Progress on top jet analysis

Daniela Dominguez, Paolo Gunnellini, Hannes Jung

Deutsches Elektronen Synchrotron

August 2017

P&C meeting

Summary presentation

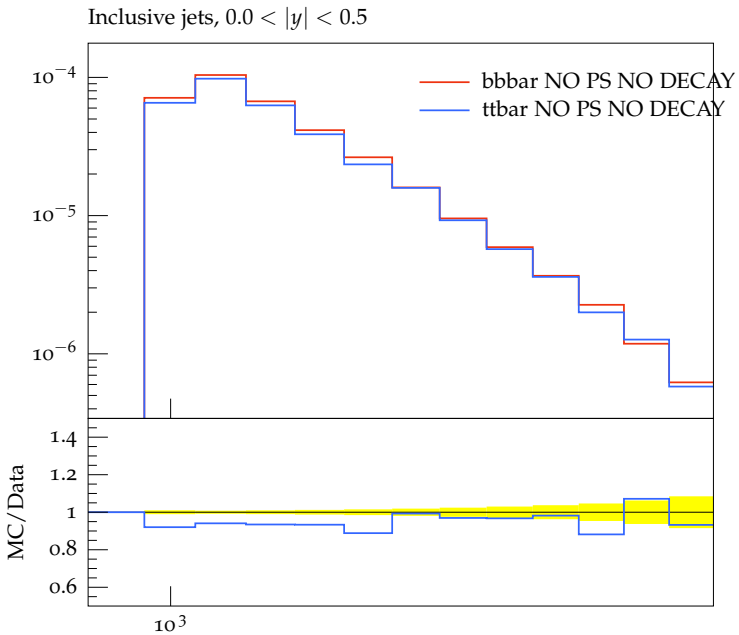
Motivation (I)

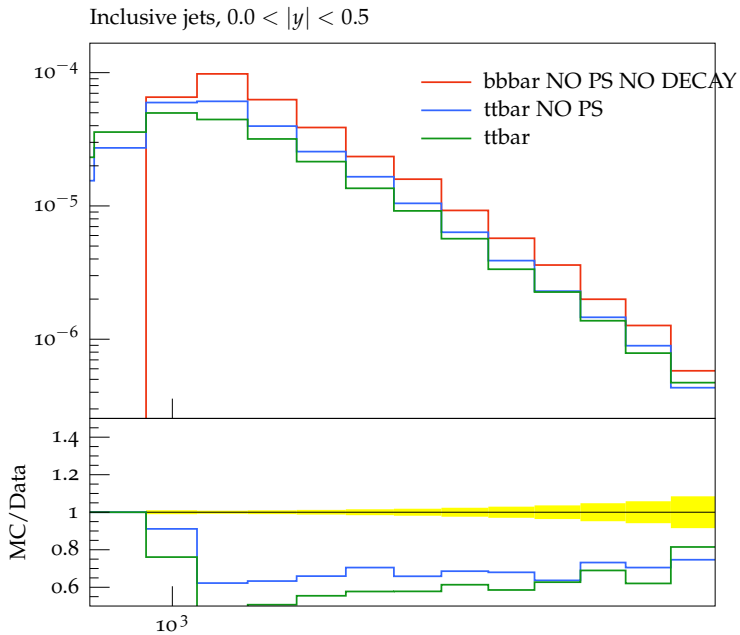
What about top jets?

- GOAL: demonstrate the flavour-blindness of QCD
- production of bottom and top quarks should be the same at high p_T (where both masses become negligible)

$\rightarrow$ First look at jet cross section clustered with large cone size (R = 0.8)

PYTHIA 8 bbbar / ttbar hard generation starting from $\hat{p}_T \sim 1$ TeV





Inclusive jet cross section in central region

Plain P8 predicts the flavour blindness but parton evolution causes out-of-cone effects $\rightarrow$ very challenging!

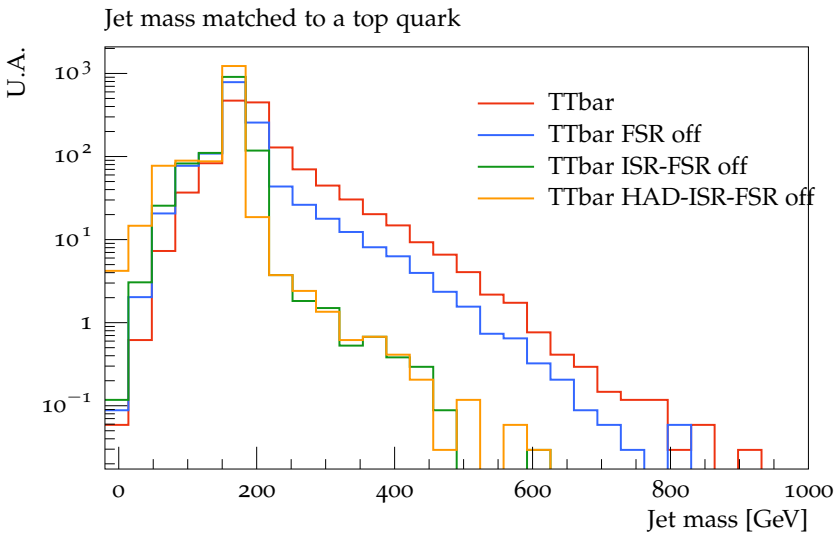
Motivation (II)

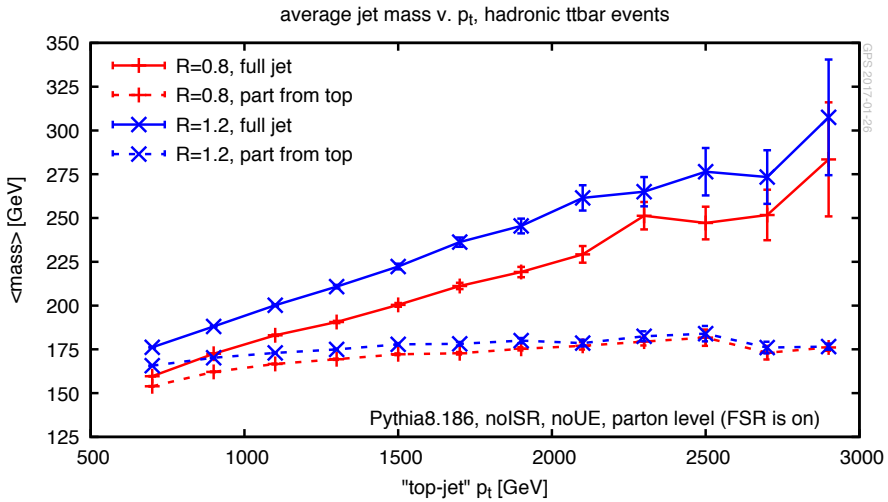
What about top jets?

- GOAL: study the QCD radiation emitted by the top
- The reconstructed top-jet mass depends on the QCD activity associated to the initial quark → boost, additional radiation, ecc.

→ First look at jet cross section clustered with large cone size (R = 0.8)

PYTHIA 8 bbbar / ttbar hard generation starting from $\hat{p}_T \sim 1$ TeV





From Gavin Salam

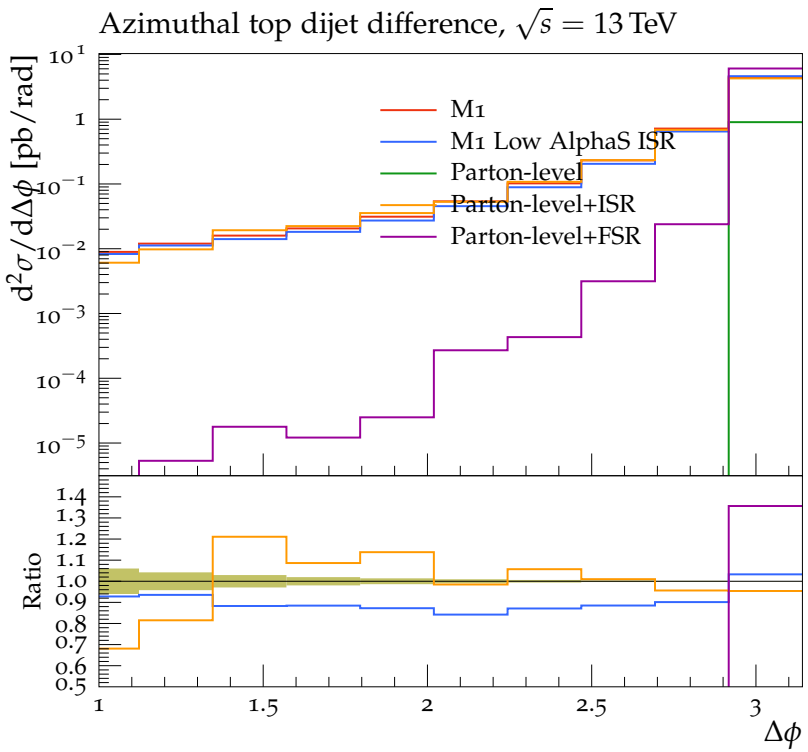
Motivation (III)

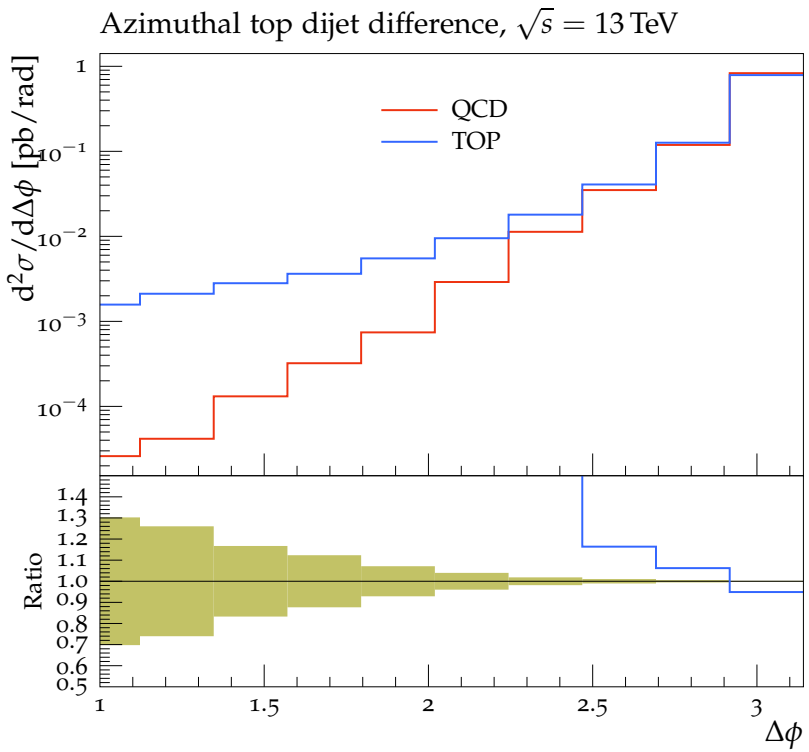
What about top jets?

- GOAL: study the correlation from top jets
- different amount of radiation between the two processes

$\rightarrow$ fat jets with $p_T > 400$ GeV in the central region

PYTHIA 8 simulation for top and QCD events





Motivation

What about top jets?

- GOAL 1: demonstrate the flavour-blindness of QCD
- GOAL 2: study the QCD radiation emitted by the top
- GOAL 3: study resummation effects in the top sector and compare them with dijet cases

Remarks:

- Regime of boosted topologies because of high p_{T}
- Considered only hadronic top decays for better p_{T} resolution
- Overwhelming QCD background to deal with

Current analysis status

DATA: RUN G only - $\sim$ 6 fb^{-1}

Two jet triggers:
HLT_AK8DiPFJet280_200_TrimMass30_BTagCSV_p20
(for 400 $<$ jet p_T $<$ 550 GeV)
HLT_AK8PFJet450 (for jet p_T $>$ 550 GeV)

Monte Carlo samples:

- TTbar POWHEG
- QCD Madgraph samples (HT-binned) [K^{NLO} = 0.65]
- Single Top (POWHEG + P8)
- DY, W (MG_aMCNLO + P8)

FIRST SELECTION:

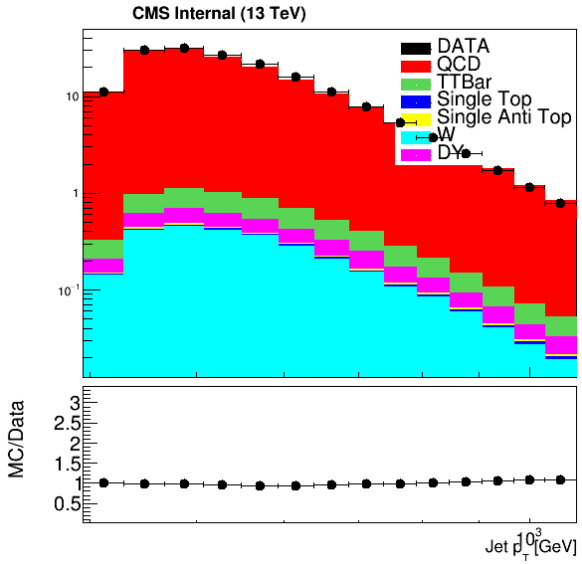
- At least two AK8 jets with $p_T > 400$ GeV in $|\eta| < 2.4$
- Looking at leading and subleading jet variables

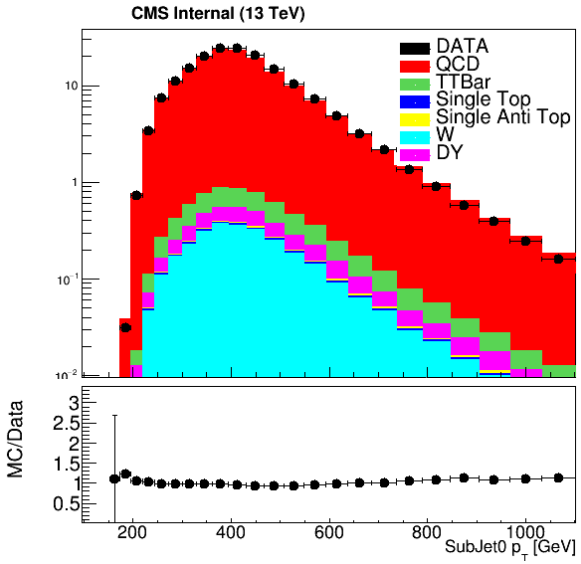
N.B. Code is taken from Kostas Kousouris (Thanks!) $\rightarrow$ TOP-16-015

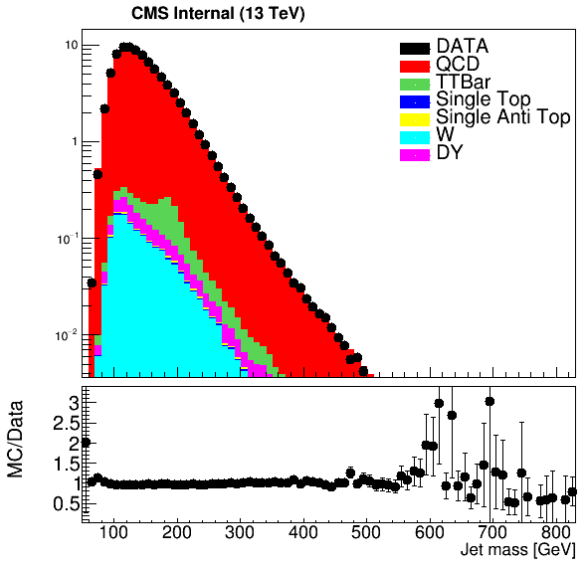
GOAL: measurement of $\Delta\phi^{t\bar{t}}$ + jet p_T cross sections

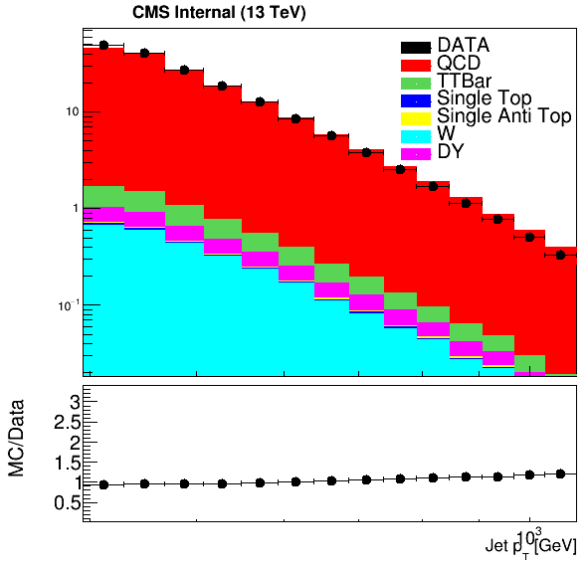
A deeper look at the data

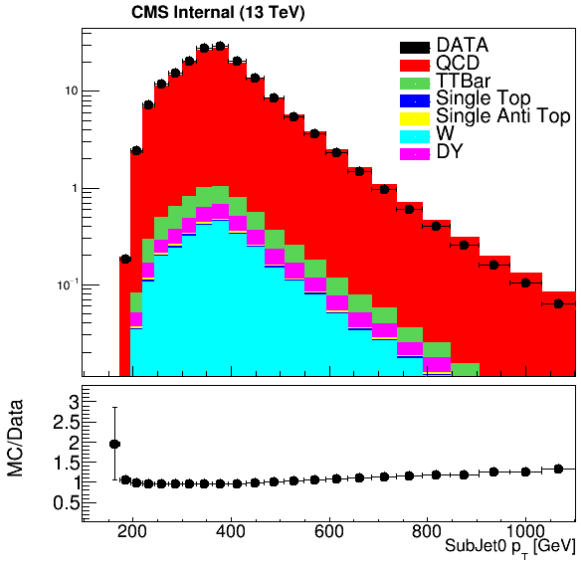
LEFT to RIGHT, TOP (leading jet) BOTTOM (subleading):
jet p_T, first subjet p_T, jet mass

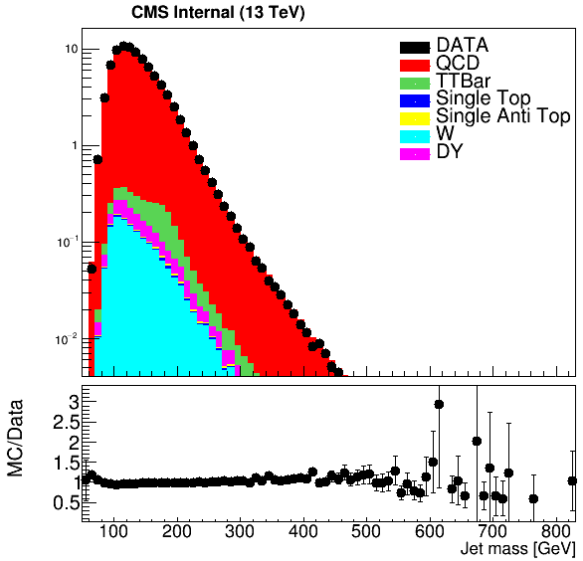






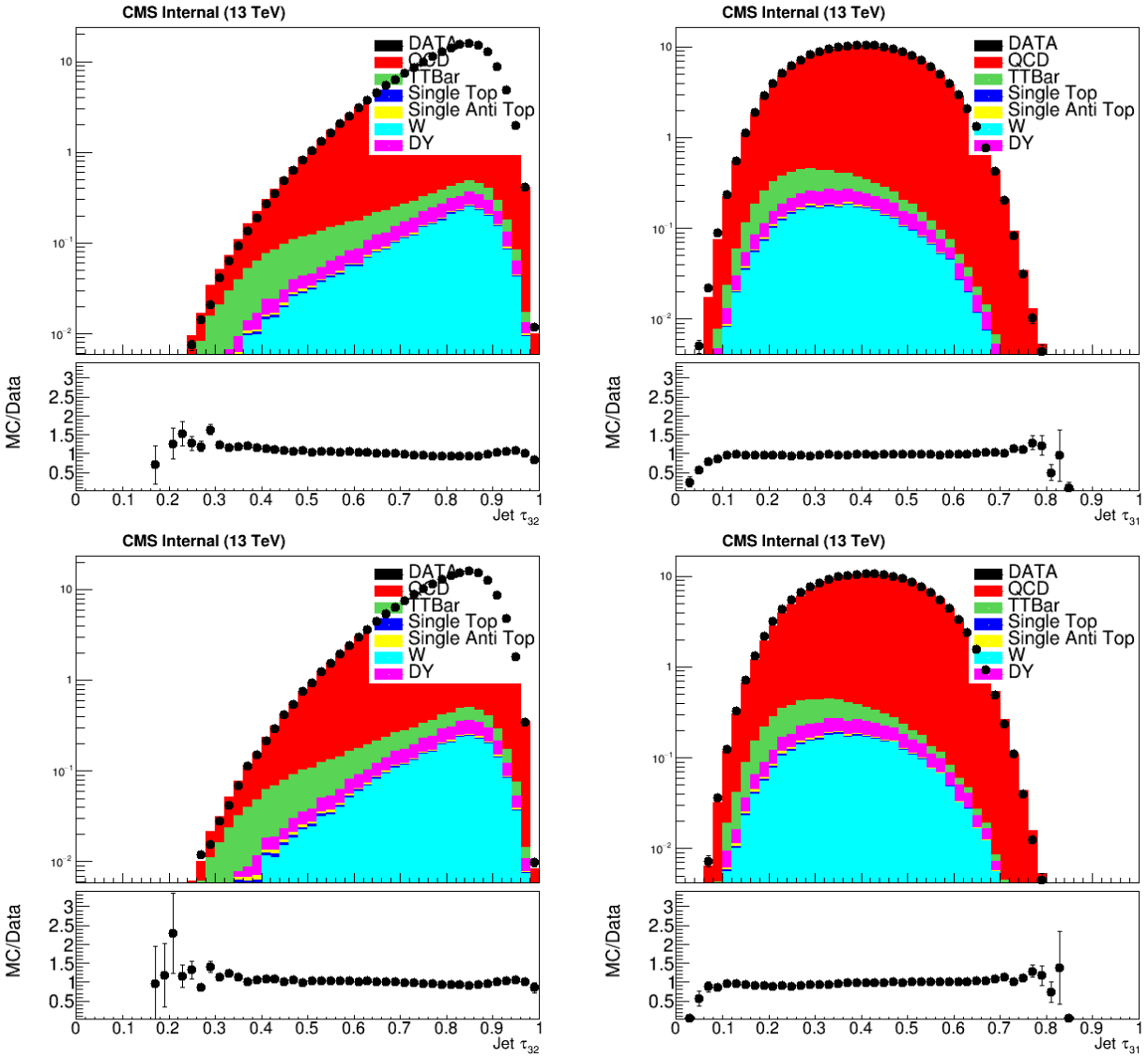






A deeper look at the data

LEFT to RIGHT, TOP (leading jet) BOTTOM (subleading):
jet τ_{32}, jet τ_{31}



Setting a multivariate analysis:

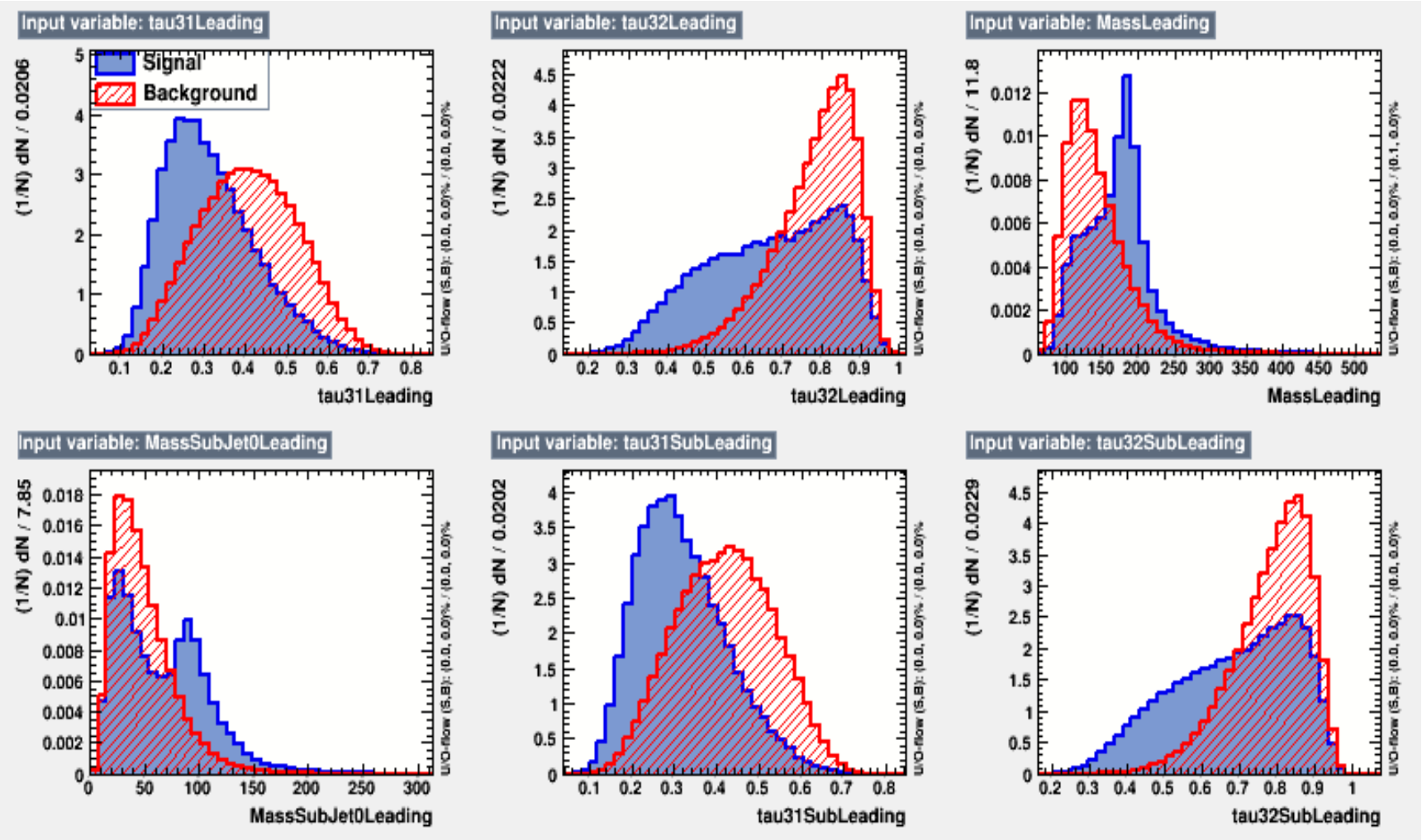
Four observables:
τ_{32} (leading)
τ_{32} (subleading)
τ_{31} (leading)
τ_{31} (subleading)

GOAL: reproduce results from TOP-16-015

S/B $\sim$ 0,012

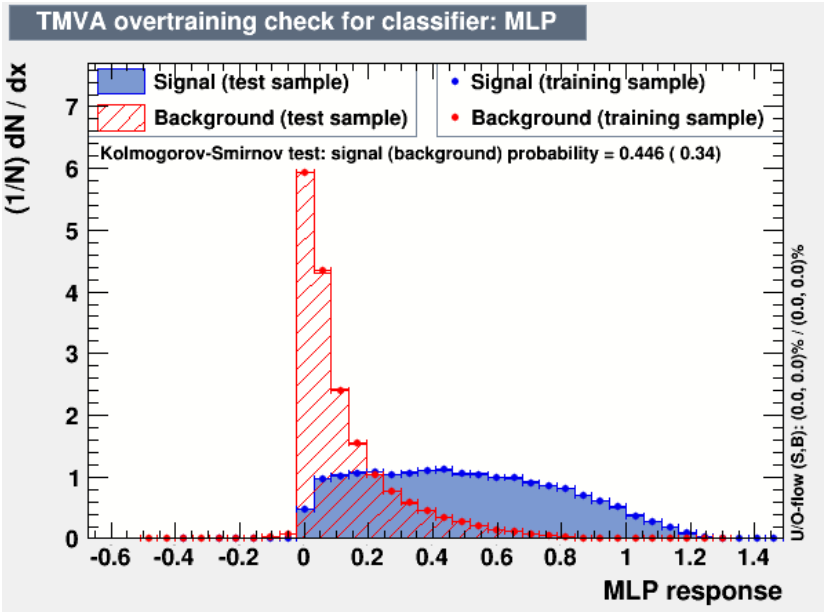
Building the Multivariate Analysis

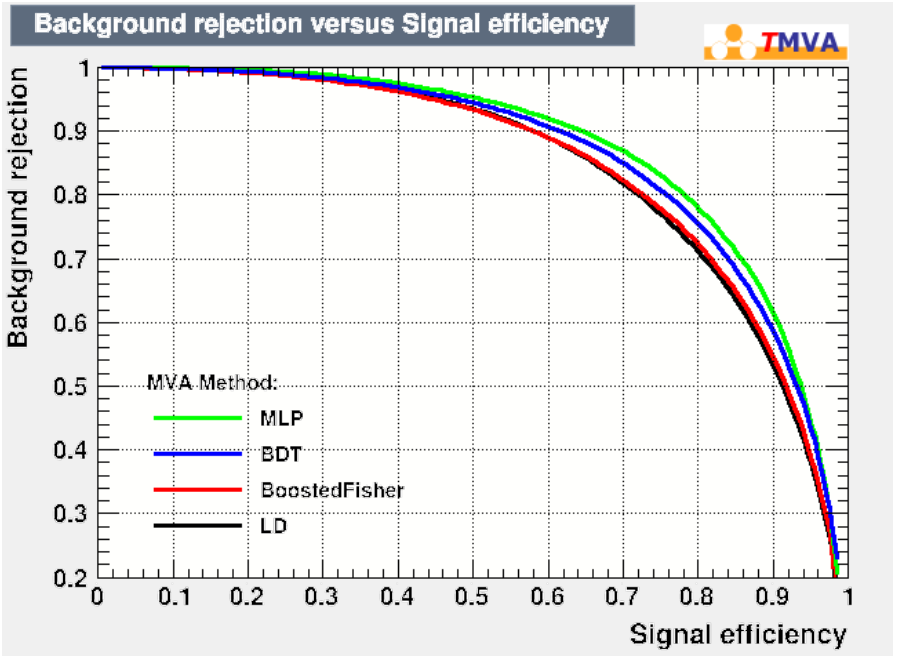
Variables seem to help on the discrimination



Multivariate analysis output

N.B. These values are based on the MC cross sections, there are no K-factors applied





Neural network (MLP) seems to be the best

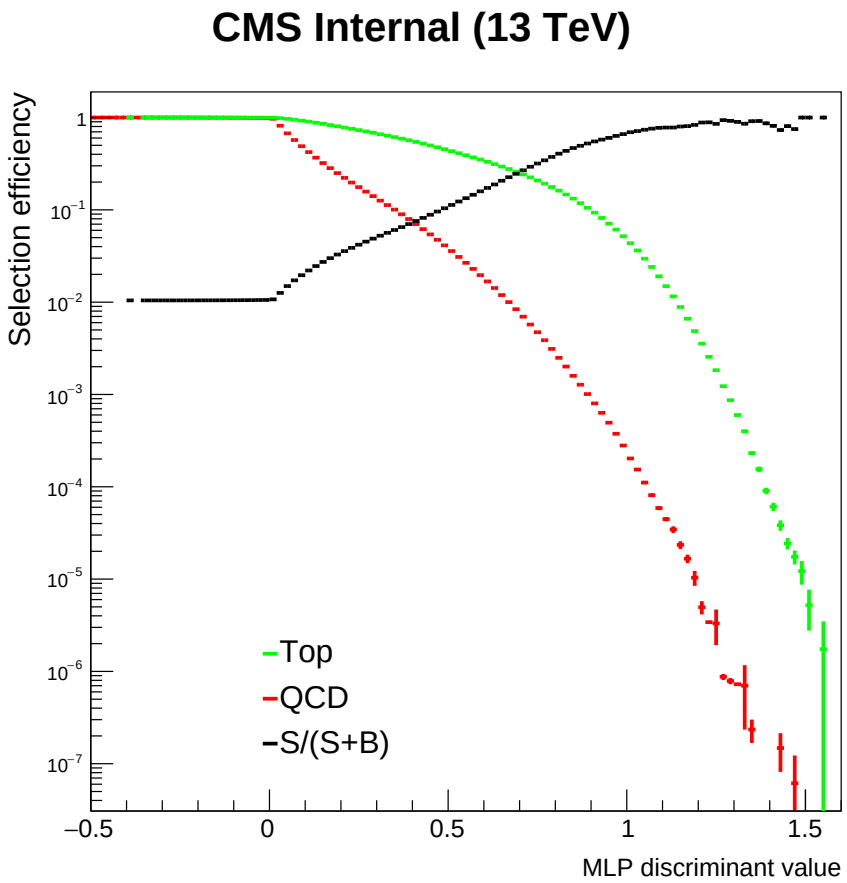
tanh activation function
600 training cycles
N+5 hidden layers (with N number of variables = 8)
Learning rate 0.02

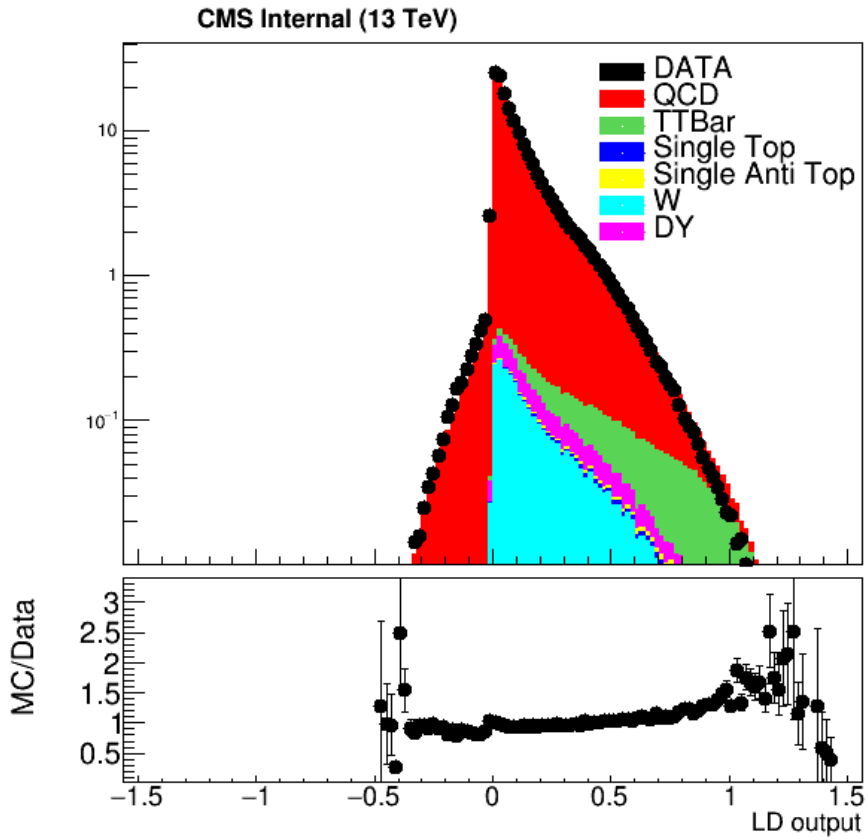
Variable importance from BDT training

```
--- BDT  : Ranking result (top variable is best ranked)
--- BDT  : ----------------------------------------------------
--- BDT  : Rank : Variable               : Variable Importance
--- BDT  : ----------------------------------------------------
--- BDT  :    1 : tau31SubLeading        : 2.310e-01
--- BDT  :    2 : tau32Leading           : 1.684e-01
--- BDT  :    3 : tau32SubLeading        : 1.577e-01
--- BDT  :    4 : tau31Leading           : 1.329e-01
--- BDT  :    5 : MassLeading            : 9.262e-02
--- BDT  :    6 : MassSubJet0SubLeading  : 9.184e-02
--- BDT  :    7 : MassSubJet0Leading     : 8.143e-02
--- BDT  :    8 : MassSubLeading         : 4.410e-02
--- BDT  : ----------------------------------------------------
```

Multivariate analysis output

MLP considered (in TOP-16-015, Fisher discr. is used)



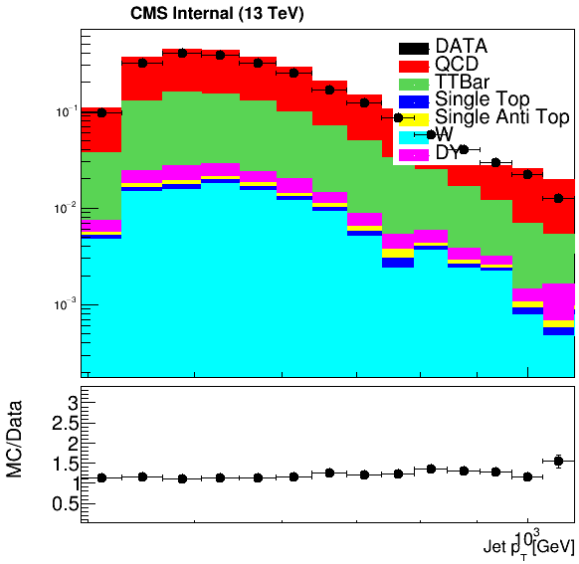


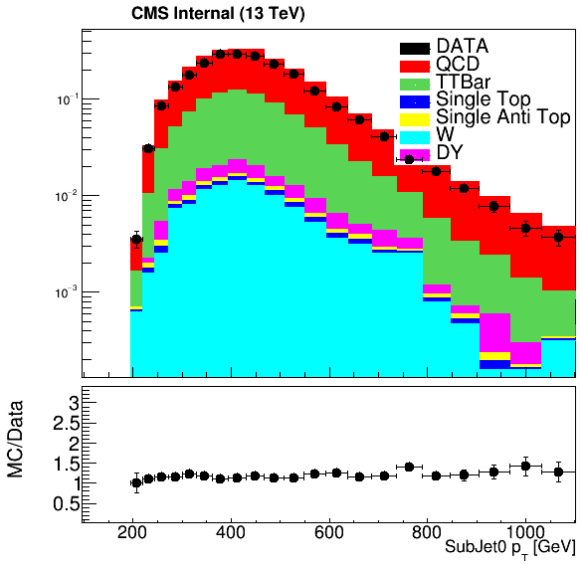
MLP output very well described by the simulation!

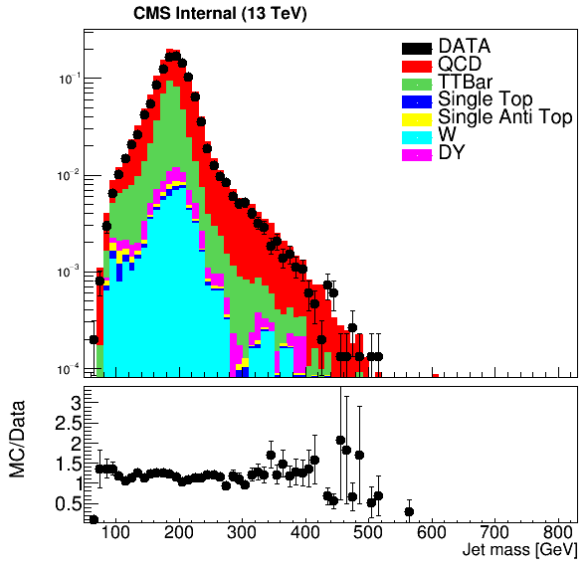
Output threshold>0.65 (sig. eff. $\sim$ 28%, bkg. rej. $\sim$ 97%)

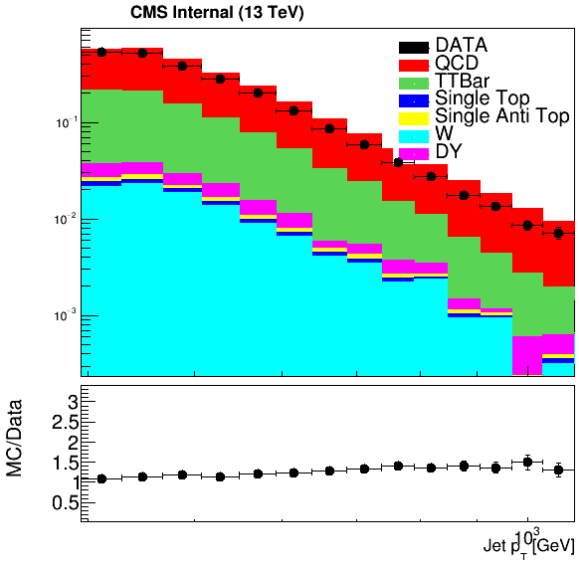
After MVA cuts

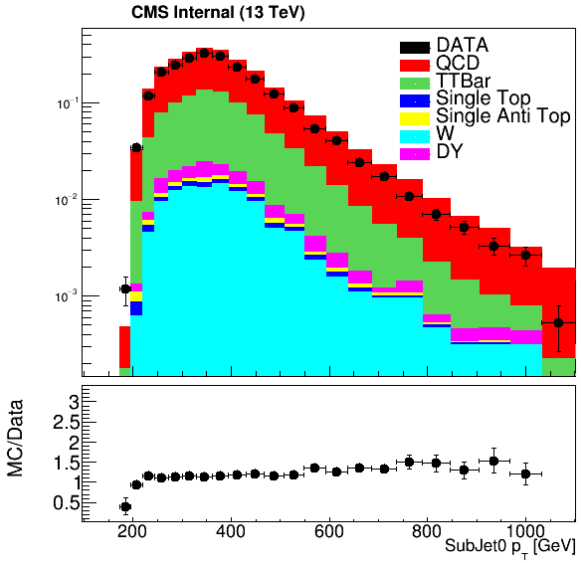
LEFT to RIGHT, TOP (leading jet) BOTTOM (subleading):
jet p_T, first subjet p_T, jet mass

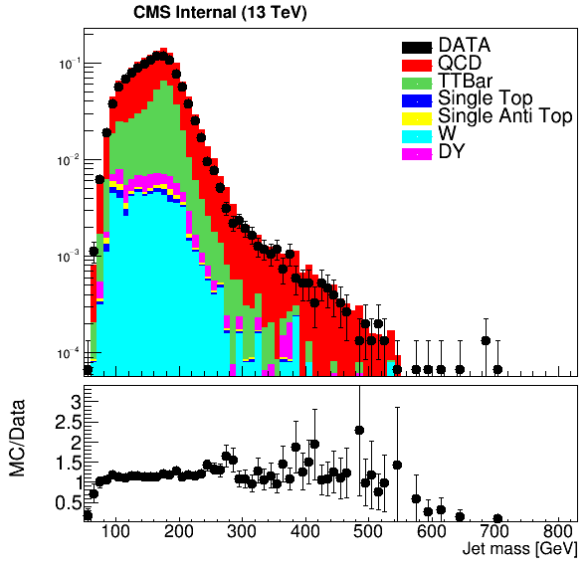












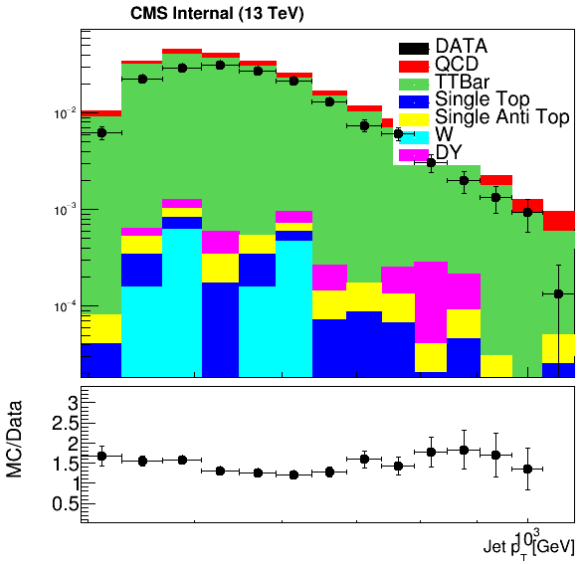
Additional requirement:

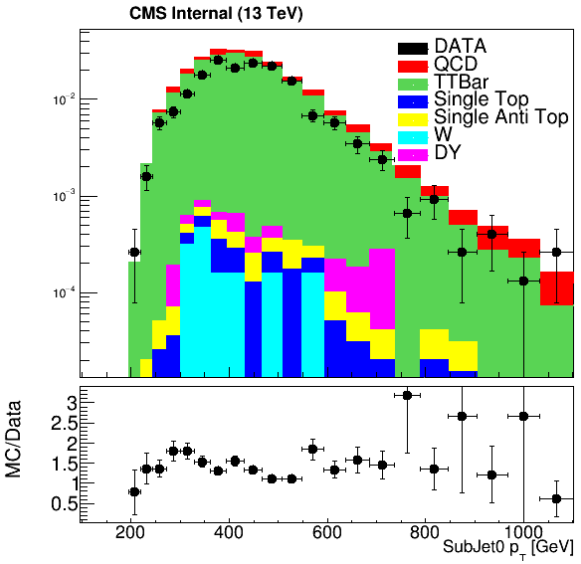
At least one of the two SD subjets needs to be tight b-tagged for both selected jets

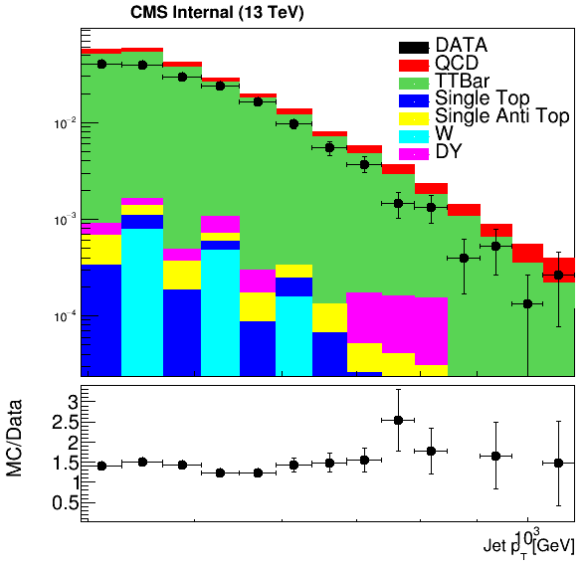
(sig. eff. $\sim$ 28% ($\cdot$ 28%), bkg. eff. $\sim$ 0.6% ($\cdot$ 3%))

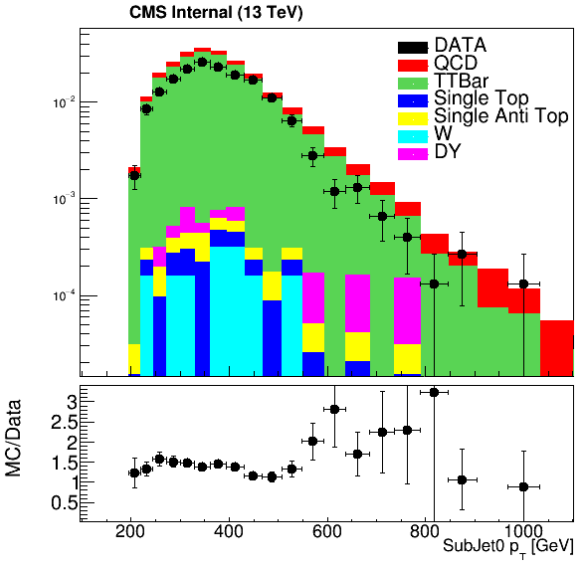
After "final" cut

LEFT to RIGHT, TOP (leading jet) BOTTOM (subleading):
jet p_T, subjet p_T









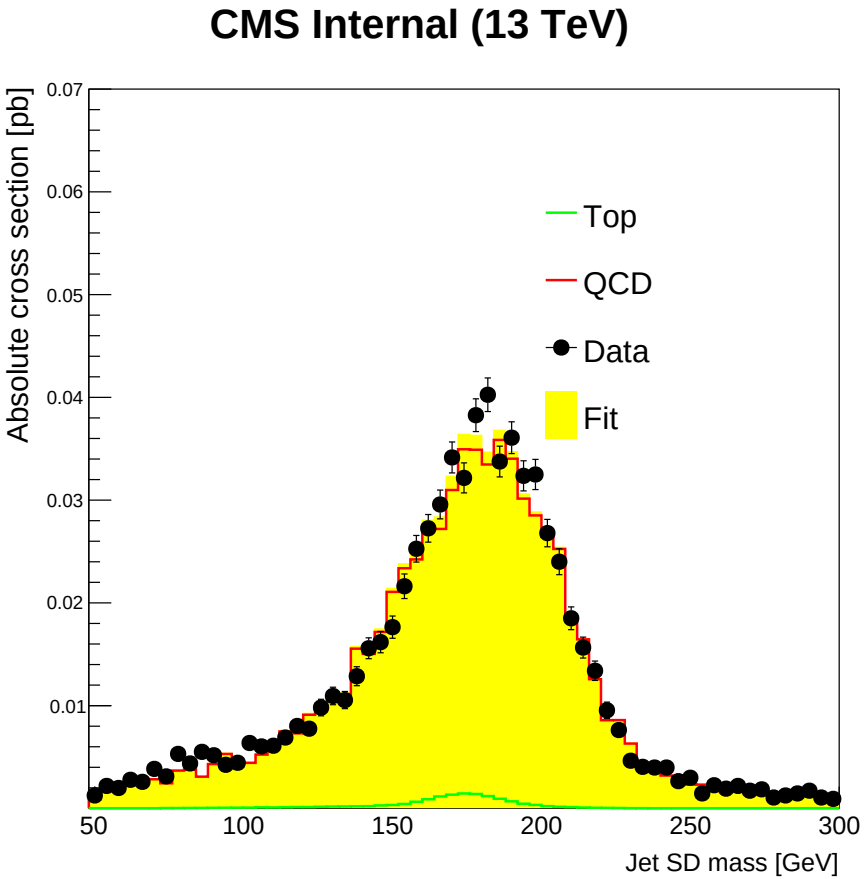
OUTCOME:

- No change in observable description
- Other background contributions become negligible
- The contribution of ttbar starts to increase (clear mass peak)

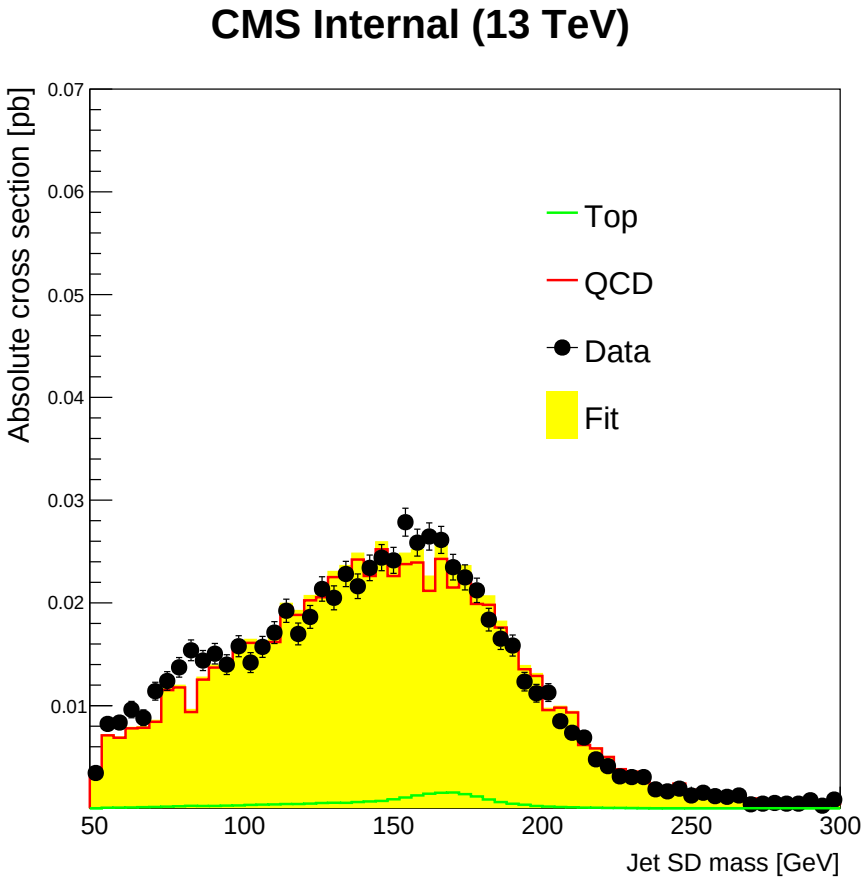
Selected events in data: 1303

Fitting the QCD normalization (data-driven)

Requiring now anti-b tag for both jets - control region
Fitting two SD jet masses (in anti b-tag region)



$\mathrm{QCD} = 0.69 \cdot \sigma_{MADGRAPH}$



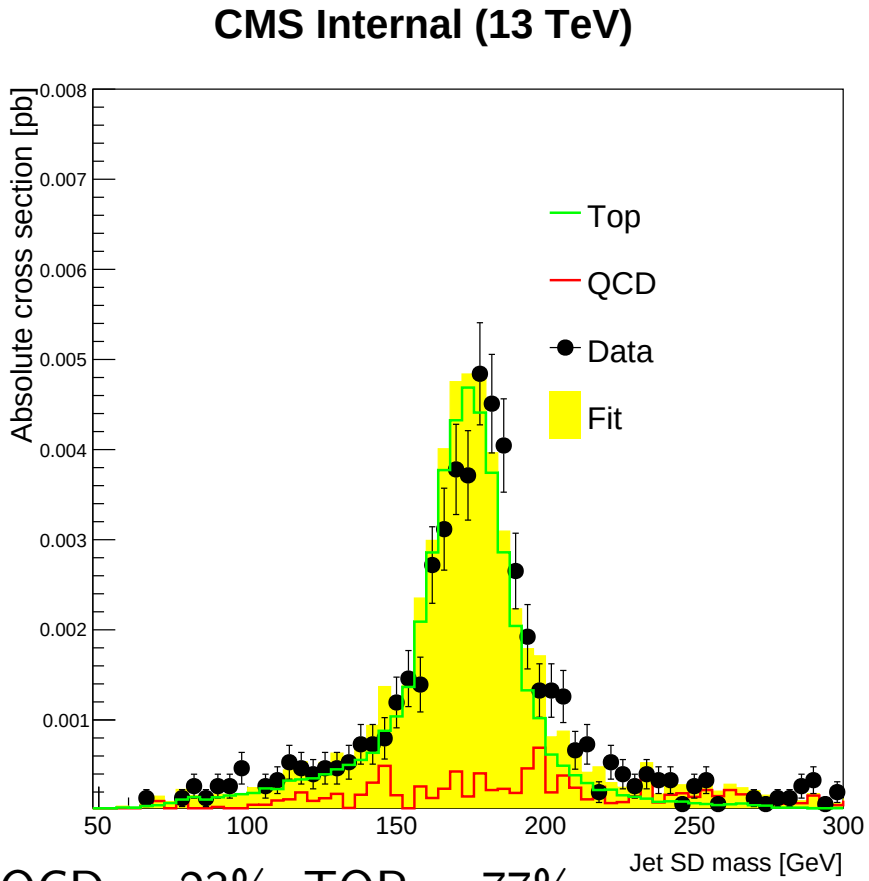
$\mathrm{QCD} = 0.67 \cdot \sigma_{MADGRAPH}$

Fitting the top signal

Fitting the top yield against soft-drop jet masses (sign. region)

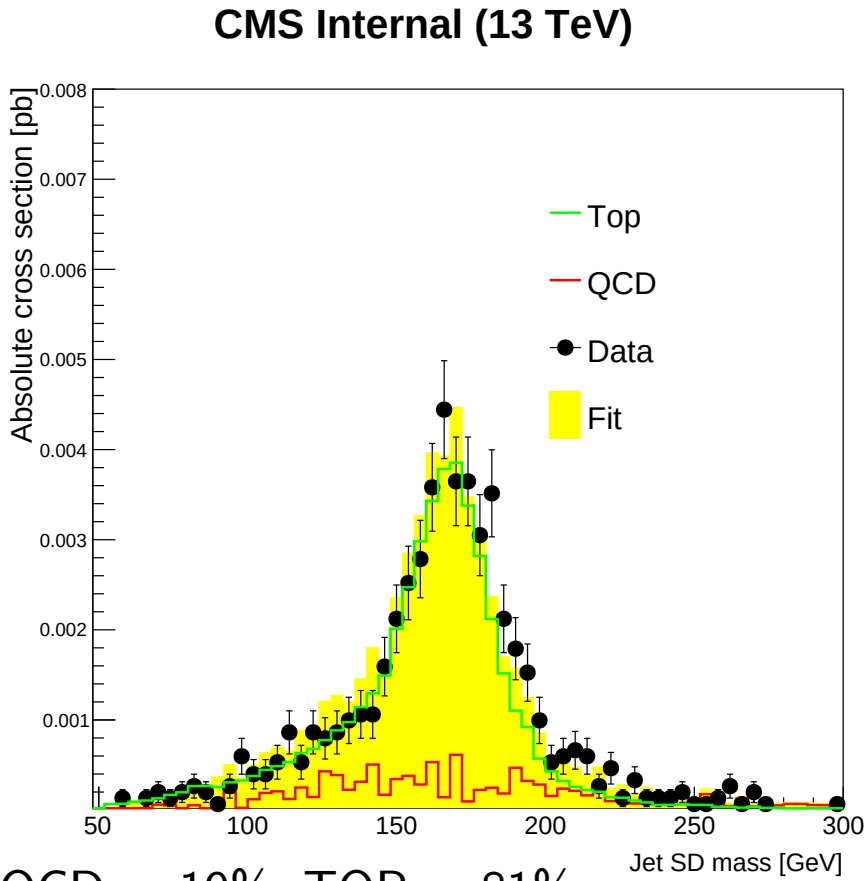
Using the QCD normalization from the control region

N.B. Absolute cross sections measured!



QCD = 23%, TOP = 77%,

χ^2/Ndf $\sim$ 91/64 S/B $\sim$ 4.76



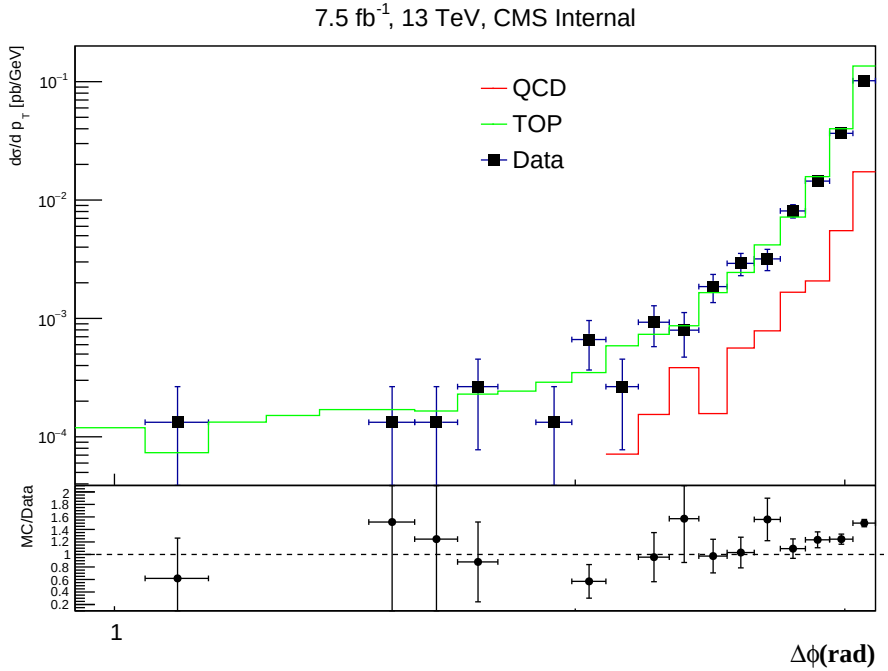
QCD = 19%, TOP = 81%,

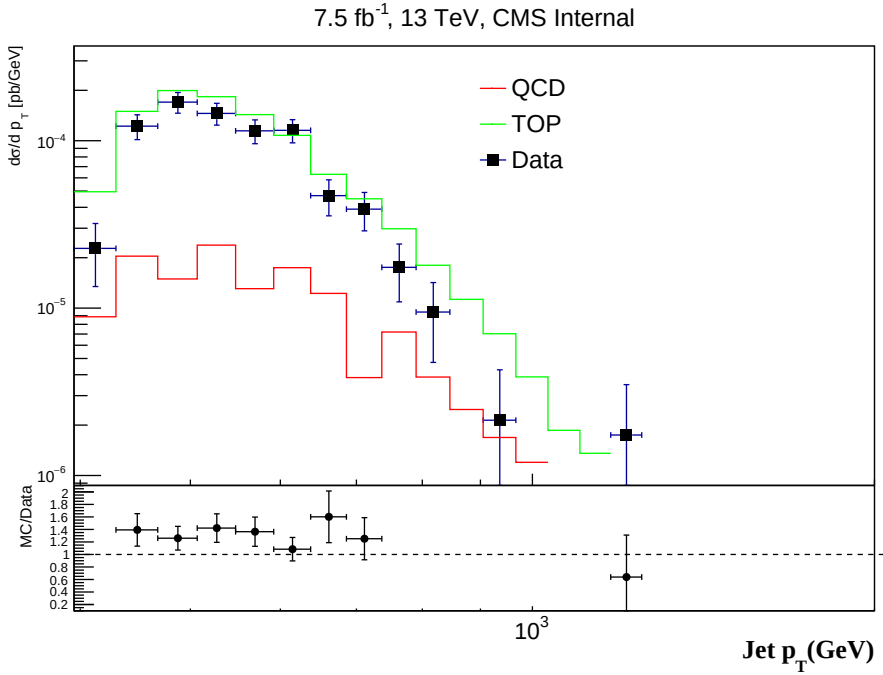
χ^2/Ndf $\sim$ 44/64 S/B $\sim$ 5.07

Observables of interest

Using the QCD norm. obtained from the fits to SD masses

N.B. Absolute cross sections measured!





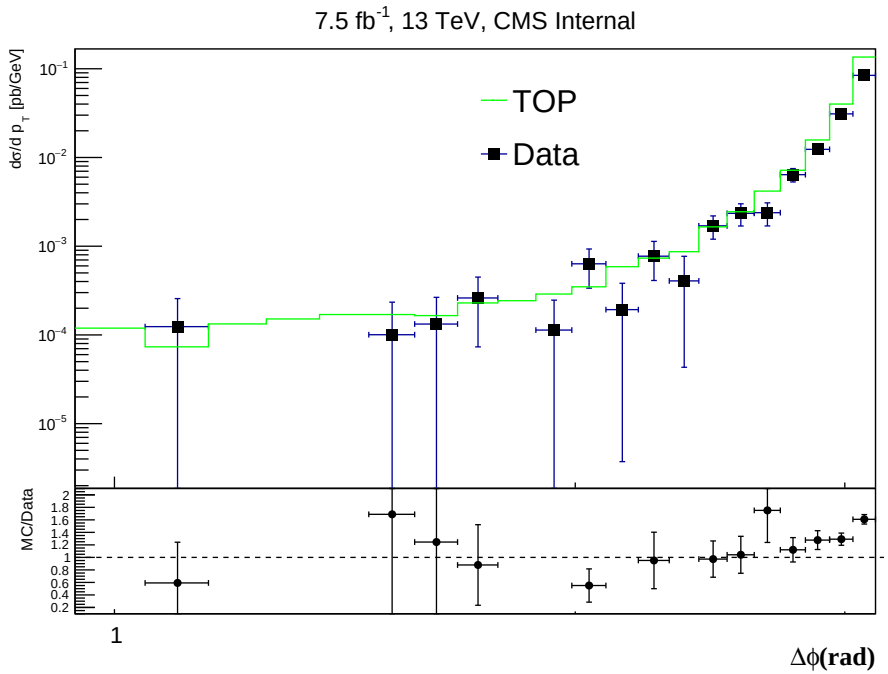
Statistics of data can be increased of a factor of 5

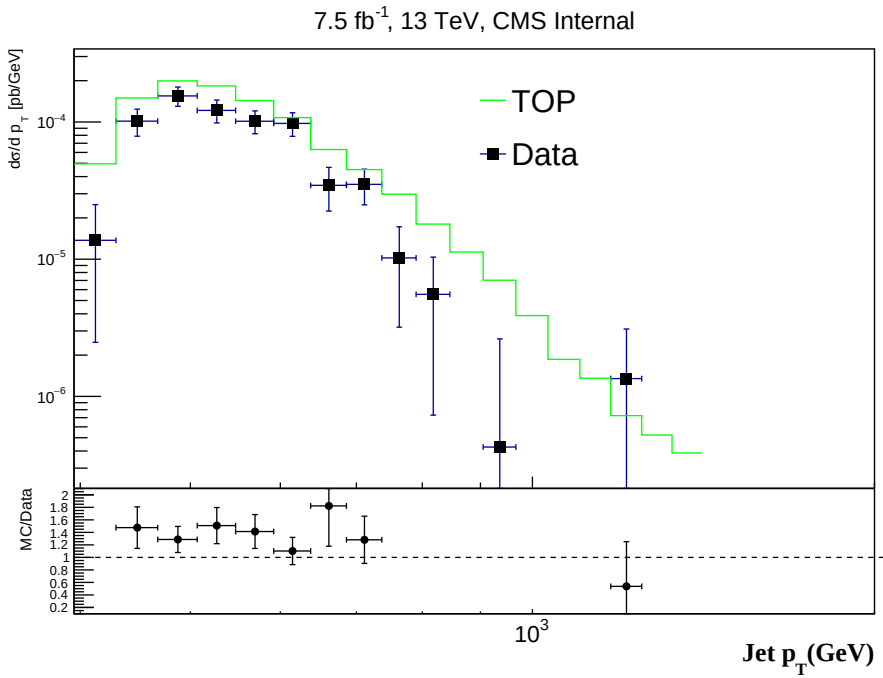
Difference in normalization between leading and subleading will be used as systematical uncertainty of background subtraction

Observables of interest

Using the QCD norm. obtained from the fits to SD masses and subtract from data

N.B. Absolute cross sections measured!



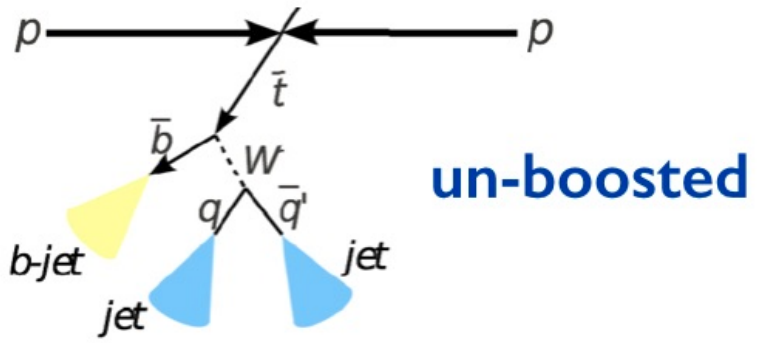


Statistics of data can be increased of a factor of 5

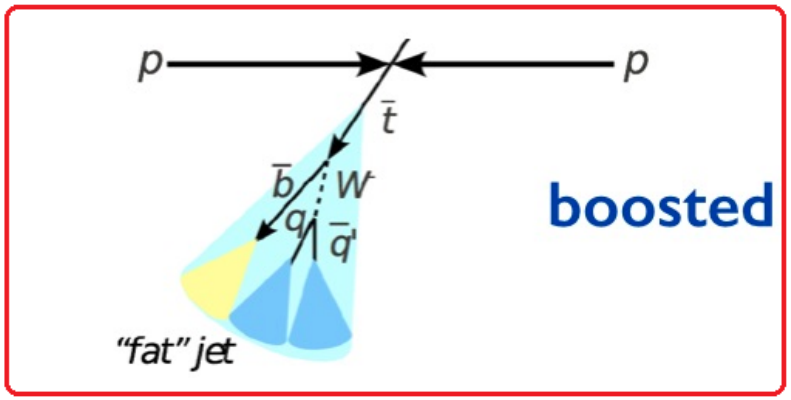
Difference in normalization between leading and subleading will be used as systematical uncertainty of background subtraction

Plans for next steps

Particle-level definition



un-boosted



boosted

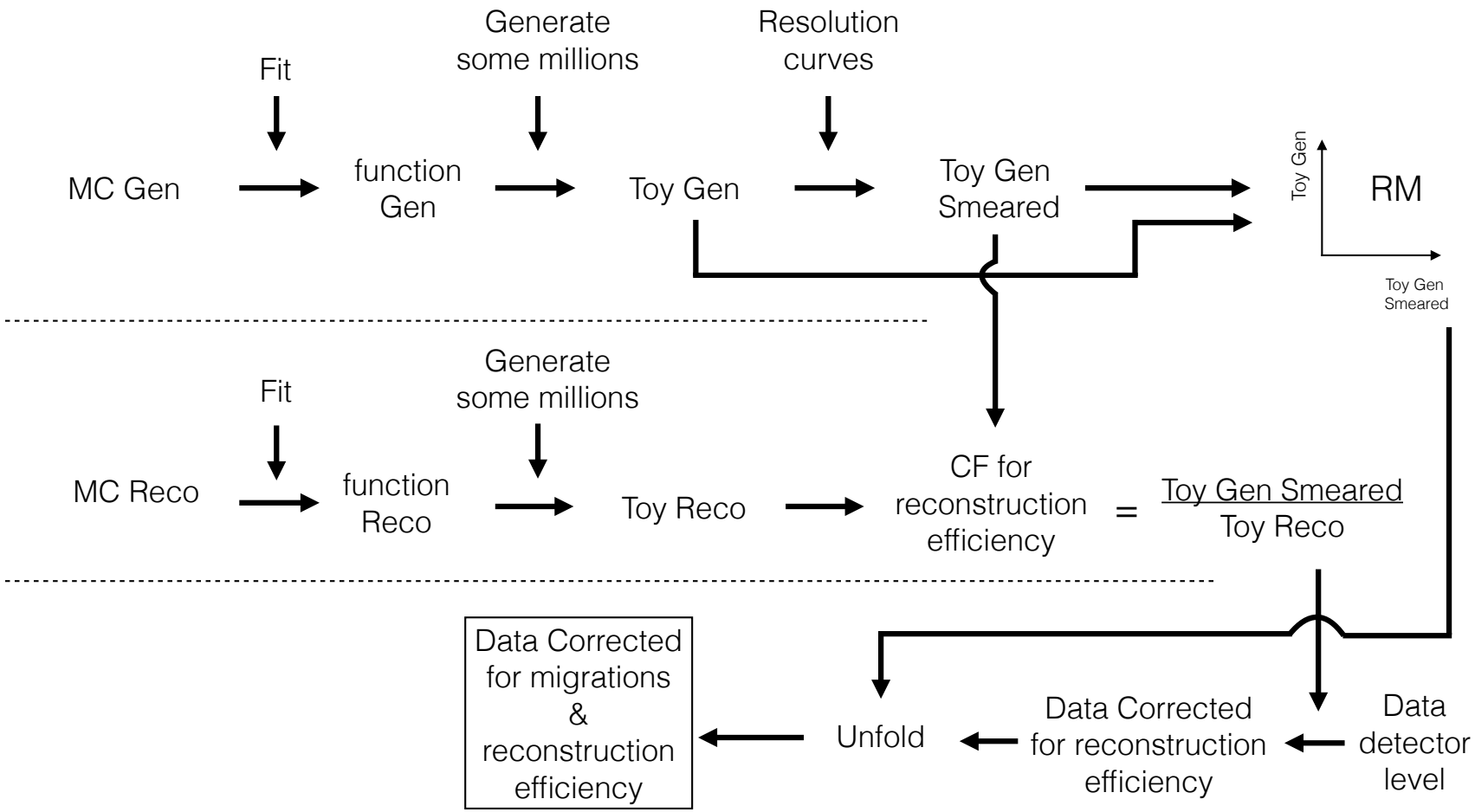
Taken from a presentation by Daniela Dominguez

$\rightarrow$ Two gen jets with $p_T > 400$ GeV in $|\eta| < 2.4$

$\rightarrow$ Both jets contain a b-hadron among the constituents

$\rightarrow$ W-boson within $\Delta R < 0.4$ with respect to the jet axis

Unfolding strategy (in sketch)



Taken from a presentation by Juan Grados

Plans for next steps

Unfolding strategy (in words)

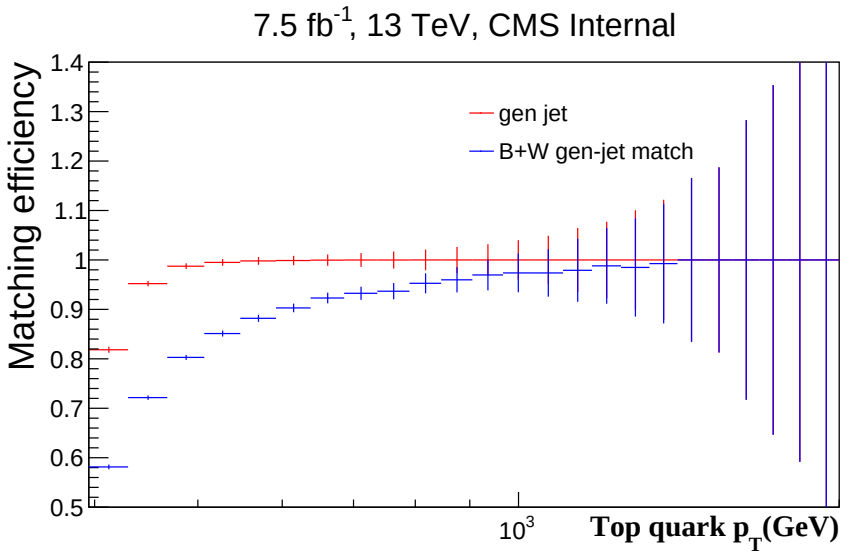
1. Fit to the GEN level distribution (particle-level cuts - previous slide)
2. Fit to the DET level distribution (analysis cuts)
3. Smear of GEN fit with a toy MC according to detector resolution $\rightarrow$ response matrix
4. Ratio between GEN-smeared and DET distributions for evaluation of acceptance efficiency
5. Correction of data through the acceptance efficiency
6. Unfolding with response matrix from toy MC

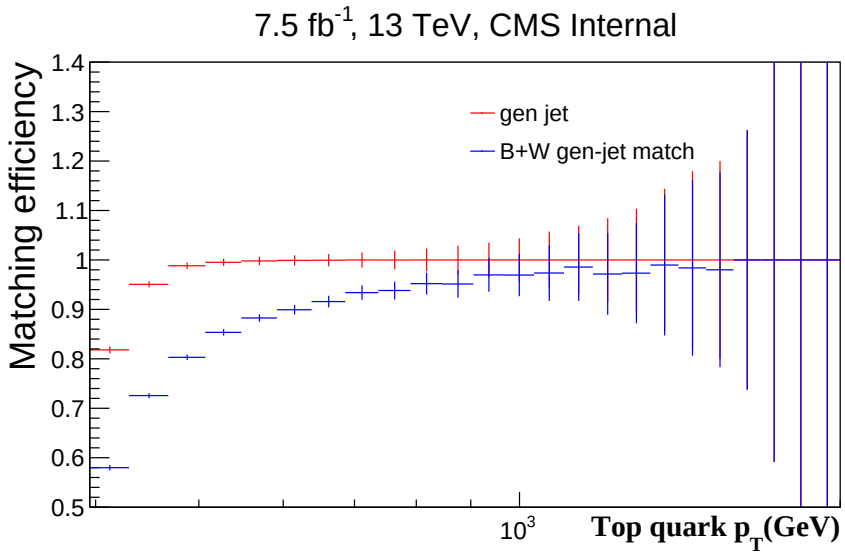
Main issue is the number of events selected in MC (34363), not sufficient for an acceptable determination of the response matrix

Matching studies

Consider the leading/subleading top-quark parton vs p_T

Match to a gen-jet ($p_T > 400$ GeV) through a $\eta - \phi$ matching algorithm: $\Delta R < 0.4$





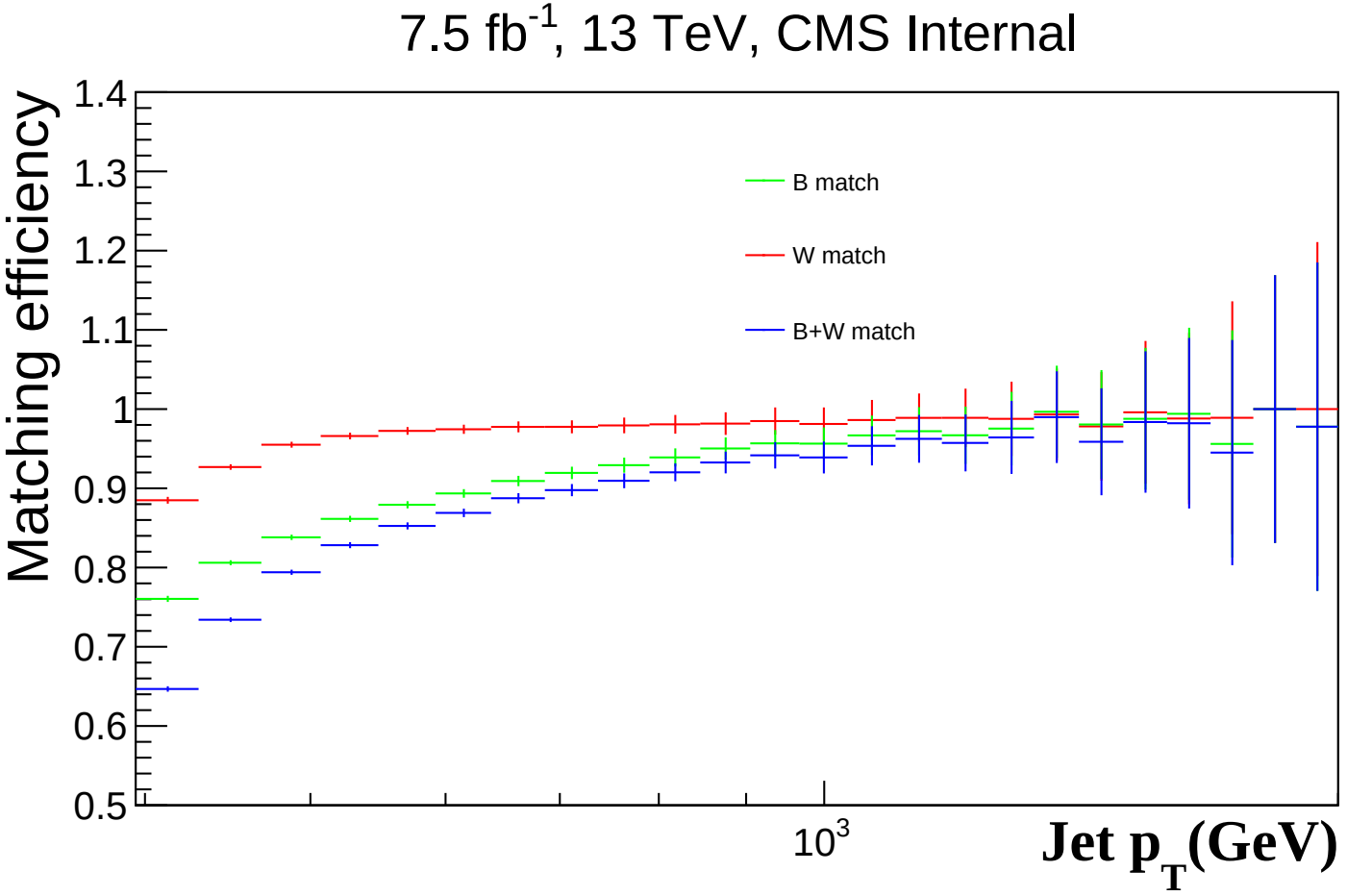
At high p_T, matching efficiency $\sim 100\%$

For b-W matching, the maximum is at $\sim 95\%$

Matching studies

Consider the a gen jet with $p_T > 400$ GeV

Look at how many times they are b- and W-tagged



Higher efficiency for W-tag, than for b-tag

The b-hadron takes little p_T and its direction is more "randomized"

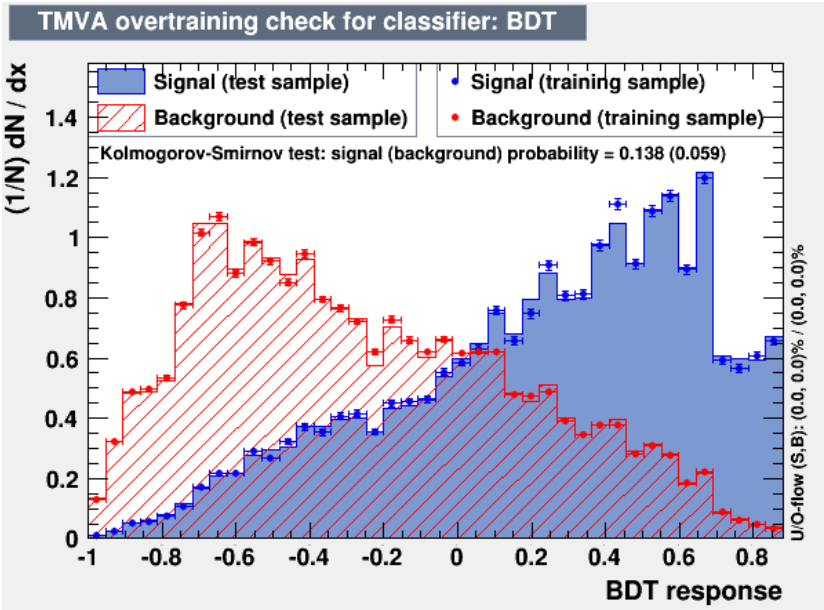
Summary

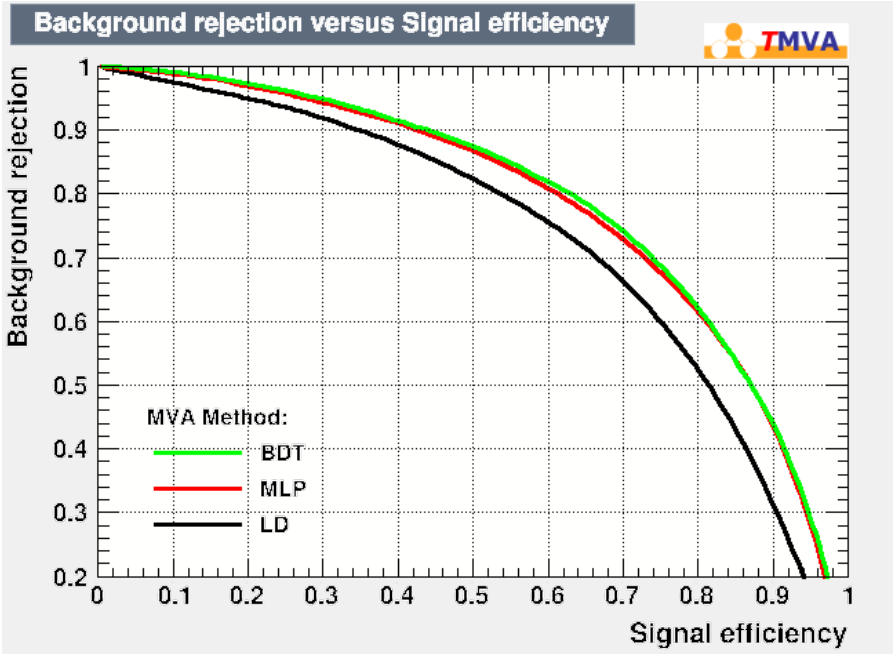
- Measuring top jets at high p_T can demonstrate the QCD flavour-blindness at high scales and probe QCD radiation as a function of transverse momentum
- Data are under study and analysis is set-up
- Selection optimization has been studied and brings to high purity of top signal
- TO DO:

1. Unfolding and systematic uncertainties

Multivariate analysis output with 4 variables (subjettiness)

Used methods without any optimization yet..using the default settings!





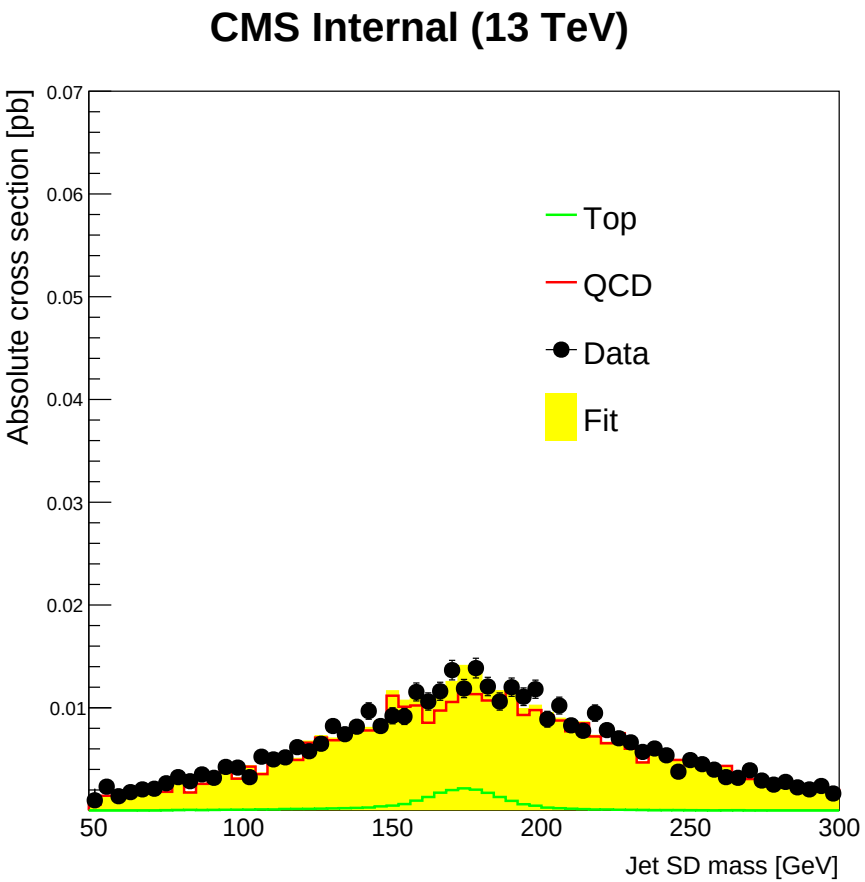
N.B. These values are based on the MC cross sections, there are no K-factors applied

No striking difference among different selection methods

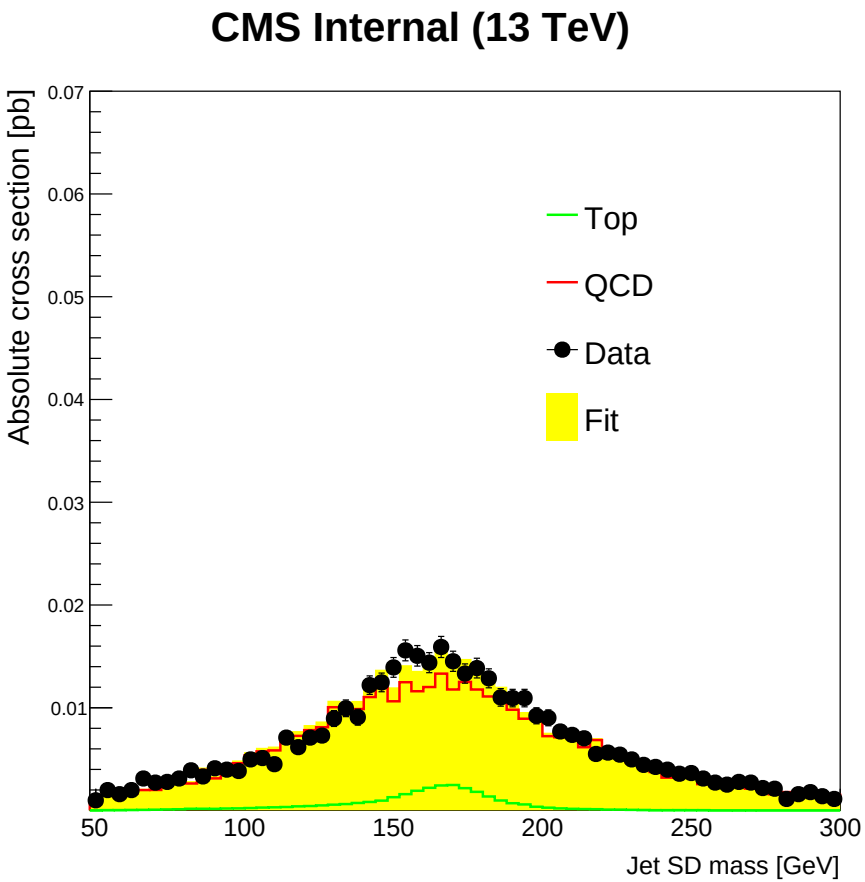
Trying to fit the relative contributions

Fitting two SD jet masses (in anti b-tag region)

N.B. Absolute cross sections measured!



$\mathrm{QCD} = 0.66 \cdot \sigma_{MADGRAPH}$



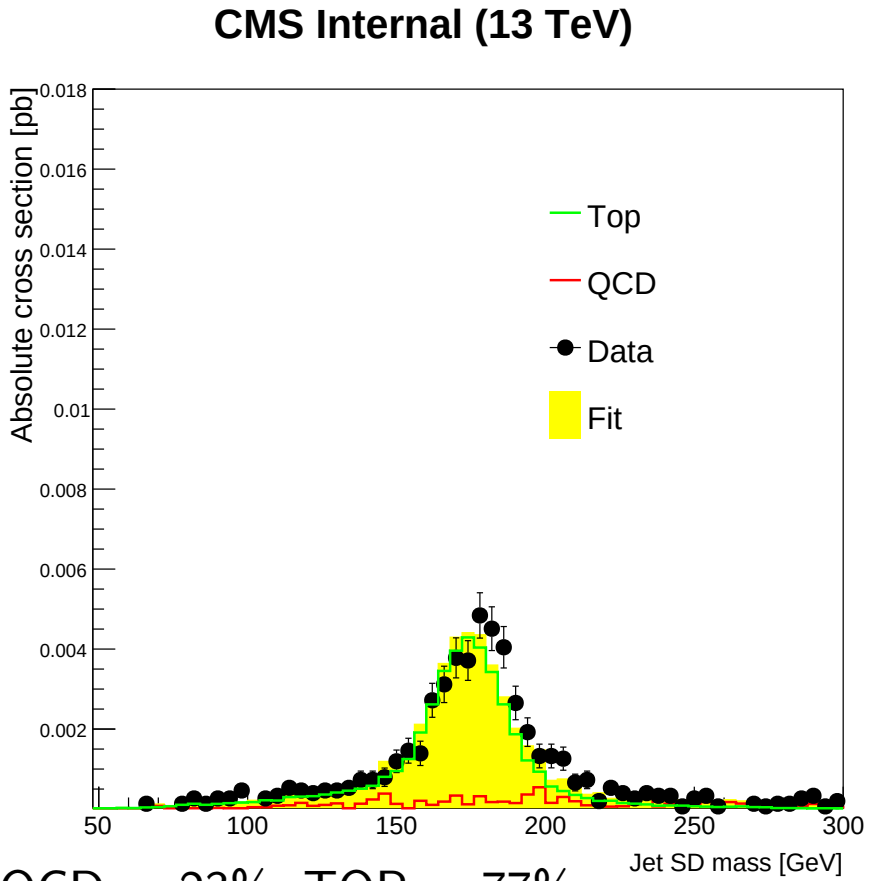
$\mathrm{QCD} = 0.63 \cdot \sigma_{MADGRAPH}$

Trying to fit the relative contributions

Fitting the top sample yield against soft-drop jet masses

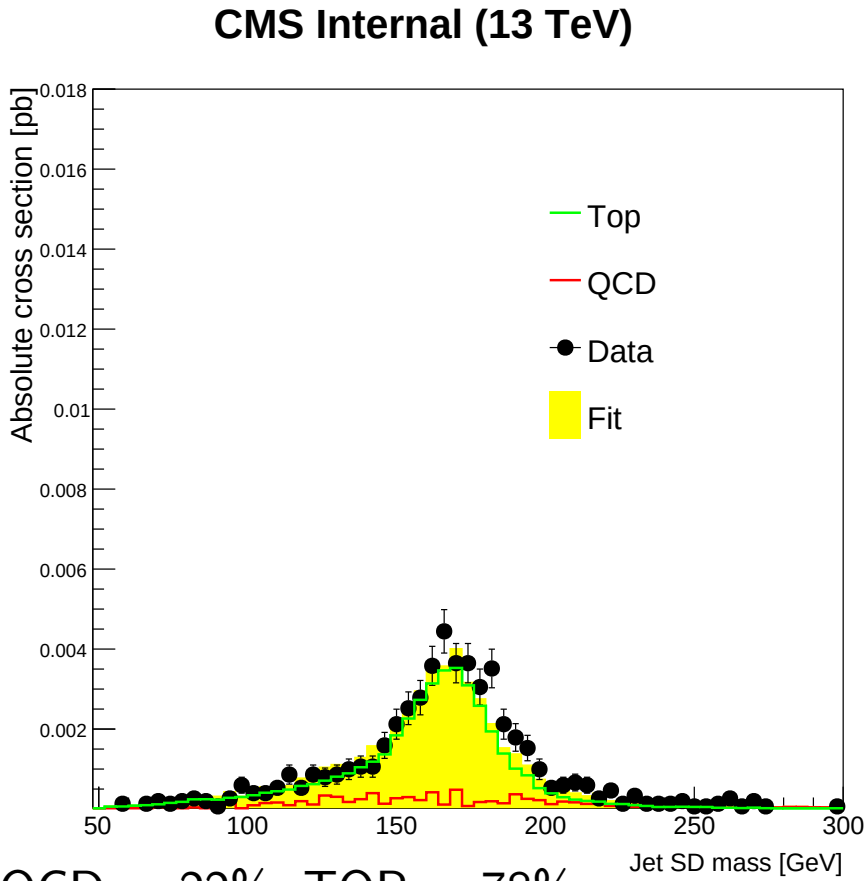
Using the QCD normalization from the control region

N.B. Absolute cross sections measured! (selected events: 806)



QCD = 23%, TOP = 77%,

χ^2/Ndf $\sim$ 68/64 S/B $\sim$ 4.73



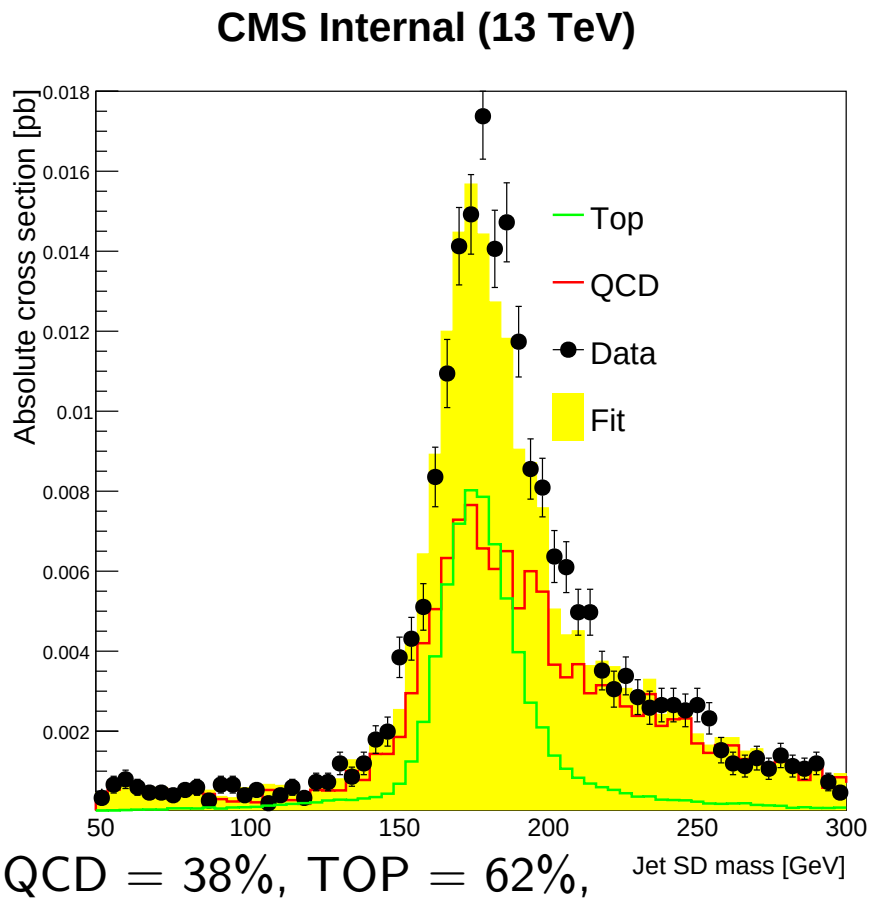
QCD = 22%, TOP = 78%,

χ^2/Ndf $\sim$ 70/64 S/B $\sim$ 4.96

MVA only with mass information

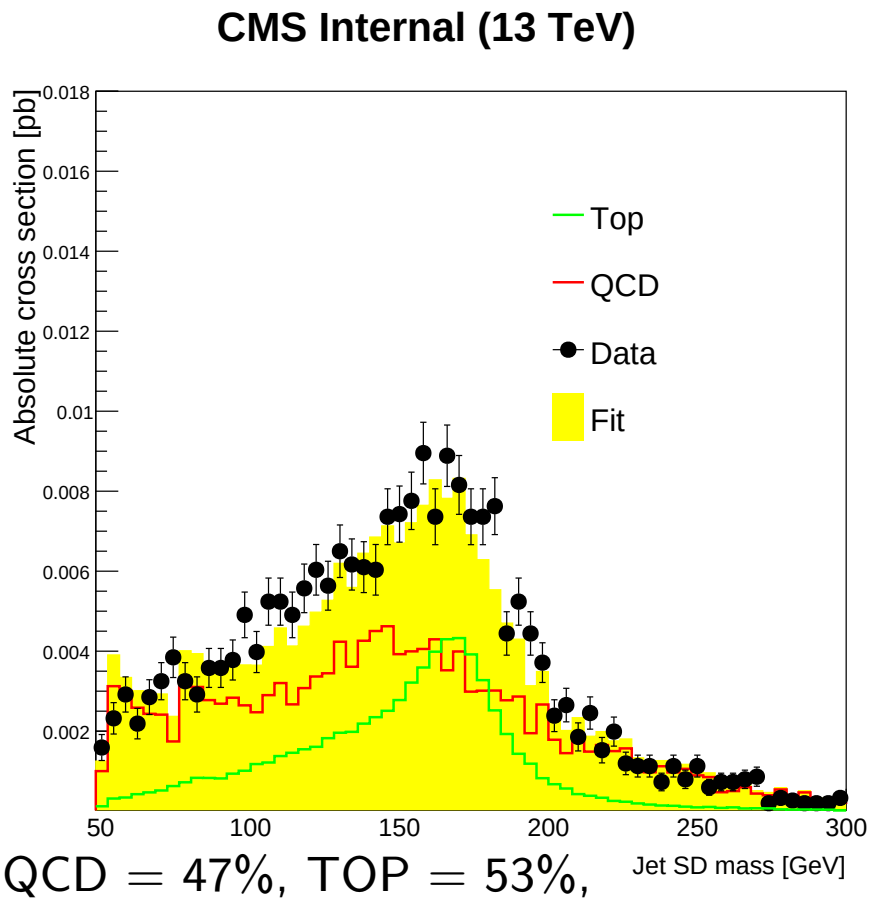
Fitting two contributions against soft-drop jet masses

N.B. Absolute cross sections measured!



QCD = 38%, TOP = 62%,

$\frac{S}{B} \sim 1.66$

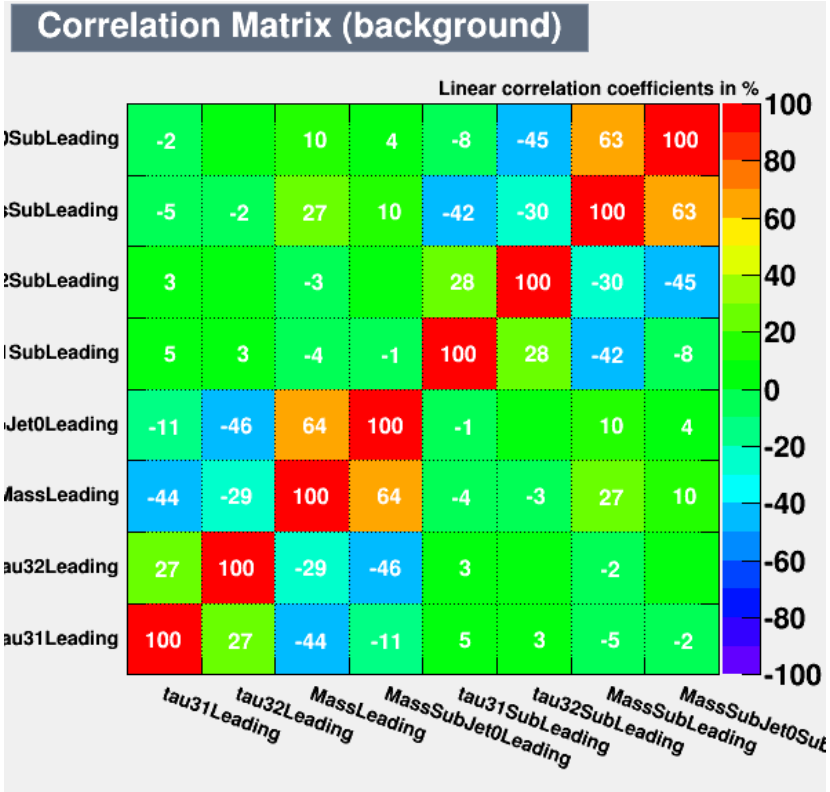


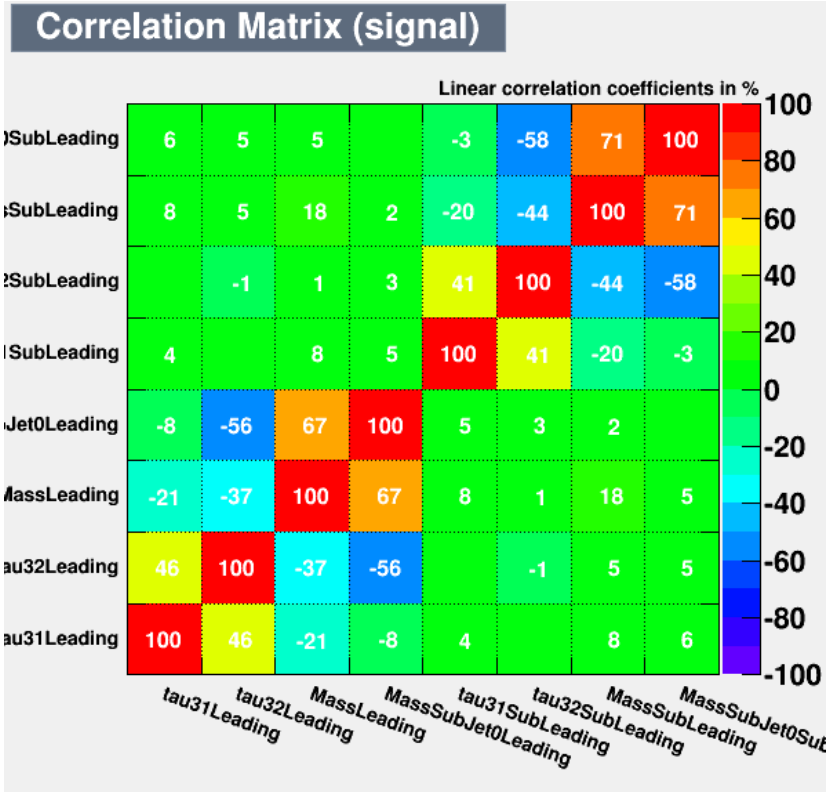
QCD = 47%, TOP = 53%,

$\frac{S}{B} \sim 1.15$

Correlations between variables

Checking the correlations between the variables (mainly different correlations between signal and background)





No correlations between variables of the leading and subleading jets