# Calibration of Q/G Jet BDT Tagger

for CLHCP


Wanyun Su[a,b,d],
Evan Saraivanov[d], Htet Aung Myin[d], Ke Li[d], Ben Nachman[c], Shih-Chieh Hsu[d], Shu Li[a,b]

a 李政道研究所 TDLI

b 上海交通大学 Shanghai Jiao Tong University

c Berkeley Lab

d University of Washington

# Introduction

- Previous calibration( Support Note ):
  - The uncertainties on Ntrack-only tagger come from a Run 1 measurement combined with Run 2 tracking uncertainties.
  - Calibration on Ntrack-only tagger has been presented in this talk.
- now use BDT :
  - Combine variables : $N_{trk}$ , $W_{trk}$ , $C_1^{track}$, $pT$ , $\eta$ . (see this talk)
    - $N_{trk}$ , $W_{trk}$ , $C_1^{track}$ are track-based variables

$$w_{\mathrm{track}} = \frac{\sum_{i\in\mathrm{Jet}} p_{\mathrm{T},i}\, \Delta R_{i,\mathrm{Jet}}}{\sum_{i\in\mathrm{Jet}} p_{\mathrm{T},i}} \text{ , tracks } i$$

$$C_{1,\mathrm{track}}^{\beta=0.2} = \frac{\sum_{i,j\in\mathrm{Jet}} p_{\mathrm{T},i}\, p_{\mathrm{T},j}\, (\Delta R_{i,j})^{\beta}}{\left(\sum_{i\in\mathrm{Jet}} p_{\mathrm{T},i}\right)^2} \text{ , tracks } i,j$$

  - The approach in this study directly use the full Run 2 data.

- Strategy :
  - Define quark-riched and gluon-riched samples.
  - Apply Matrix Method to extract quark and gluon distributions.
  - Compare data and mc BDT distribution to obtained the scale factor for quark and gluon jets.
    - Use Sherpa as nominal.

# Matrix Method

1) The matrix method can extract pure quark or gluon jets under the assumption that the two sample have the same shape of the variables in each of quark and gluon jets.

$$X_1 = f_q^1 X_q + (1 - f_q^1) X_g$$
$$X_2 = f_q^2 X_q + (1 - f_q^2) X_g,$$

2) The method is performed for each pT range.

$X_1$ - quark-riched sample

$X_2$ - gluon-riched sample

$f_q$ - quark fraction (from MC)

$X_q$ - extracted quark sample

$X_g$ - extracted gluon sample

3) Here the BDT are  chosen as our tagging variables.By solving this equations in each BDT bin, we can get the BDT distribution of quark/gluon in each pT range.

# Matrix Method

- Quark vs gluon jet tagging is calibrated using two samples with different quark fractions $(f_q)$. -- see backup slides
  - **Method 1 - dijet only (pT> 500 GeV):** the two samples both come from multijet but used $\eta$ to separate the leading two jets as higher ($X_1$, quark-like) and lower ($X_2$, gluon-like) jets.
  - **Method 2 - gamma+jet & (prescaled)dijet (pT<600 GeV):**
    - X1= leading jet from Gamma Jet
    - X2= leading jet and subleading jet from multijet.
    - f1q= fraction of quark jet in gamma jet divided by total quark jets in gamma jet (MC)
    - f2q= fraction of quark jet in multijet divided by total quark in multijet (MC)

  - **Method 3 - gamma+2jet & (unprescaled) trijet (Validation):**
    - X1= leading jet and subleading from gamma+2jet. (claimed to be more quark like, labeled as ‘higher’)
    - X2= leading 3 jets from multijet (claimed to be more gluon like, labeled as ‘lower’)
      f1q= fraction of quark jet in gamma2jet divided by total quark jets in gamma jet (MC)
    - f2q= fraction of quark jet in multijet divided by total quark in multijet (MC)

$$X_1 = f_q^1 X_q + (1 - f_q^1) X_g$$
$$X_2 = f_q^2 X_q + (1 - f_q^2) X_g,$$

# Method 1 (dijet - only): Samples and Selection

Multi-jets mc16 a+d+e :

- 3646[77-85].Sherpa_*.DAOD_JETM1.*   (nominal)

data 15-18  (140 $fb^{-1}$)

## Object definition :

- Gluon: partonLabel = 21
- Quark: 0 < partonLabel < 5
- Other: else..
- PFLow jet

- (for dijet) Higher/lower η jet : if | j1_η | >  | j2_η | :
  j1 is the Higher η jet (quark-riched) ,
  j2 is the Lower η jet (gluon-riched).

## Event Selection

- trigger "pass_HLT_j420"
- 1st Jet:  j1_pT > 500 GeV
- j1_pT / j2_pT < 1.5
- | j1_η | < 2.1,  | j2_η | < 2.1

Using  Insitu package

# Method 2 (gammajet+dijet): Samples and Selection

Gamma+jets :mc16 a+d+e :

- 36454[2..7].Sherpa_222_NNPDF30NNLO_SinglePhoton_pty*.deriv.DAOD_JETM4.* (nominal)
- data 15-18 (140 $fb^{-1}$)

## Event Selection (gamma+jet(quark-riched))

- Trigger: ”2015:HLT_g120_loose || HLT_g200_etcut
  !2015: HLT_g140_loose || HLT_g300_etcut”
- photon selection
  - ph_pT > 125 GeV
  - ID: Tight
  - ISO: FixedCutTight
  - |ph_η| <2.37
- jet selection
  - j1_pT > 40 GeV
  - | j1_η | < 2.1

## Event Selection (dijet(gluon-riched))

- Trigger: “HLT_j15” || “HLT_j25” || “HLT_j35” || “HLT_j45” || “HLT_j60” || “HLT_j110” || “HLT_j175” || “HLT_j260” || “HLT_j360” || “HLT_j420”
- j1_pT > 40 GeV
- j1_pT / j2_pT < 1.5
- | j1_η | < 2.1
- | j2_η | < 2.1

# Method 3 (gamma2jet+trijet): Samples and Selection

Samples: same as method 1 (multijet) and 2 (gamma+jet).

## Event Selection (gamma+2jets(quark-riched))

- Trigger: ”2015:HLT_g120_loose || HLT_g200_etcut
  !2015: HLT_g140_loose || HLT_g300_etcut”
- photon selection
  - ph_pT > 125 GeV
  - ID: Tight
  - ISO: FixedCutTight
  - |ph_η| <2.37
- jet selection
  - j1_pT > 40 GeV
  - j2_pT > 20 GeV
  - | j1_η | < 2.1
  - | j2_η | < 2.1

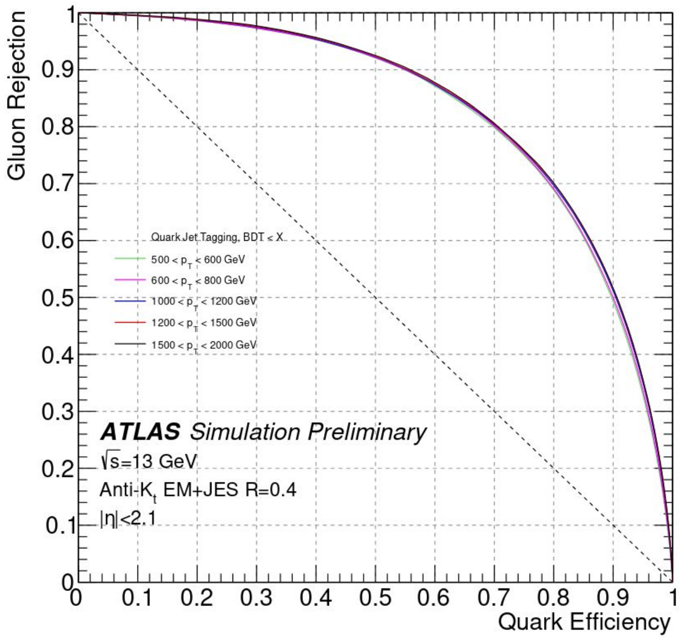
## Event Selection (Trijet(gluon-riched))

- Trigger: “HLT_j420”

- j1_pT > 500 GeV
- j1_pT / j2_pT < 1.5
- j3_pT > 20 GeV
- | j1_η | < 2.1
- | j2_η | < 2.1
- | j3_η | < 2.1

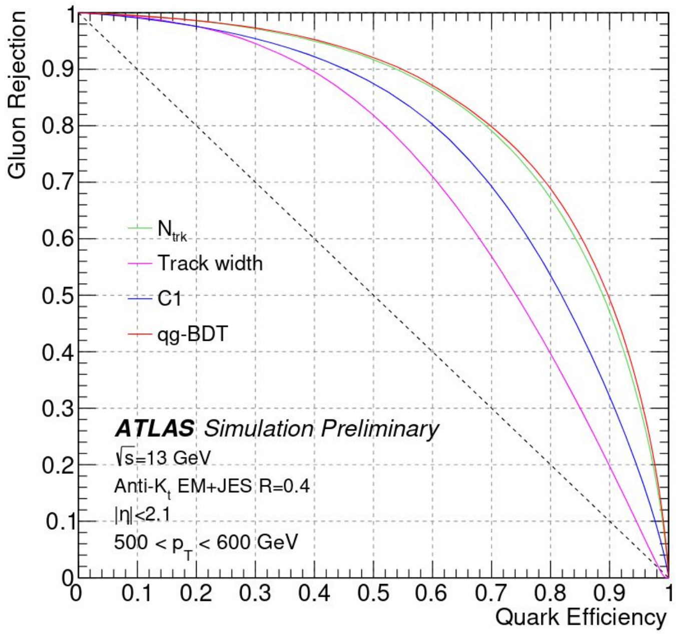
# Method 1 (dijet - only): ROC curves

ROC curve for BDT in different pT range




Mostly the same for dijet-only sample

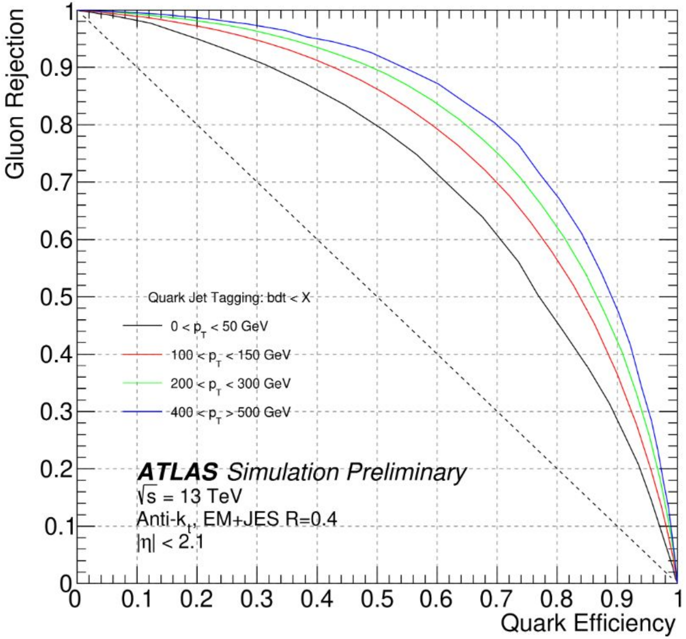
ROC curve for different variables in pT: 500~600 GeV




BDTin high pT behaves a bit better than others

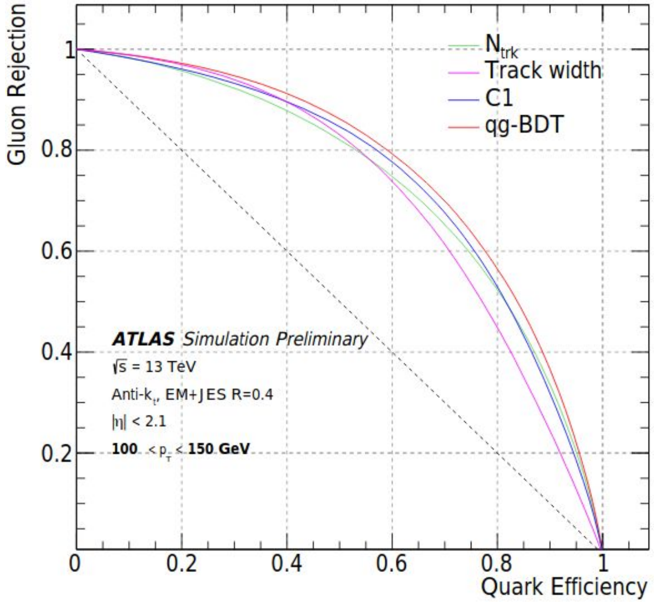
# Method 2 (gamm+jet) : ROC curves

ROC curve for BDT in different pT range



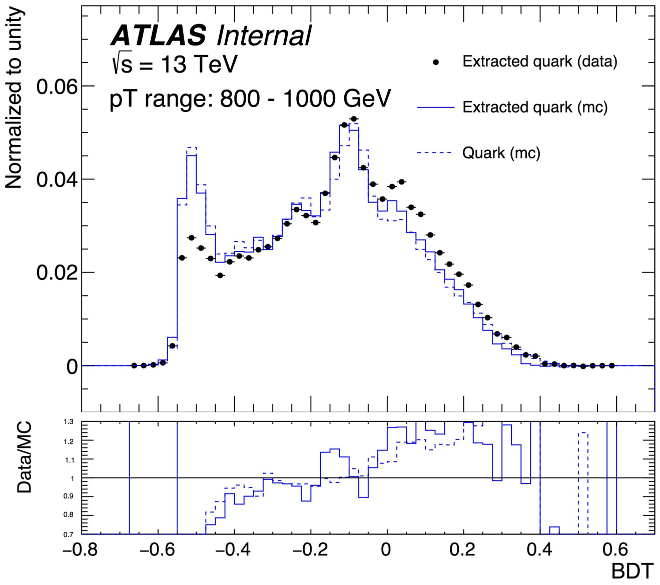

Better in higher pT in gamma-jet samples

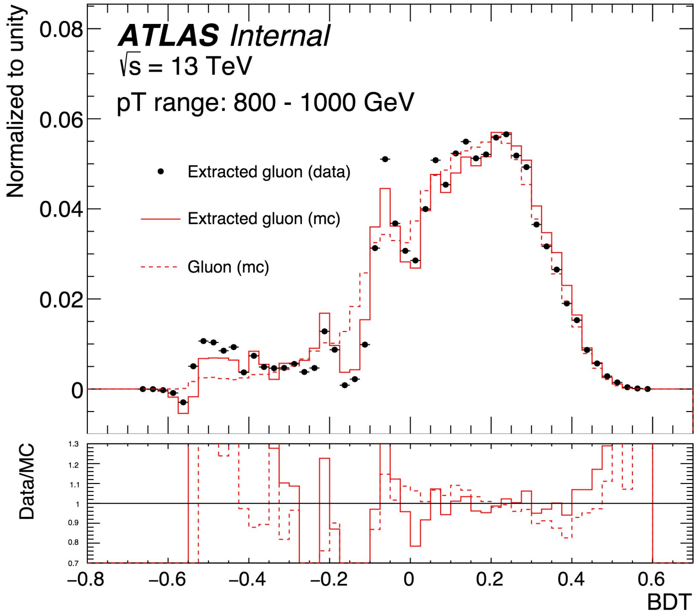
ROC curve for different variables in pT: 100~150 GeV




BDT in low pT behaves better than others

# Scale factors in BDT distribution in Method 1 (dijet - only)



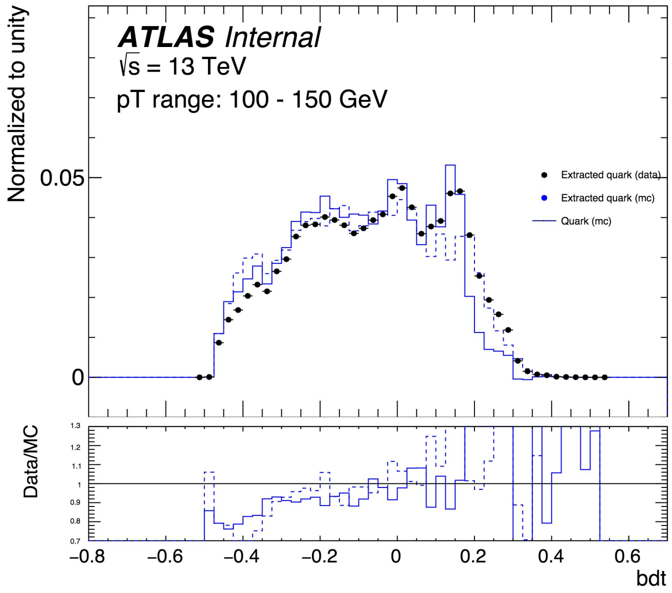

**Scale factor:**

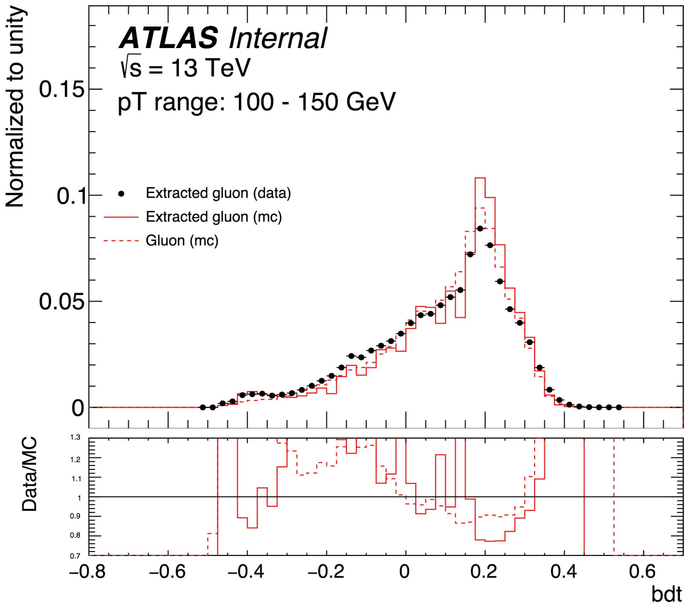
$$SF_{\mathrm{Q/G}}\left(x;p_{\mathrm{T,j}}\right)=\frac{p_{\mathrm{Q/G,\ Ext.Data}}\left(x;p_{\mathrm{T,j}}\right)}{p_{\mathrm{Q/G,\ Ext.MC}}\left(x;p_{\mathrm{T,j}}\right)},$$

SFs in quark and gluon are 20 ~ 30%

(bkg is negligible (see backup))

# Scale factors in BDT distribution in Method 2 (gammajet+dijet)




SFs in quark are ~20%, in gluon are ~ 30%

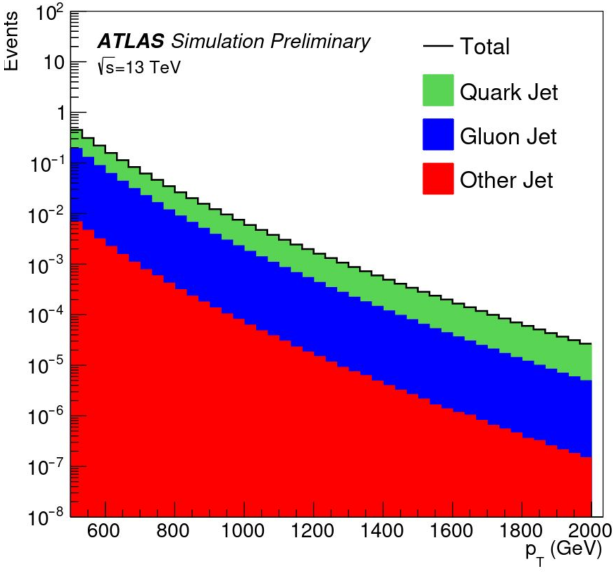
# Summary:

- Using Matrix Method to extract quark / gluon in Data and MC
    - Get the SFs
- Method 3 (using gamma+2jets & Trijet) for validation is ongoing.
- Systematic uncertainties are ongoing.
- Working points needed to be decided.
    - In R20.7 there is 60% quark jet efficiency working point.
    - We plan to define 60% WP unless there are other requestes.
- SFs in other pT range and Data/MC comparison of the input observable distributions in backup slides.

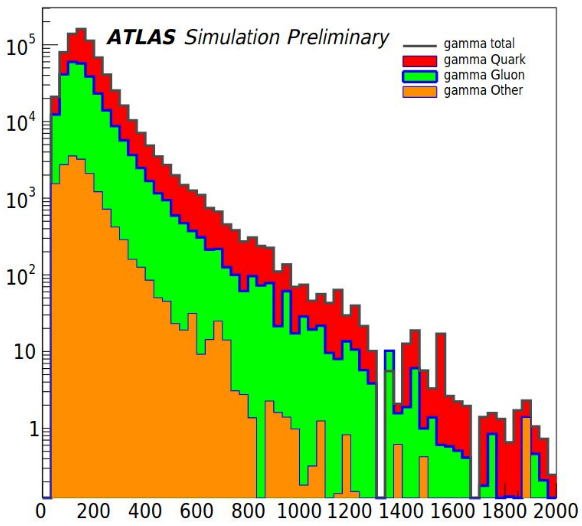
***Thank you***

# *Backup*

# Leading jet pT spectrum



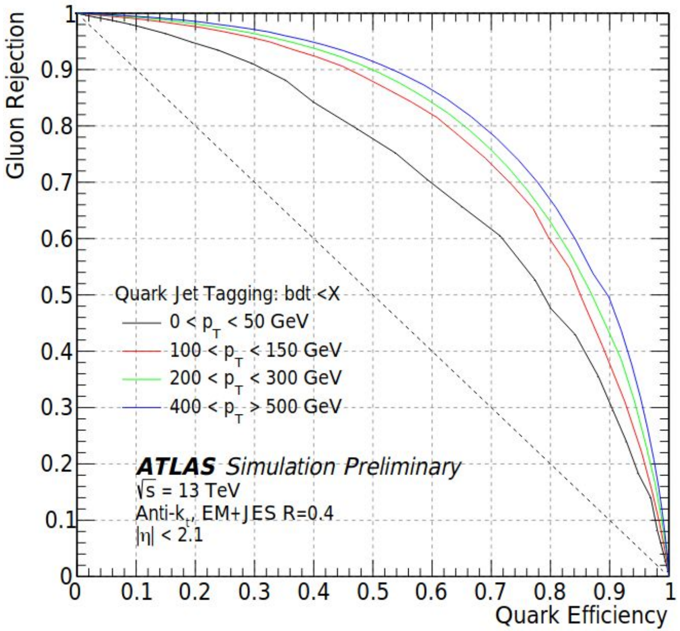

Dijet Sherpa




Gammajet Sherpa

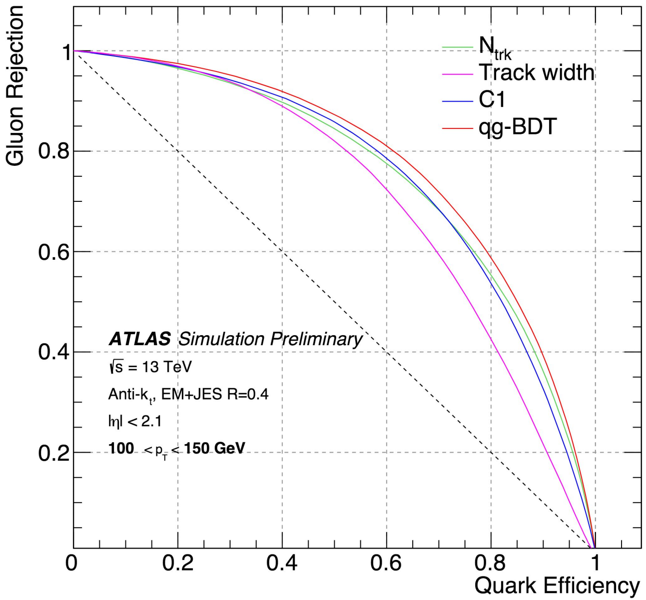
# Method 2 (dijet) : ROC curves

ROC curve for BDT in different pT range



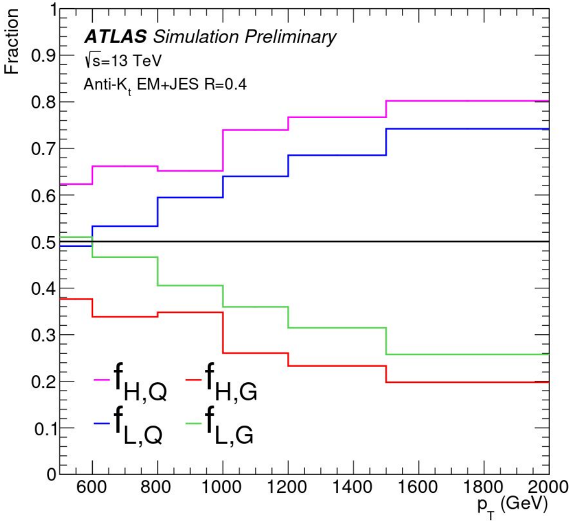

Better in higher pT

ROC curve for different variables in pT: 100~150 GeV

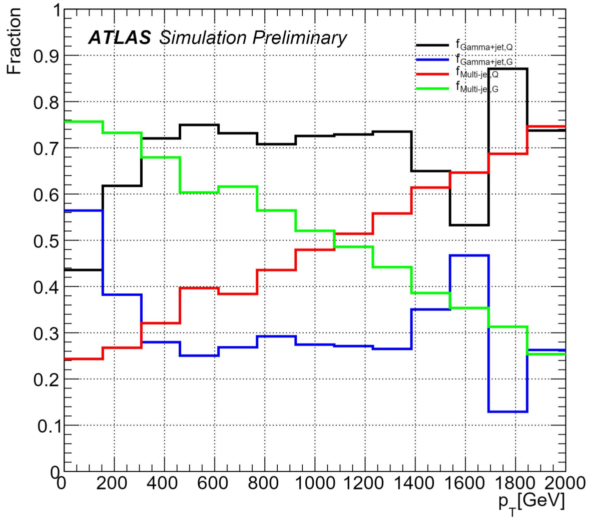



BDT in low pT behaves better than others

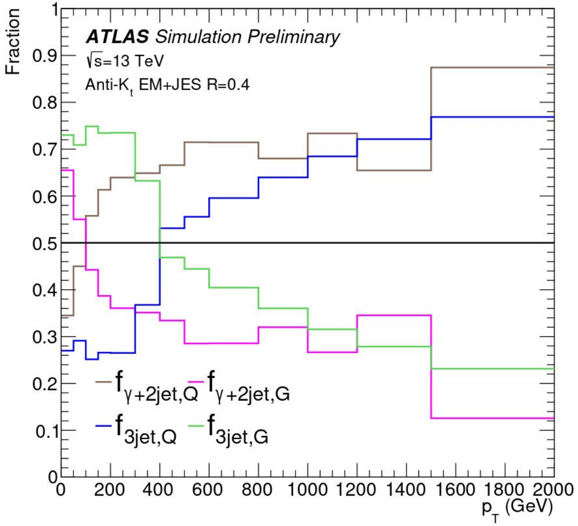
# Q/G fration in pT




method1: dijet-only

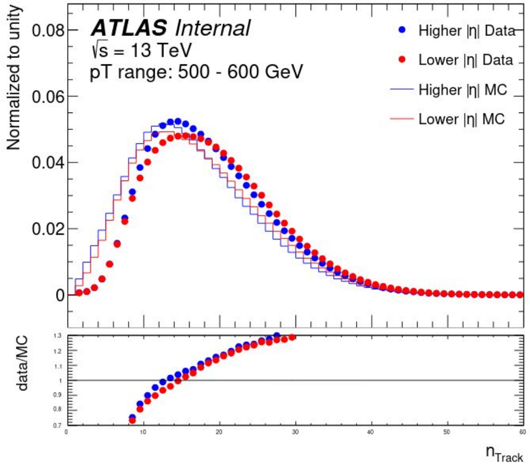


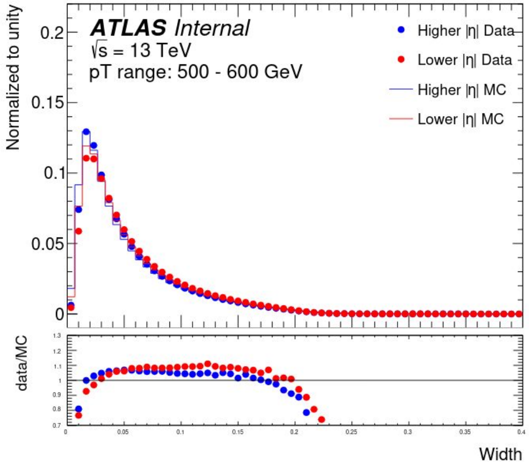

method2:  gamma+jet & dijet

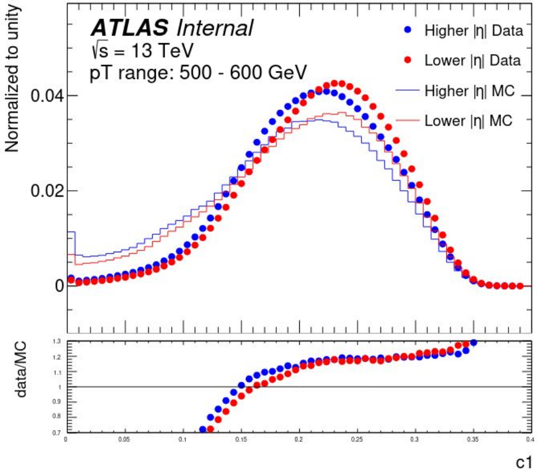



method3:      gamma+2jet & trijet

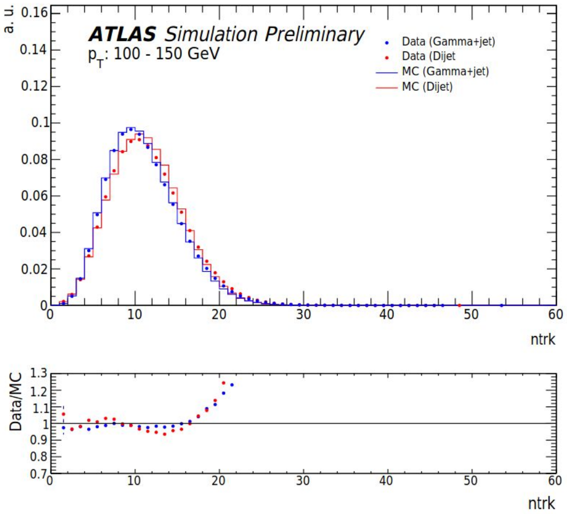
# Track-based variables of BDT (mothod1:dijet)

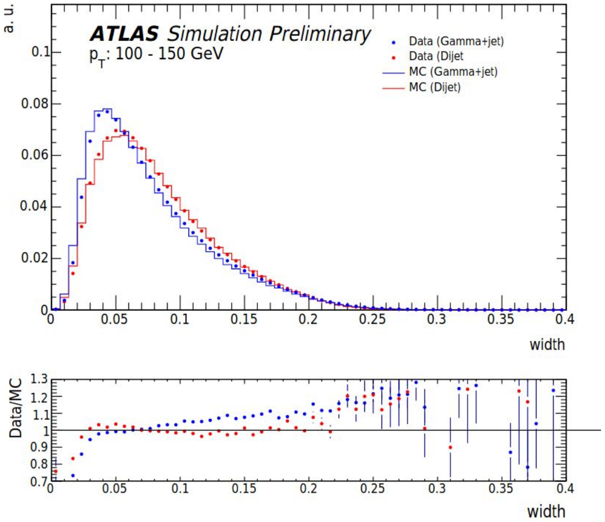


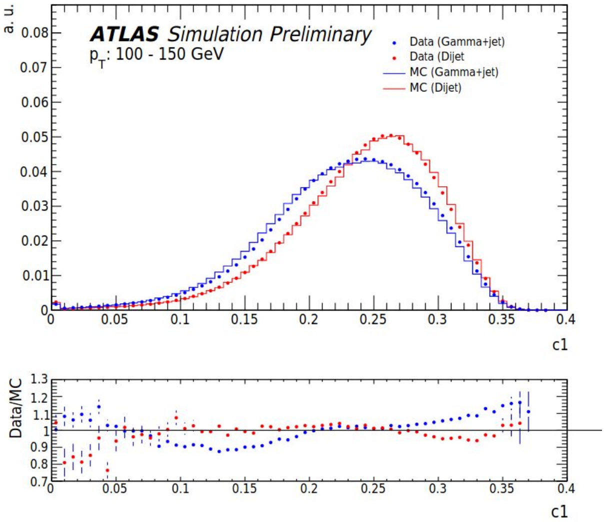



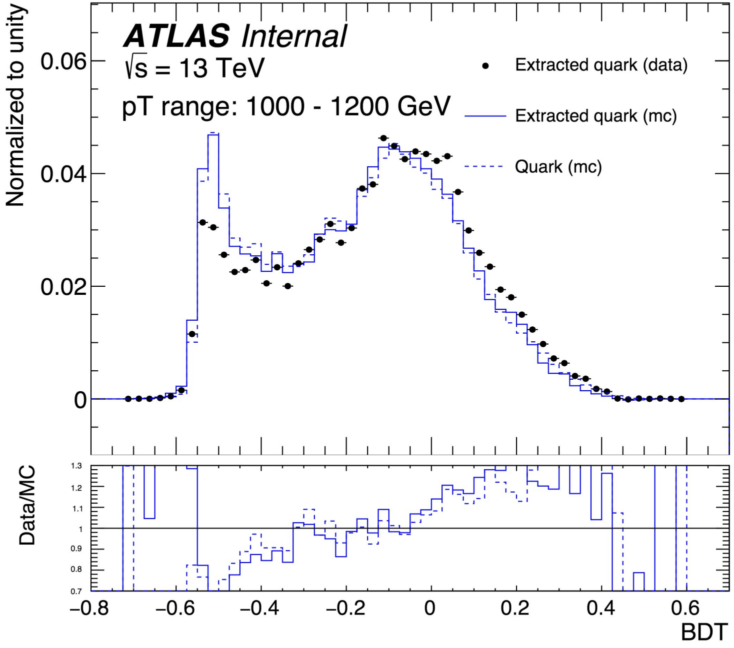
# Track-based variables of BDT（mothod2:Gammajet+dijet)

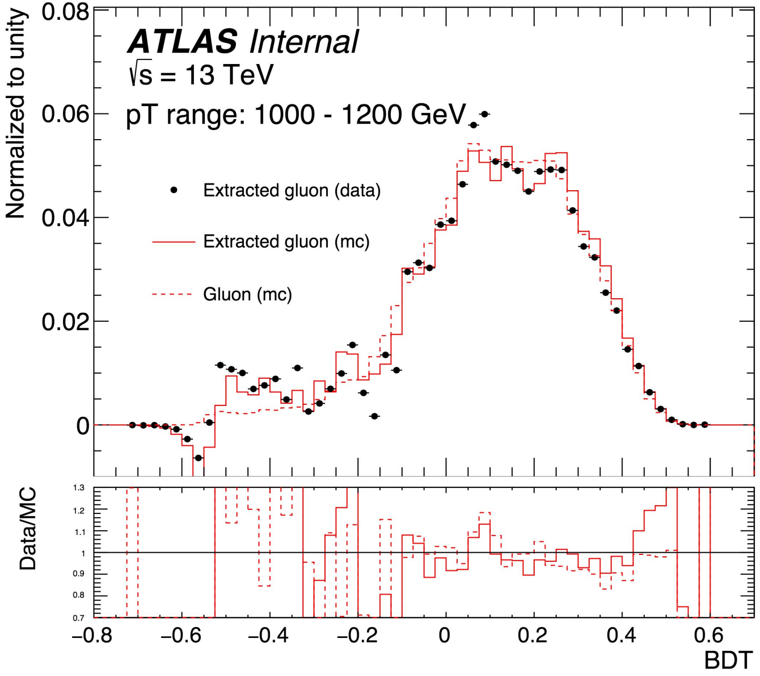


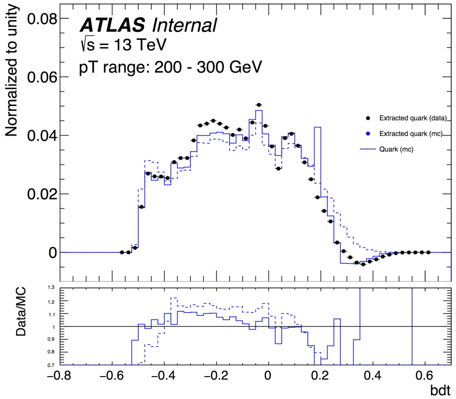



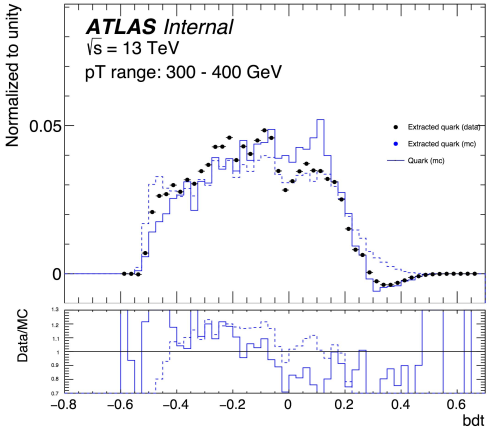
# Scale factors in BDT distribution in Method 1 (dijet - only)

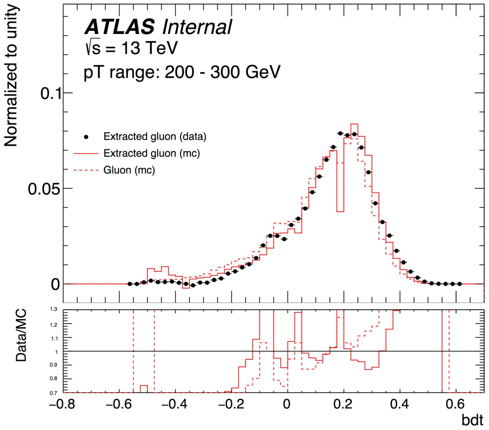


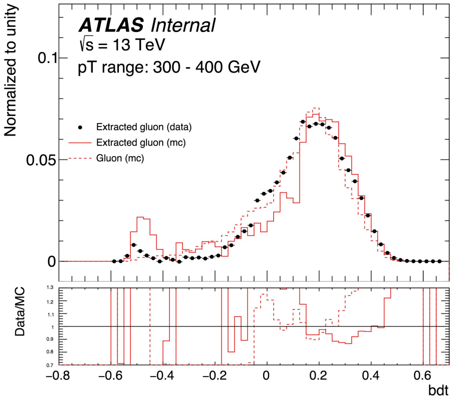



# Scale factors in BDT distribution in Method 2 (gammajet+dijet)

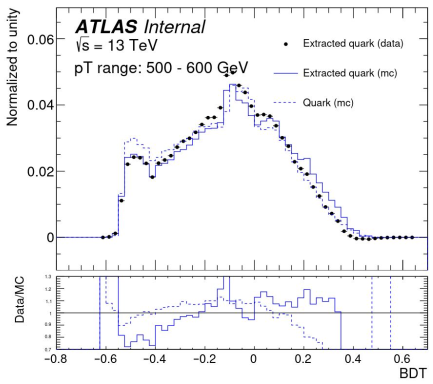


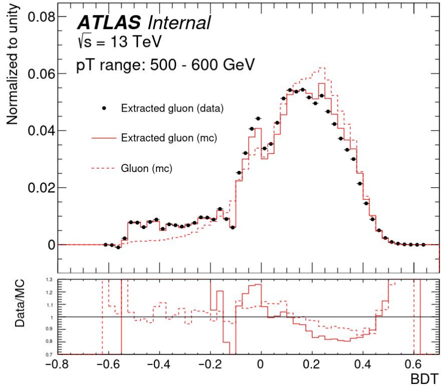



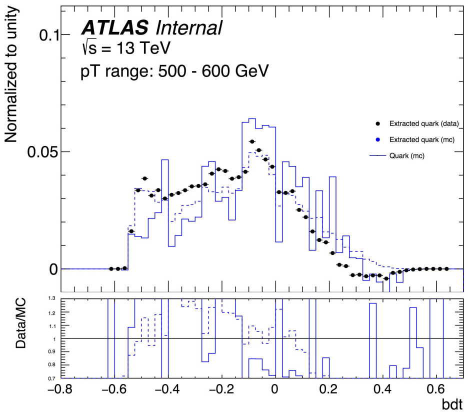



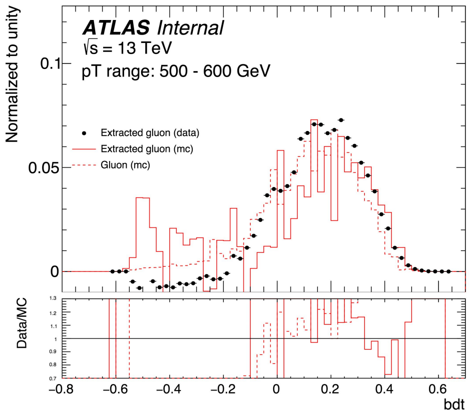
# Overlap pT range in Method 1 and 2




Method 1

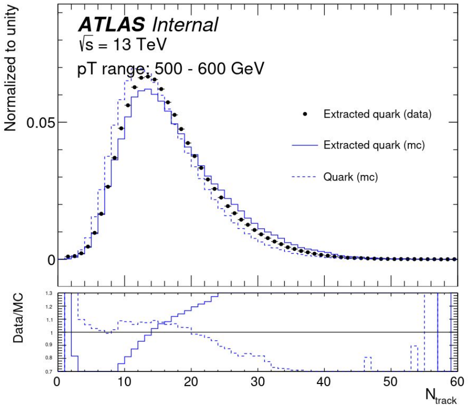


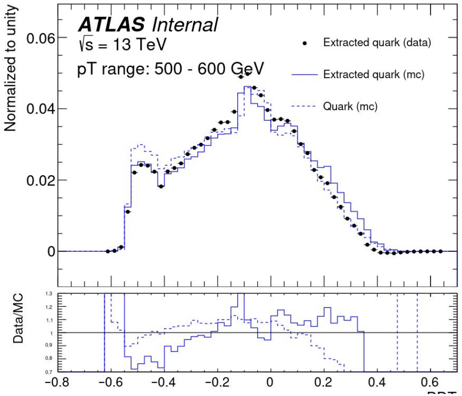



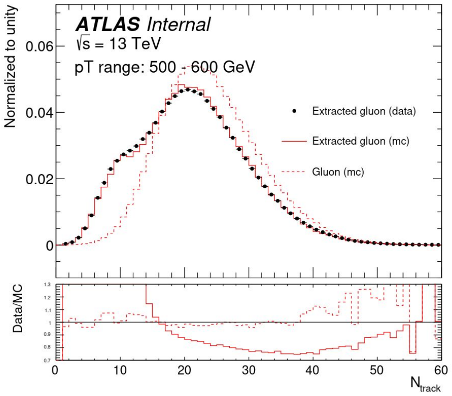

Method 2

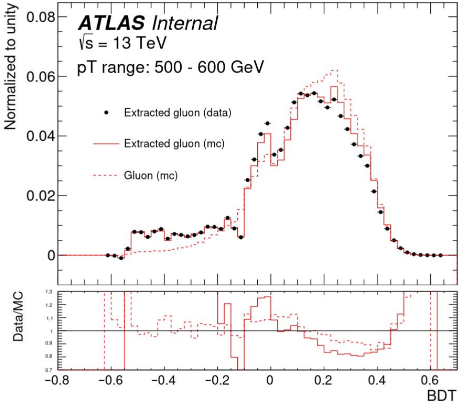
SFs in Method 2 is larger

# Comparison between BDT and NTrack tagger (Method 1)



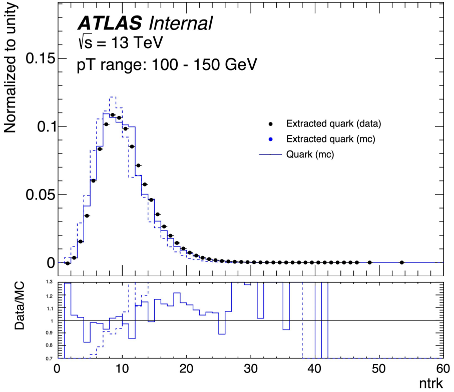

Possible solutions:

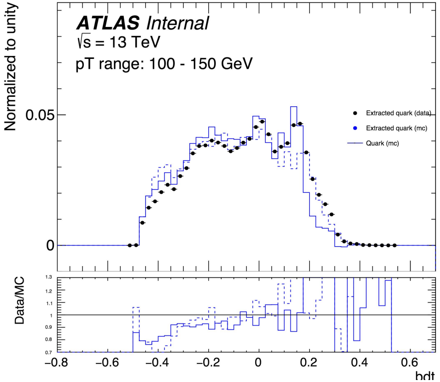
Since tracking efficiency has η-dependence, one can reweight quark or gluon jets in quark-enriched samples by multiplied $w_i$ :

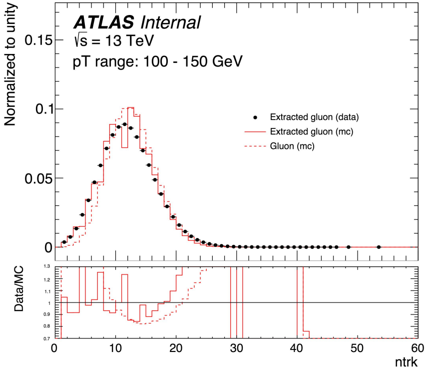
factor: $w_i = n^{(gluon\text{-}enriched)}_i / n^{(quark\text{-}enriched)}_i$

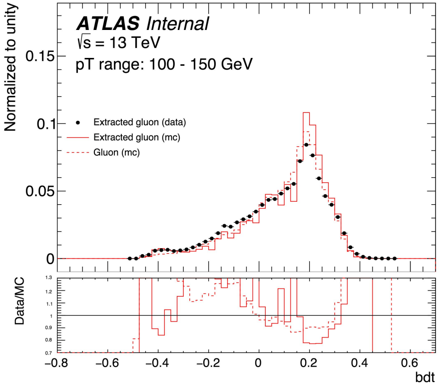
$i = 1,2,.....,N_{bins}$

# Comparison between BDT and NTrack tagger (Method 2)




SFs of BDT are smaller than Ntrk in both quark and gluon.

# Systematic Uncertainties Plan for method-1

- MC closure
- PDF
- Statistics
- Shower scale
- Parton shower
- hadronization
- Matrix Elements

for each BDT bin in each pT range:

$\boldsymbol{\sigma}_{\mathbf{ToT}}$ = RMS (all $\boldsymbol{\sigma}$ listed below)

$\boldsymbol{\sigma}_{\mathbf{mc\ closure}} = \mathrm{binval}_{\mathrm{extracted\ mc}} - \mathrm{binval}_{\mathrm{truth\text{-}labeled\ mc}}$

$\boldsymbol{\sigma}_{\mathbf{pdf}} = \mathrm{RMS}(\mathrm{binval}_{\mathrm{sherpa[i]}})$

$\boldsymbol{\sigma}_{\mathbf{stat}}$ = from bootstrapping:

repeat this many times and take the RMS: for the higher and lower data, replace the data bin content by a random variable samples from a Poisson distribution with mean given by the nominal value. Then extract the quark and gluon bin content.

(binval : quark (or gluon) bin value)

# Systematic Uncertainties Plan for method-2

- MC closure
- PDF
- Statistics
- Fake photon estimation

other uncertainties are included in Pythia - Sherpa differences due to lack of samples.