NTUA Top Tagger

Tag & Probe methodology

G. Bakas, K. Kousouris, I. Papakrivopoulos, G. Tsipolitis

Motivation

- The main background for this analysis is QCD
- A data driven method is used for subtracting it
- The method relies on the assumption that by inverting the b-tagging requirement in the signal region (SR) we can have the shape of the QCD contribution
- The tagger is required **not to** use b-tagging information and the full range of the subjet mass
  - An in-house BDT was developed to overcome this limitation
  - DeepAK8 uses b-tagging
  - Older top taggers use jet mass cuts

CMS top tagger $R = 0.8$	HEPTopTagger $R = 1.5$		OptimalR $R = 0.5$–1.5
$\delta_p > 0.05$	$f_{\mathrm{drop}} = 0.8$	$m_{23}/m_{123} > 0.35$	same as HEPTopTagger
$A = 0.0004$	$m_{\mathrm{cut}} = 30\,\mathrm{GeV}$	$0.2 < \arctan\frac{m_{13}}{m_{12}} < 1.3$	$\Delta R = 0.1$
$N_{\mathrm{sub}} \geq 3$	$R_{\mathrm{filt}}^{\mathrm{max}} = 0.3$	$f_W = 0.15$	$m_{\mathrm{rec}}^{1.5} - m_{\mathrm{rec}} > 0.2 m_{\mathrm{rec}}^{1.5}$
$m_{\mathrm{min}} > 50\,\mathrm{GeV}$	$N_{\mathrm{filt}} = 5$	$140 < m_{123} < 220\,\mathrm{GeV}$	$\Delta R_{\mathrm{opt}} < 0.5$
$140 < m_{\mathrm{jet}} < 220\,\mathrm{GeV}$	$p_{T,\mathrm{sub}} > 30\,\mathrm{GeV}$		

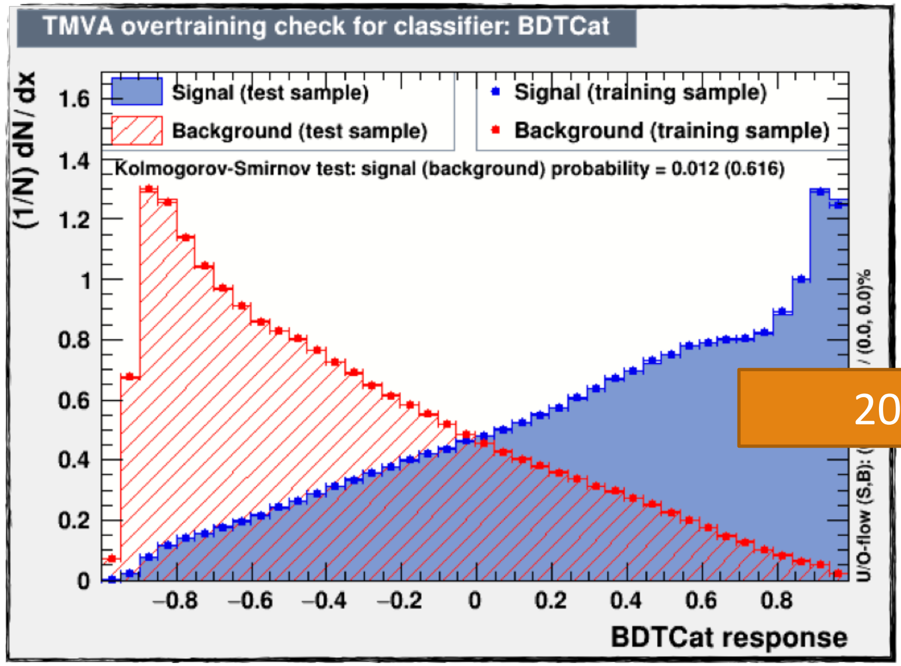
BDT Output

- In house developed top tagger, for top candidate jets
  - BDT based
  - Input variables:
    - N-subjetiness: $\tau 1$, $\tau 2$, $\tau 3$
    - Energy correlation functions (ECF) ECFB1N2, ECFB1N3, ECFB2N2, ECFB2N3
    - Soft drop mass of the leading and subleading subjets
    - Fraction of the jet over the of all the jets in the event
    - No b-tagging requirements

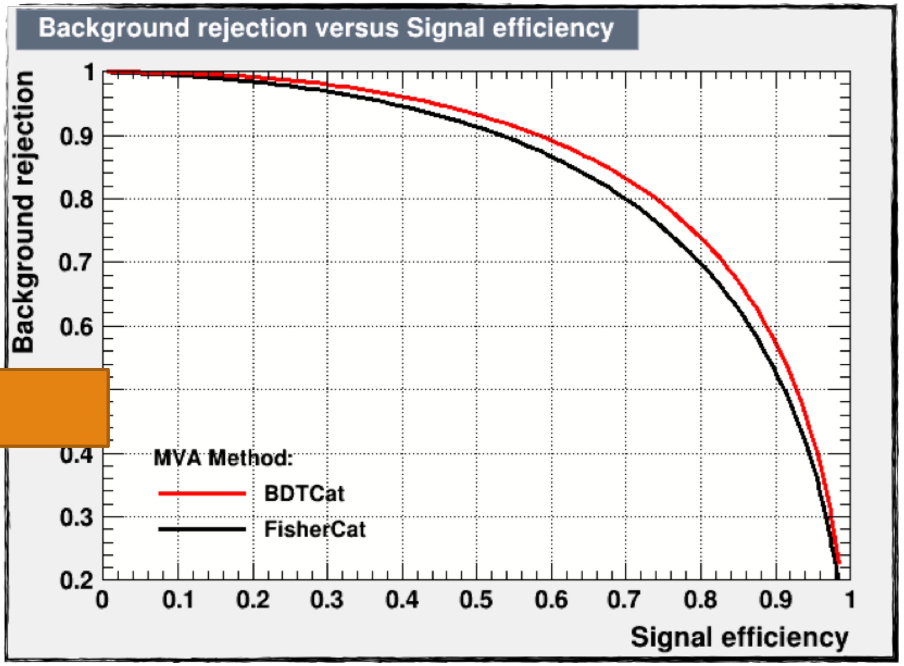
- Phase space split in categories based on the pt of the jet:
  - [400, 600) GeV
  - [600, 800) GeV
  - [800, 1200) GeV
  - [1200, Inf) GeV

The use of DeepAK8 was investigated but it uses b-tagging so it is not applicable in our use case

- Different training and working point for each year (Signal) and QCD (Bkg) samples used in the training



2018



Overview

- BDT Input and Output in the SR_B Region
  - SR_B : Baseline selection + tight Mass Cut (120,220) GeV, **no TopTagger** Selection
  - Leading + subleading in different pT regions:
    - [400,600], [600,800], [800, Inf)
    - [400,500], [500,600], [600, Inf)
  - **Find** Data vs MC Input and Output for UL our Analysis here

Fraction of events used in the cross section measurement that are also used in the Top Tagger SF measurement is of the order of 35%

- Top Tagger Scale Factors
  - Data is subtracted QCD and Subdominant bkgs (MC) so that the data sample is pure

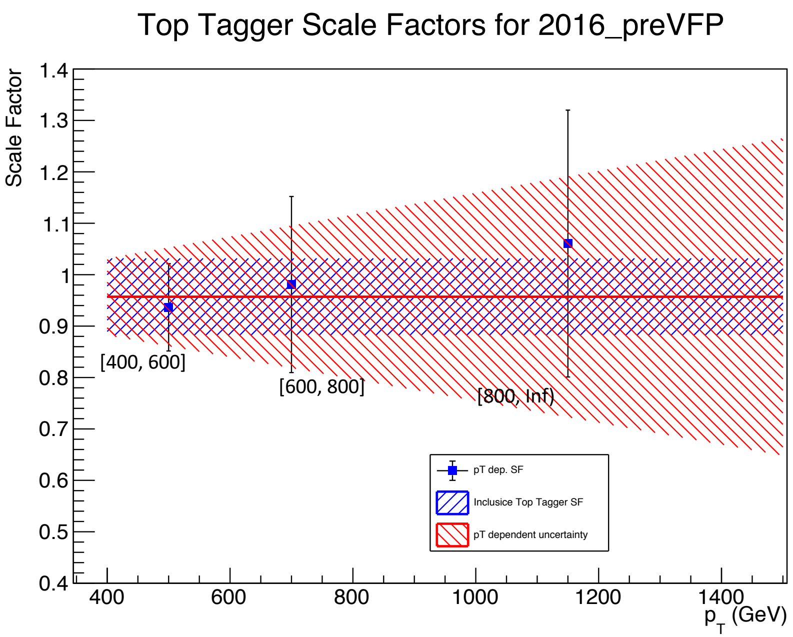
$$efficiency = \frac{Tight\ \&\ SR}{Tight\ \&\ Probe} = \frac{\#\ (1\ jet\ pass\ baseline + Tight\ TopTagger\ Cut\ AND\ 1\ jet\ pass\ SR)}{\#\ (1\ jet\ pass\ baseline + Tight\ TopTagger\ Cut\ AND\ 1\ jet\ pass\ only\ baseline)}$$

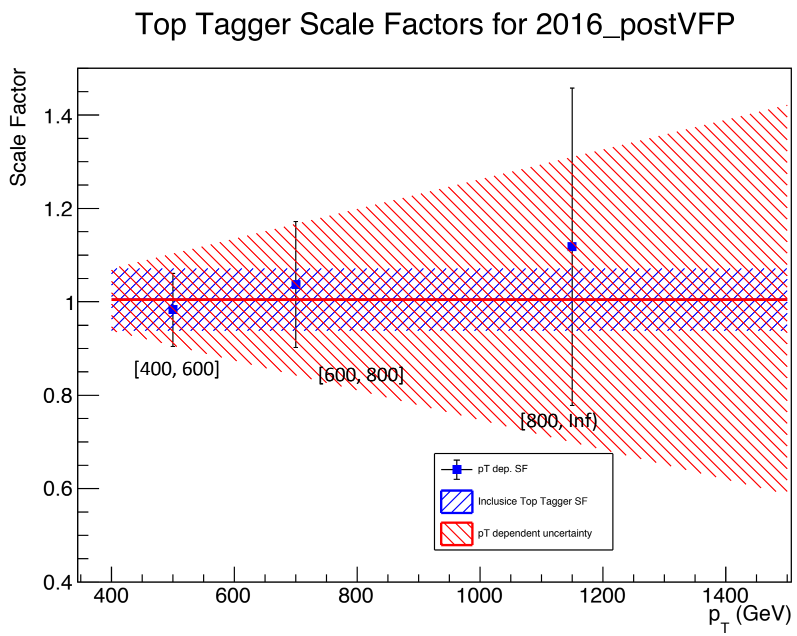
  - Implemented Randomization (check random jet) to fill histogram to avoid pT bias
  - Divide the phase space into pT regions: [400-600]GeV, [600-800]GeV, [800-Inf]GeV
  - For the QCD estimation, we perform a fit in both regions (Tight & Probe, Tight & SR):
    - Shape of QCD is estimated from Data while inverting btagging requirement
    - # QCD events in each region is calculated from fit using the Leading JetMassSoftDrop variable
    - To scale the ttbar → fit the Leading JetMassSoftDrop in each region and get the signal strength
    - For the evaluation of Signal distribution from data, we do the following:

$$\forall\ region: S(x) = D(x) - N_{QCD} d_0(x) - Sub.Bkg(x)$$

Scale Factors

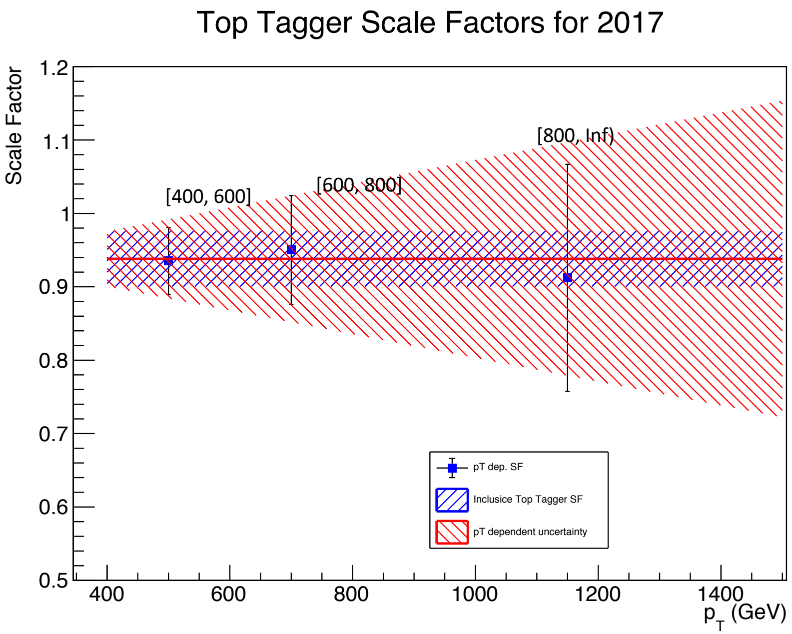
2016 preVFP

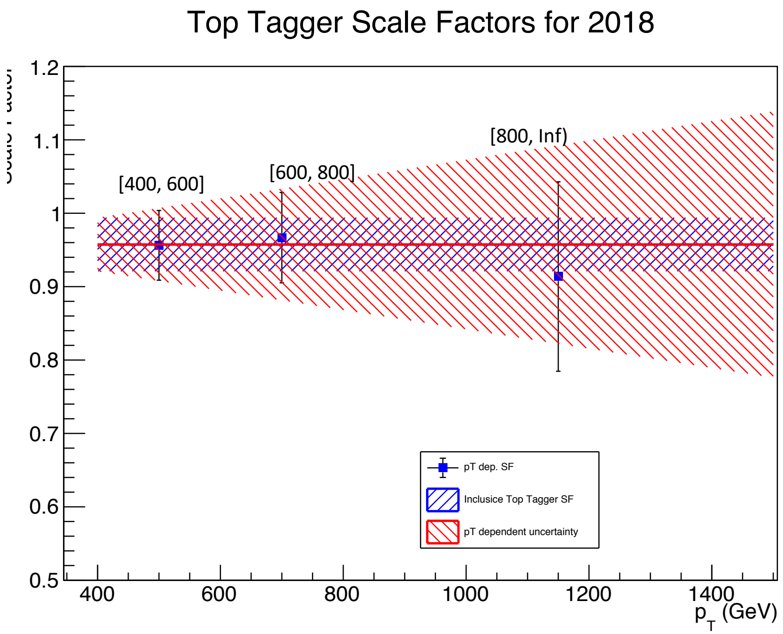




2016 postVFP

2017

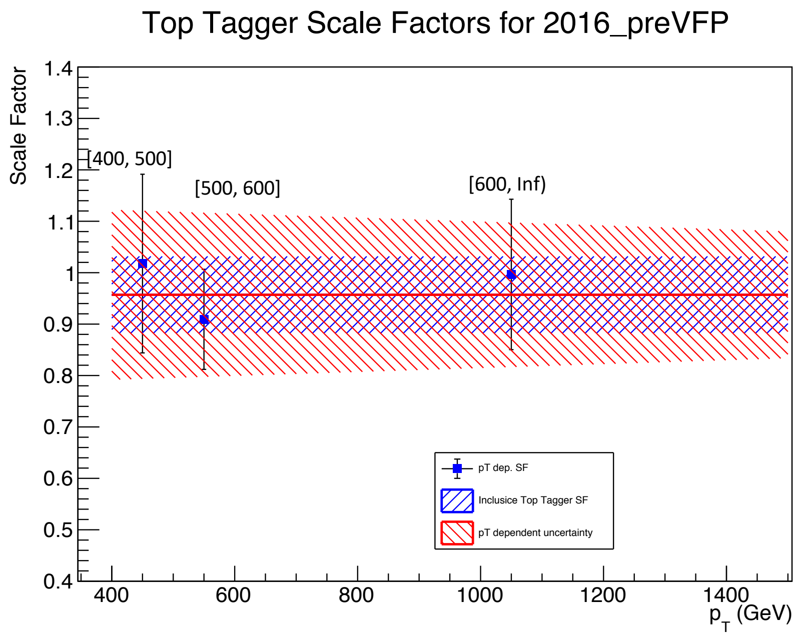


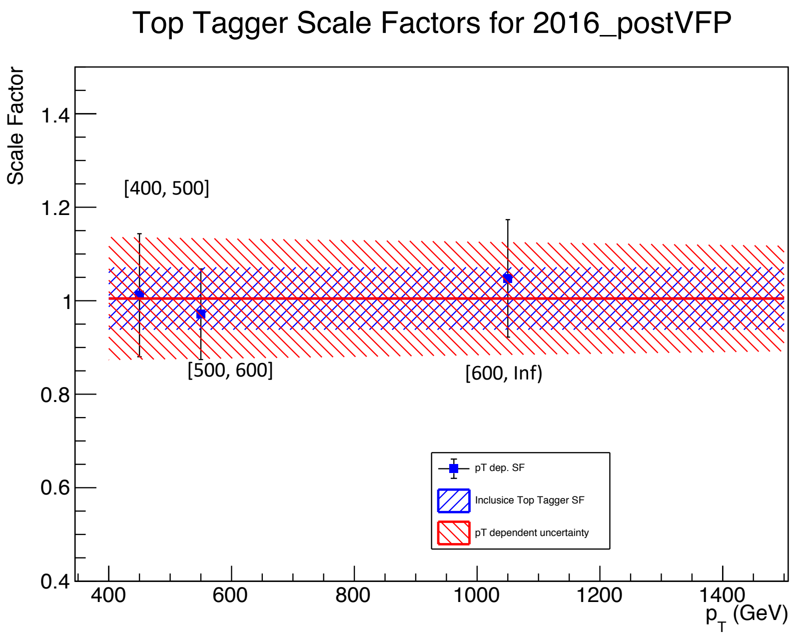


2018

Scale Factors

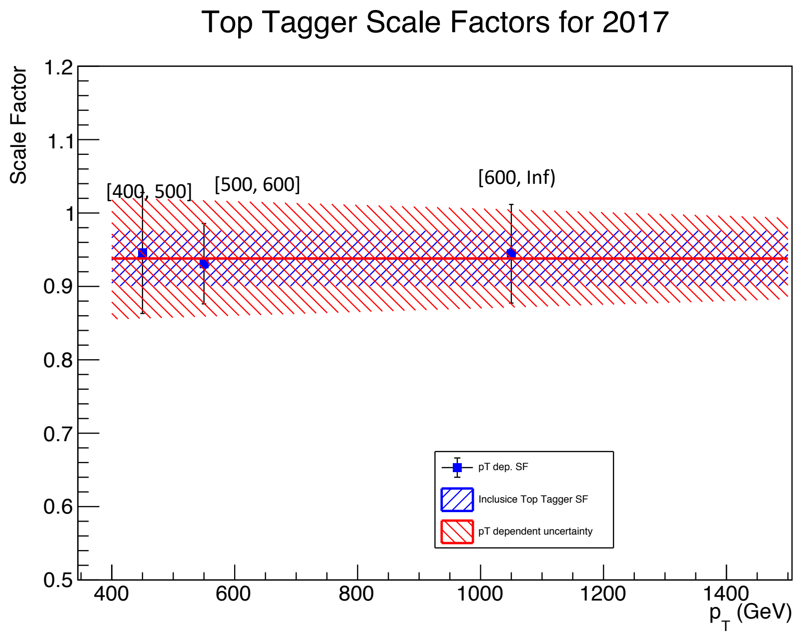
2016 preVFP

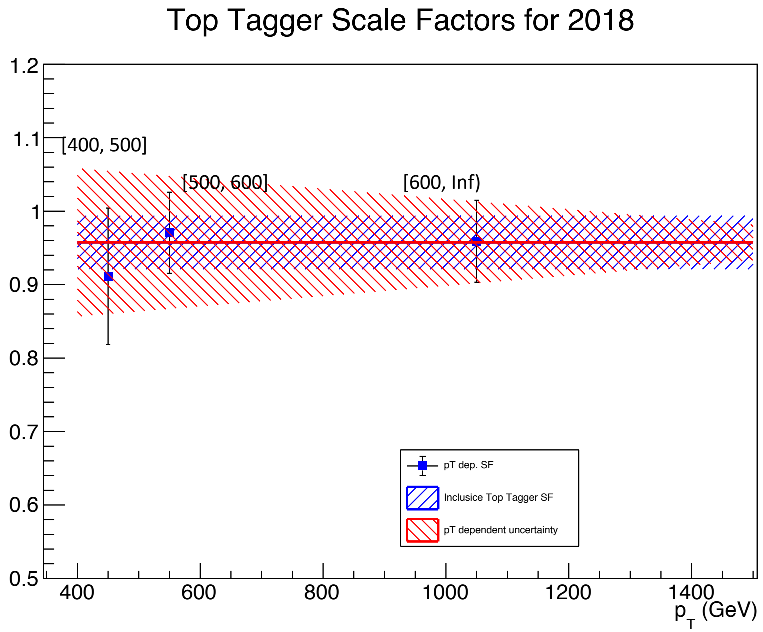




2016 postVFP

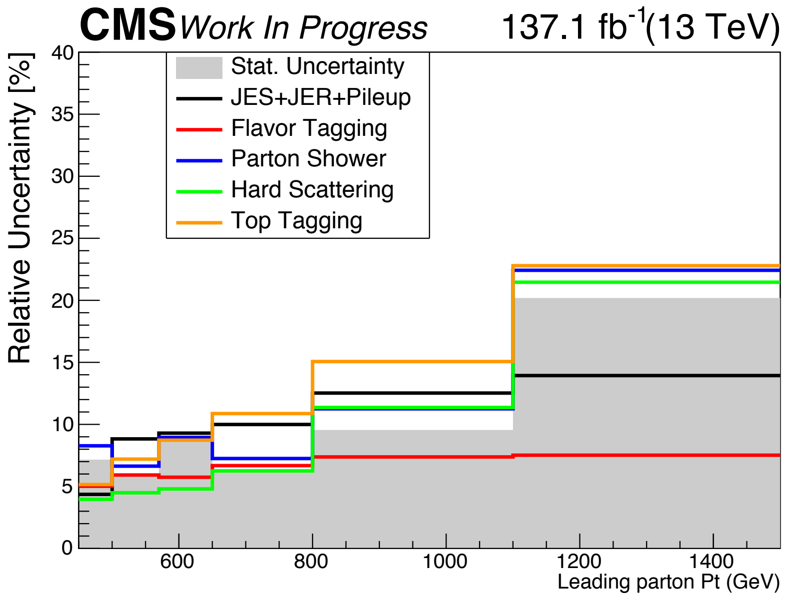
2017

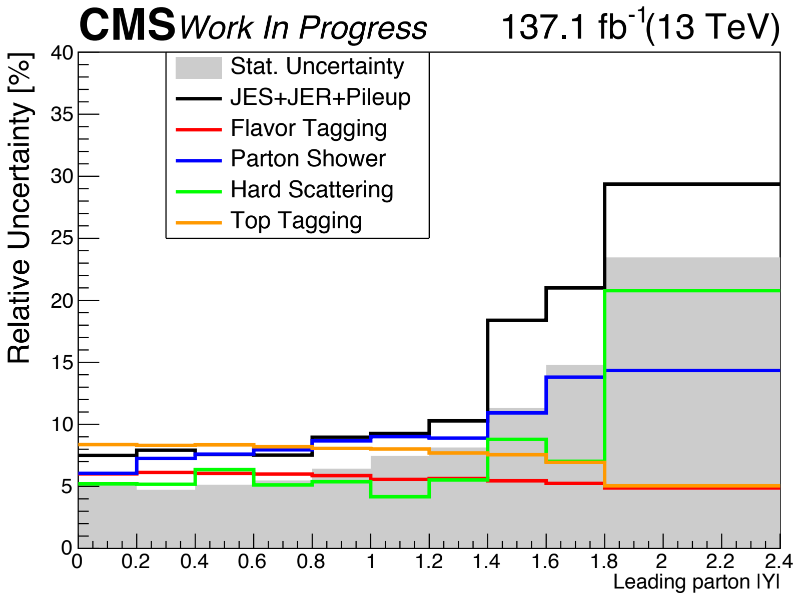


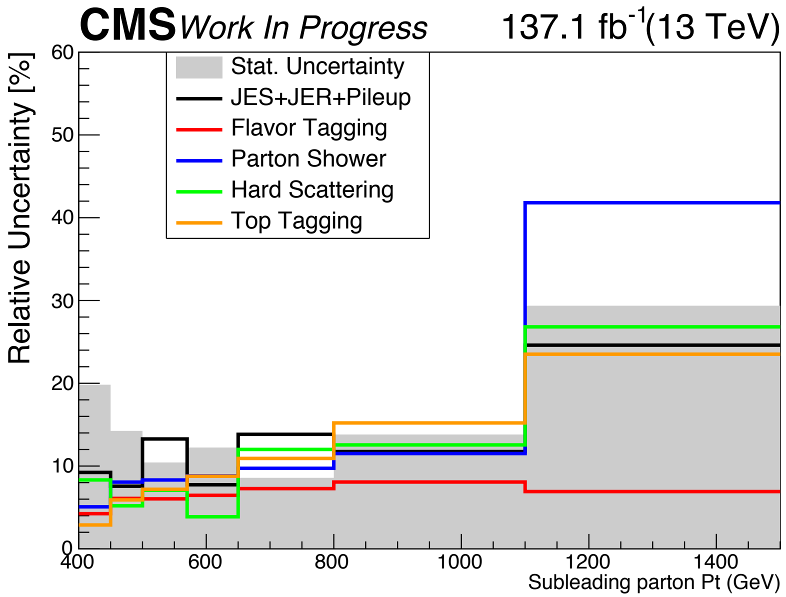


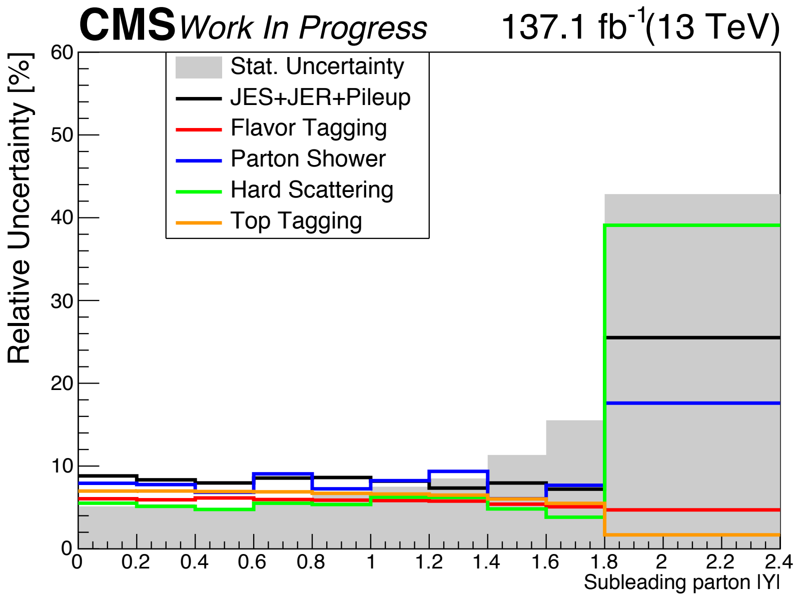
2018

Systematic Uncertainties Breakdown

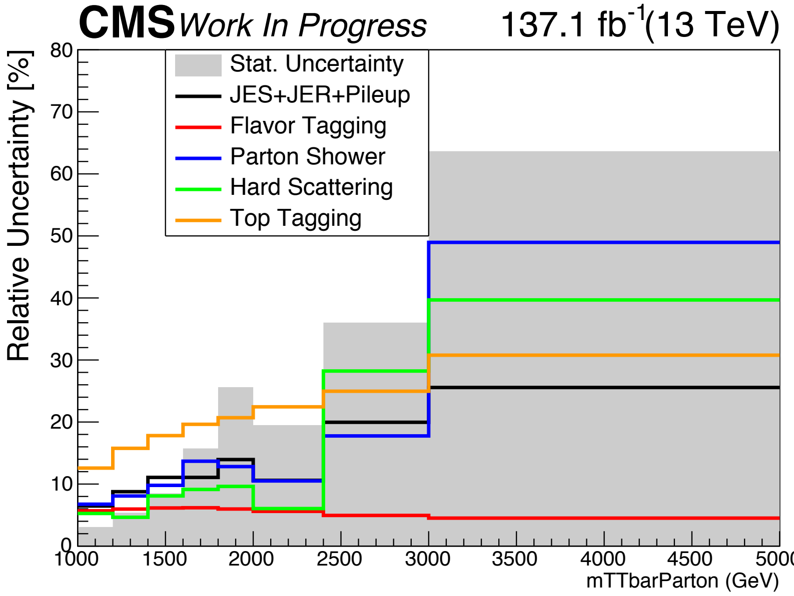


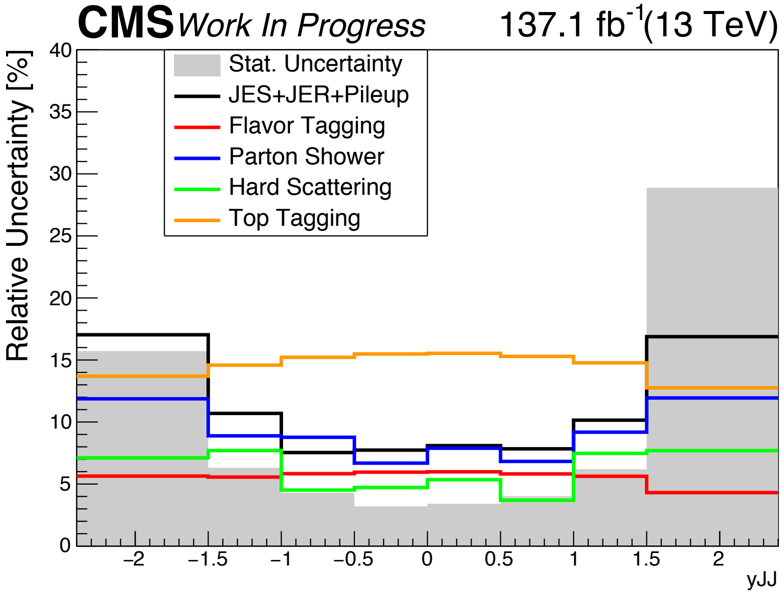


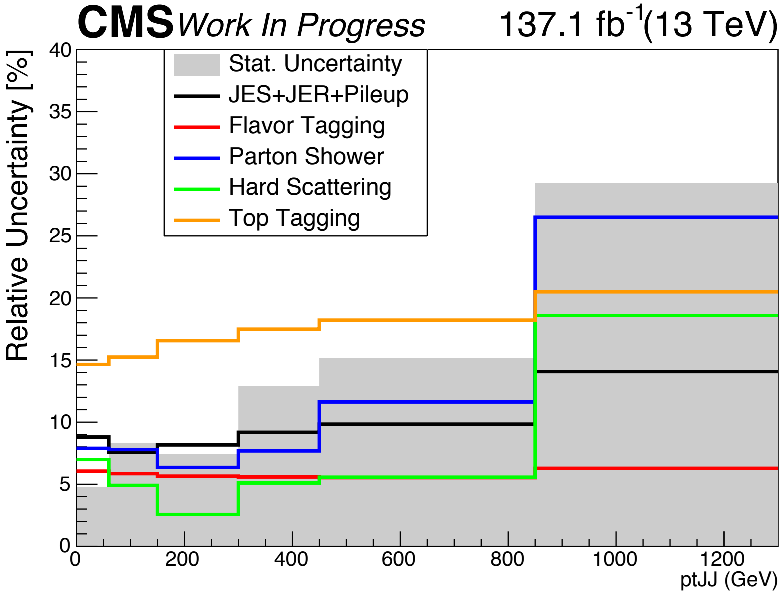


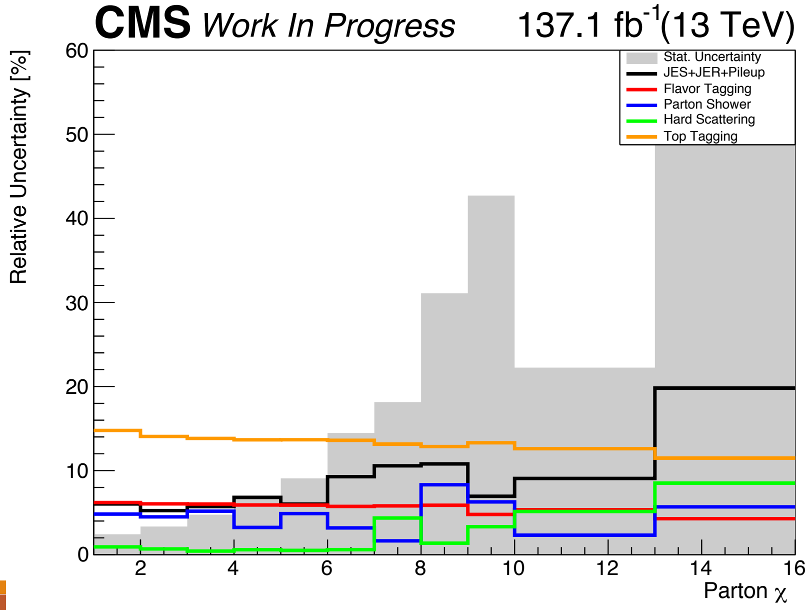


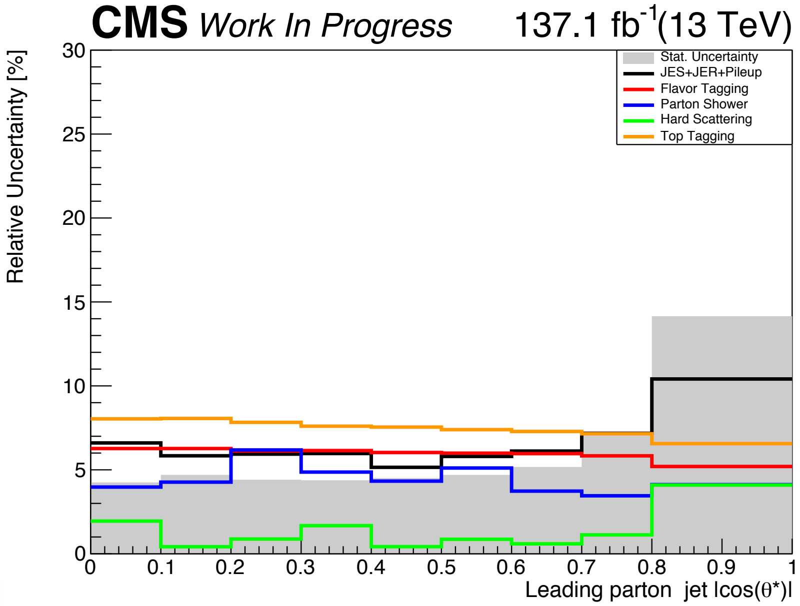
Systematic Uncertainties Breakdown







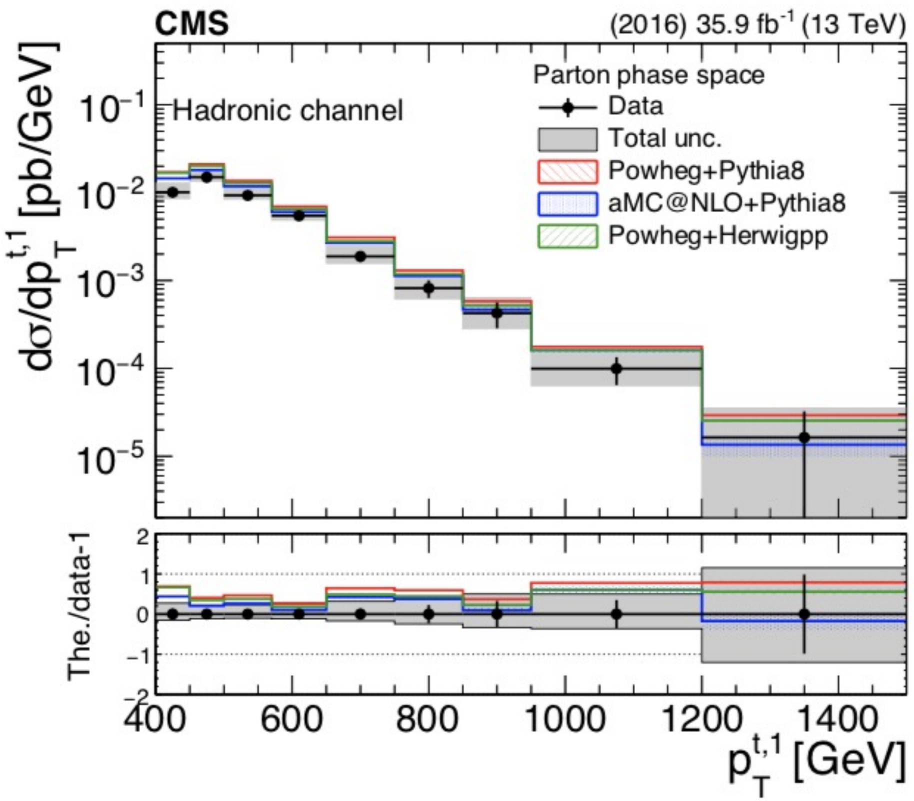


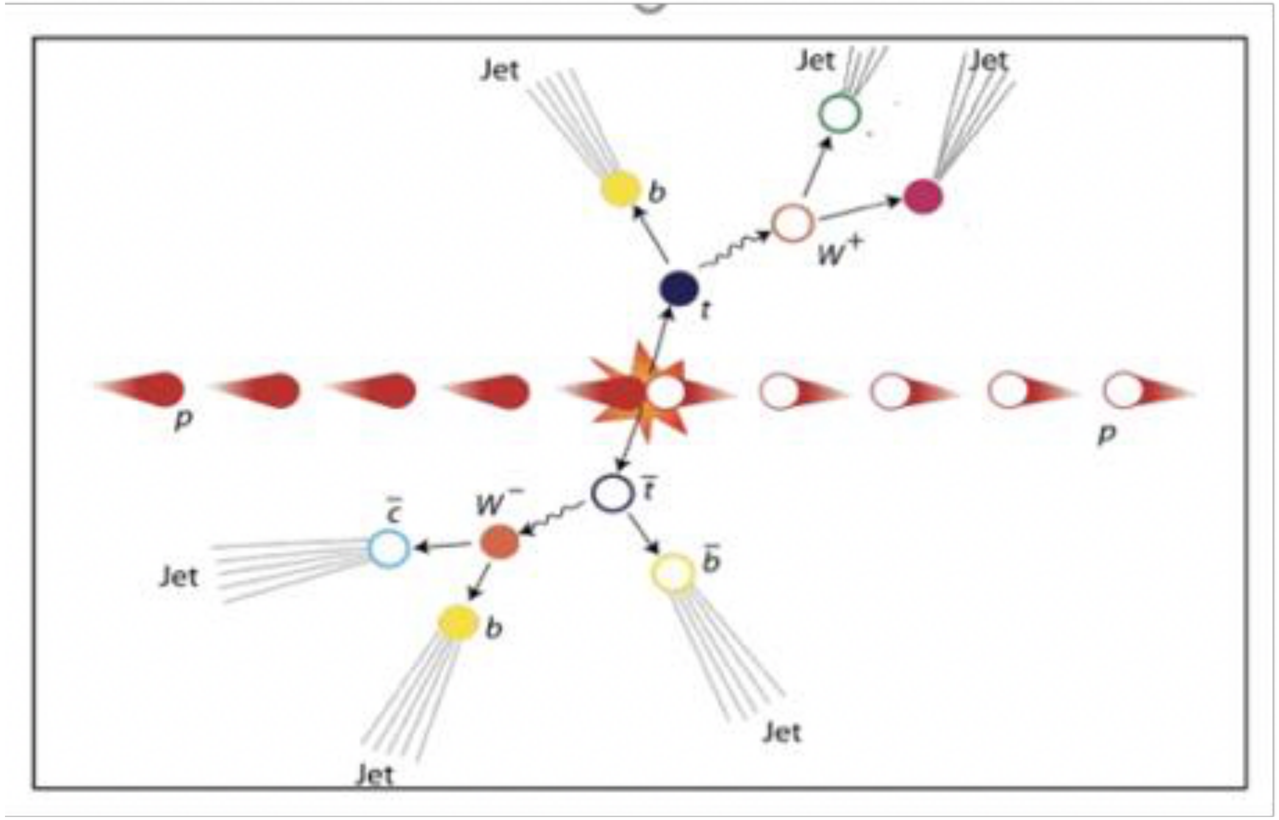


Backup

Analysis Overview

- Differential cross section for boosted ttbar pair fully hadronic final state
- Trying to identify two big jets that contain the products of the top/anti-top decay.





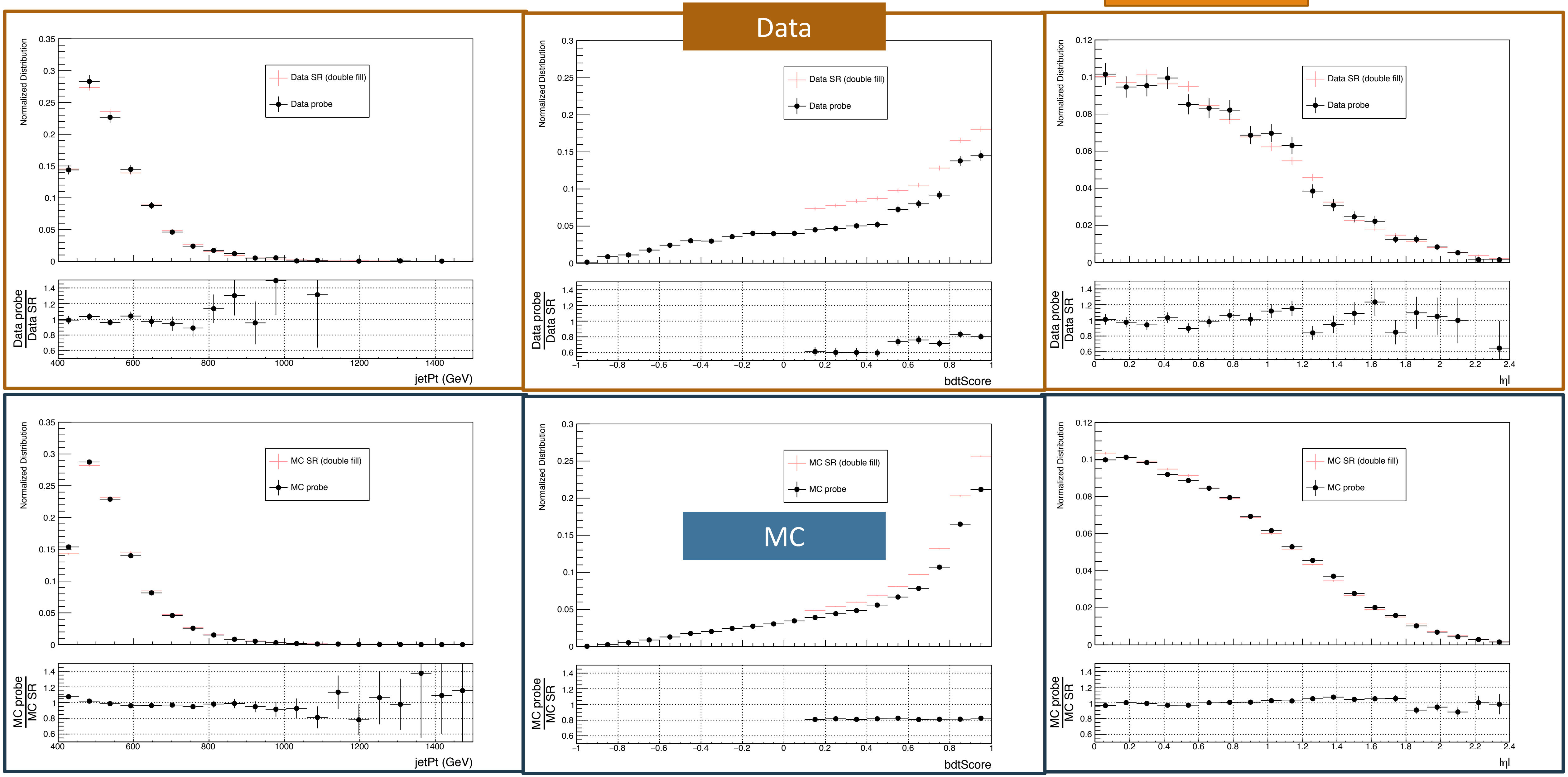
- A NN, for tagging ttbar events was used in TOP-18-013
- A BDT for tagging jets as tops is used in this analysis

Motivation

- The main background for this analysis is QCD
- A data driven method is used for subtracting it
- The method relies on the assumption that by inverting the b-tagging requirement in the signal region (SR) we can have the shape of the QCD contribution
- This can be verified by a set of closure tests
- The tagger is required to not use b-tagging information
- The use of DeepAK8 was investigated but it uses b-tagging so it is not applicable in our use case
- An in-house BDT was developed to overcome this limitation

Shape Comparison of the probe jet and the jets used for the measurement (double Filled)

2018



Signal Selection

Variables	Selected Cut
pT leading jets	> 450 GeV
pT 2nd leading jets	> 400 GeV
Njets	> 1
N leptons	= 0
|eta| (both leading jets)	< 2.4
mJJ	> 1000 GeV
jetMassSoftDrop (only for fit)	(50,300) GeV
Top Tagger	> 0.2
B tagging (2 btagged jets)	> Medium WP
Signal Trigger	

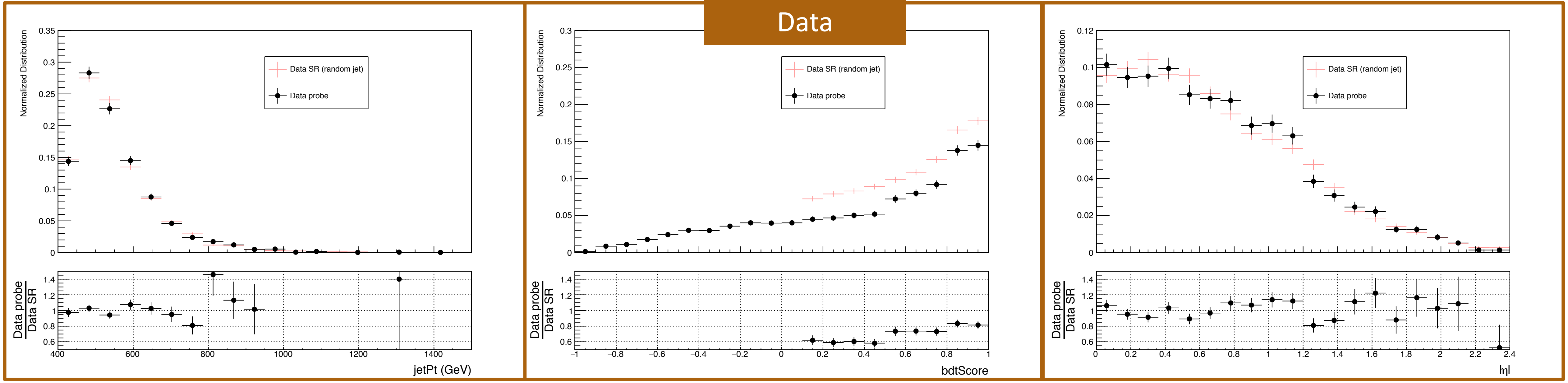
Control Region Selection

Variables	Selected Cut
pT leading jets	> 450 GeV
pT 2nd leading jets	> 400 GeV
N leptons	= 0
|eta| (both leading jets)	< 2.4
mJJ	> 1000 GeV
jetMassSoftDrop (only for fit)	(50,300) GeV
Top Tagger	> 0.2
B tagging (0 btagged jets)	< Medium WP
Control Trigger	

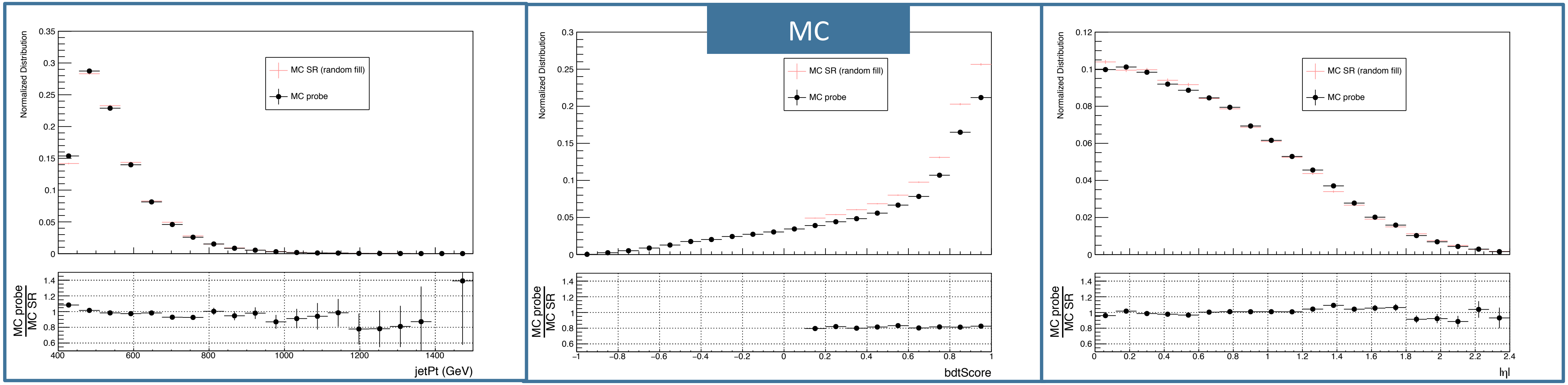
Shape Comparison of the probe jet and the jets used for the measurement

2018

Data



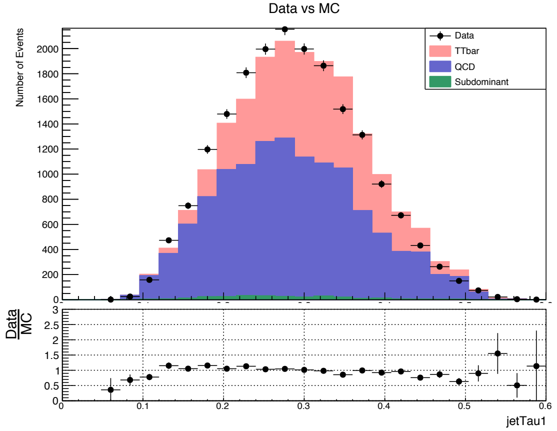
MC

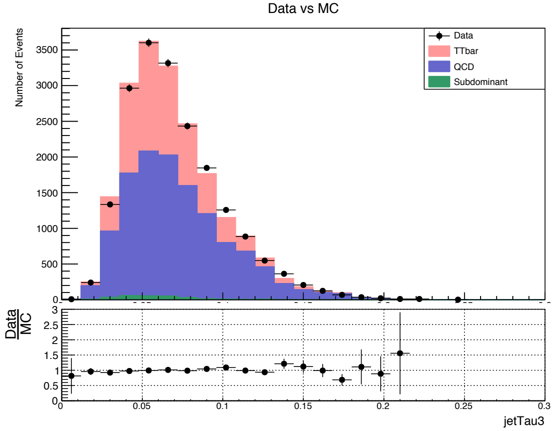


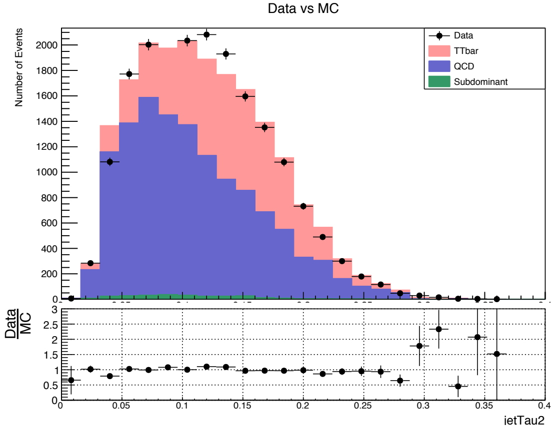
Top Tagger Input Variables

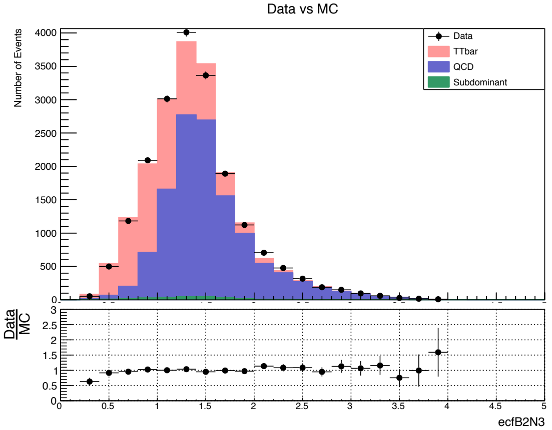
2018

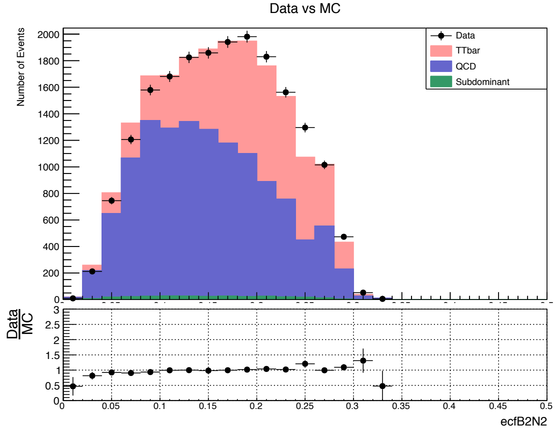
Leading jet

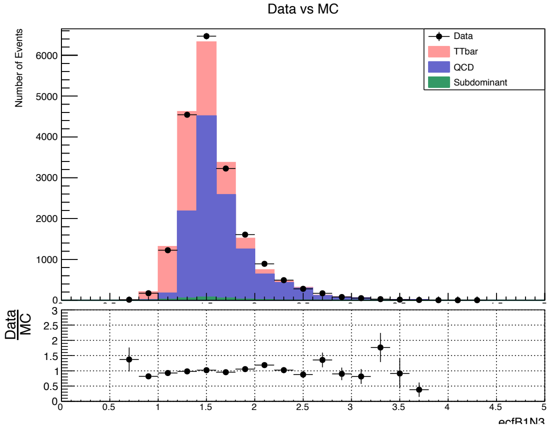


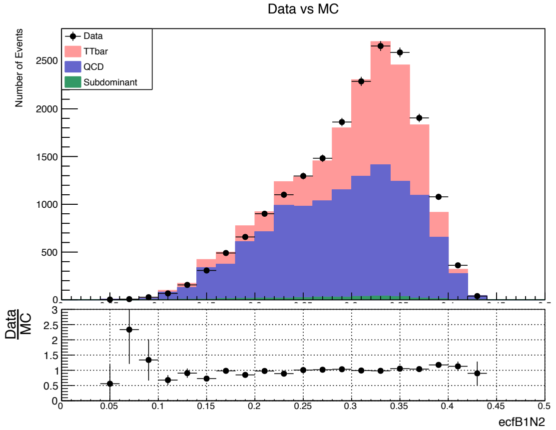


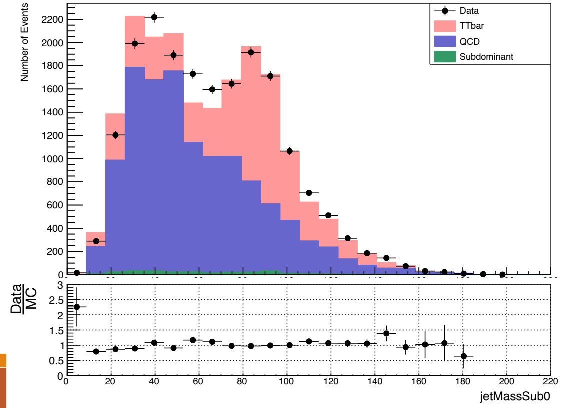


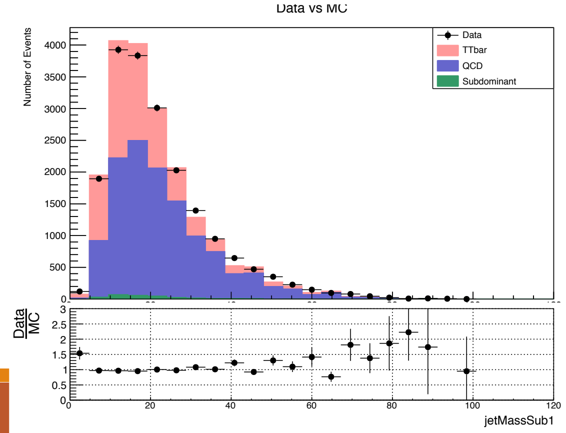


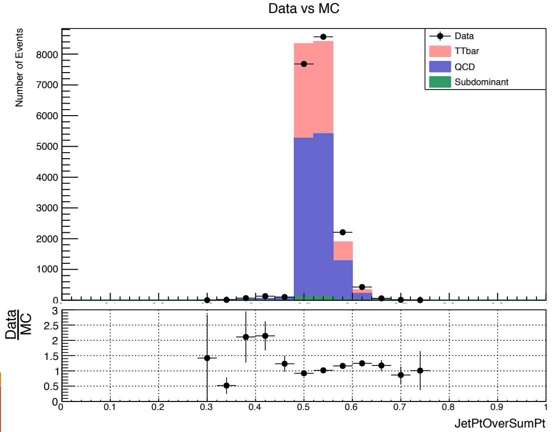








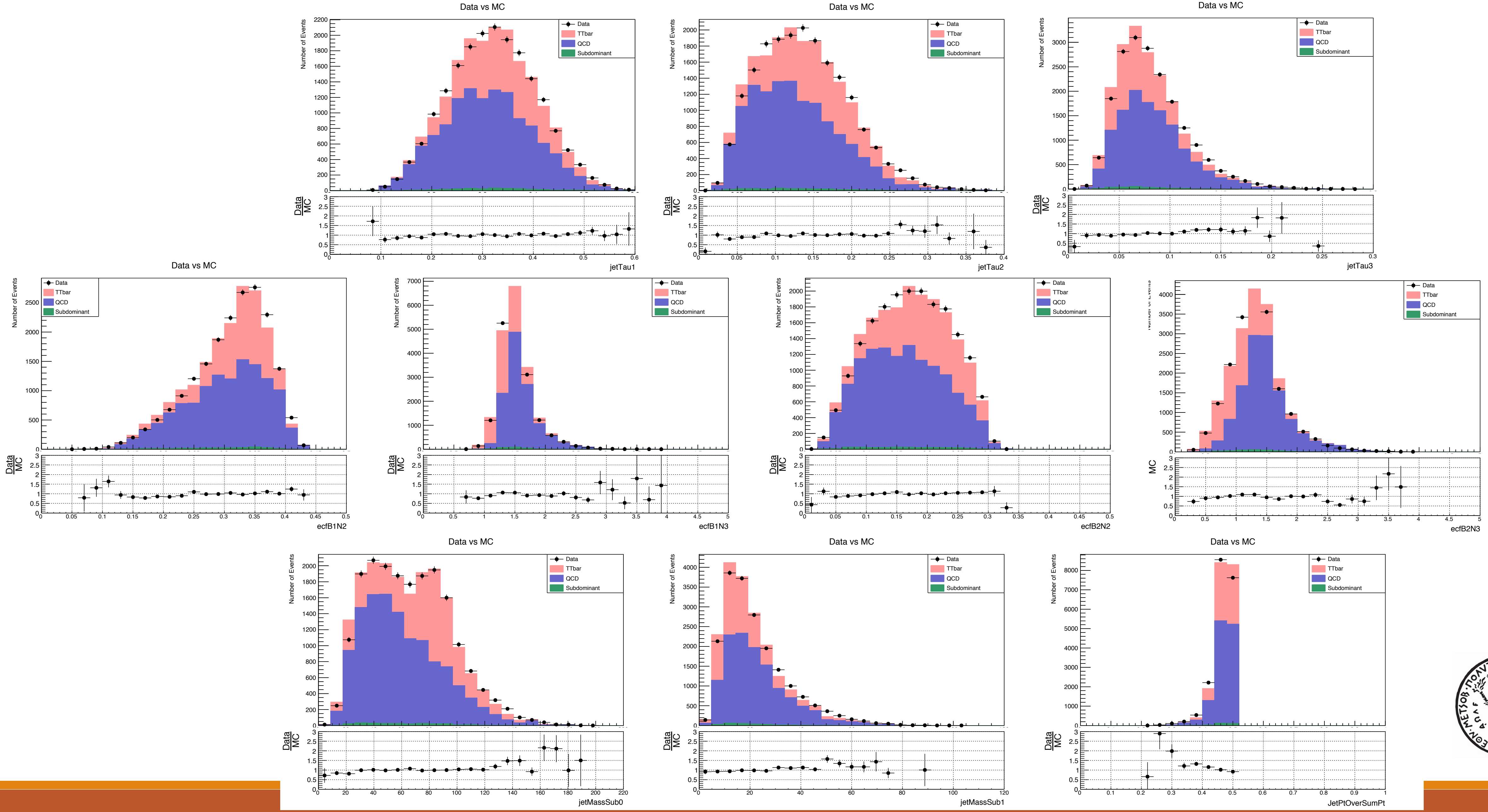




Top Tagger Input Variables

2018

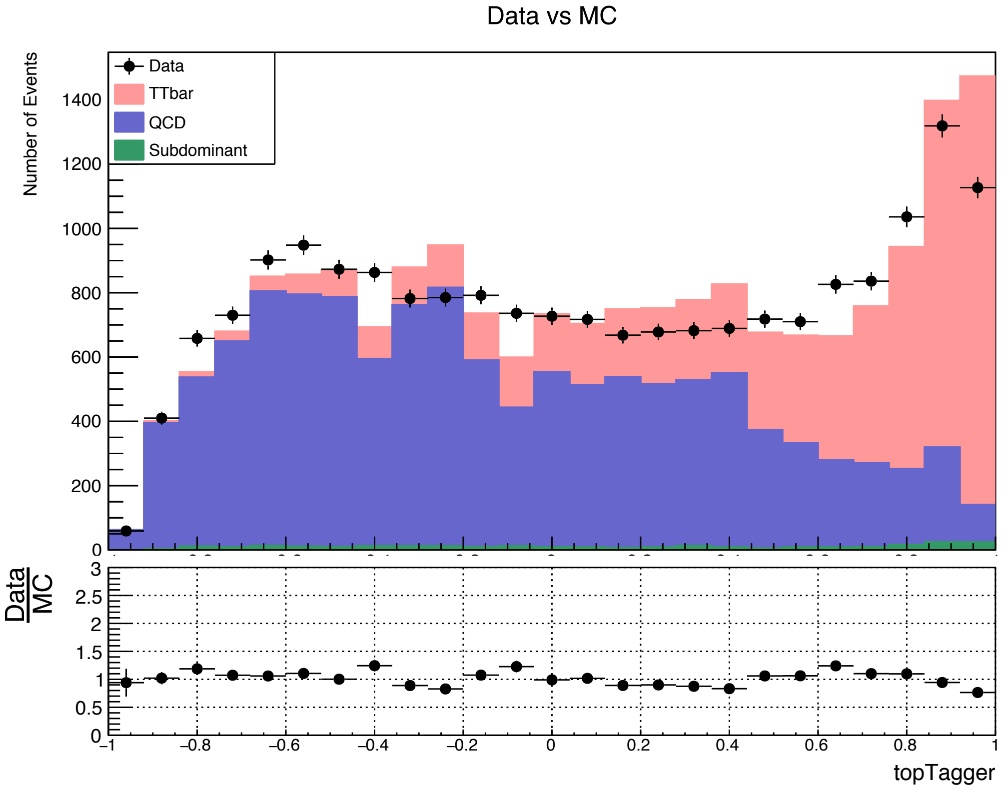
Second Leading Jet



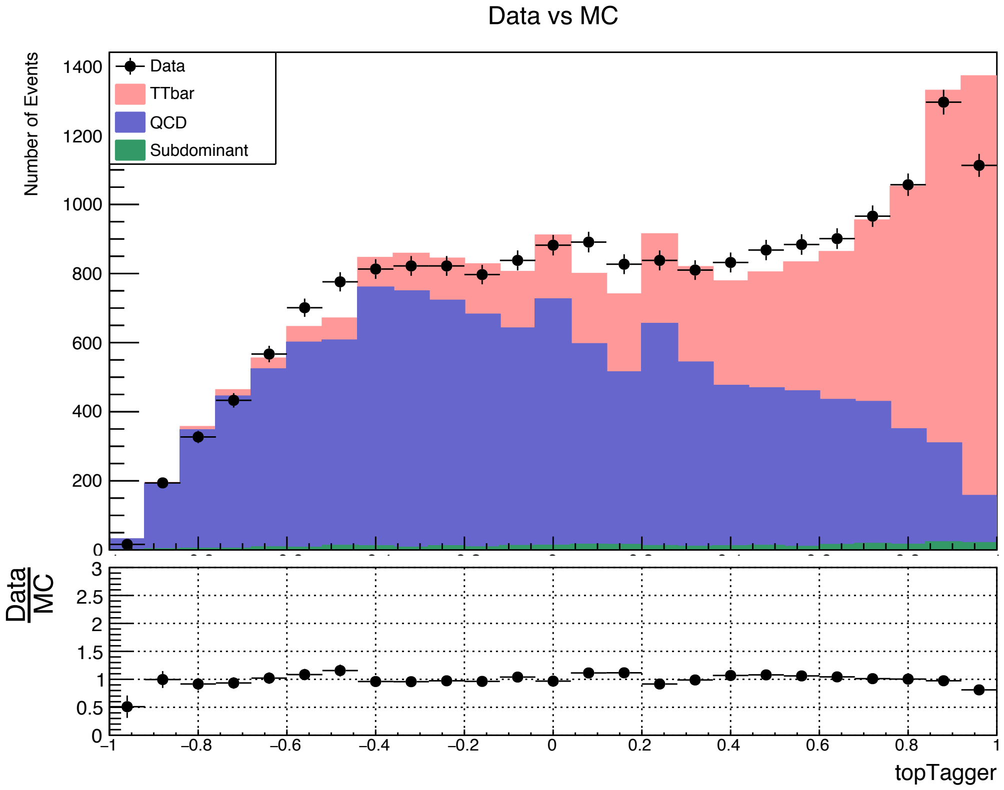
Top Tagger Output

2018

Leading Jet



Second Leading Jet



Top Tagger Efficiencies

Table 27: Top Tagger efficiency Values for 2016 preVFP.

Eff. Type	Eff. Data (stat)	Eff. $t\bar{t}$ (stat)	Eff. $t\bar{t}$ (stat + systematic)
Inclusive	0.757 ± 0.058	0.791 ± 0.009	0.791 ± 0.01
p_T [400, 600]GeV	0.742 ± 0.067	0.793 ± 0.011	0.793 ± 0.021
p_T [600, 800]GeV	0.774 ± 0.134	0.79 ± 0.016	0.79 ± 0.025
p_T [800, Inf)GeV	0.824 ± 0.198	0.777 ± 0.037	0.777 ± 0.047

Table 28: Top Tagger efficiency Values for 2016 preVFP using JMAR proposed p_T regions.

Eff. Type	Eff. Data (stat)	Eff. $t\bar{t}$ (stat)	Eff. $t\bar{t}$ (stat + systematic)
Inclusive	0.757 ± 0.058	0.791 ± 0.009	0.791 ± 0.01
p_T [400, 500]GeV	0.806 ± 0.136	0.792 ± 0.021	0.792 ± 0.031
p_T [500, 600]GeV	0.721 ± 0.076	0.793 ± 0.013	0.793 ± 0.022
p_T [600, Inf)GeV	0.785 ± 0.114	0.787 ± 0.014	0.787 ± 0.024

Table 29: Top Tagger efficiency Values for 2016 postVFP.

Eff. Type	Eff. Data (stat)	Eff. $t\bar{t}$ (stat)	Eff. $t\bar{t}$ (stat + systematic)
Inclusive	0.79 ± 0.052	0.786 ± 0.008	0.786 ± 0.011
p_T [400, 600]GeV	0.776 ± 0.061	0.79 ± 0.01	0.79 ± 0.021
p_T [600, 800]GeV	0.81 ± 0.104	0.781 ± 0.015	0.781 ± 0.024
p_T [800, Inf)GeV	0.861 ± 0.259	0.77 ± 0.035	0.77 ± 0.046

Table 30: Top Tagger efficiency Values for 2016 postVFP using JMAR proposed p_T regions.

Eff. Type	Eff. Data (stat)	Eff. $t\bar{t}$ (stat)	Eff. $t\bar{t}$ (stat + systematic)
Inclusive	0.79 ± 0.052	0.786 ± 0.008	0.786 ± 0.011
p_T [400, 500]GeV	0.782 ± 0.1	0.773 ± 0.018	0.773 ± 0.029
p_T [500, 600]GeV	0.774 ± 0.076	0.8 ± 0.012	0.8 ± 0.02
p_T [600, Inf)GeV	0.817 ± 0.097	0.779 ± 0.013	0.779 ± 0.025

Table 31: Top Tagger efficiency Values for 2017.

Eff. Type	Eff. Data (stat)	Eff. $t\bar{t}$ (stat)	Eff. $t\bar{t}$ (stat + systematic)
Inclusive	0.814 ± 0.032	0.868 ± 0.006	0.868 ± 0.009
p_T [400, 600]GeV	0.81 ± 0.04	0.867 ± 0.008	0.867 ± 0.017
p_T [600, 800]GeV	0.827 ± 0.063	0.871 ± 0.012	0.871 ± 0.021
p_T [800, Inf)GeV	0.793 ± 0.132	0.869 ± 0.029	0.869 ± 0.037

Table 32: Top Tagger efficiency Values for 2017 using JMAR proposed p_T regions.

Eff. Type	Eff. Data (stat)	Eff. $t\bar{t}$ (stat)	Eff. $t\bar{t}$ (stat + systematic)
Inclusive	0.814 ± 0.032	0.868 ± 0.006	0.868 ± 0.009
p_T [400, 500]GeV	0.808 ± 0.069	0.854 ± 0.014	0.854 ± 0.023
p_T [500, 600]GeV	0.812 ± 0.047	0.872 ± 0.009	0.872 ± 0.018
p_T [600, Inf)GeV	0.822 ± 0.058	0.870 ± 0.011	0.870 ± 0.019

Table 33: Top Tagger efficiency Values for 2018.

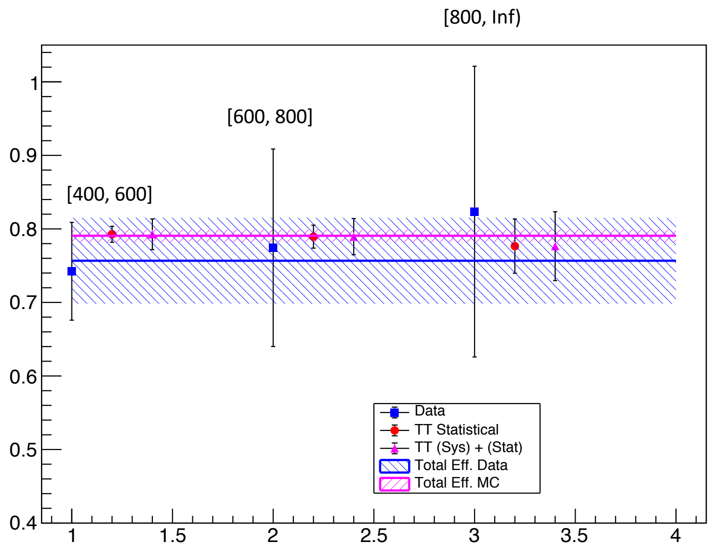
Eff. Type	Eff. Data (stat)	Eff. $t\bar{t}$ (stat)	Eff. $t\bar{t}$ (stat + systematic)
Inclusive	0.792 ± 0.03	0.827 ± 0.005	0.827 ± 0.008
p_T [400, 600]GeV	0.789 ± 0.039	0.825 ± 0.006	0.825 ± 0.014
p_T [600, 800]GeV	0.805 ± 0.051	0.833 ± 0.01	0.833 ± 0.02
p_T [800, Inf)GeV	0.752 ± 0.104	0.822 ± 0.024	0.822 ± 0.037

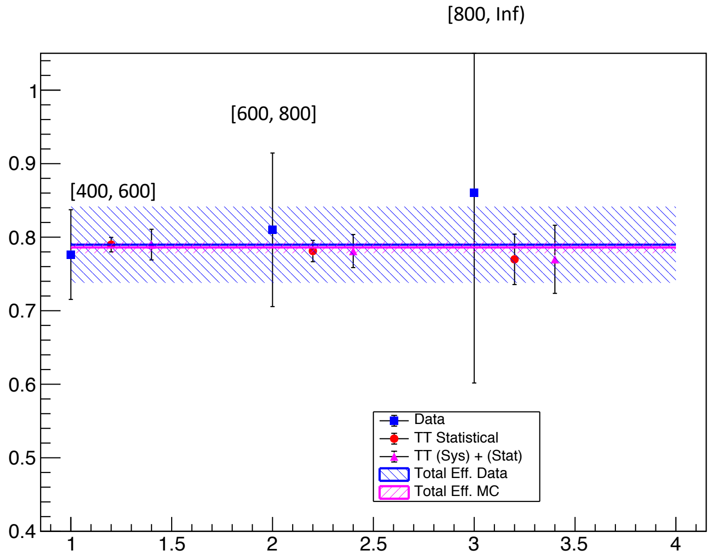
Table 34: Top Tagger efficiency Values for 2018 using JMAR proposed p_T regions.

Eff. Type	Eff. Data (stat)	Eff. $t\bar{t}$ (stat)	Eff. $t\bar{t}$ (stat + systematic)
Inclusive	0.792 ± 0.03	0.827 ± 0.005	0.827 ± 0.008
p_T [400, 500]GeV	0.739 ± 0.074	0.811 ± 0.011	0.811 ± 0.019
p_T [500, 600]GeV	0.807 ± 0.045	0.832 ± 0.007	0.832 ± 0.018
p_T [600, Inf)GeV	0.797 ± 0.046	0.832 ± 0.009	0.832 ± 0.021

TagAndProbe Efficiency per Pt region

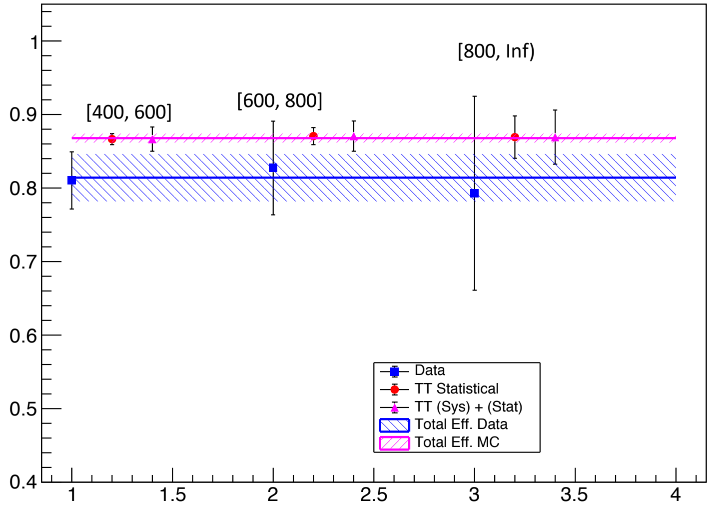
2016 preVFP

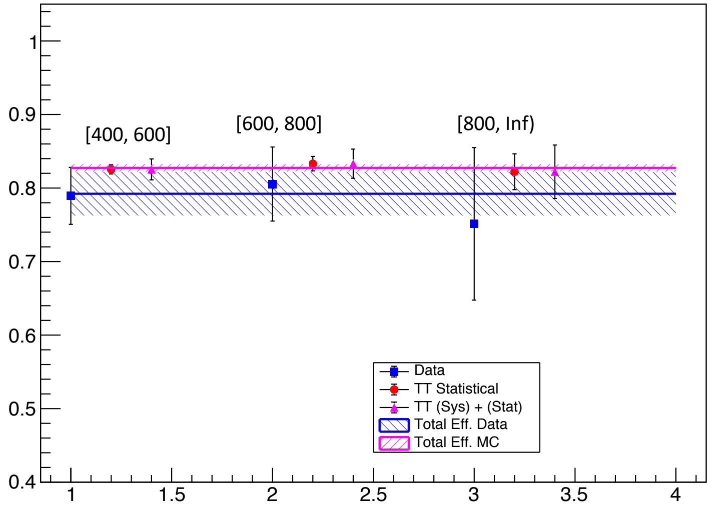




2016 postVFP

2017

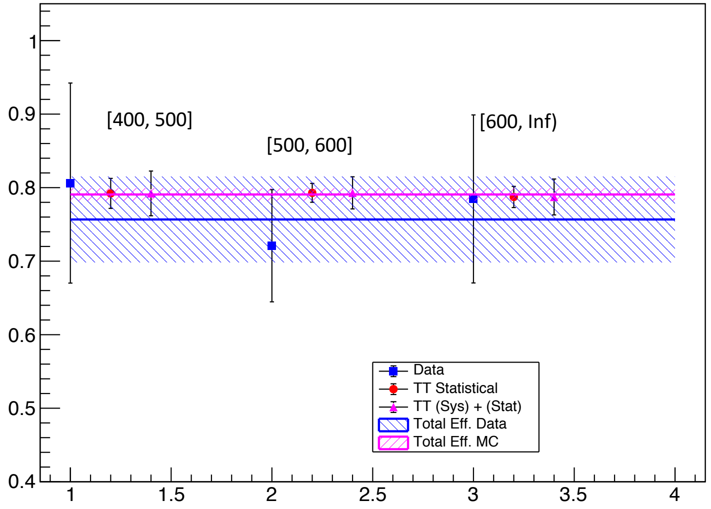


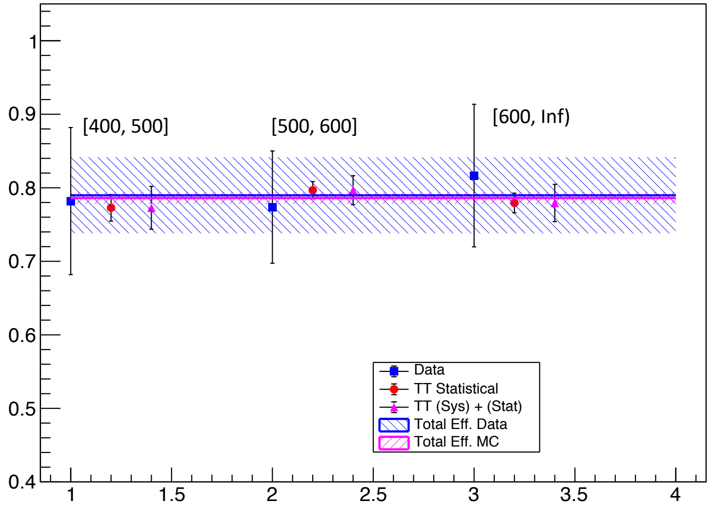


2018

TagAndProbe Efficiency per Pt region (JMAR regions)

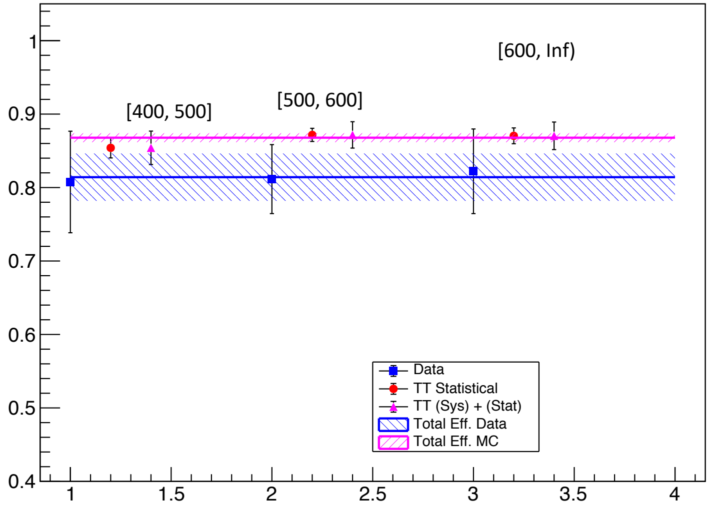
2016 preVFP

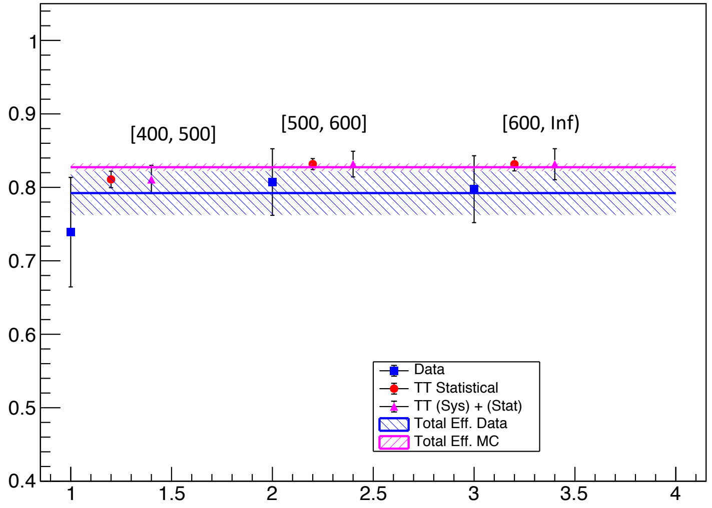




2016 postVFP

2017





2018

Scale Factor Values

Table 35: Top Tagger SF Values for 2016 preVFP.

SF Type	Value ± error
Inclusive	0.957 ± 0.074
p_T [400, 600]GeV	0.937 ± 0.085
p_T [600, 800]GeV	0.981 ± 0.17
p_T [800, Inf)GeV	1.06 ± 0.26

Table 36: Top Tagger SF Values for 2016 preVFP using JMAR proposed p_T regions.

SF Type	Value ± error
Inclusive	0.957 ± 0.074
p_T [400, 500]GeV	1.02 ± 0.173
p_T [500, 600]GeV	0.91 ± 0.097
p_T [600, Inf)GeV	0.997 ± 0.15

Table 37: Top Tagger SF Values for 2016 postVFP.

SF Type	Value ± error
Inclusive	1.01 ± 0.067
p_T [400, 600]GeV	0.983 ± 0.078
p_T [600, 800]GeV	1.04 ± 0.135
p_T [800, Inf)GeV	1.12 ± 0.34

Table 38: Top Tagger SF Values for 2016 postVFP using JMAR proposed p_T regions.

SF Type	Value ± error
Inclusive	1.01 ± 0.067
p_T [400, 500]GeV	1.01 ± 0.132
p_T [500, 600]GeV	0.971 ± 0.097
p_T [600, Inf)GeV	1.05 ± 0.13

Table 39: Top Tagger SF Values for 2017.

SF Type	Value ± error
Inclusive	0.938 ± 0.038
p_T [400, 600]GeV	0.935 ± 0.046
p_T [600, 800]GeV	0.95 ± 0.059
p_T [800, Inf)GeV	0.912 ± 0.155

Table 40: Top Tagger SF Values for 2017 using JMAR proposed p_T regions.

SF Type	Value ± error
Inclusive	0.938 ± 0.038
p_T [400, 500]GeV	0.946 ± 0.082
p_T [500, 600]GeV	0.931 ± 0.055
p_T [600, Inf)GeV	0.945 ± 0.068

Table 41: Top Tagger SF Values for 2018.

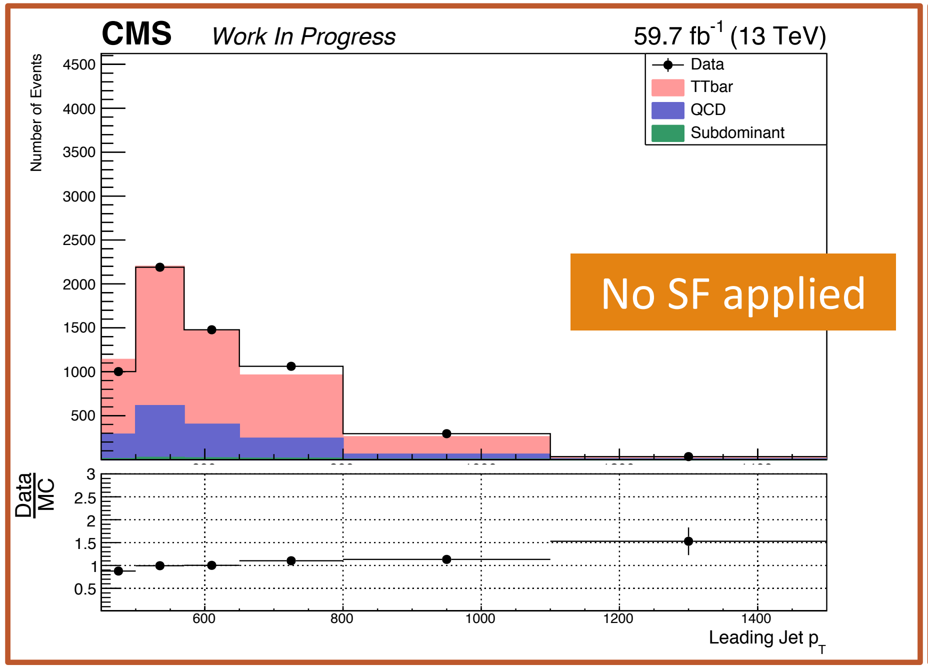
SF Type	Value ± error
Inclusive	0.958 ± 0.037
p_T [400, 600]GeV	0.956 ± 0.048
p_T [600, 800]GeV	0.967 ± 0.062
p_T [800, Inf)GeV	0.914 ± 0.13

Table 42: Top Tagger SF Values for 2018 using JMAR proposed p_T regions.

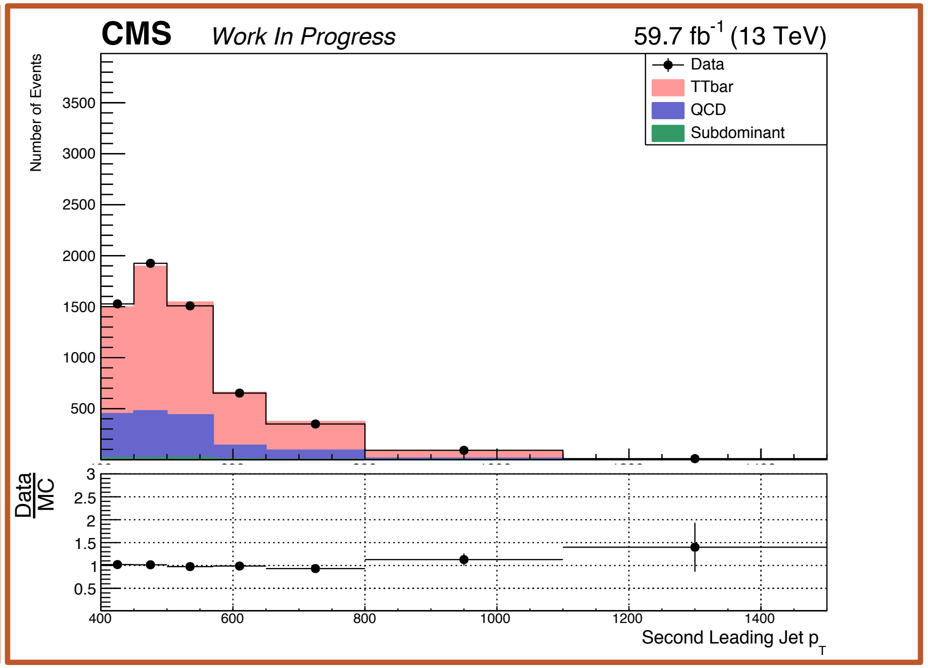
SF Type	Value ± error
Inclusive	0.958 ± 0.037
p_T [400, 500]GeV	0.912 ± 0.093
p_T [500, 600]GeV	0.971 ± 0.055
p_T [600, Inf)GeV	0.959 ± 0.056

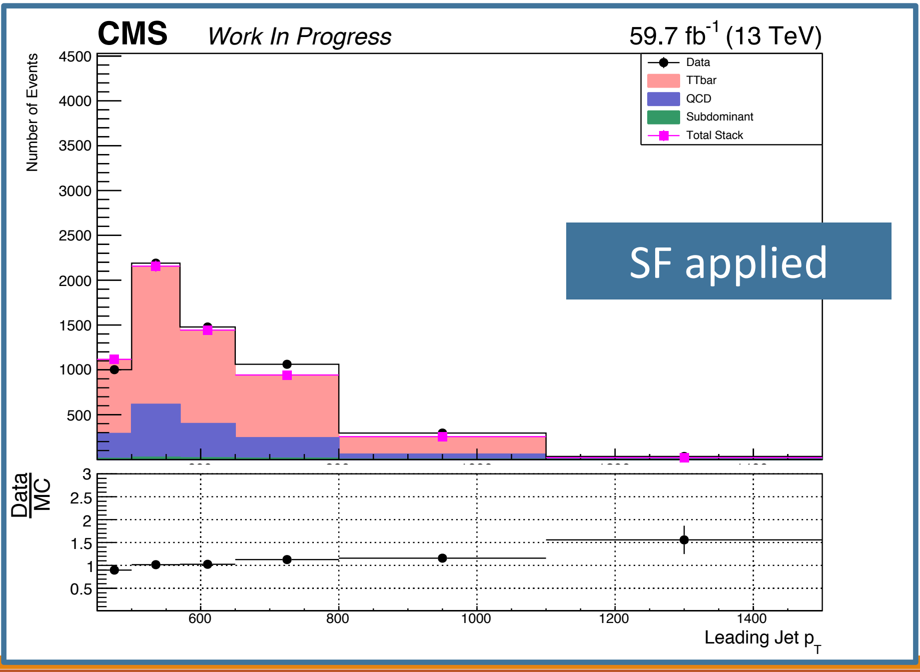
Data vs MC

2018

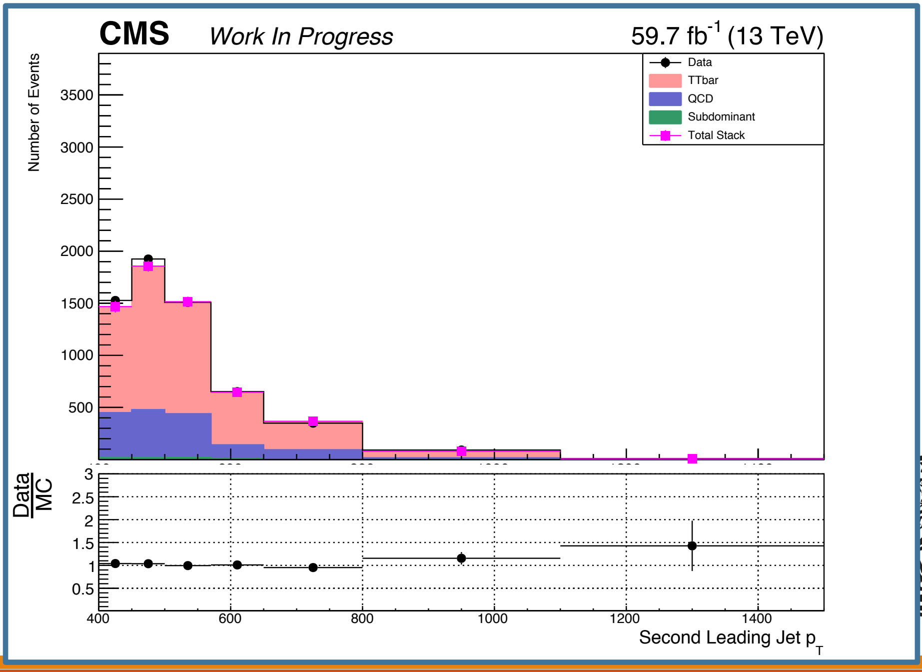


No SF applied





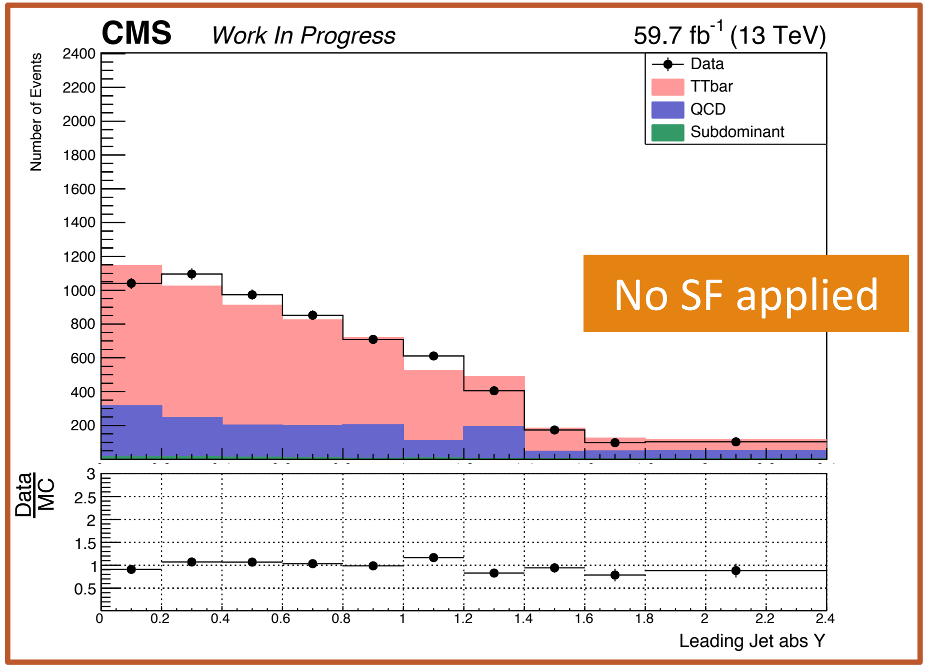
SF applied

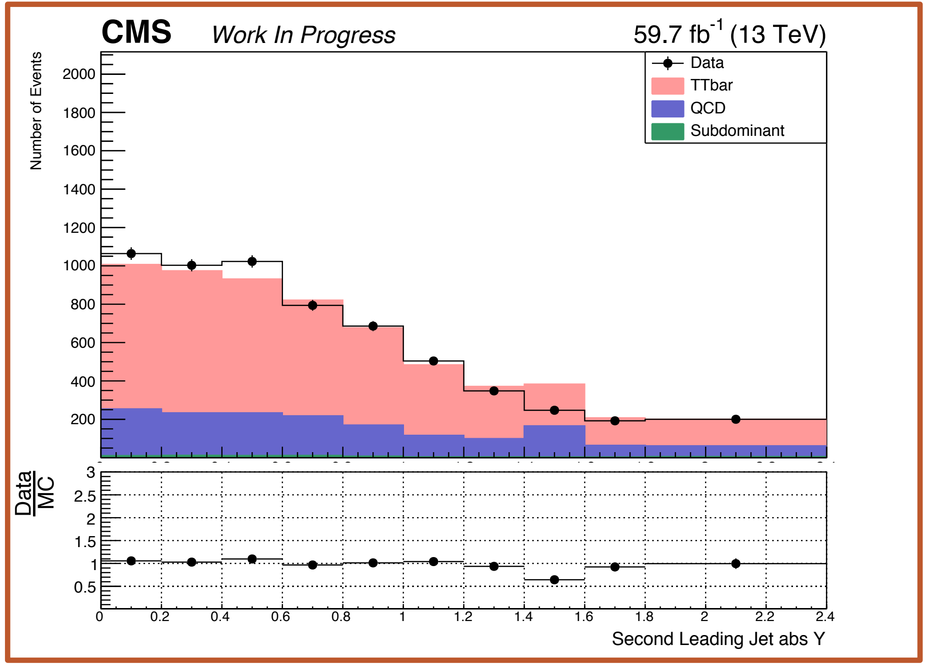


Data vs MC plots

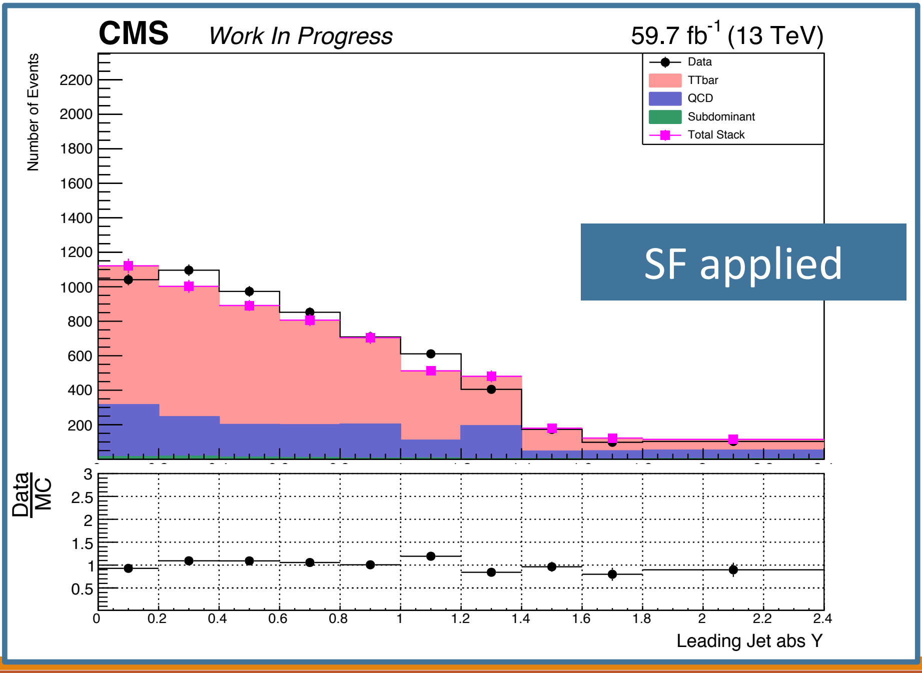
2018

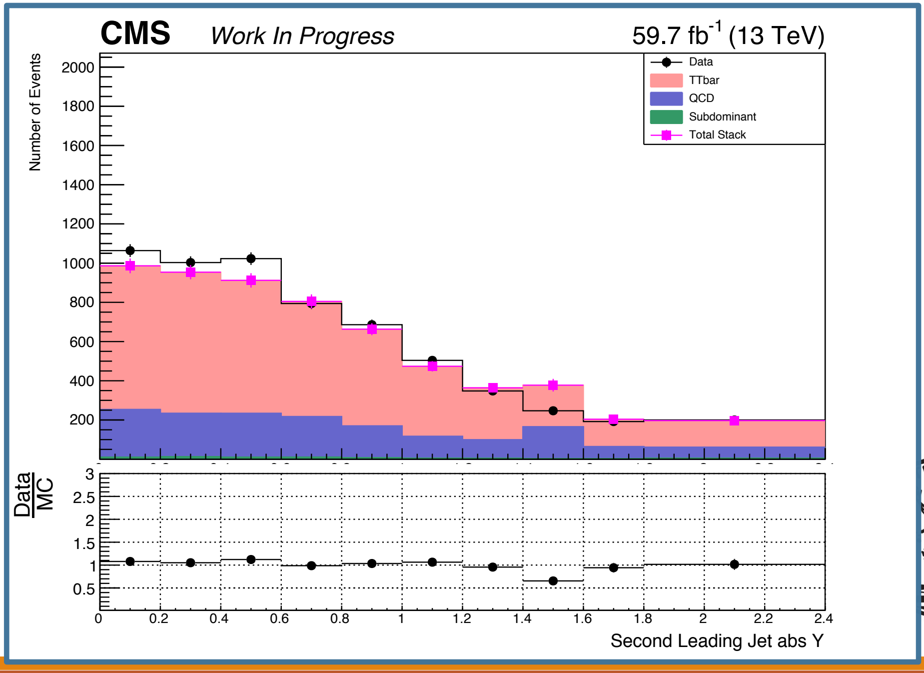
No SF applied





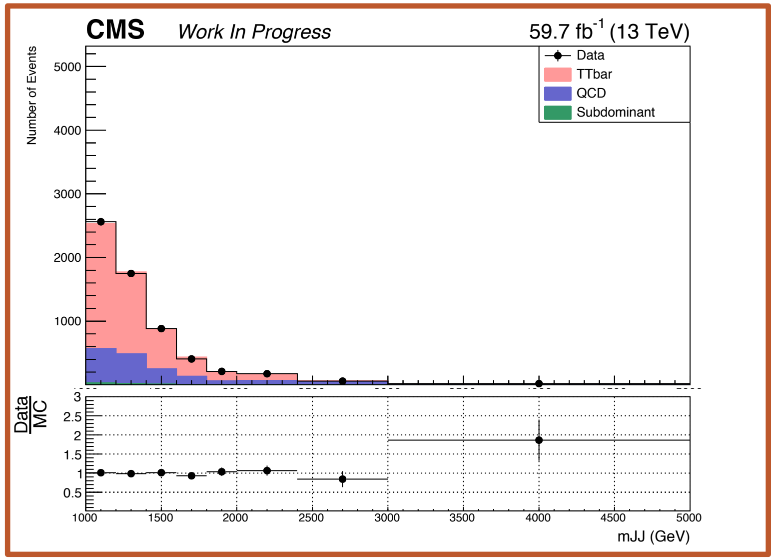
SF applied

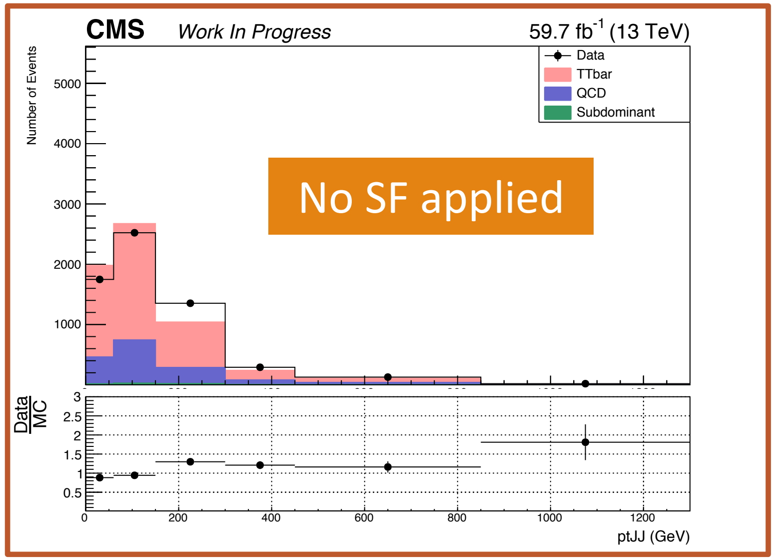




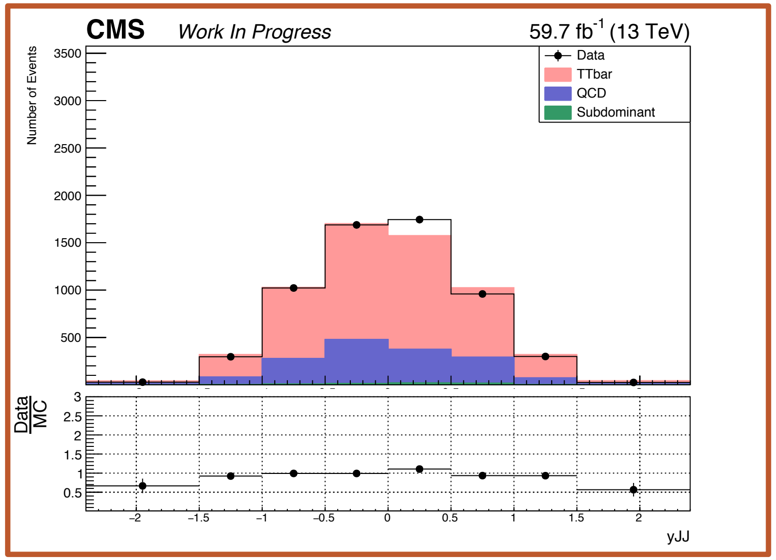
Data vs MC plots

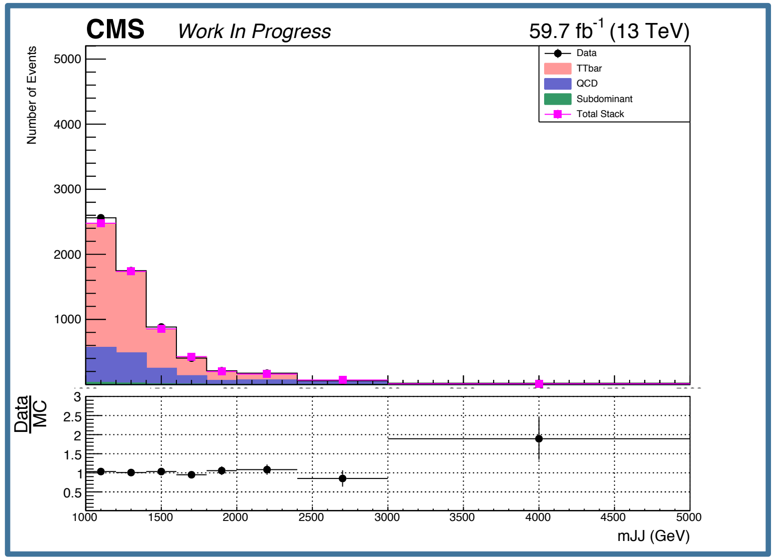
2018

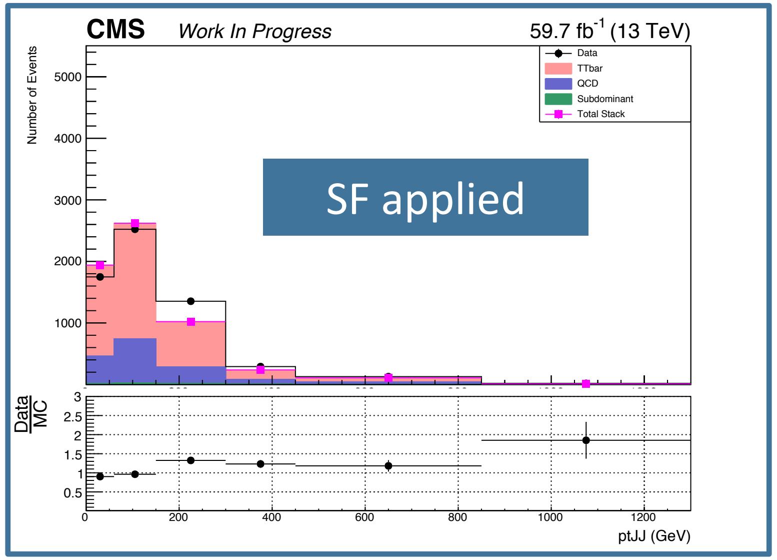




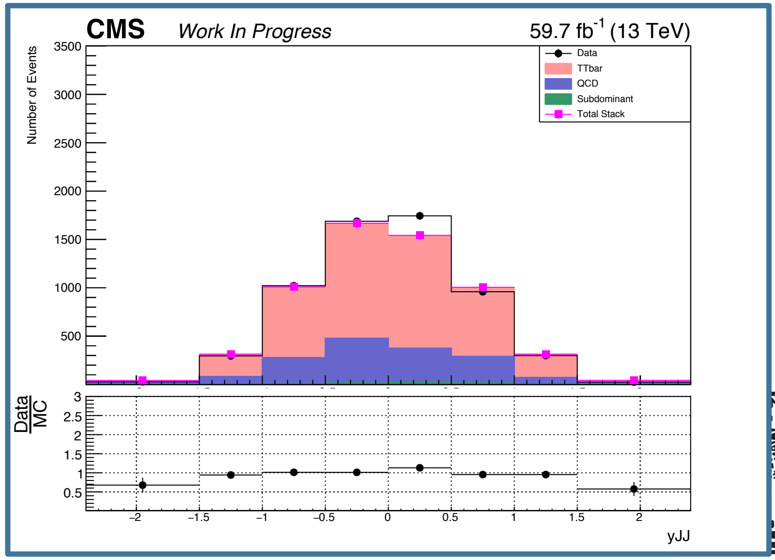
No SF applied







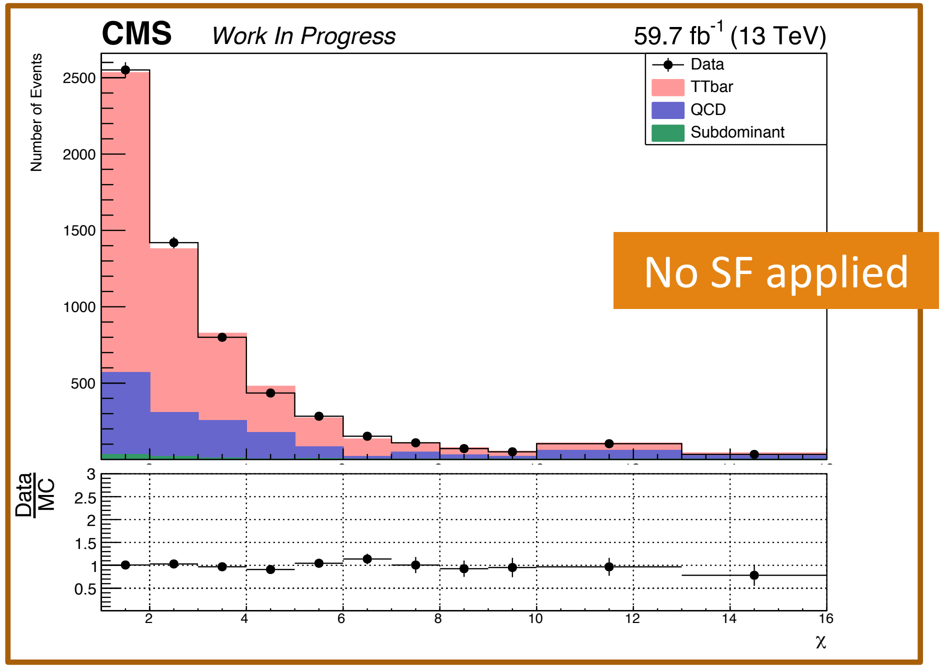
SF applied

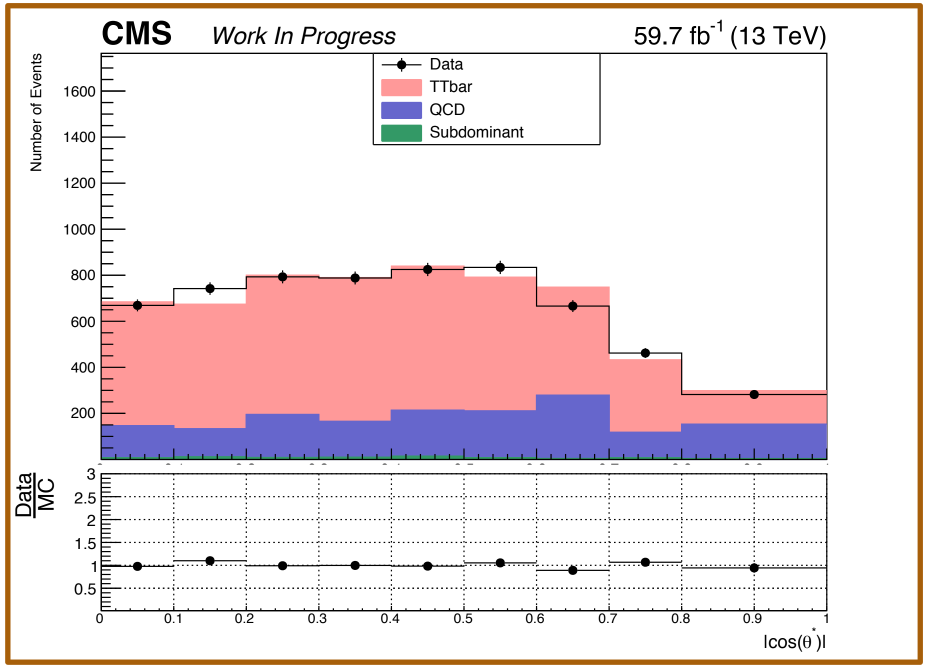


Data vs MC plots

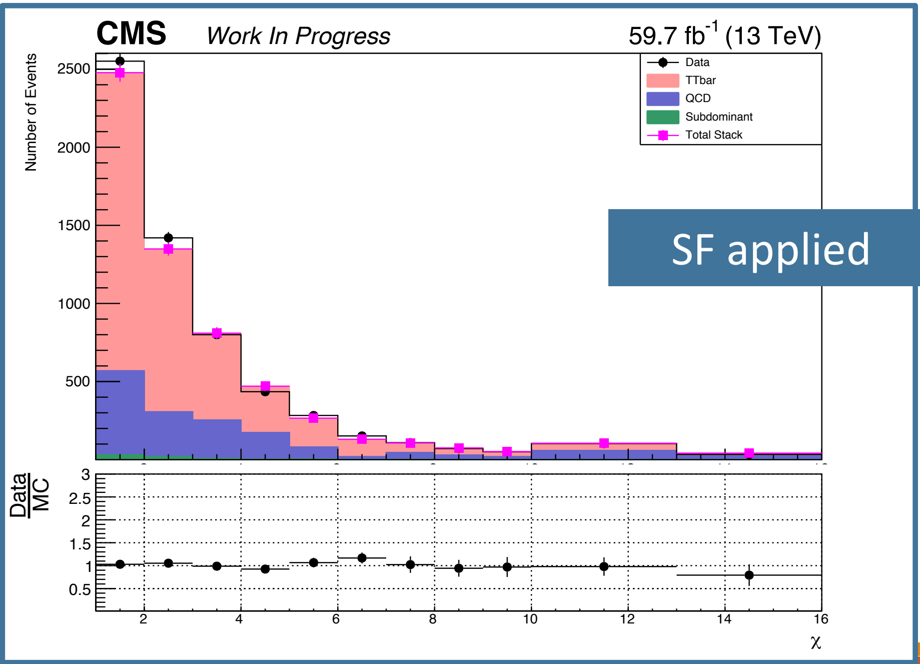
2018

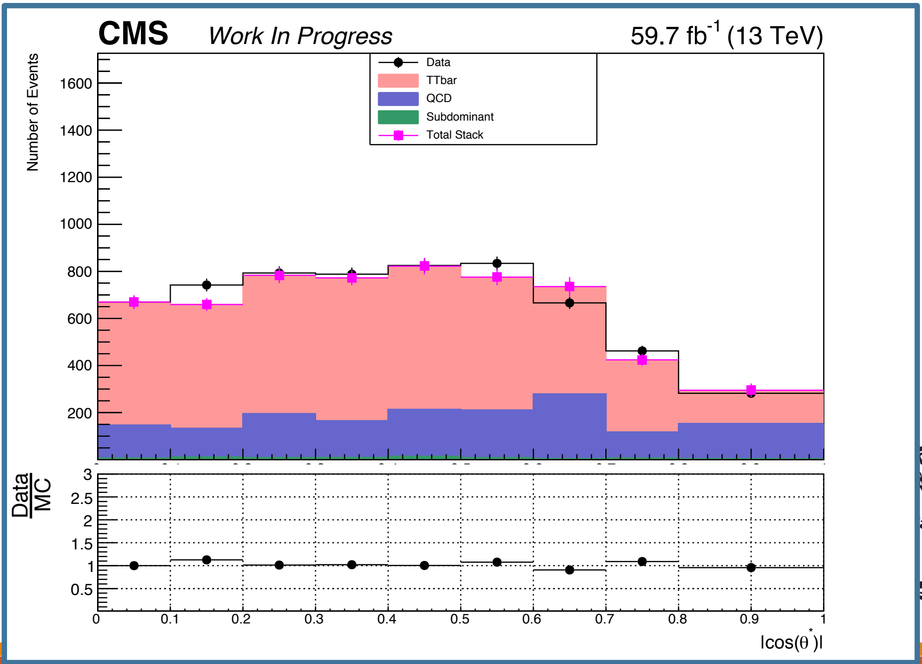
No SF applied





SF applied

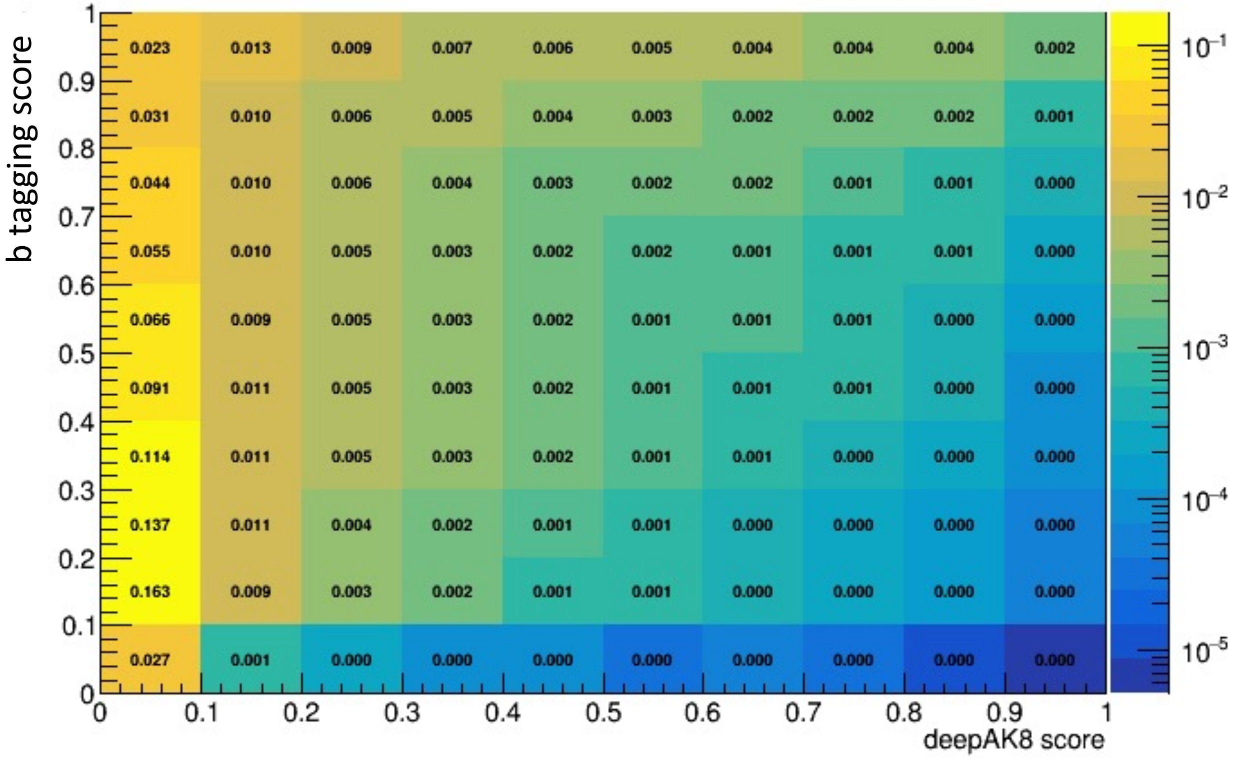




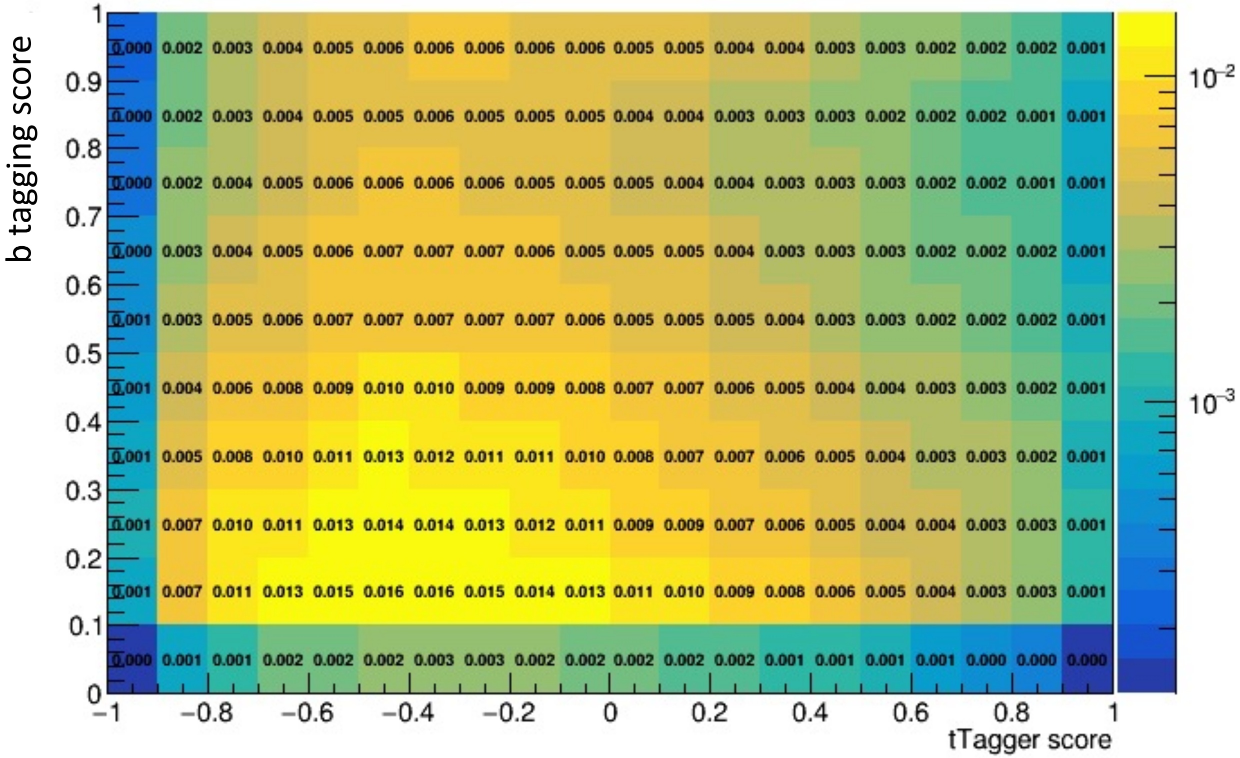
DeepAK8 backup

Btagging Correlation to DeepAK8 or TopTagger

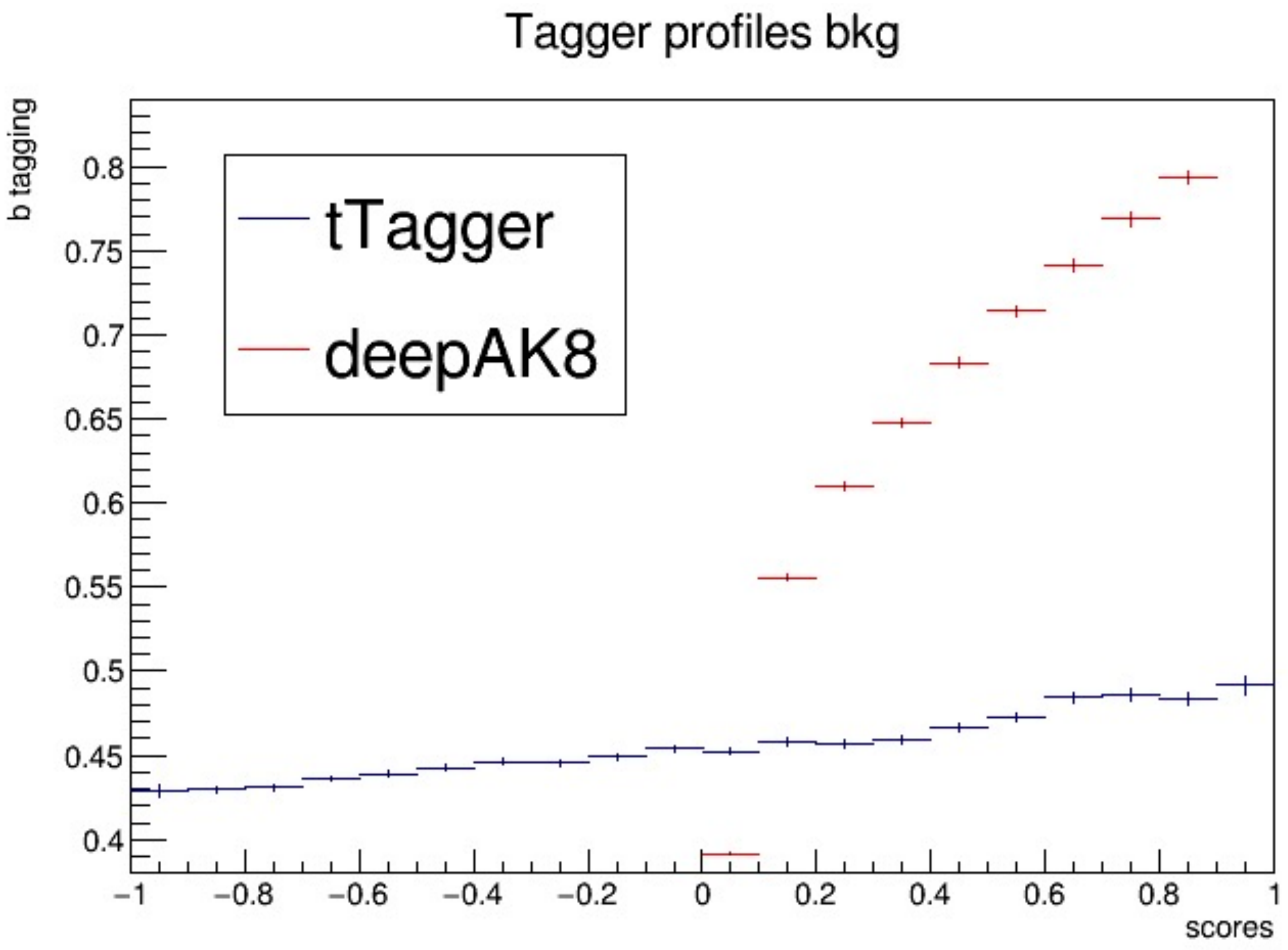
deepAK8 b tagging correlation bkg sample



tTagger b tagging correlation bkg sample



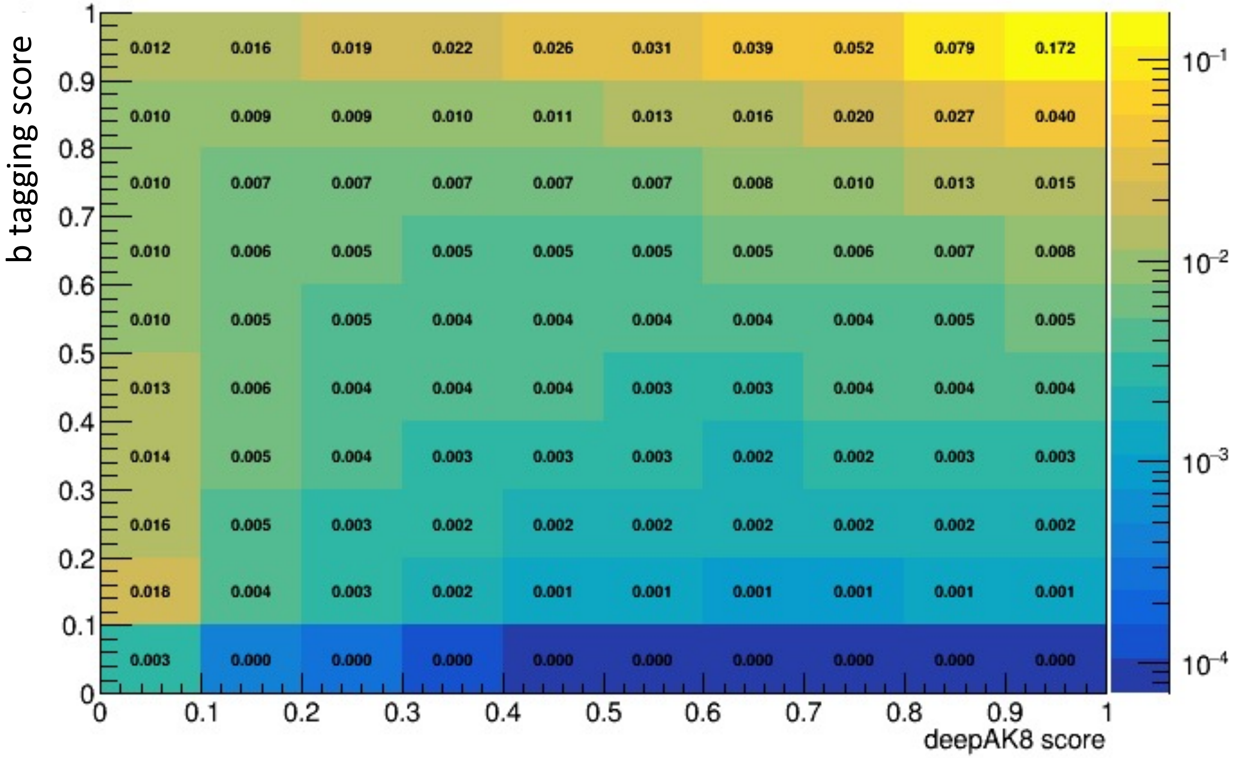
Btagging Correlation to DeepAK8 or TopTagger profiles



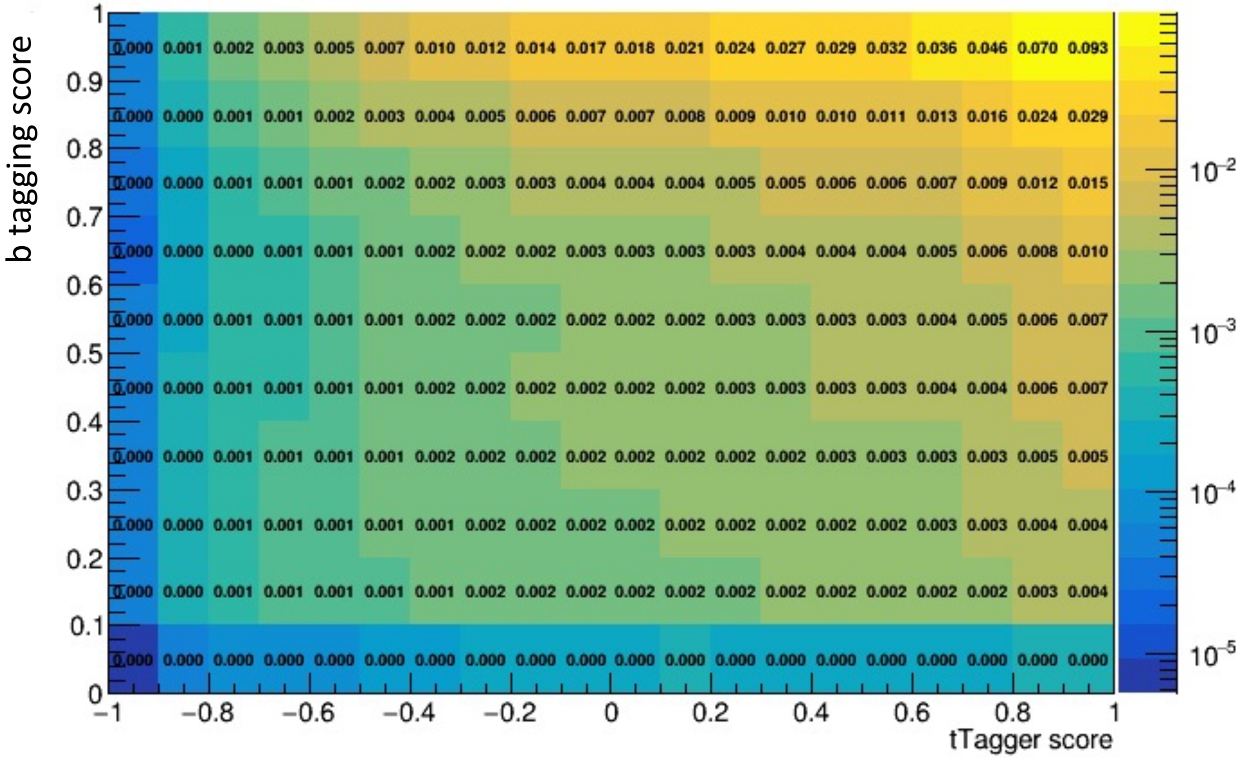
b-tagging mean value
vs
top tagging values

Btagging Correlation to DeepAK8 or TopTagger

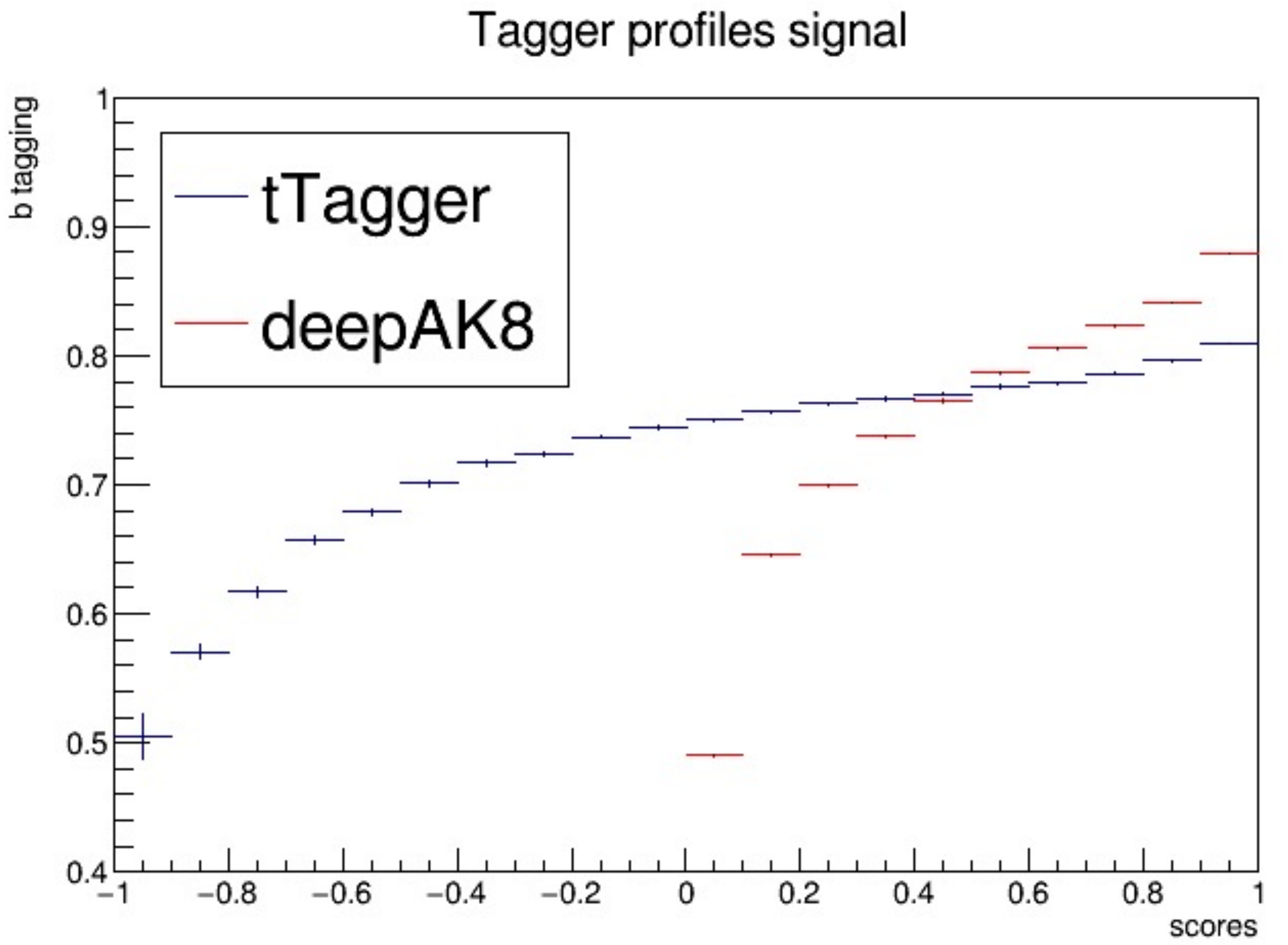
deepAK8 b tagging correlation signal sample



tTagger b tagging correlation signal sample



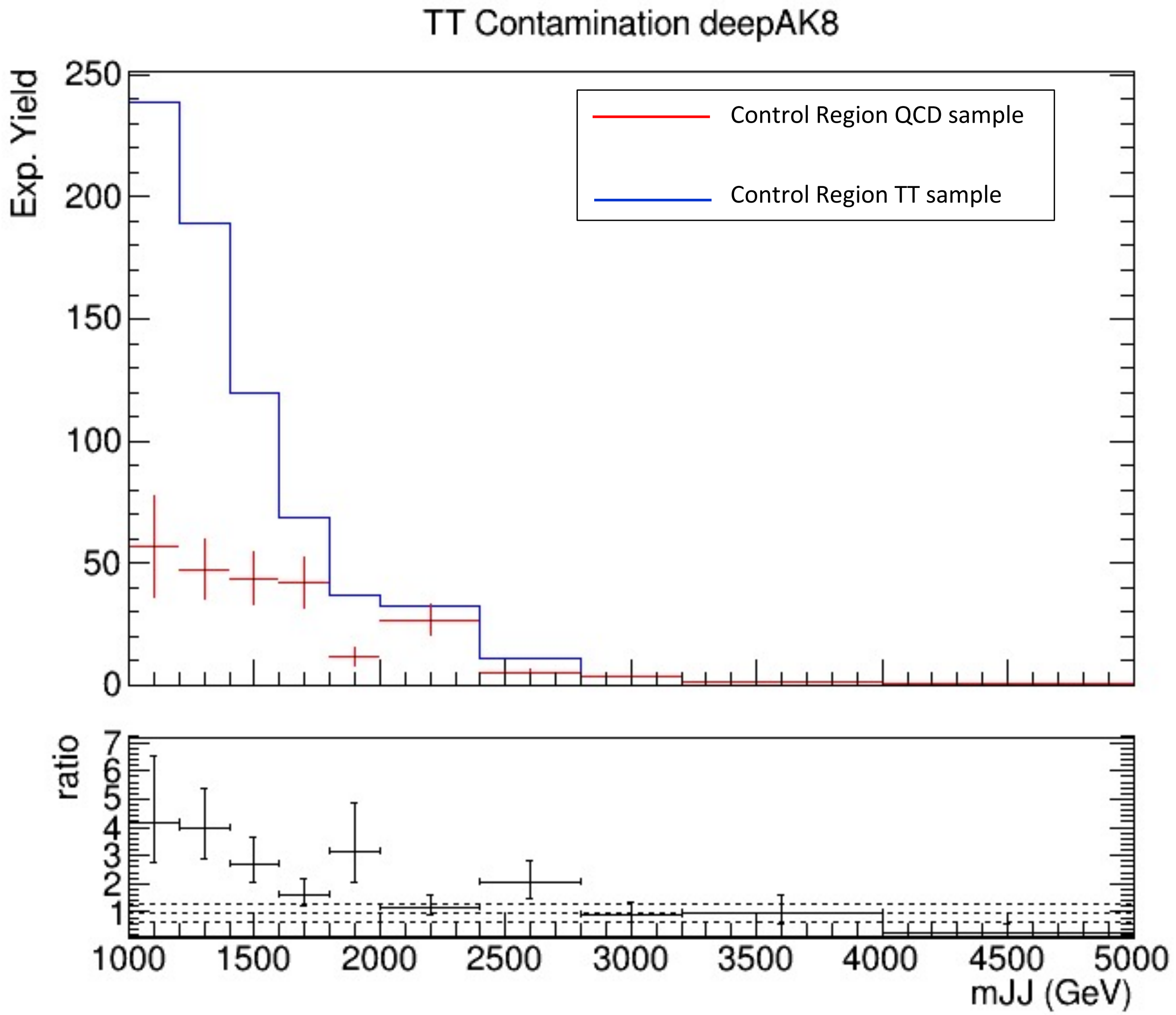
Btagging Correlation to DeepAK8 or TopTagger profiles

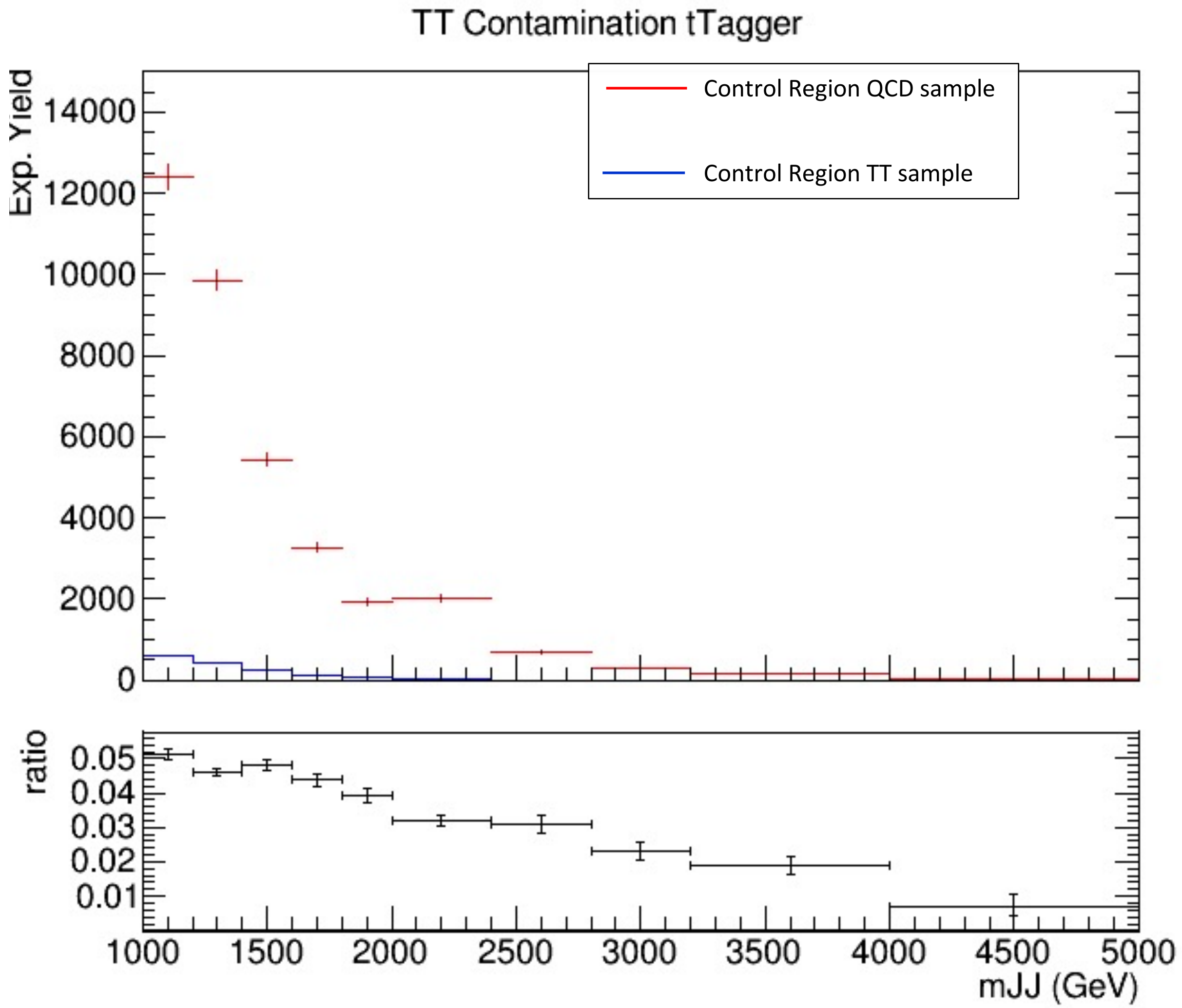


b-tagging mean value
vs
top tagging values

Control Region Contamination

- Expected yield from QCD Bkg samples and TT Signal sample in the CR (btag == 0) vs mJJ
- The QCD contribution is used to get the QCD shape





Control Region Contamination

- Expected yield from QCD Bkg samples and TT Signal sample in the CR (btag == 0) vs jetPt
- The QCD contribution is used to get the QCD shape