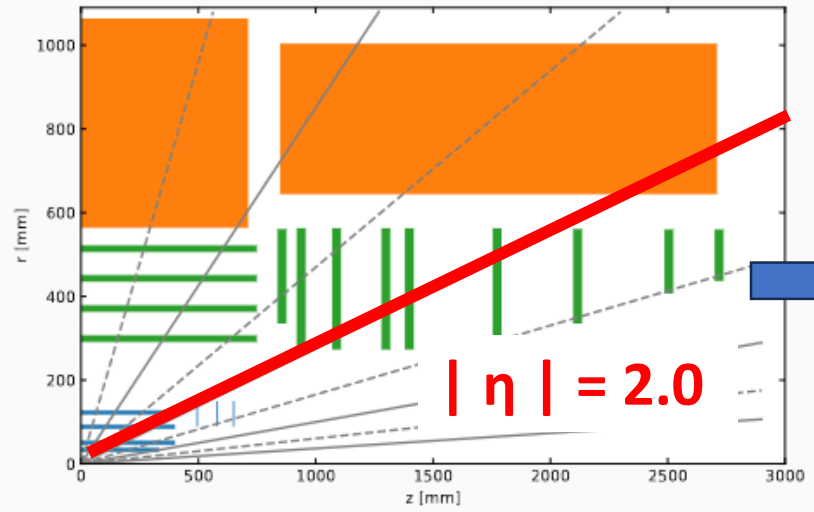
Quark gluon tag: First look in HL-LHC

Florencia Castillo

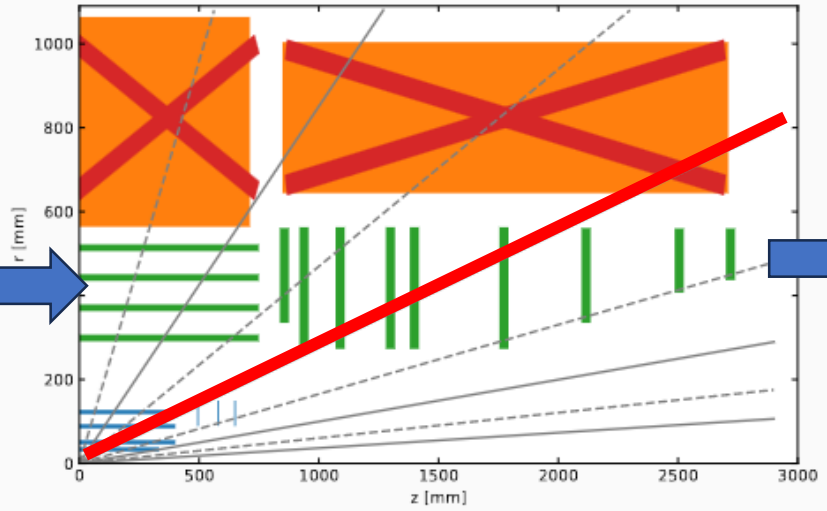
13/02/2024

Motivation: Tracker extended coverage

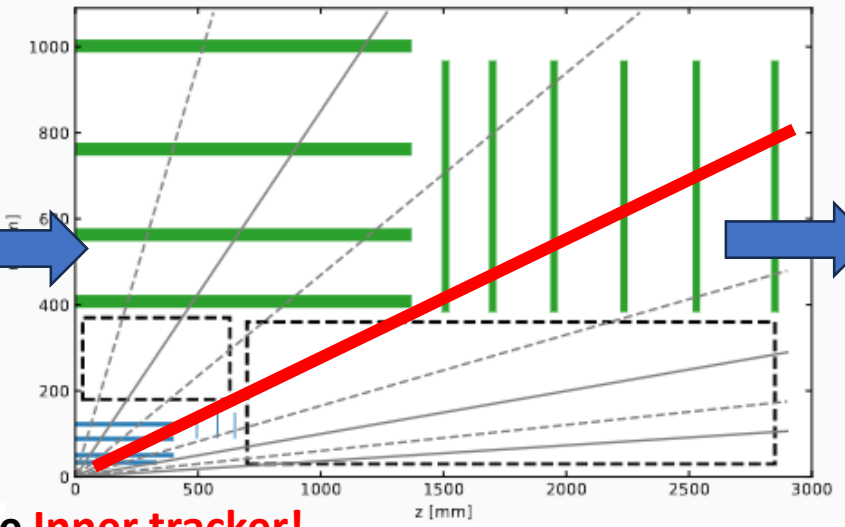
Inner detector



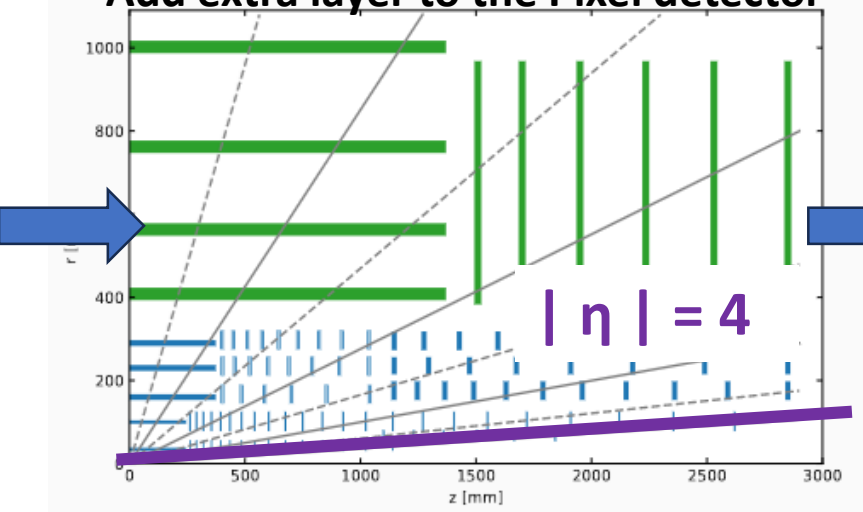
Drop the Transition Radiation Tracker (TRT)



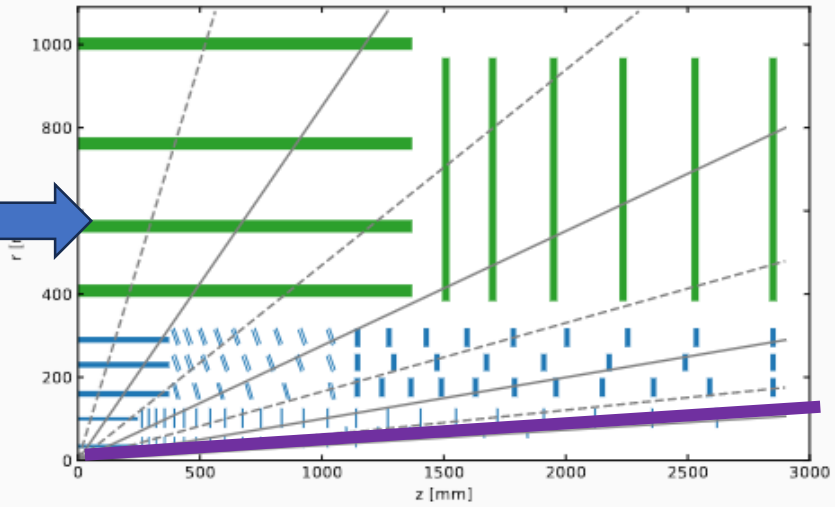
New Strip detector with larger radius

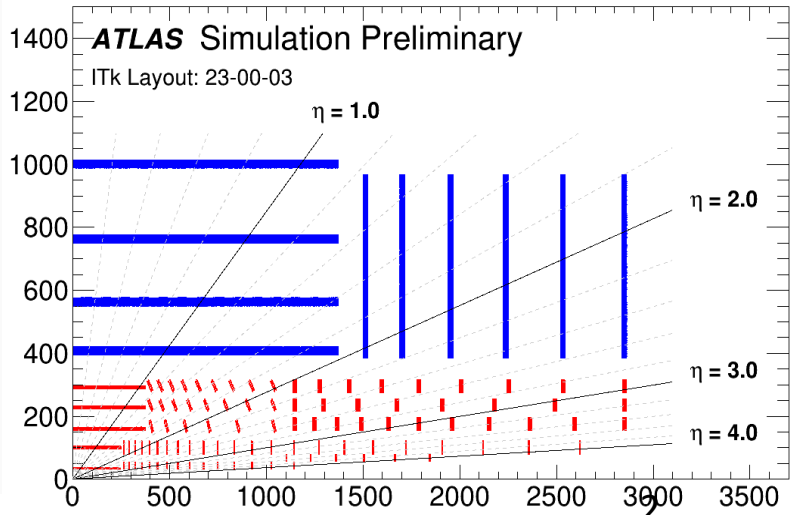


Add extra layer to the Pixel detector



Incline the modules! Now we have the Inner tracker!

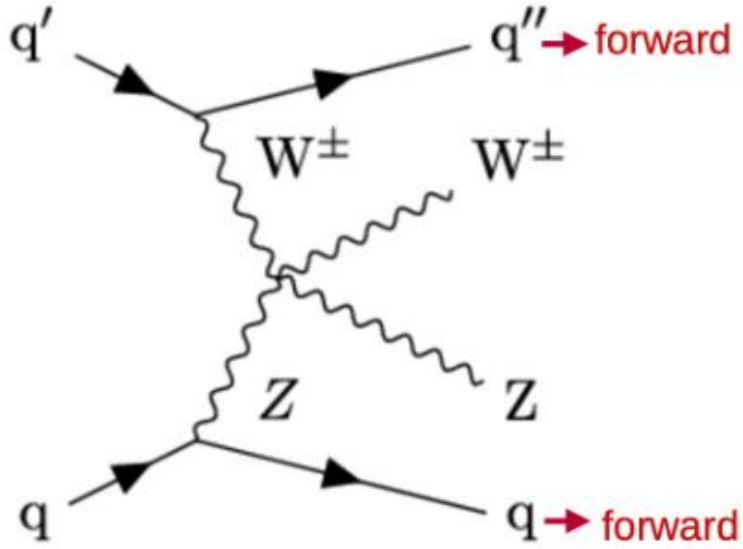




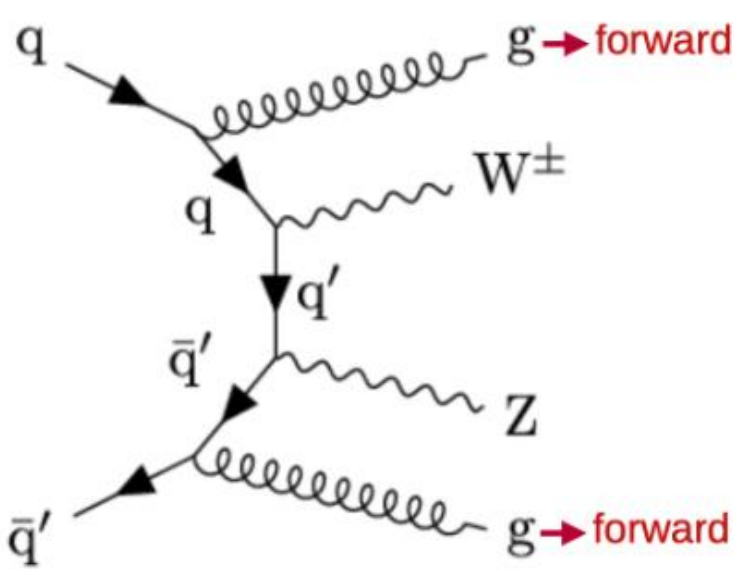
Figures from [Paul Gessinger - Expected tracking performance of the ATLAS Inner Tracker Upgrade for Phase-II]

Motivation: Physics analyses

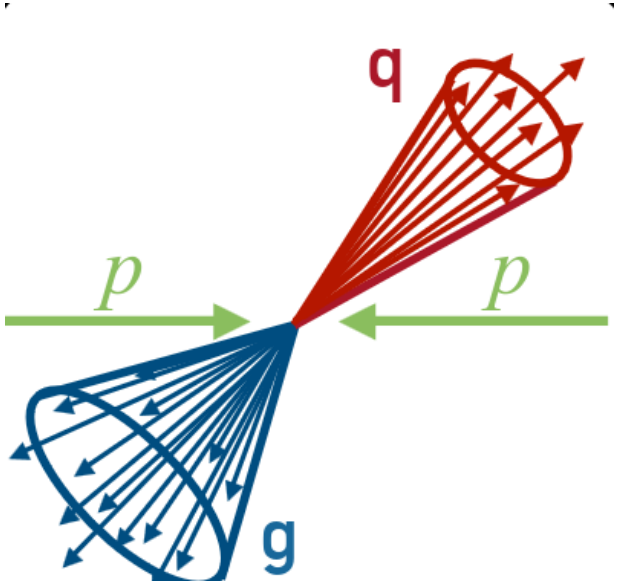
- Quark-gluon identification is an important tool to separate signal and background for VBS processes (among others, VBF, SUSY, Resonances)
  - Signal and background can be distinguish from the tag identification of the (forward) jet
  - **GOAL:** Check the status of the current quark-gluon tag and study possible improvements



WZ-EW: **Signal**
(forward quark)



WZ-QCD: **Main background**
(forward gluon)



Status as today of the quark-gluon tag in ATLAS

OLD

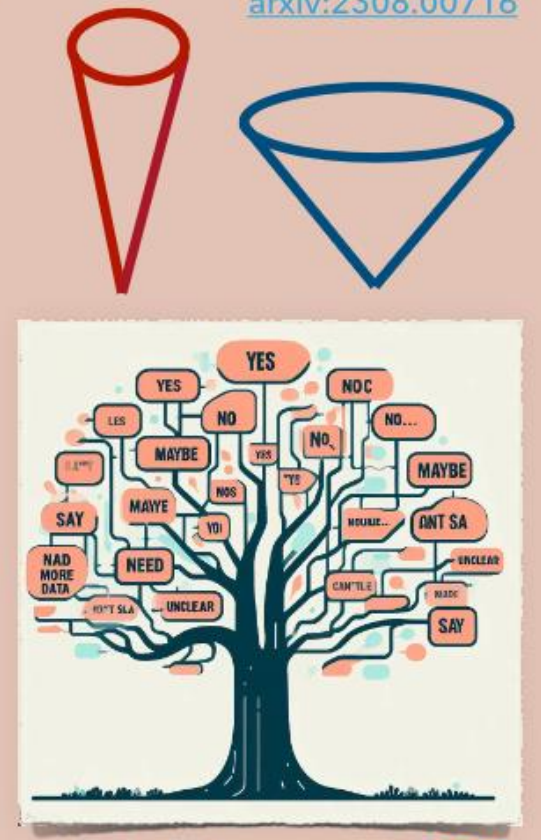
NEW

First look of these today in the HL-LHC phase

arxiv:1405.6583

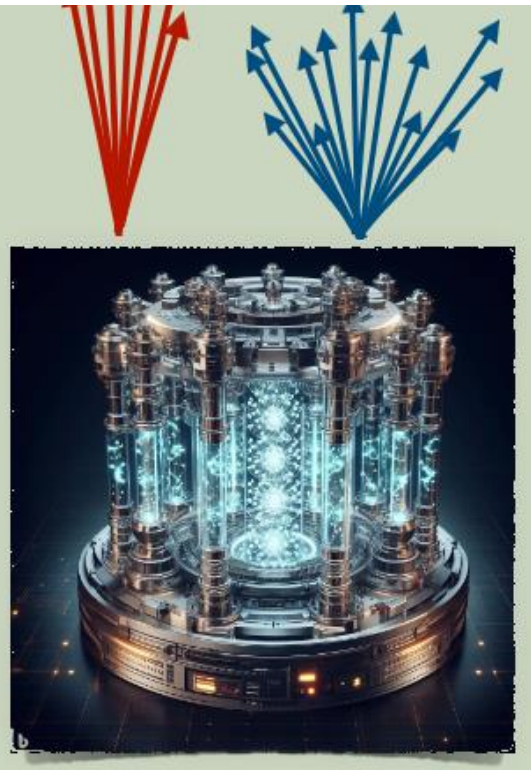
DALL-E's impression of **cut-based** tagger

arxiv:2308.00716



DALL-E's impression of **BDT**

Future plan:
Evaluate the transformers tagger



DALL-E's impression of **Transformer**

[Samuel Jankovych - Constituent based Quark/Gluon Jet Tagging]

Input samples

- **Running over di-jet production**
  - mc21_14TeV.801165.Py8EG_A14NNPDF23LO_jj_JZ0.deriv.DAOD_JETM1.e8481_s4038_r14365_r14367_p5742
  - mc21_14TeV.801166.Py8EG_A14NNPDF23LO_jj_JZ1.deriv.DAOD_JETM1.e8481_s4038_r14365_r14367_p5742
  - mc21_14TeV.801167.Py8EG_A14NNPDF23LO_jj_JZ2.deriv.DAOD_JETM1.e8481_s4038_r14365_r14367_p5742
  - mc21_14TeV.801168.Py8EG_A14NNPDF23LO_jj_JZ3.deriv.DAOD_JETM1.e8481_s4038_r14365_r14367_p5742
  - mc21_14TeV.801169.Py8EG_A14NNPDF23LO_jj_JZ4.deriv.DAOD_JETM1.e8481_s4038_r14365_r14367_p5742
  - mc21_14TeV.801170.Py8EG_A14NNPDF23LO_jj_JZ5.deriv.DAOD_JETM1.e8481_s4038_r14365_r14367_p5742
- **TOTAL OF EVENTS**: 6 samples * 10 root files of 10.000 events each = 600.000 events

- **Creation of a minitree algorithm**
  - https://gitlab.cern.ch/fcastill/hl-lhc-tagger
  - Running using: asetup AnalysisBase,main,latest

Jets containers status

- Default option is implemented (same used for HL-LHC analyses: gitlab):
  - HLLHC/JES_MC16_HLLHC_June2021.config
- PFlow not ready yet
  - They are very close to get the samples
    - https://its.cern.ch/jira/browse/ATLGBLCONDTAGS-135
- Unified Flow Objects (UFOs): It is developed to improve the mass resolution throughout a wide pT range by combining tracking information and calorimeter data
  - Once the PFlow is commissioned it will start the production of UFO jets

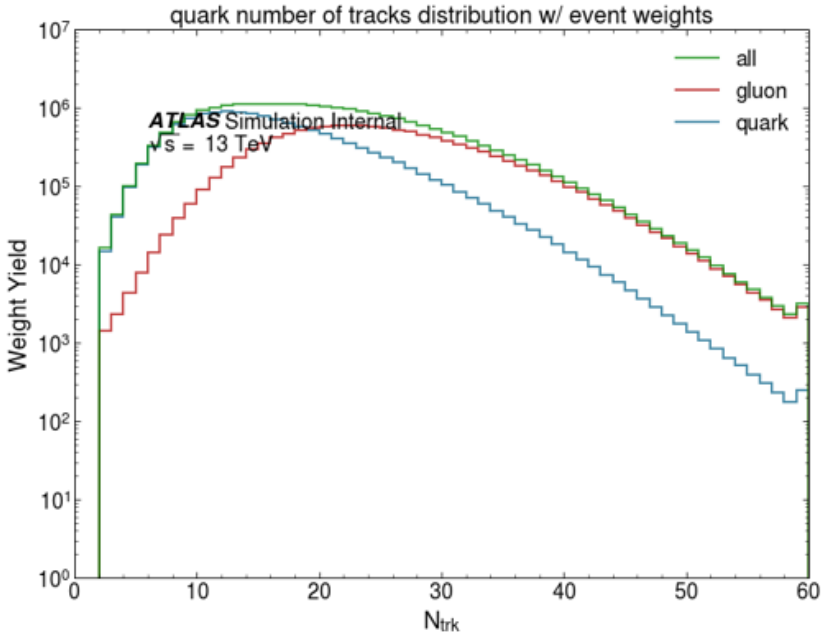
BDT performance of HL-LHC

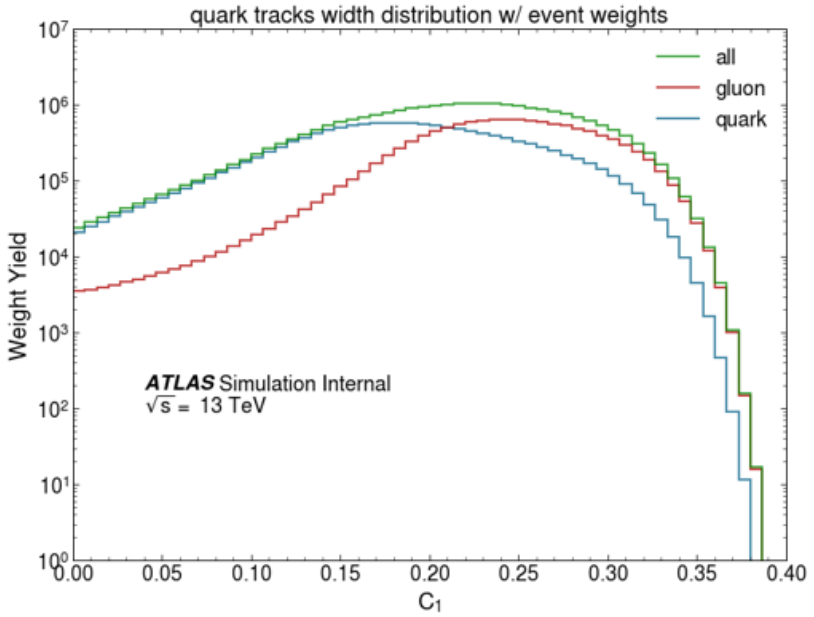
- **Input variables**: Ntracks, Track C1 (energy correlation), Track With and flat pT
- **Jets**: Calibrated EM topo jets
- **Cuts**: pT > 500 GeV and abs(eta)<2.1, Number of jets > 1 and event_weight <100
- **Weights** (PU not included/it was in the run 2 version):

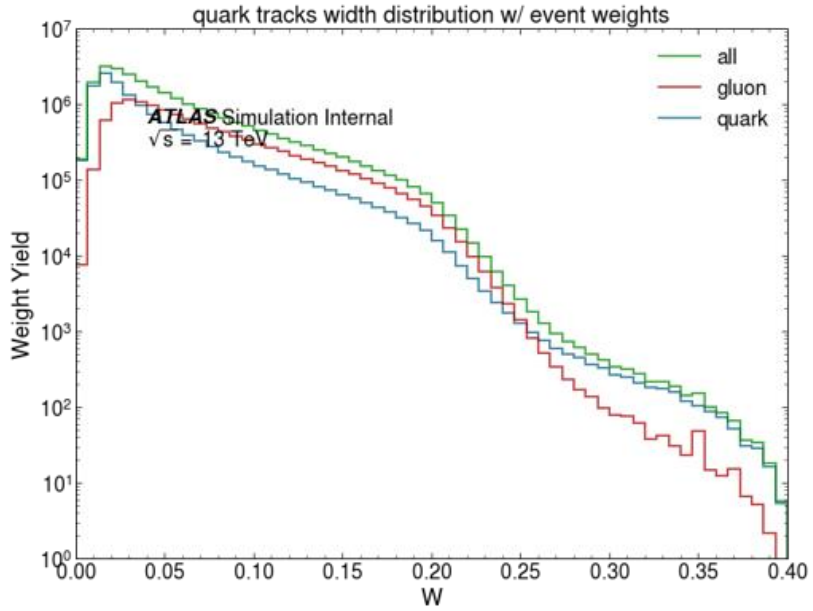
  Event_weight = xsec * mconly_weight / sumweight

  sumweight = Sum of mconly_weight

Input variables







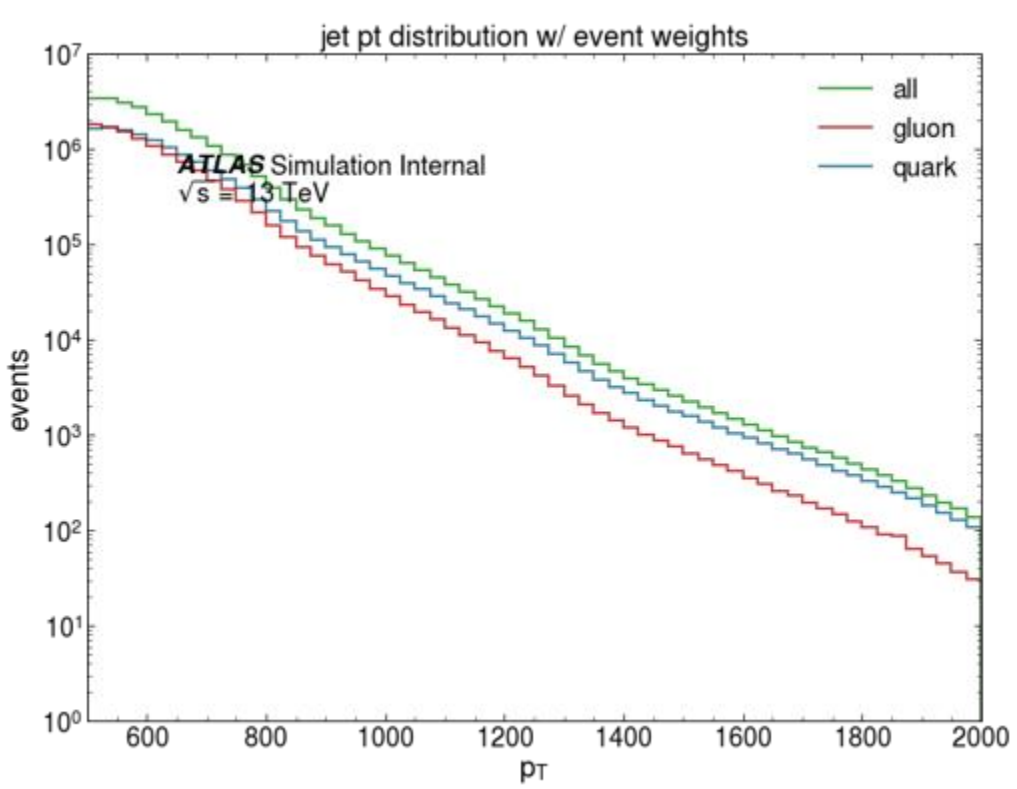
$$n_{\text{track}} = \sum_{\text{trk}\in\text{jet}}$$

$$C_1^{\beta=0.2} = \frac{\sum_{i,j\in\text{jet}}^{i\neq j} p_{\text{T},i} p_{\text{T},j} \left(\Delta R_{i,j}\right)^{\beta=0.2}}{\left(\sum_{\text{trk}\in\text{jet}} p_{\text{T}}^{\text{track}}\right)^2}$$

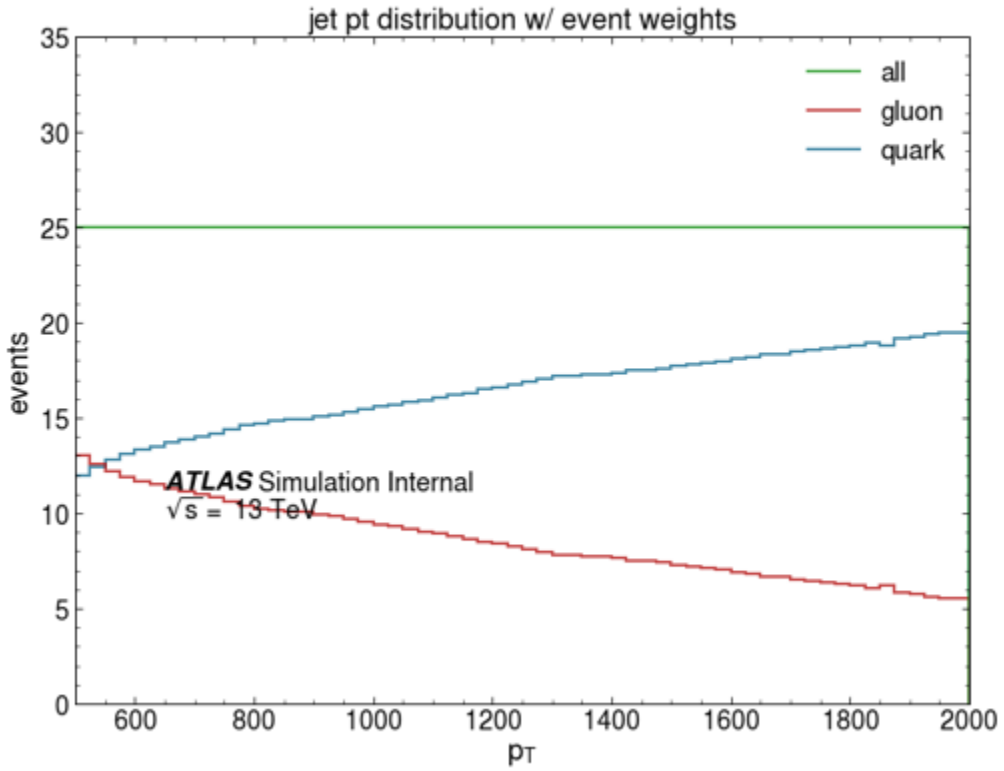
$$w^{\text{track}} = \frac{\sum_{\text{trk}\in\text{jet}} p_{\text{T}}^{\text{track}} \Delta R_{\text{trk,jet}}}{\sum_{\text{trk}\in\text{jet}} p_{\text{T}}^{\text{track}}}$$

Input variables weighted for the BDT training

PT weighted and flat distribution

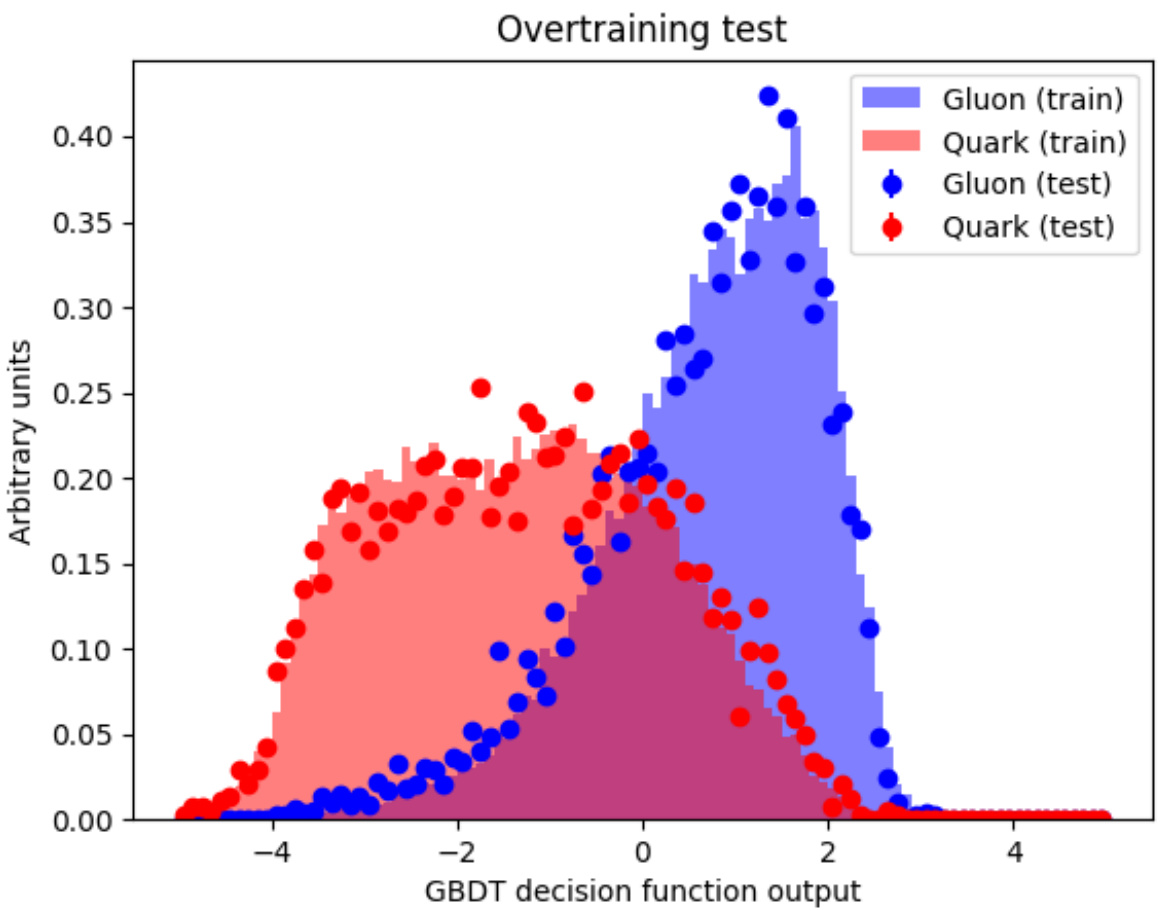


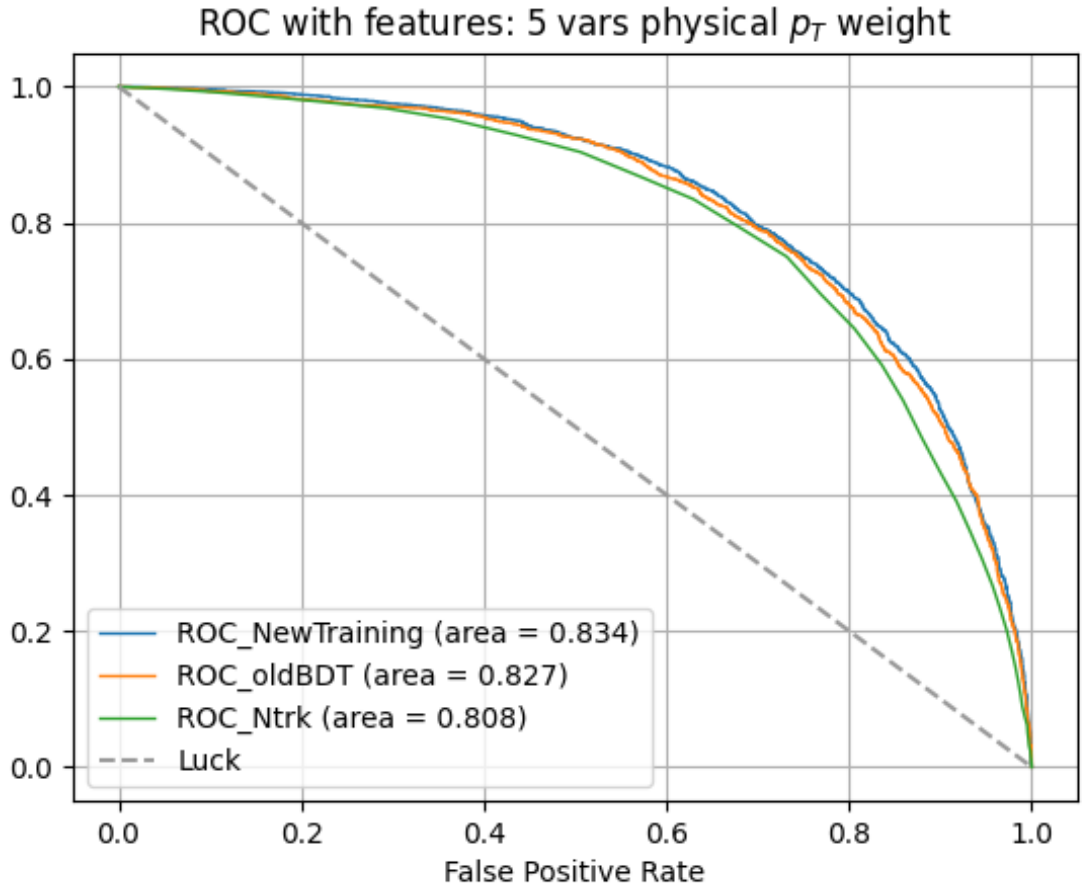
weighted pT -> not an input variable



Input variable flat pT

BDT performance (HL-LHC)



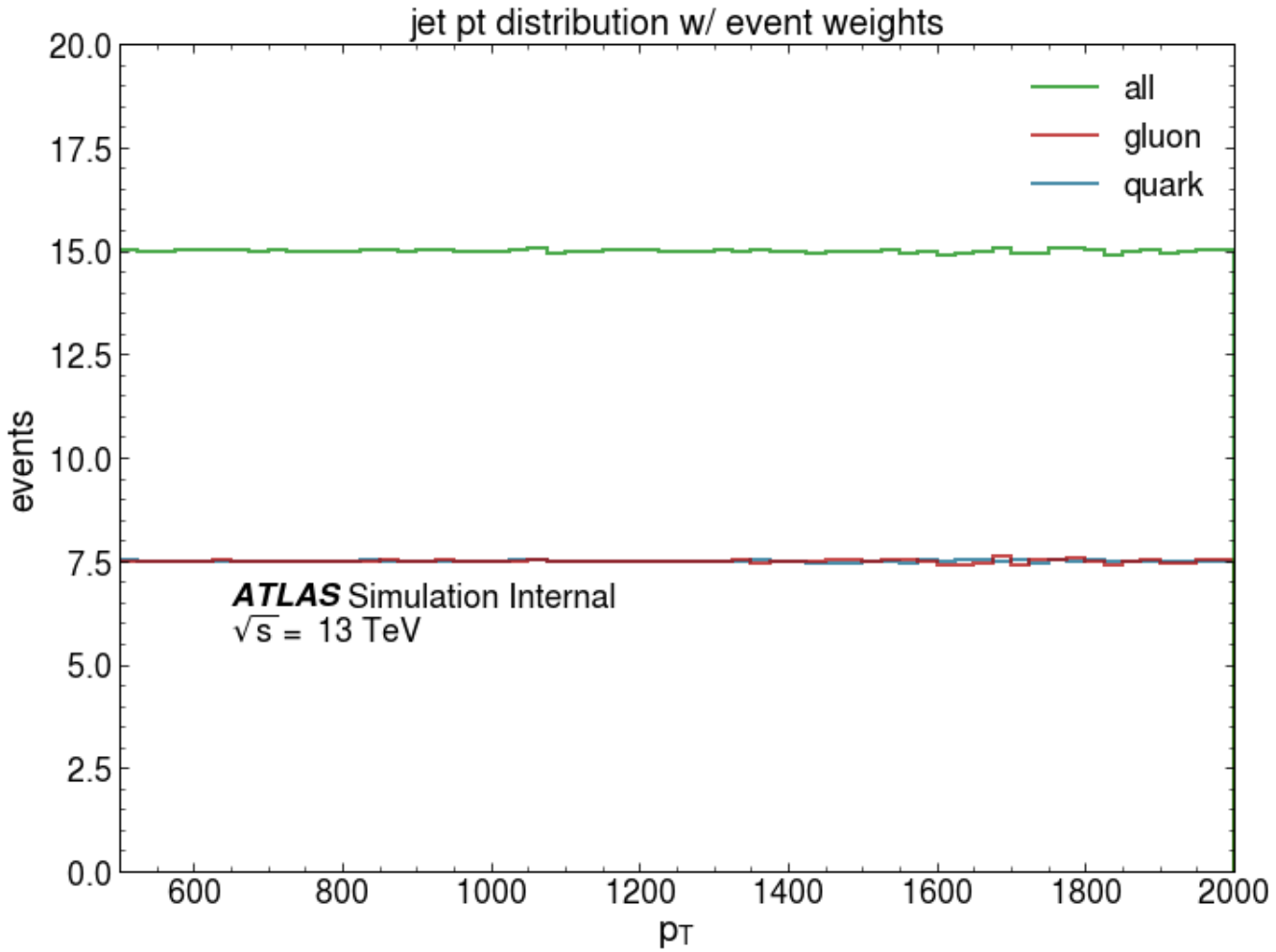


Low stats, need to request more samples (Run 2 Roc=0.830 see backup)

Size comparison

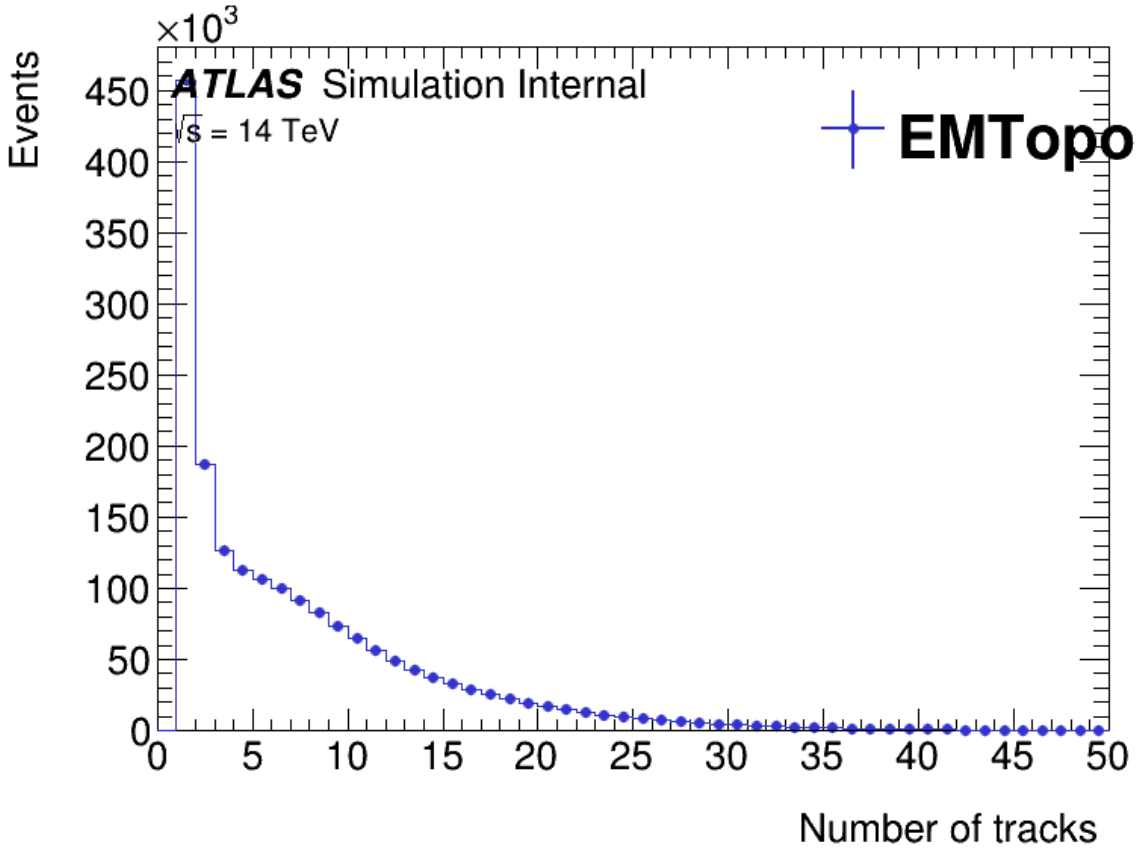
- **Run 2**
  - After cut applied --> pt > 500 GeV and pt < 2000 GeV, abs(eta_jet1) < 2.1, abs(eta_jet2)<2.1, event_weight < 100 and number of jets >1 & ptjet1/ptjet2 < 1.5 & nTracks > 1
  - Memory usage: 17.3 GB
    - Training data: 4.0 GB, 108.049.937 entries (80%)
    - Testing data: 515.2 MB, 13.506.243 entries (10%)
    - Validation data: 515.2 MB, 13.506.243 entries (10%)
- **HL-LHC**
  - After cut applied --> pt > 500 GeV and pt < 2000 GeV, abs(eta_jet1) < 2.1, abs(eta_jet2)<2.1, event_weight < 100 and number of jets > 1 & nTracks > 1
  - Memory usage: 16 MB
    - Training data: 5 MB, 109.684 entries (jets) (80%)
    - Testing data: 642.7 KB, 13.711 entries (10%)
    - Validation data: 642.7 KB, 13.711 entries (10%)
- HL-LHC input entries (without any cut) -> 6 samples * 10 root files of 10.000 events each = 600.000 events
- After pt, eta, number of jet requirements -> 65.000 events

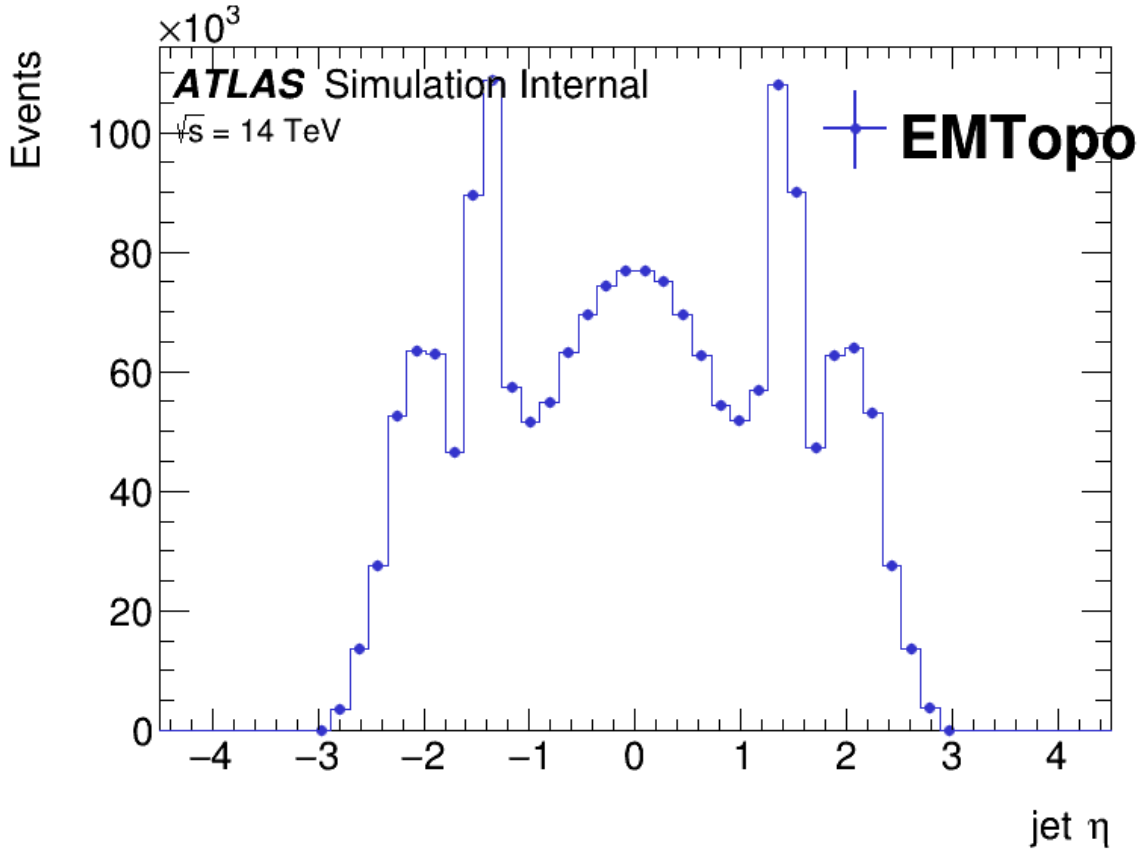
New try: not adding the topology. Still working on it.



Caveat: nTracks > 0. No forward jets

https://acode-browser1.usatlas.bnl.gov/lxr/source/athena/Reconstruction/Jet/JetMomentTools/Root/JetQGTaggerVariableTool.cxx





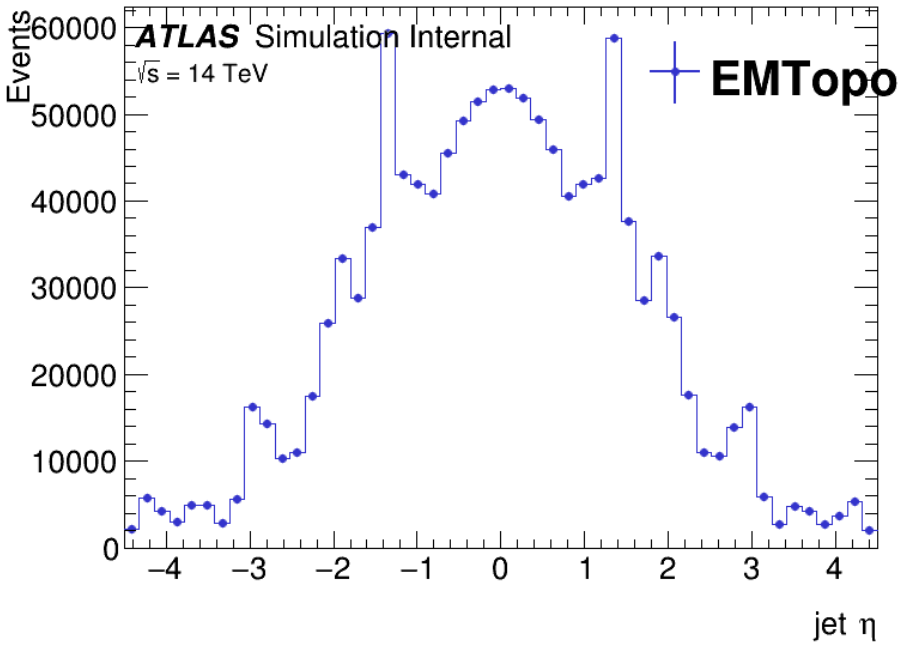
Input variables from getAttribute. Code:

```
jet->getAttribute(xAOD::JetAttribute::NumTrkPt500, nTrkVec);
jet->getAttribute("DFCommonJets_QGTagger_TracksWidth", trkWidth);
jet->getAttribute("DFCommonJets_QGTagger_TracksC1", trkC1);
```

Find a way to have tracks and input variables defined in the forward region

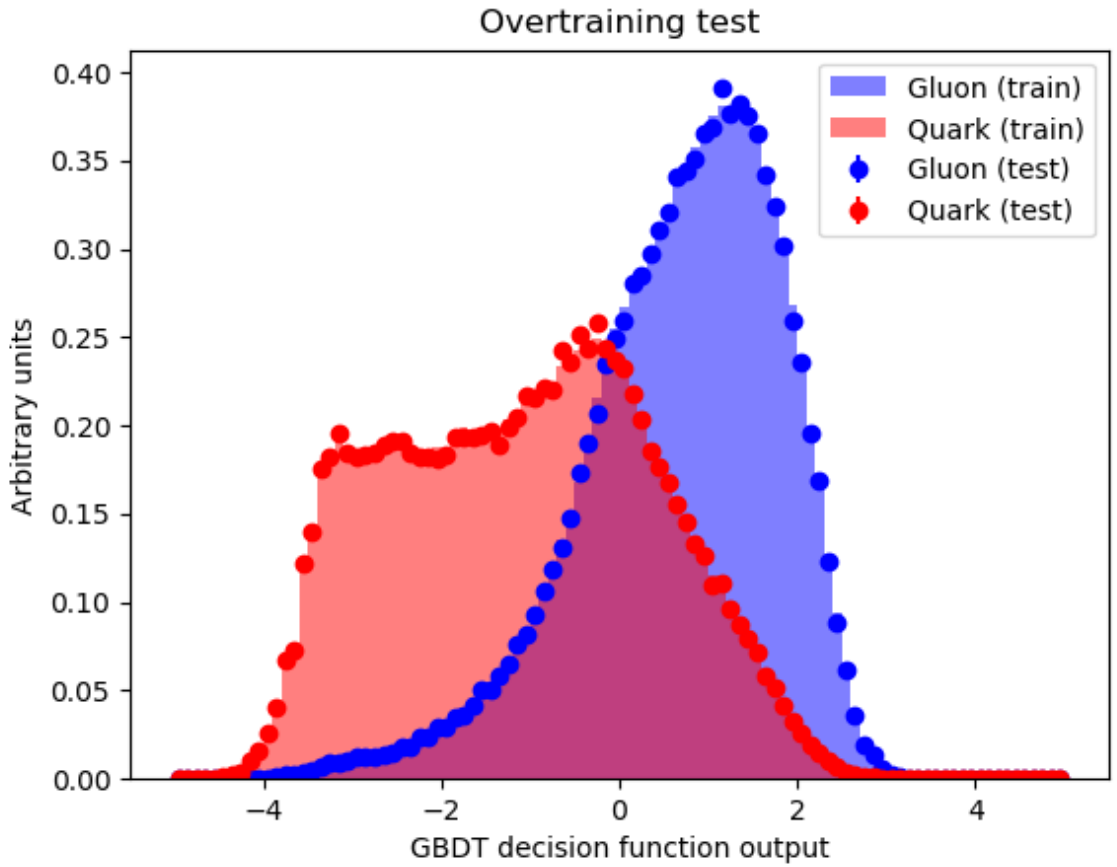
Outlook & Next steps

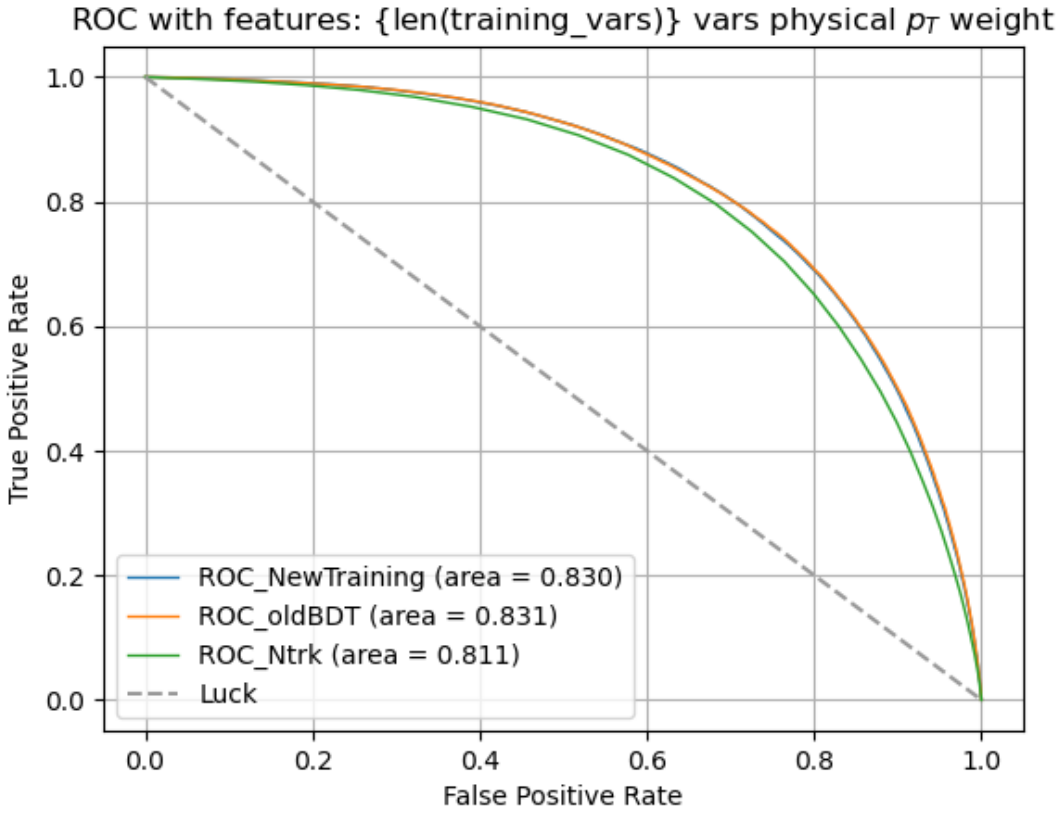
- First look to BDT in HL-LHC environment
- More stats needed it
  - Samples with more population of forward jets
    - Next HL-LHC sample production can be found at https://its.cern.ch/jira/browse/ATLMCPROD-10818 This includes 5 million events for each dijet slice
- Compare input variables HL-LHC vs Run 2
- Explore samples populated of jets in the forward region and request derivation



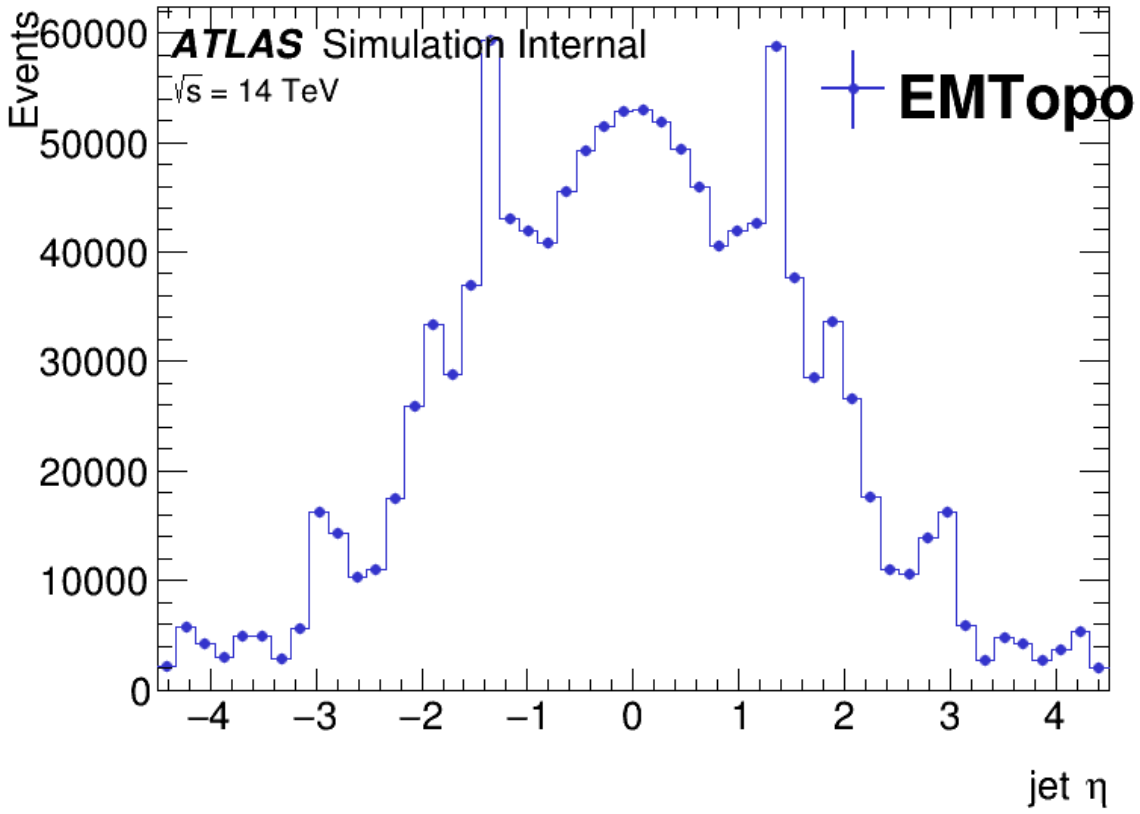
BACKUP

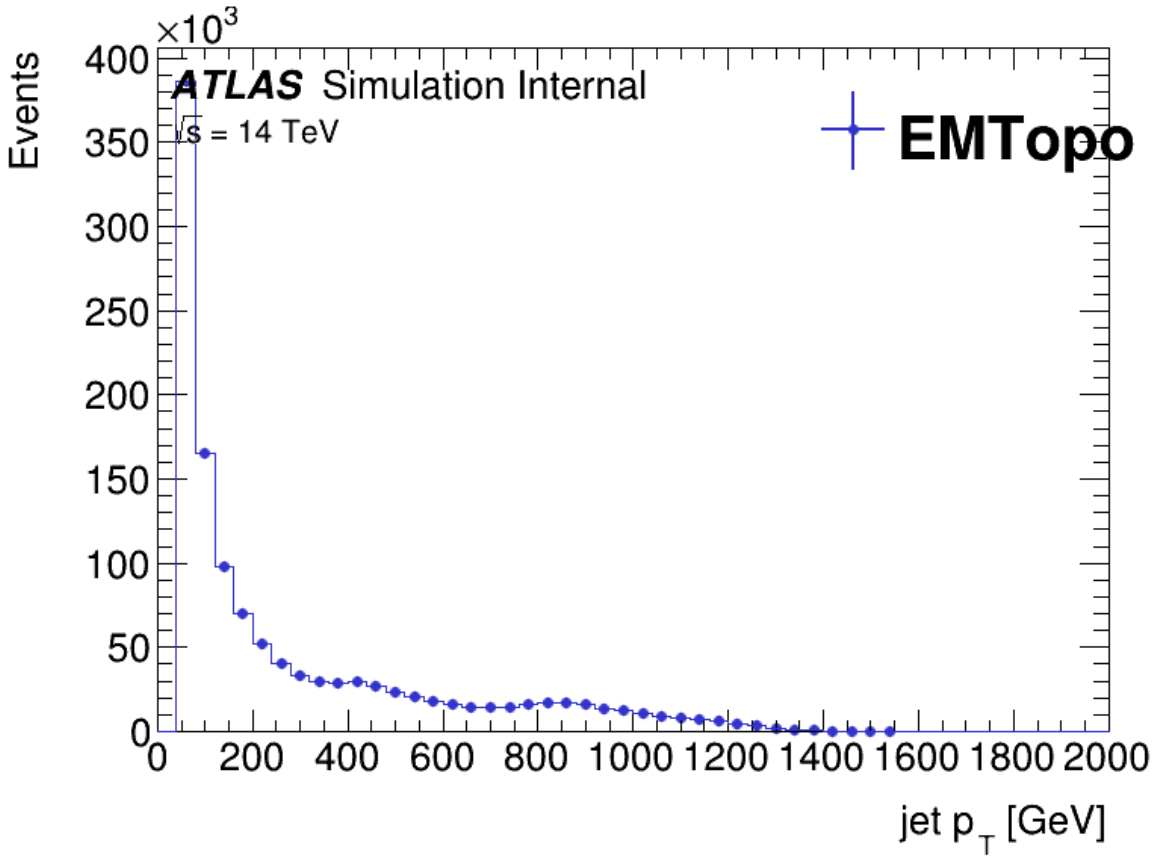
BDT perfomance (Run 2, my run)



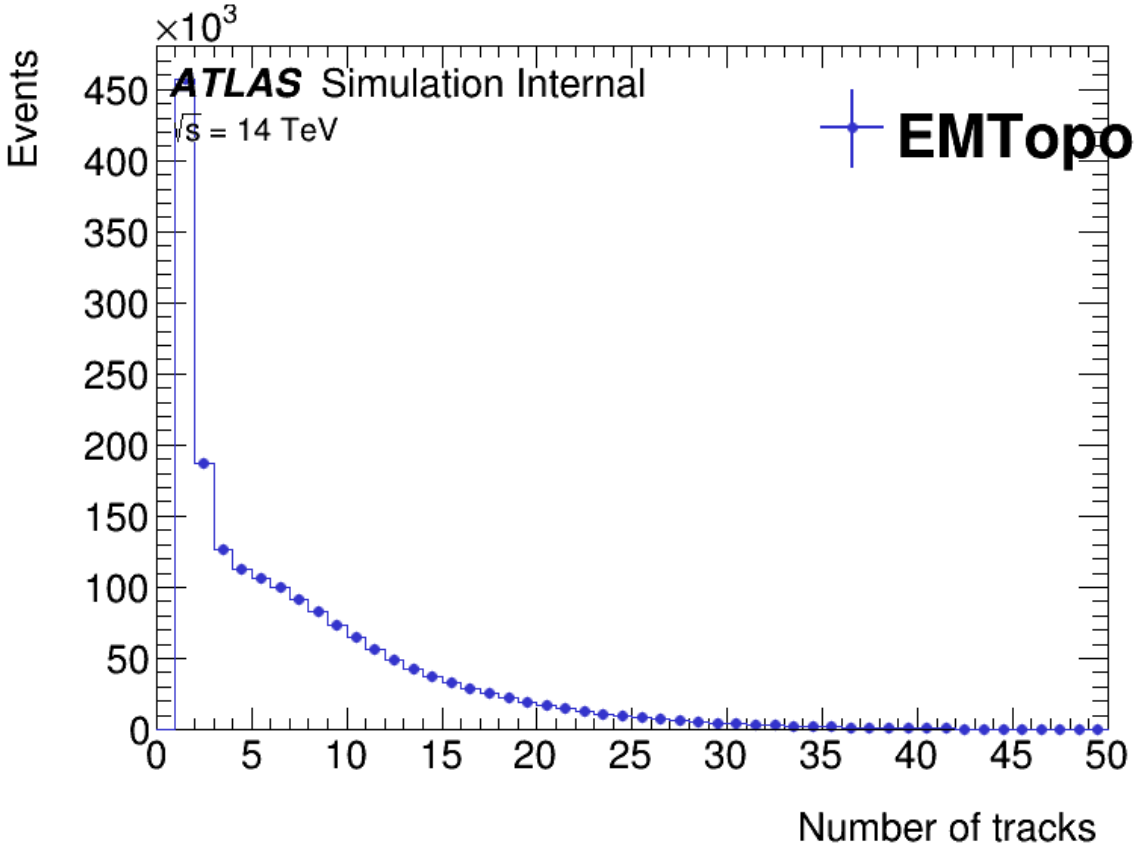


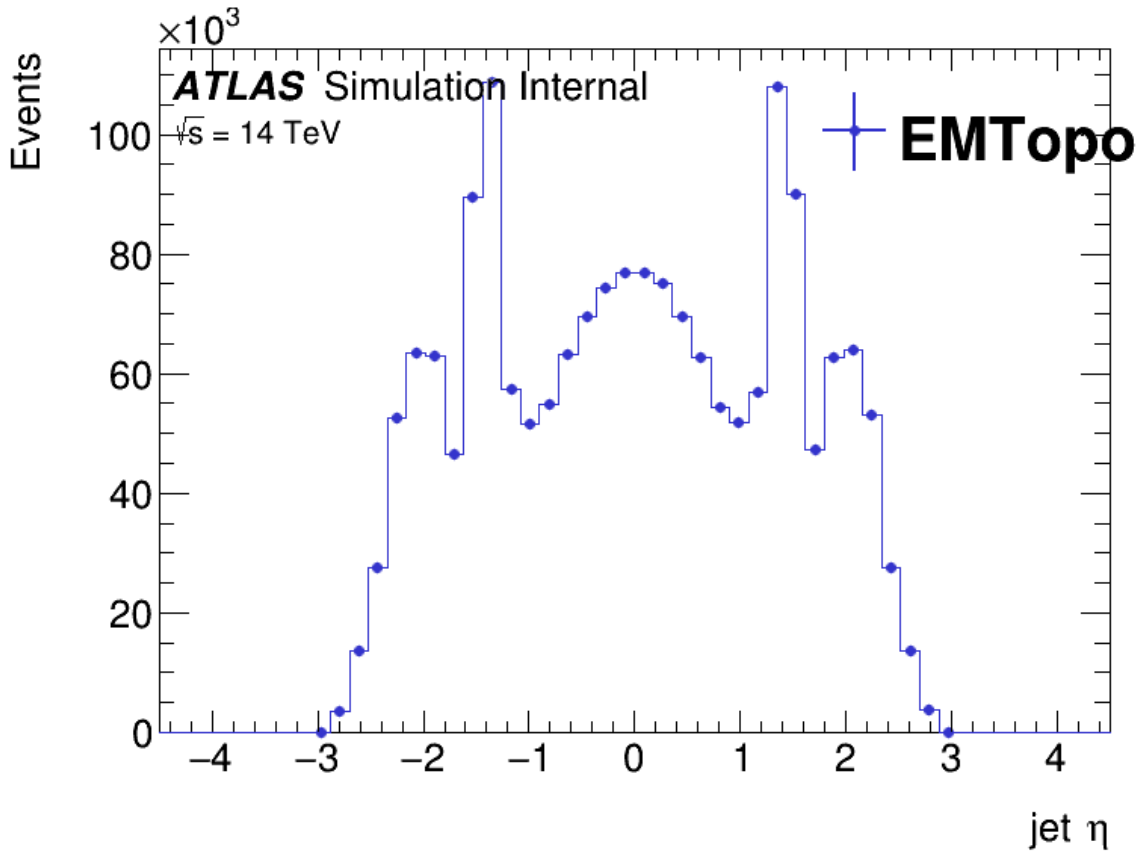
Pt > 50 GeV and Number of jets > 1



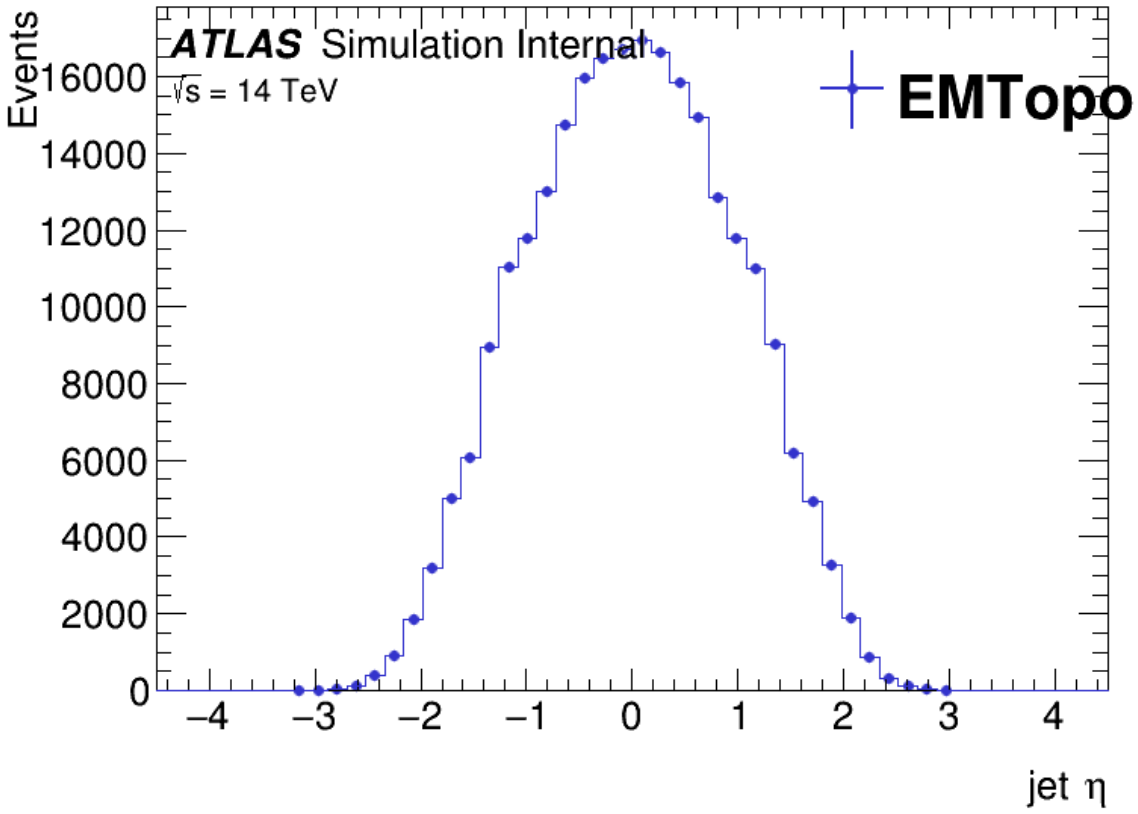


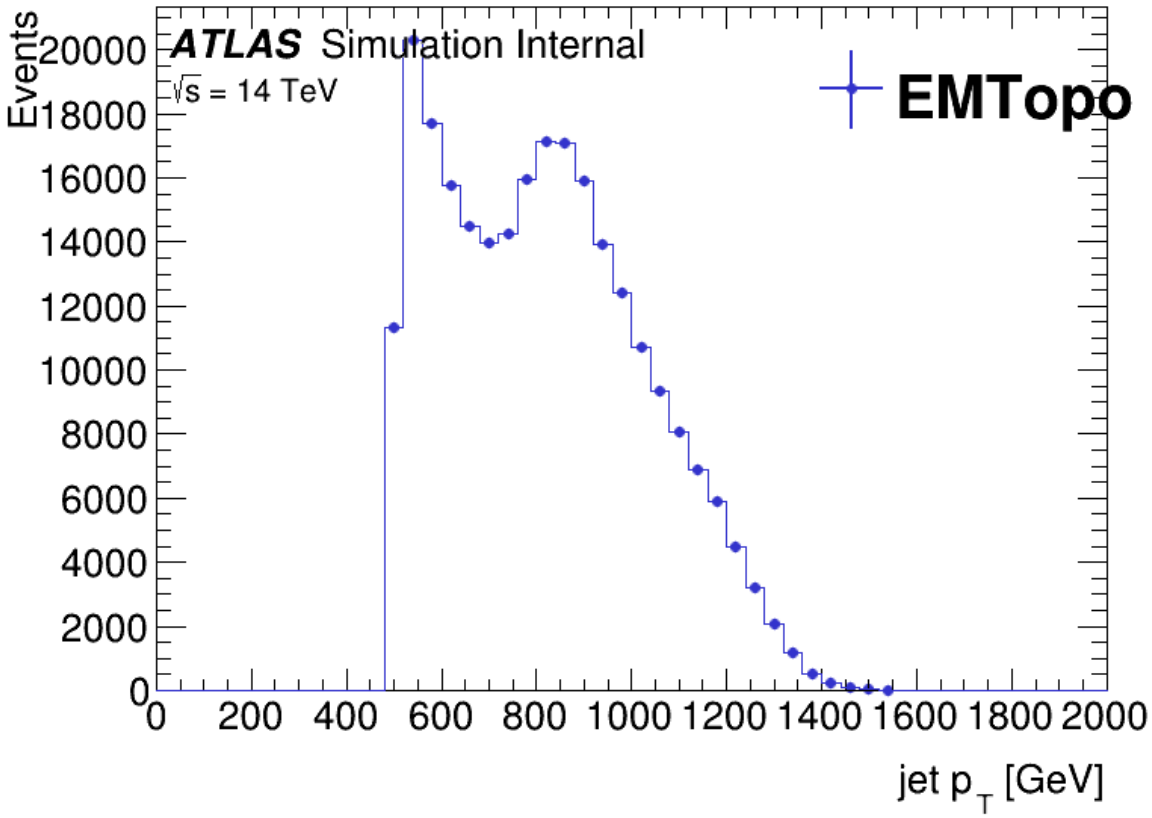
EM topo nTracks > 0





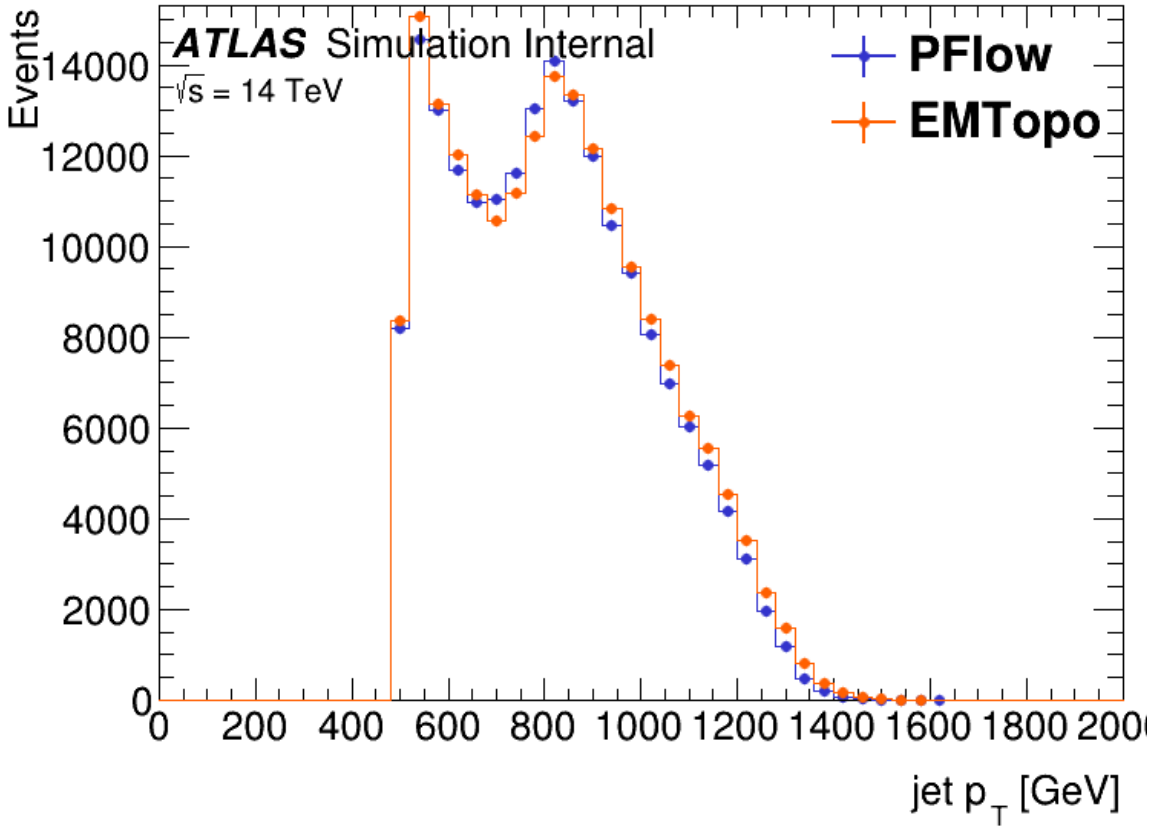
Pt > 500 GeV and Number of jets > 1

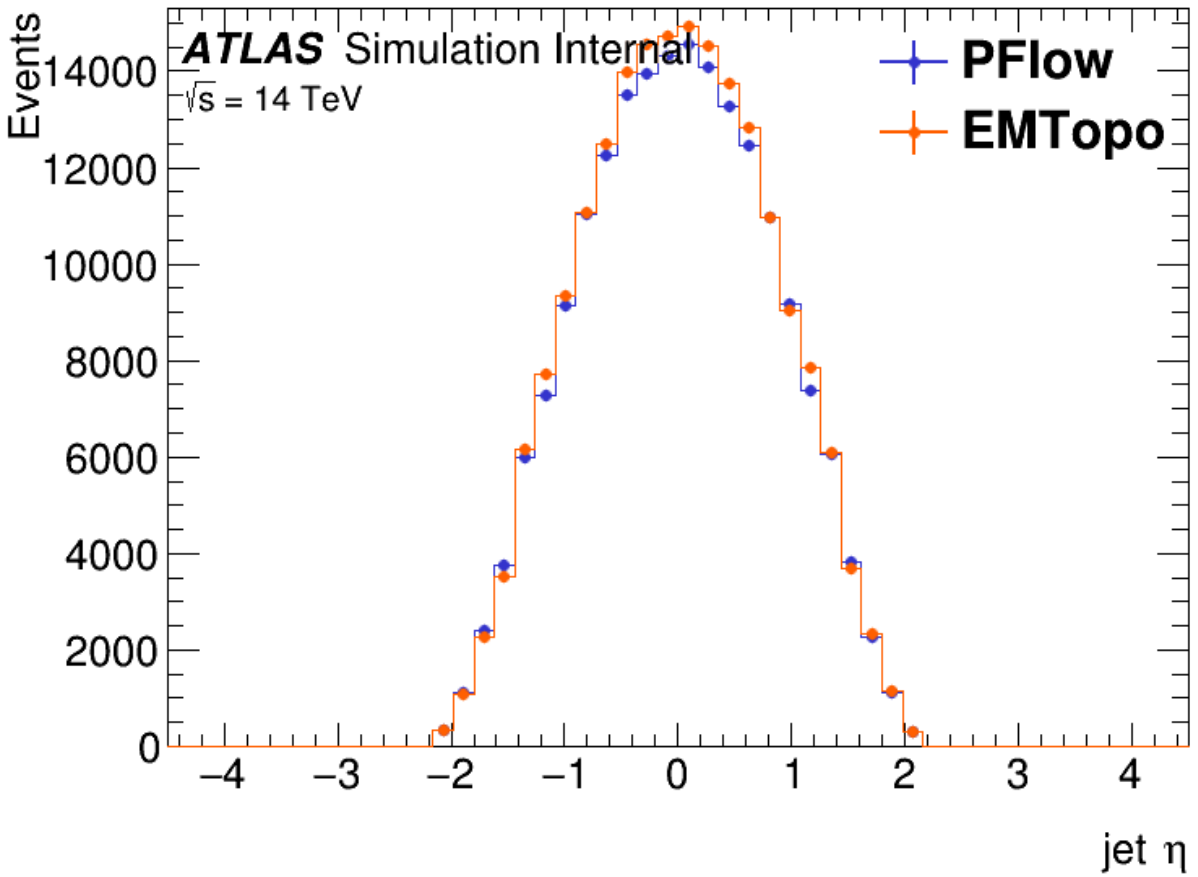




Jets -> EMTopo vs PFlow (no calibrated)

- Cuts: pt > 500 GeV, abs(eta)<2.1. JEM1 deriv.

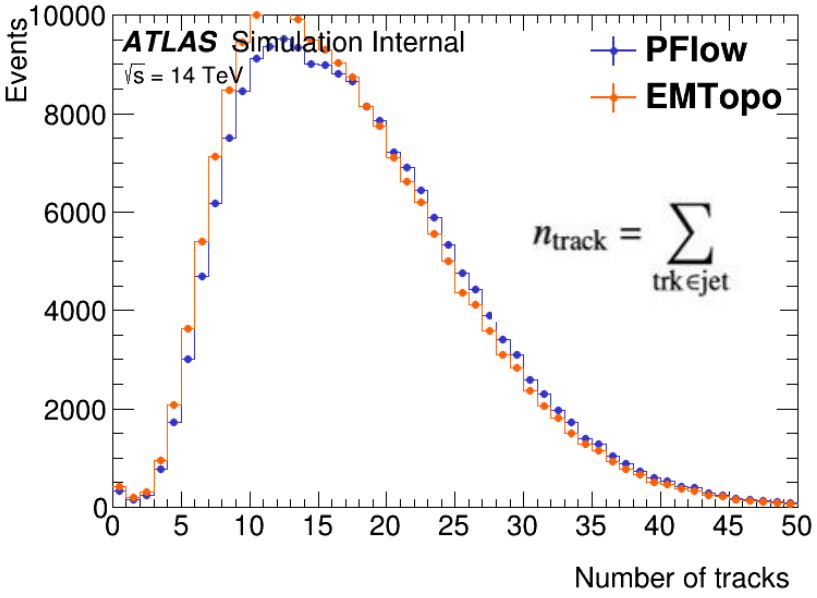


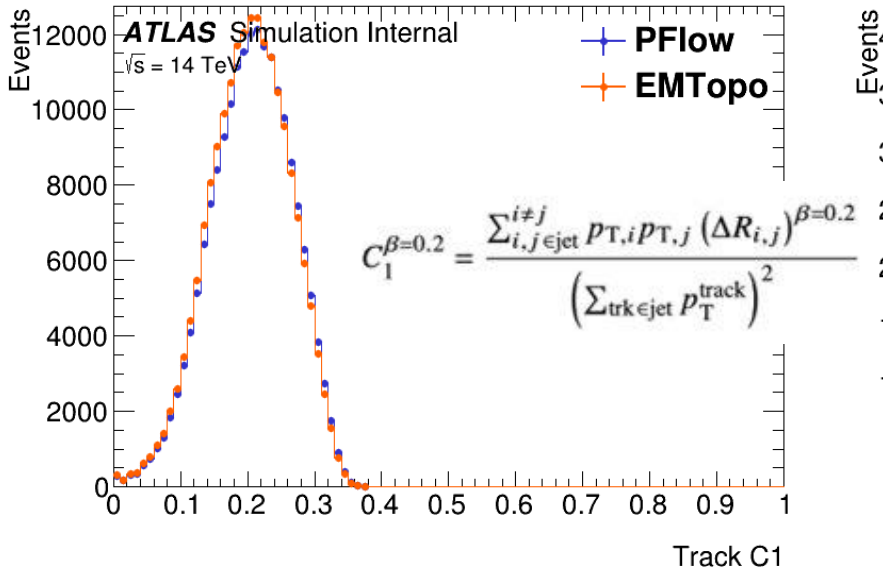


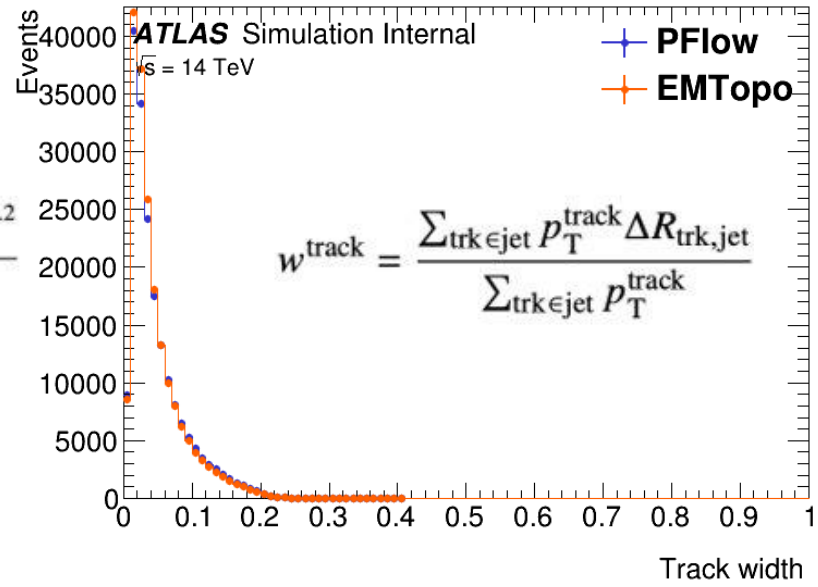
- Not big discrepancy in the central region for high pt jets

Jets -> EMTopo vs PFlow (no calibrated)

- Cuts: pt > 500 GeV, abs(eta)<2.1.



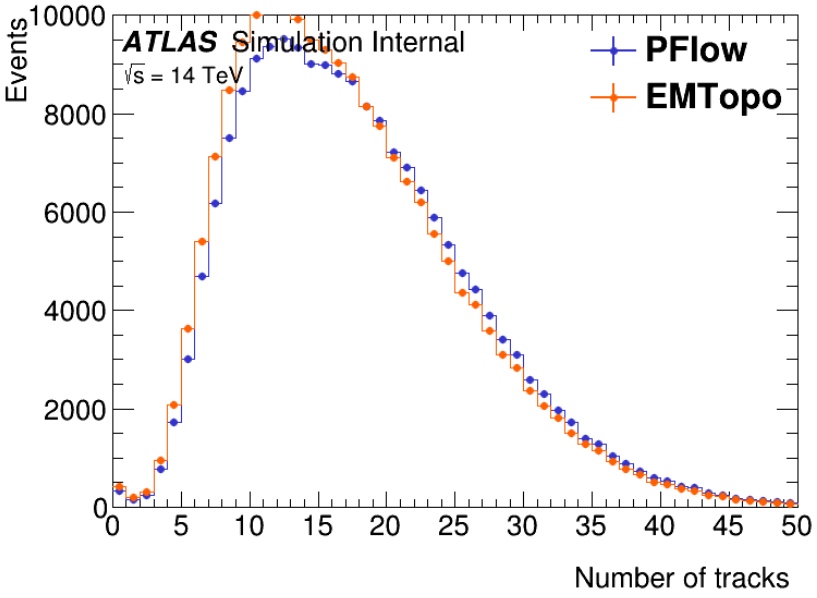


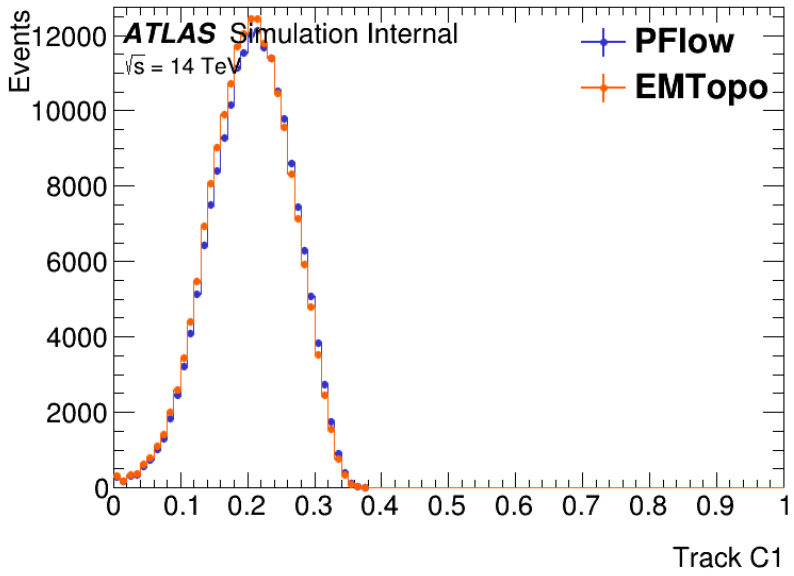


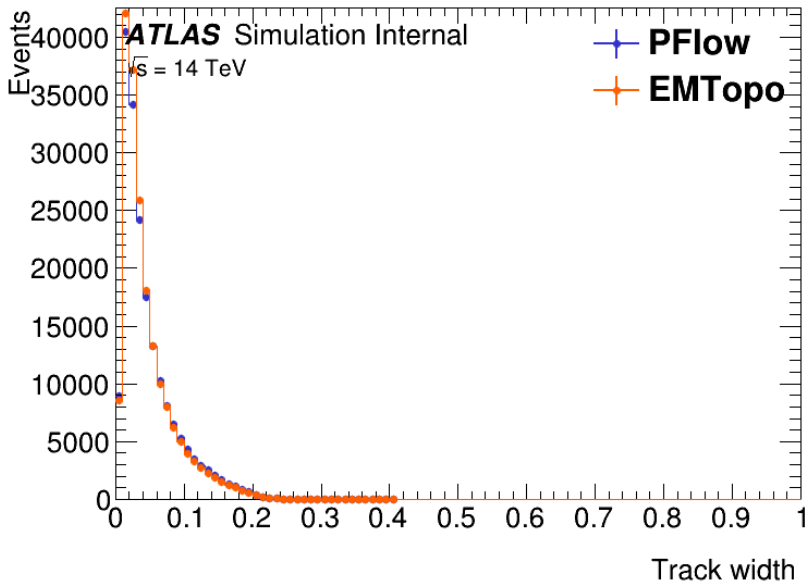
- bdt input variables (along with flat pt)

EMTopo calibrated vs PFlow uncalibrated

- Cuts: pt > 500 GeV, abs(eta)<2.1.

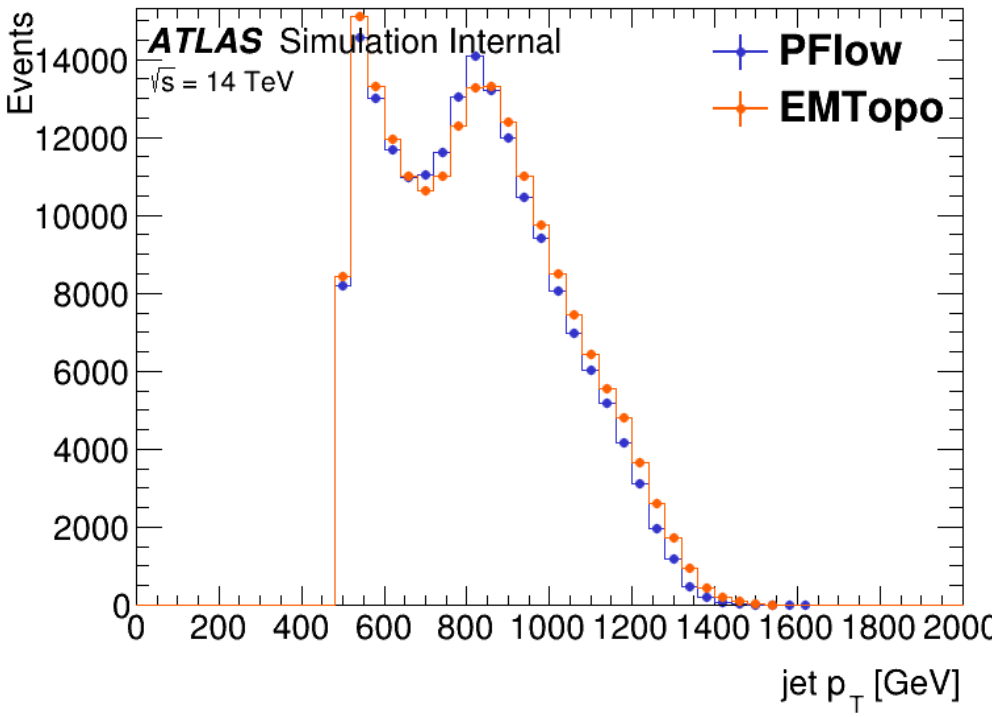


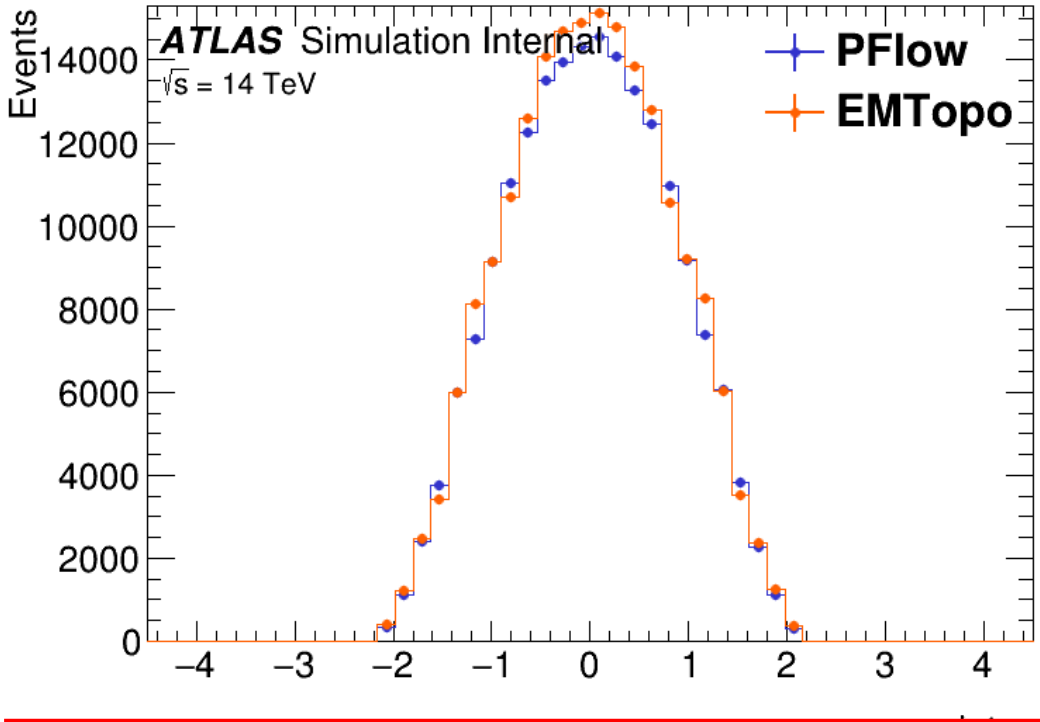




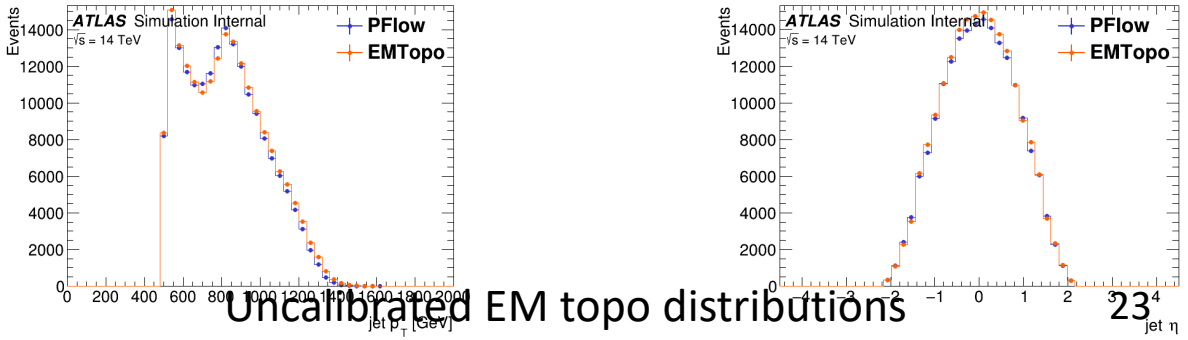
- bdt input variables (along with flat pt)

EMTopo calibrated vs PFlow uncalibrated



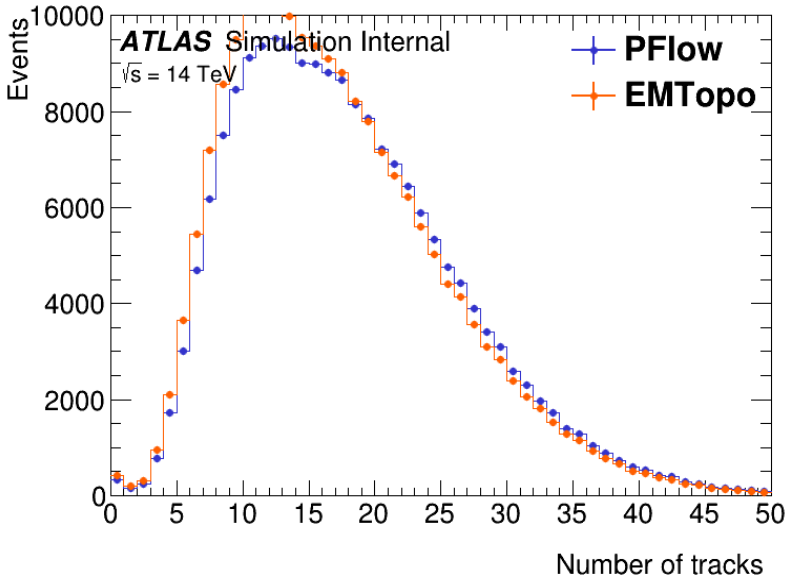


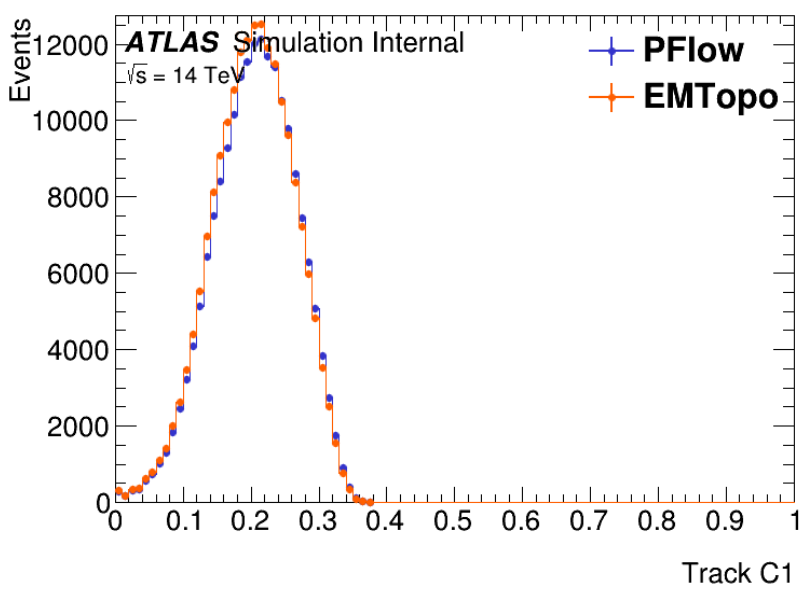
Discrepancies appears from EMTopo compared to PFlow after calibration

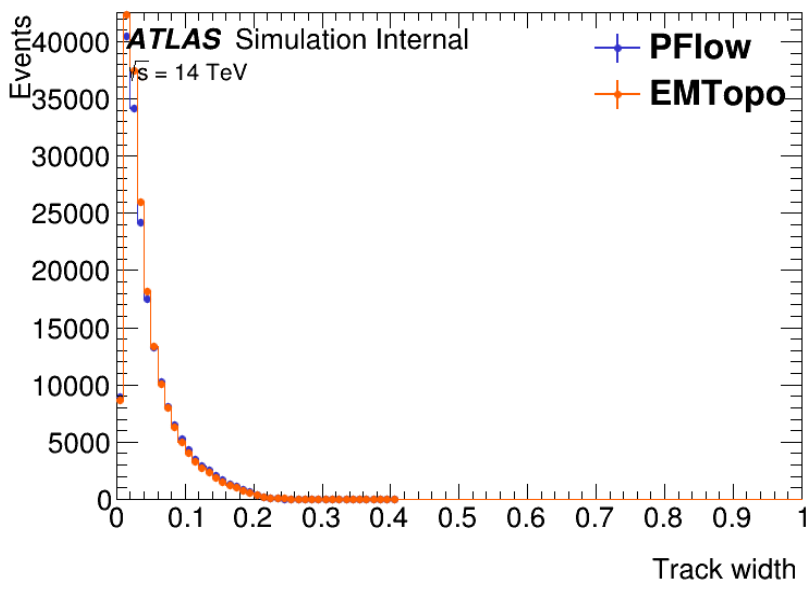


Uncalibrated EM topo distributions

EMTopo calibrated vs PFlow uncalibrated







First test of BDT tagger for HL-LHC is done with calibrated EM Topo jets