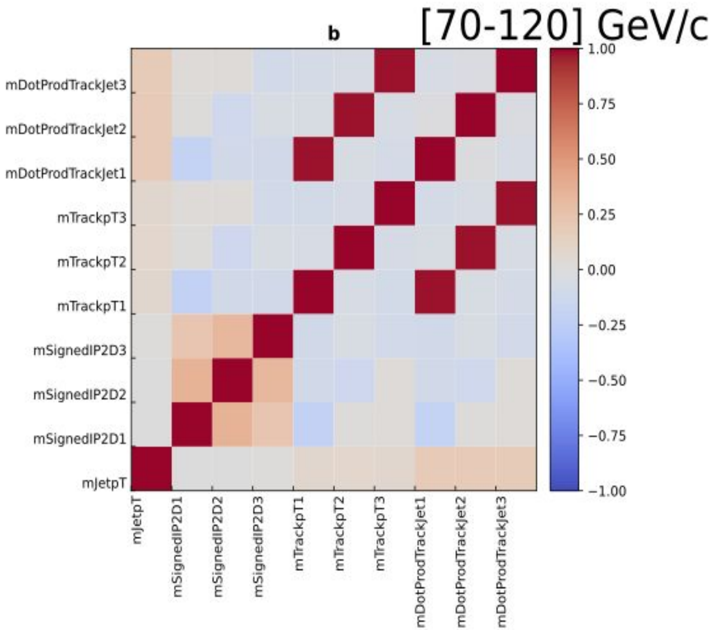
3. Choosing the score cuts based on our choices of efficiency and purity

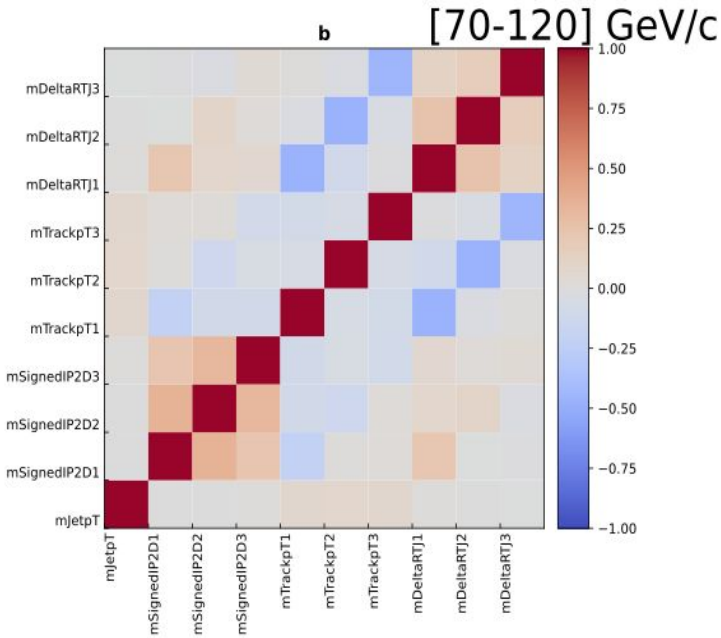
4. Applying the BDT to data when the testing is optimal

Jet Tagging - Track counting and **BDTs**



Jet Tagging - Track counting and **BDTs**





Strong correlation = red
Strong anti-correlation = bleu

Jet Tagging - Track counting and **BDTs**

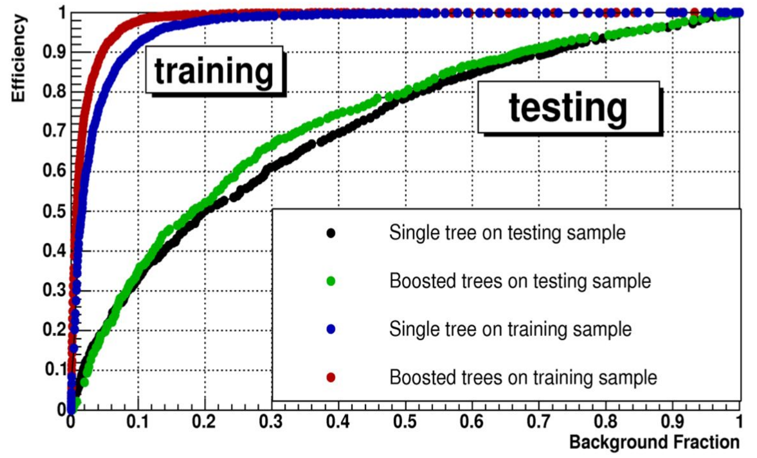
Classification with Boosted Decision Trees (BDTs)

1. Finding the right input (importance and correlations)

2. **Training and testing on Monte Carlo**
   a. **Separating the Monte Carlo dataset in two**
   b. **Training on one set**
   c. **Testing on the other set**

3. Choosing the score cuts based on our choices of efficiency and purity

4. Applying the BDT to data when the testing is optimal



Example of efficiency : DTs vs. BDTs

Jet Tagging - Track counting and **BDTs**

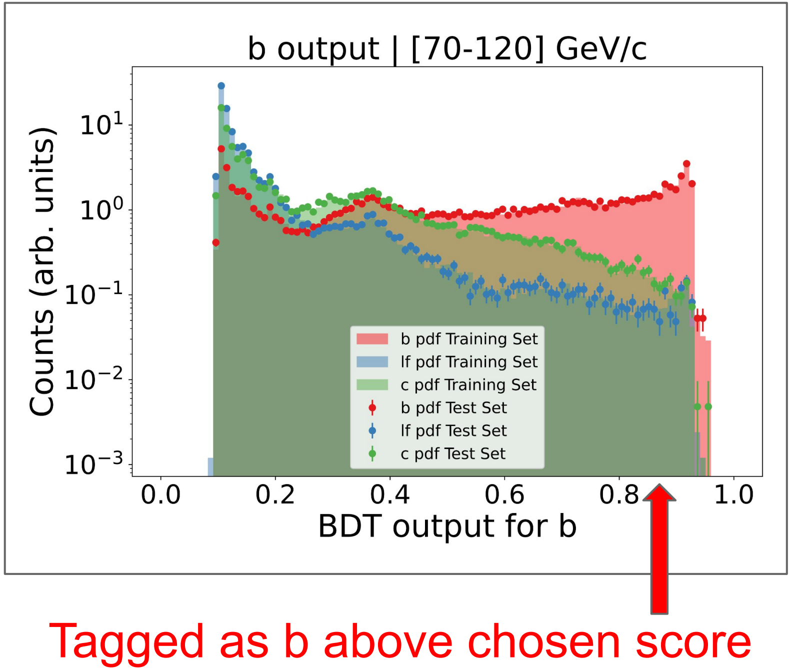
<u>**Classification with Boosted Decision Trees**</u> (BDTs)

1. **Finding the right input (importance and correlations)**

2. Training and testing on Monte Carlo
   a. Separating the Monte Carlo dataset in two
   b. Training on one set
   c. Testing on the other set

3. **Choosing the score cuts based on our choices of efficiency and purity**

4. Applying the BDT to data when the testing is optimal



Jet Tagging - Track counting and **BDTs**

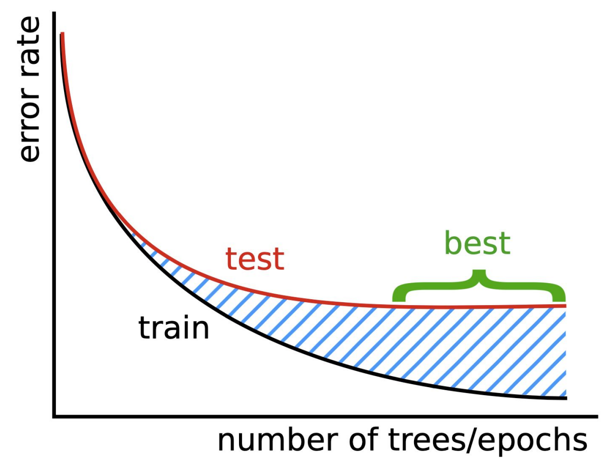
<u>Classification with Boosted Decision Trees</u> (BDTs)

1. Finding the right input (importance and correlations)

2. Training and testing on Monte Carlo
    a. Separating the Monte Carlo dataset in two
    b. Training on one set
    c. Testing on the other set

3. Choosing the score cuts based on our choices of efficiency and purity

4. **Applying the BDT to data when the testing is optimal**

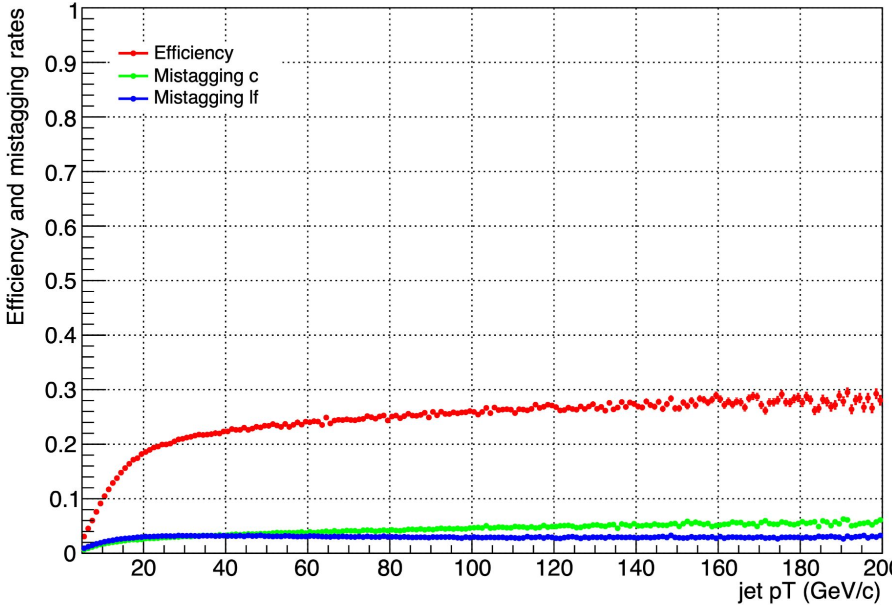


What to expect for BDTs

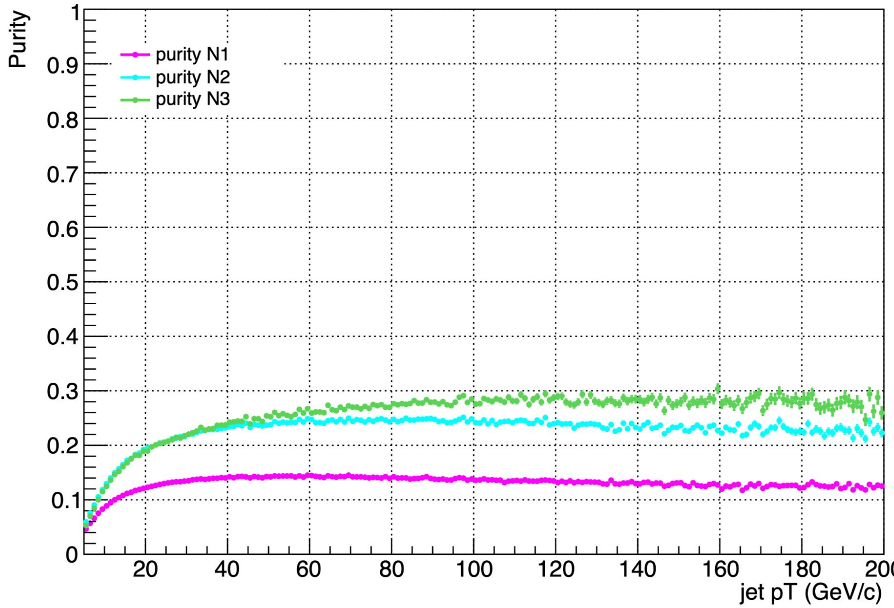
Jet Tagging - Track Counting performances with IP

Chosen threshold : 0.008cm

Efficiency with IP of the 3rd largest IPs (red)



Purity with IP of the 3rd largest IPs (green)



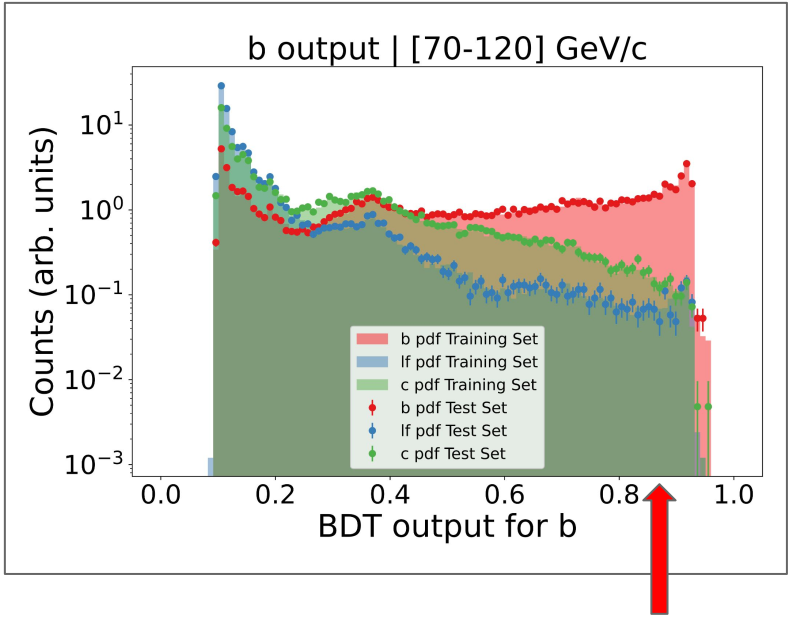
Jet Tagging - Track counting and **BDTs**

Choosing the score cuts based on our choices of efficiency and purity

(1) Choose score cut on b: (> than chosen limit)
high score cut = low efficiency, high purity

(2) Choose score cut on lf: (< than chosen limit)
low score cut = lower efficiency, higher purity

In practice: scan all the score combinations to find the cuts that match our needs the best in terms of efficiency and purity



Tagged as b above chosen score

Jet Tagging - BDTs (4 inputs) performances

First analysis with 4 inputs: jet pT, IP of 1st, 2nd and 3rd largest IPs

Choice of cuts on scores: maximizing the 3 following quantities and matching the efficiency to the Track Counting method by scanning the BDT scores

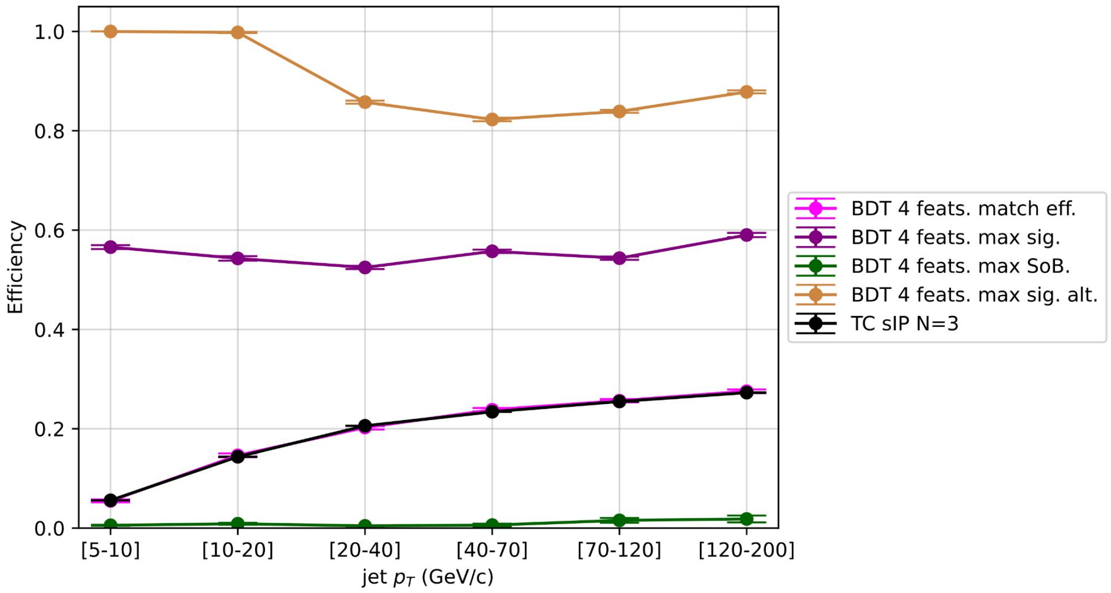
$$\text{Significance} = \frac{\text{Number of selected b-jets}}{\sqrt{\text{Total number of selected jets (b,c,lf)}}}$$

$$\text{Significance alternative} = \frac{\text{Number of selected b-jets}}{\sqrt{\text{Number of selected b- and c-jets}}}$$

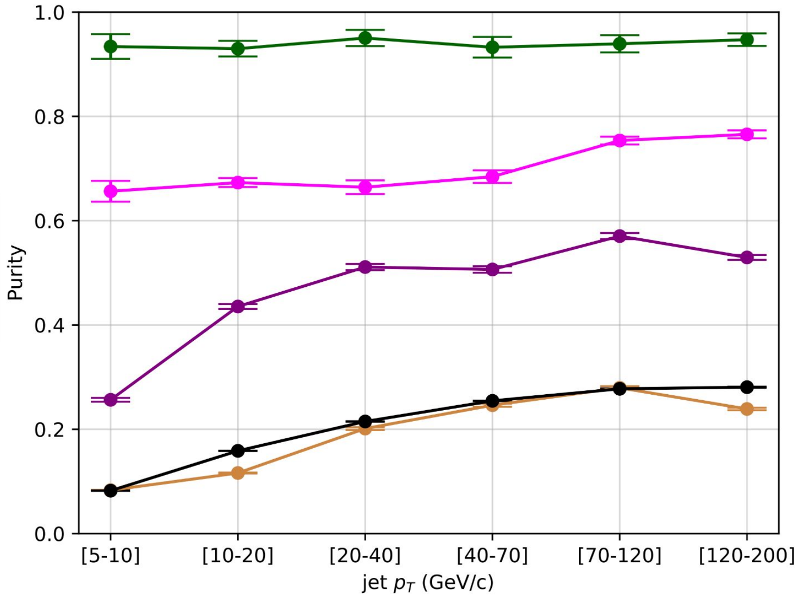
$$\text{Signal over Background} = \frac{\text{Number of selected b-jets}}{\text{Number of selected lf- and c-jets}}$$

Jet Tagging - BDTs (4 inputs) performances

Efficiency (Track Counting and efficiency matching overlap)



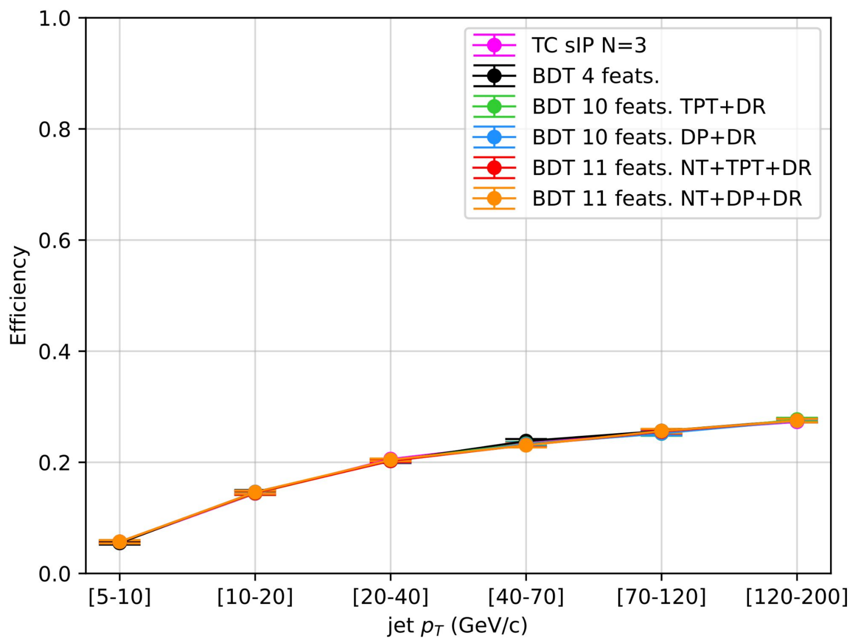
Purity



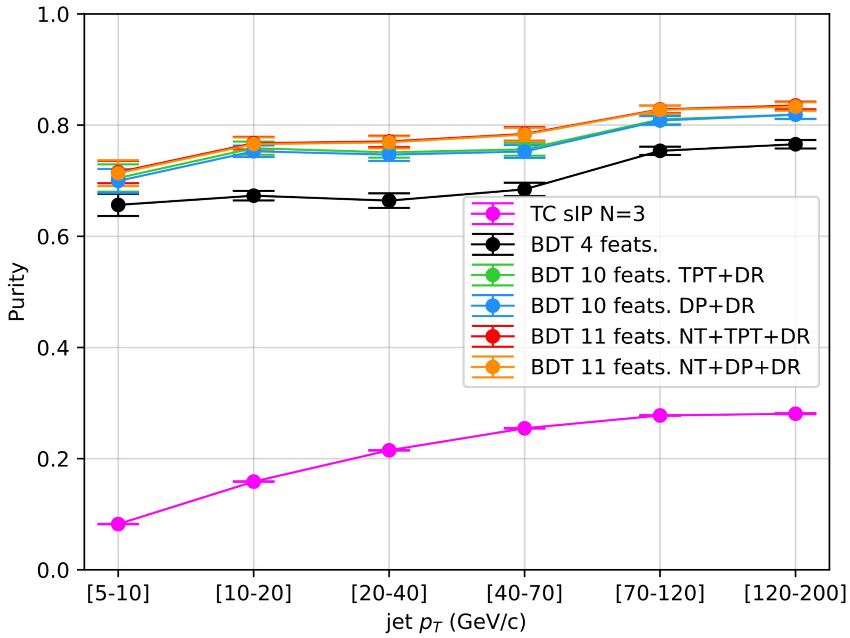
Chosen cuts on BDT scores : significance maximization and efficiency matching to Track Counting

Jet Tagging - Efficiency matching - 4, 10 and 11 inputs

Efficiency matched to Track Counting



Purity



Here we chose cuts on scores to match the efficiency obtained with the Track Counting method

Jet Tagging - Efficiency matching - 4, 10 and 11 inputs

Different combinations of inputs were tested:

(1) BDTs with 4 inputs :

jet pT + IP of 1st, 2nd, 3rd largest IPs

(2) BDTs with 10 inputs :

4 first inputs + Track pT 1,2,3 + Delta R 1,2,3

4 first inputs + Dot Product Track-Jet 1,2,3 + Delta R 1,2,3

(3) BDTs with 11 inputs :

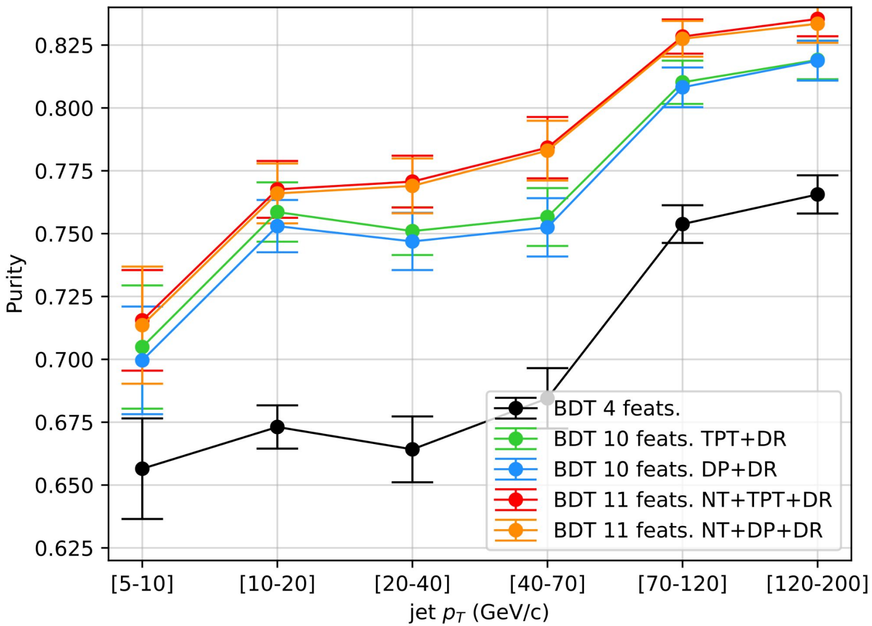
10 green inputs + Number of tracks in the jet

10 blue inputs + Number of tracks in the jet

Results:

- **similar purity** for the 2 combinations
- a **limit** on maximum purity may exist

Purity zoomed in on BDTs



All BDTs were set to the same efficiency

$$\Delta R = \sqrt{\Delta\phi^2 + \Delta\eta^2}$$: Delta R between track and jet

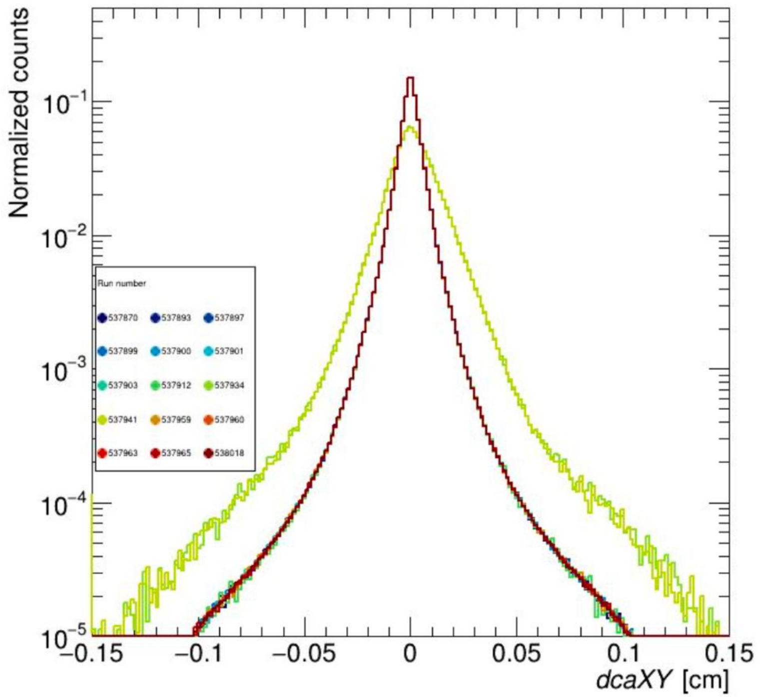
My service work for the ALICE collaboration

Service work - Track Properties and Tracking QC

Quality Checks of datasets, comparison between periods and between dataset and Monte Carlo anchored to it

Personal contribution to the analysis code

Wrote a full tutorial for my peers who took over from me



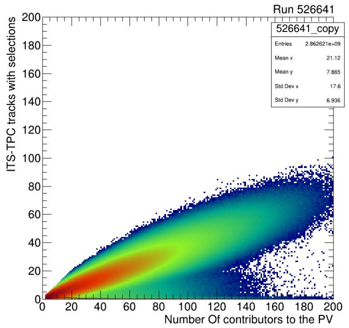
Example of problematic Monte Carlo DCA XY

Service work - Track Properties and Tracking QC

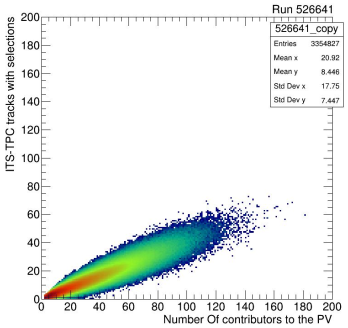
Quality Checks of datasets, comparison between periods and between dataset and Monte Carlo anchored to it

Personal contribution to the analysis code

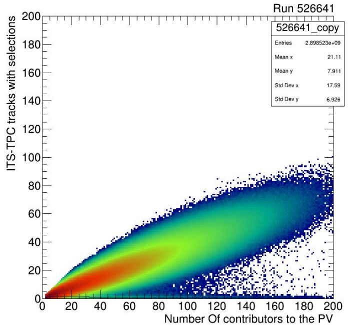
Wrote a full tutorial for my peers who took over from me



Without new cut



Anchored MC



With new cut

Number of contributors to the PV
vs.
Tracks with selection

Service work - Track Properties and Tracking QC

Quality Checks of datasets, comparison between periods and between dataset and Monte Carlo anchored to it

Personal contribution to the analysis code

Wrote a full tutorial for my peers who took over from me

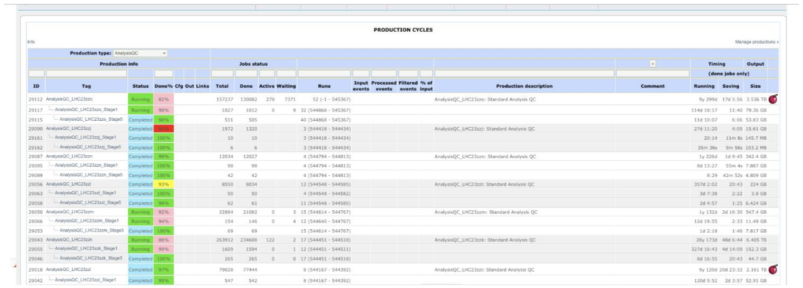


Figure 18: AnalysisQC page on MonALISA

We will use this path as <input_dir> : /alice/data/2023/LHC23zzh/. For <recopass> (reconstruction) we will use 'cpass2' in this example and for <analysis_name> we choose 'EventTrackQA'. Now we write (on a single line) the following command in the terminal :

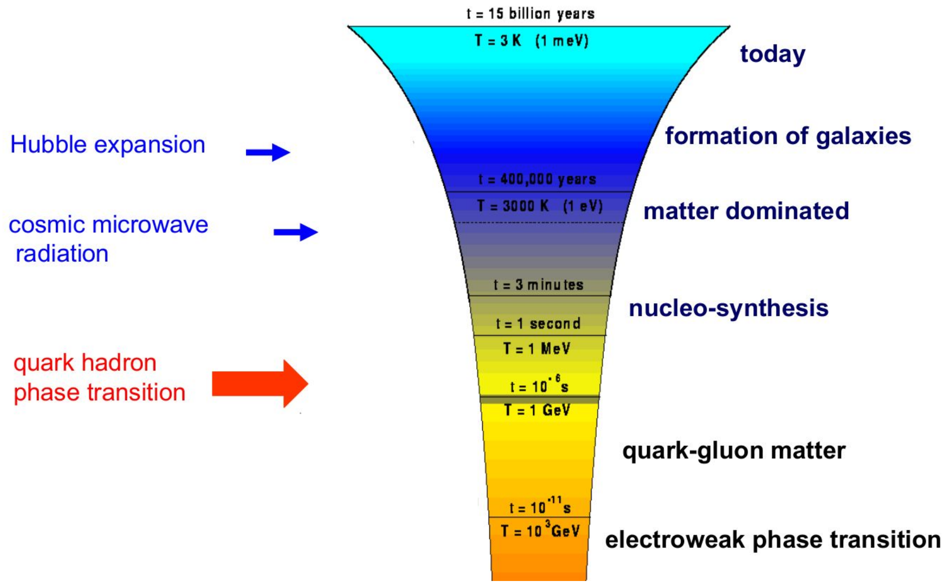
```
./download_files.py /alice/data/2023/LHC23zzh/ cpass2 EventTrackQA
localcopy/alice/data/2023/LHC23zzh_cpass2
```

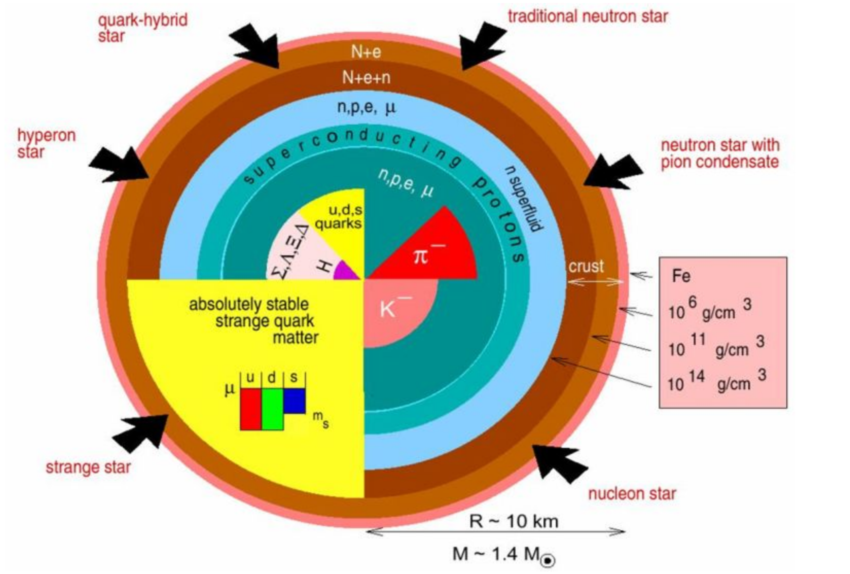
Picture of the tutorial

Thank you for listening

BACK UP

Primordial Universe and Neutron Stars

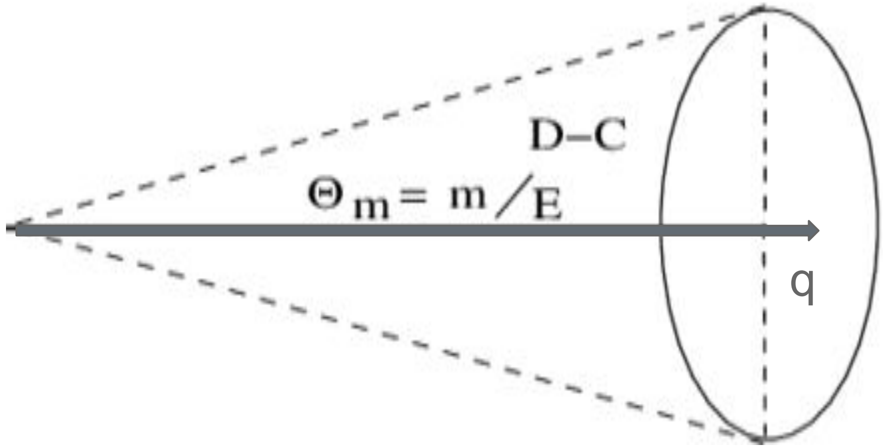




Dead Cone effect

Radiative energy loss suppressed in the “Dead Cone”

Stronger with b quarks than with c quarks because of larger masse



Jet algorithms

Jet reconstruction algorithms are either “sequential clustering algorithm” (SCA) or “Cone Algorithms”

Used anti-kT algorithm to reconstruct jets (SCA) which is IRC safe

IRC safe: InfraRed and Collinear - guarantees that the measured jet can be linked to a theoretical observable

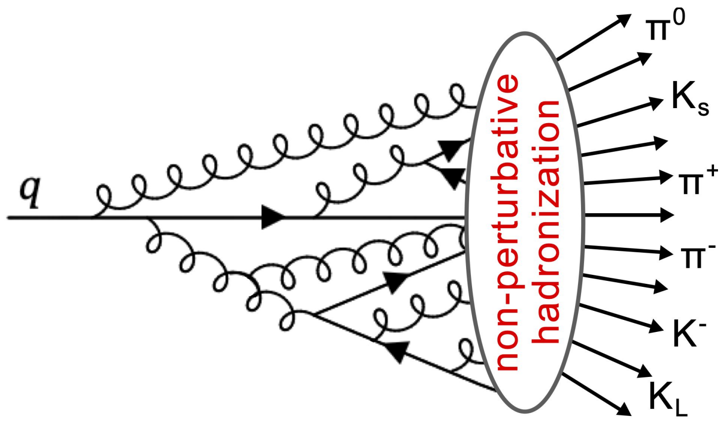


Illustration of jet production

Boosting (combination of Decision Trees)

(1) Train tree T1 on N events

(2) Train second tree T2 on new N' events, half of which was misclassified by T1

(3) Build third tree T3 on events where T1 and T2 disagree

(4) The boosted classifier takes the majority vote from (T1,T2,T3)

BDTs - Python Packages

Matplotlib to produce plots (version 3.9.2)

Pandas for data analysis and data manipulation (version 2.2.3)

NumPy for scientific computing (version 1.24.4)

Hipe4ml for link between **ROOT** and **Python** : TTree manipulation in Python, handling BDT models and visualization (like correlation plots) (version 0.0.15)

Scikit-Learn for creation of the classifier (version 1.3.0)

XGBoost for gradient boosting (version 1.7.6)

Optuna to optimize the hyper-parameters of the model (version 4.1.0)

Hipe4ml_converter to convert the BDT model to **ONNX** format (version 0.0.7)

BDTs - hyper-parameters optimization with Optuna

Hyper-parameters optimized with the **Optuna** package:

- max_depth: maximum depth of a tree
- learning_rate: step size of the gradient descent
- n_estimators: number of trees
- min_child_weight: minimum sum of instance weight needed in a child
- subsample: subsample ratio for the training process
- colsample_bytree: subsample ratio of columns when constructing each tree

Non optimized hyper-parameters:

- n_jobs: number of parallel threads used to run XGBoost
- tree_method: exact, approx or hist (hist was chosen). Specifies which tree method to use. hist is a fast approximated solution

Jet Tagging - MC studied and cuts

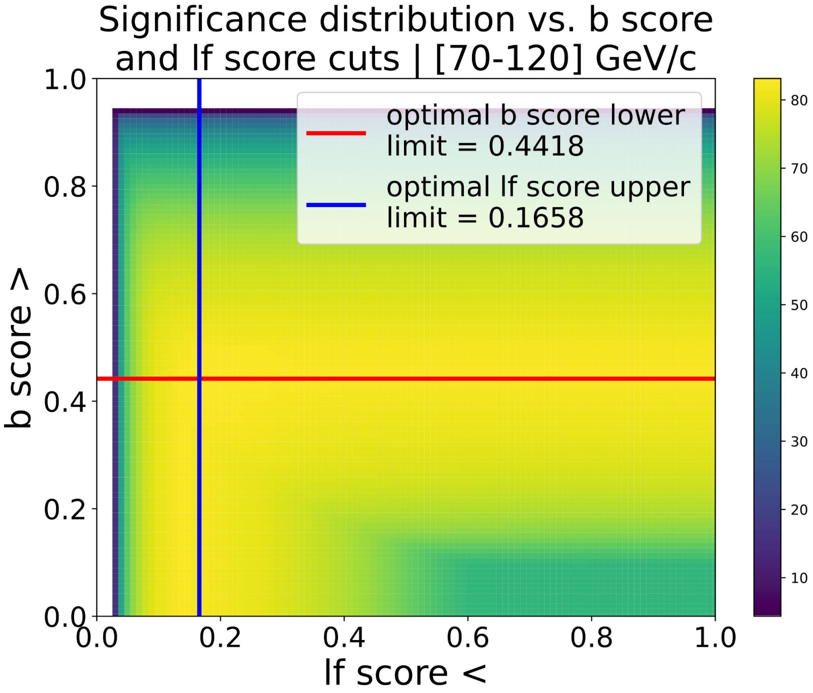
MC dataset: proton-proton @ 13.6 TeV jet-jet oversampled

Cuts on MC data:

- jet $|\eta| < 0.9$
- $5 <$ jet pT < 200 GeV/c
- jet radii = 0.4
- track $|\eta| < 0.9$
- $0.5 <$ track pT < 200 GeV/c

Jet Tagging - Scanning cuts on scores - example

Example of a full scan on b and lf scores to maximize the significance



Jet Tagging - All BDTs with 10 inputs

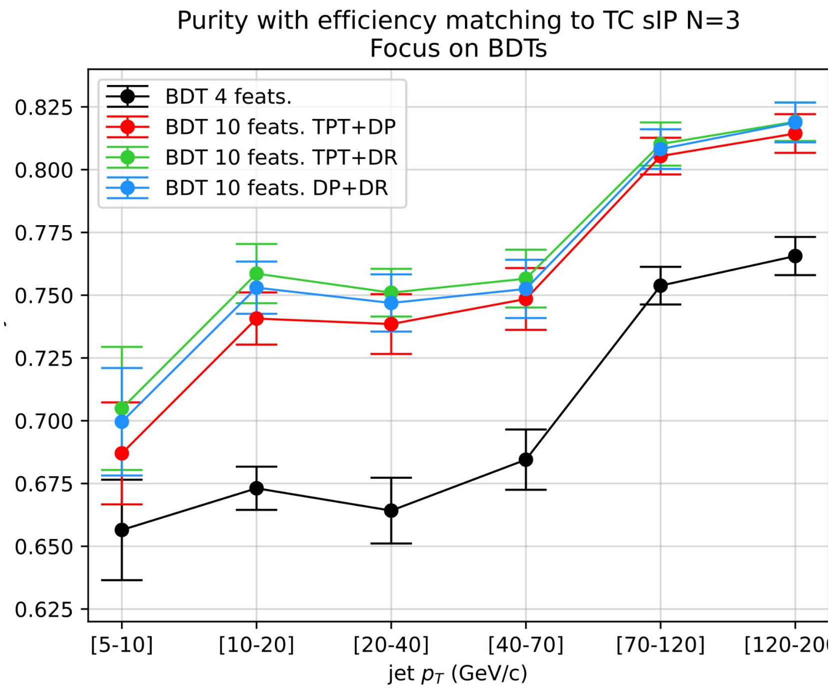
<u>All 10 input configurations :</u>

Track pT + Dot Product Track-Jet

Track pT + Delta R

Dot Product Track-Jet + Delta R

The combination TPT + DP gives lower purity than the other combinations



$\Delta R = \sqrt{\Delta\phi^2 + \Delta\eta^2}$: Delta R between track and jet

Error bars

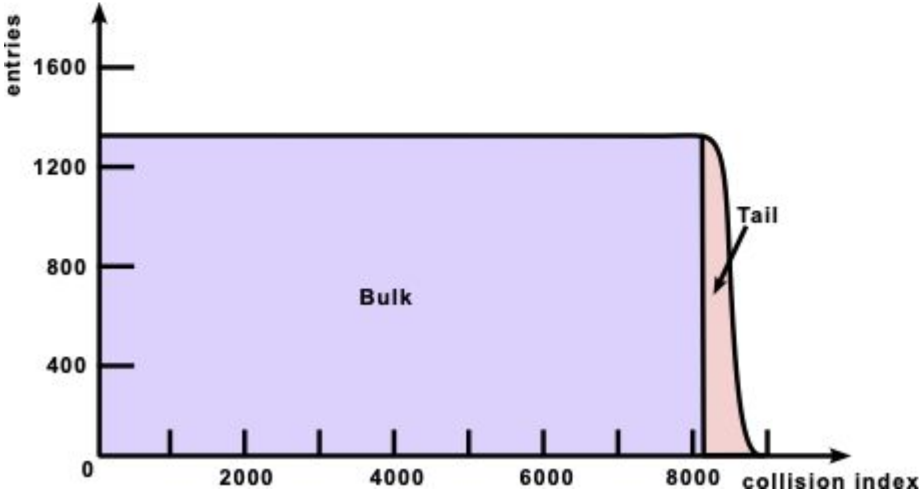
For Track Counting: Statistical error bars

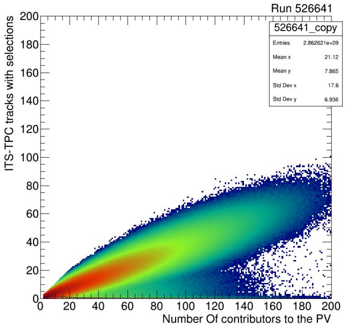
For BDTs:

- 110,000 jets for each flavor in each jet pT bin
- 80% for training, 20% for testing
- Error bars obtained by training-testing 20 times on shuffled sets of events randomly drawn from a larger pool (> 110,000 jets)

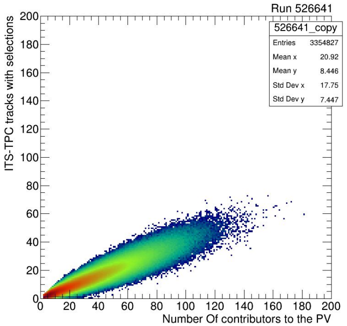
Timeframe cut

The drop in entries in the tail creates a horizontal population in the plots on the right. The added cut removes the tail and keeps the bulk

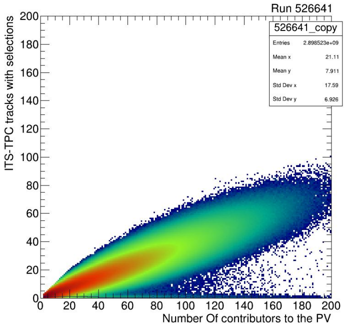




Without new cut



Anchored MC



With new cut

Number of contributors to the PV vs. Tracks with selection

End of BACK UP