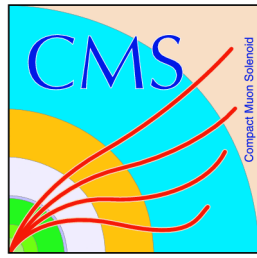




Universität Hamburg

DER FORSCHUNG | DER LEHRE | DER BILDUNG



V + top tagging in CMS

BOOST2014, London

Tobias Lapsien (University of Hamburg)
on behalf of the CMS collaboration

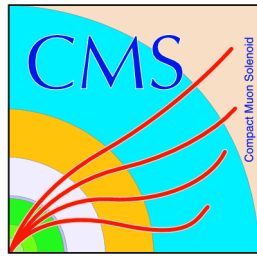
Content

- Top Tagging
 - Introduction
 - Performance in Simulation
 - Performance in Data
- V Tagging
 - Discriminating variables
 - Resolved jets
 - Unresolved jets
- Summary



Universität Hamburg

DER FORSCHUNG | DER LEHRE | DER BILDUNG

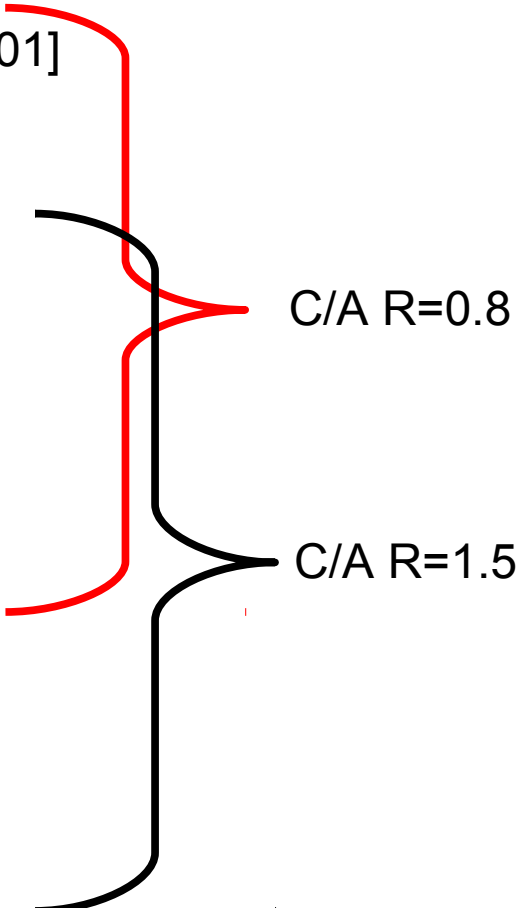


Top Tagging

CMS PAS JME-13-007

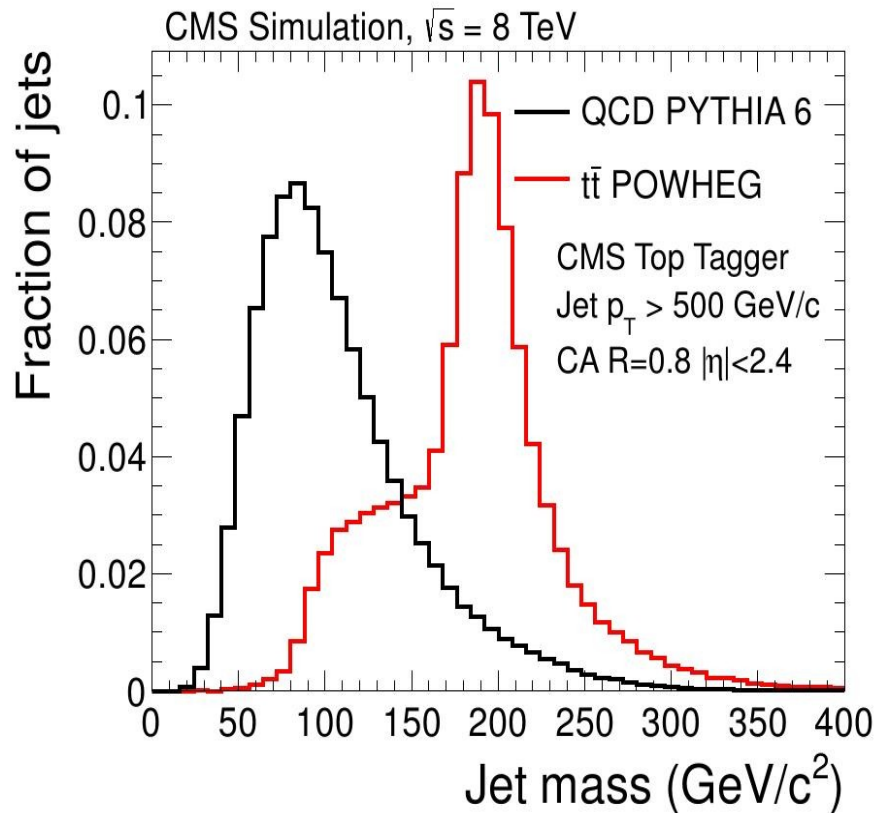
Top tagging algorithms



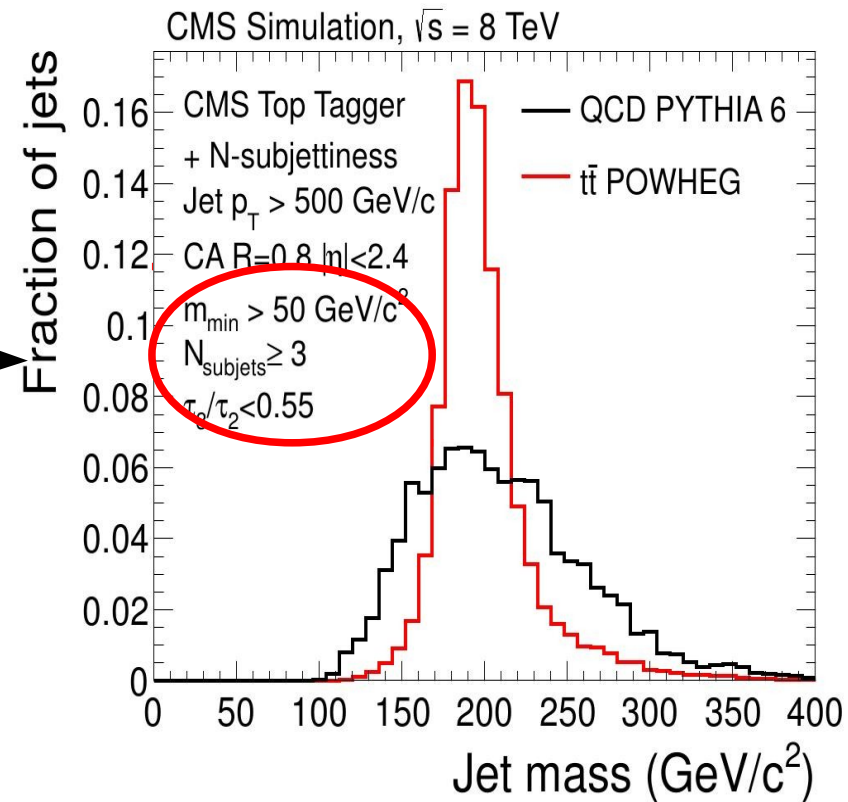
- CMS top tagger
[D. E. Kaplan et al.: Phys. Rev. Lett. 101 (2008) 142001]
 - N-subjettiness
[J. Thaler and K. Van Tilburg: JHEP 1103 (2011) 015]
 - Subjet b-tagging
[CMS Collaboration, CMS PAS BTV-13-001]
 - Shower deconstruction
[D. E. Soper, M. Spannowsky: arXiv:1211.3140v1]
 - HEP top tagger
[T. Plehn et al., JHEP 1010 (2010) 078]
 - MultiR HEP top tagger
[Plehn et al.]
- 
- C/A R=0.8
- C/A R=1.5

Discriminating variables for the CMS top tagger

Jet mass

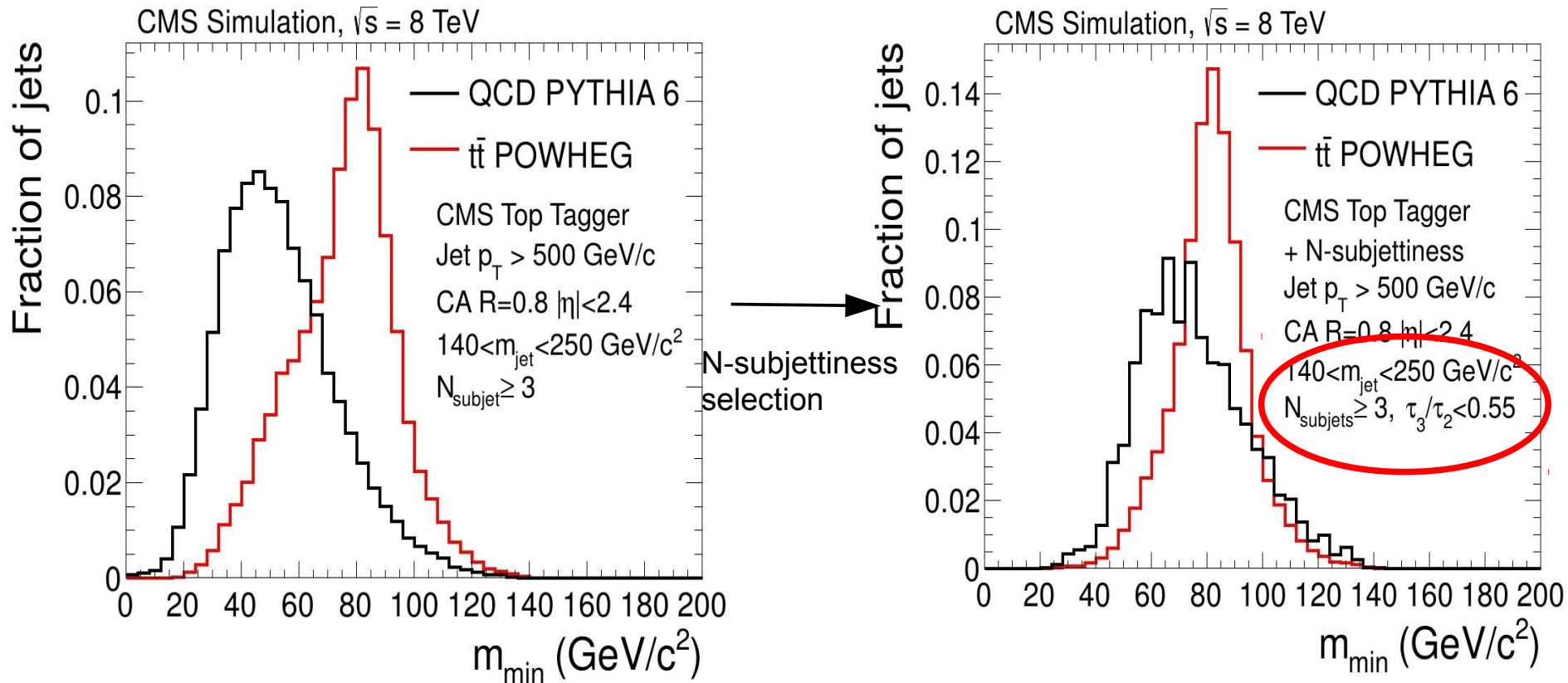


Tagging selection



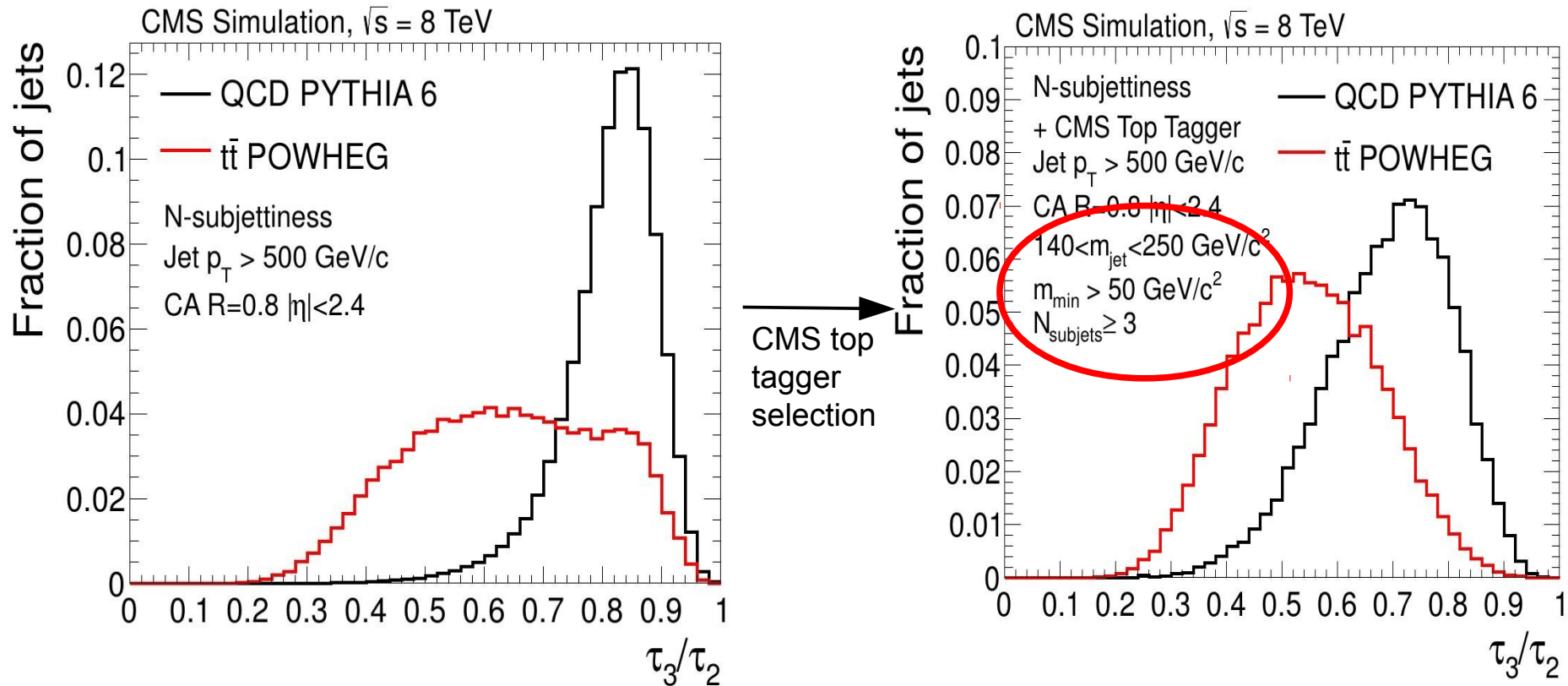
Discriminating variables for the CMS top tagger

Minimum pairwise mass

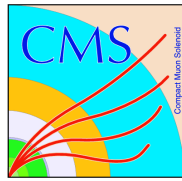


Discriminating variables for the CMS top tagger

N subjettiness

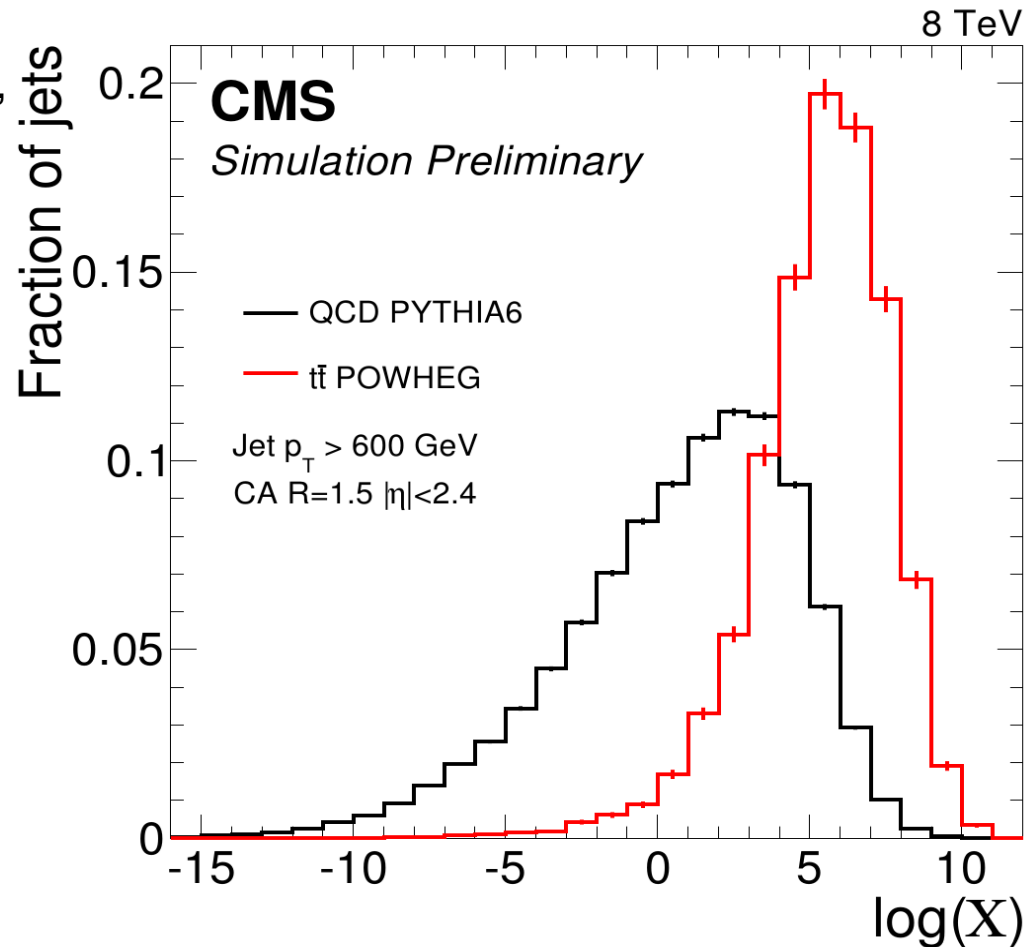


Shower deconstruction



- variable χ : probability quotient, a set of microjets in a fatjet were created by the decay of a top quark, divided by the probability that they were created by light quarks and gluons [1]:

$$X(\{p\}_N) = \frac{P(\{p\}_N|S)}{P(\{p\}_N|B)}$$



[1] Davison E. Soper, Michael Spannowsky "Finding top quarks with shower deconstruction" (arXiv:1211.3140v1)

Shower deconstruction



- Microjets: clustering the jet constituents of the fat jet to smaller jets with cone size of $R\{0.1, \dots, 0.3\}$ with the k_T -algorithm
- microjets with $p_T > 10\text{GeV}$
- Different microjet cone sizes are used for different fat jet p_T regions (see table)
- Two versions of the shower deconstruction tagger available (C/A 8, C/A 15)

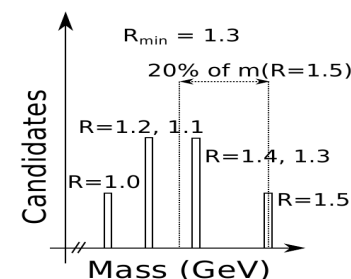
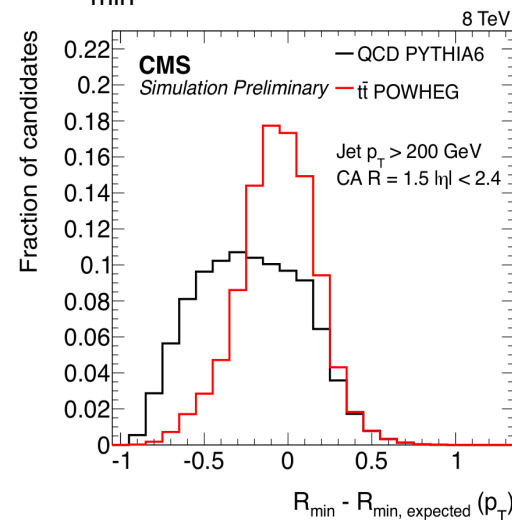
Parameter	Value
microjet R (k_T jets)	p_T dependent
max. number of leading microjets	9
microjet p_T	$> 10\text{ GeV}$
Fatjet R	0.8 / 1.5
Fatjet p_T	$> 200\text{ GeV}$
W mass	80.4 GeV
W window	$\pm 12\text{ GeV}$
Top mass	160 GeV
Top mass window	$\pm 25\text{ GeV}$
NBreitWigner	2

p_T range [GeV]	microjet cone size
0 - 500	0.3
500 - 700	0.2
700 - ∞	0.1

MultiR HEP top tagger



- Improved version of the HEPTopTagger, taking into account information at multiple cone sizes (see talk of Torben Schell)
- MultiR-Algorithm [Also documented in upcoming note by Plehn et al]:
 - Start with C/A, $R=1.5$ seed fat-jet
 - Perform unclustering to identify small fat-jets with $R=0.5$ to $R=1.5$ (in steps of 0.1) and run HEPTopTagger on each of them
 - Calculate: R_{\min} = Smallest cone size for which the mass differs by less than 20% from the mass at $R=1.5$
 - Calculate expected $R_{\min, \text{expected}}$: Expected R_{\min} for a signal jet as function of the filtered fat-jet p_T
 - Variables:
 - Top candidate mass: $m(R=R_{\min})$
 - W / top mass ratio: $f_W(R=R_{\min})$
 - R_{\min} difference: $R_{\min} - R_{\min, \text{expected}}$

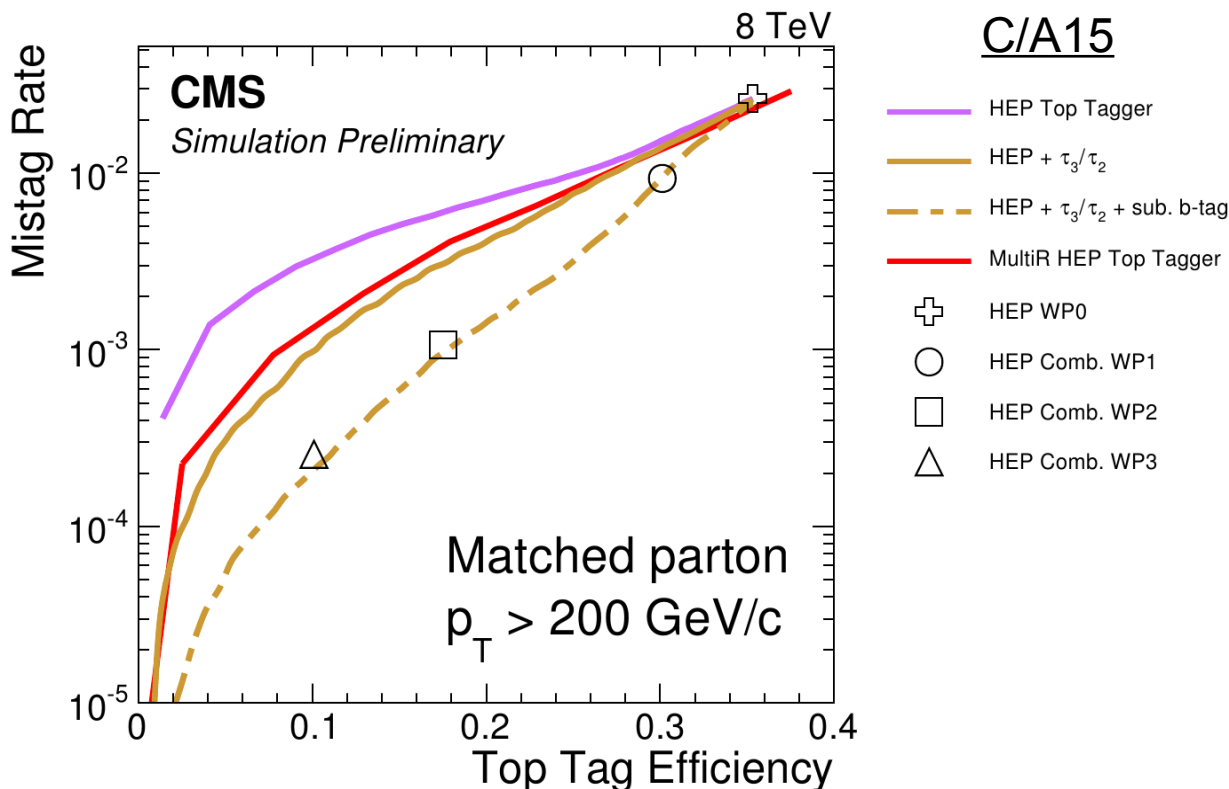


Definition of efficiency and mistag rate:

$$\epsilon = \frac{\text{tagged matched jets}}{\text{matched jets}}$$

For background jet matched to gluon/quark

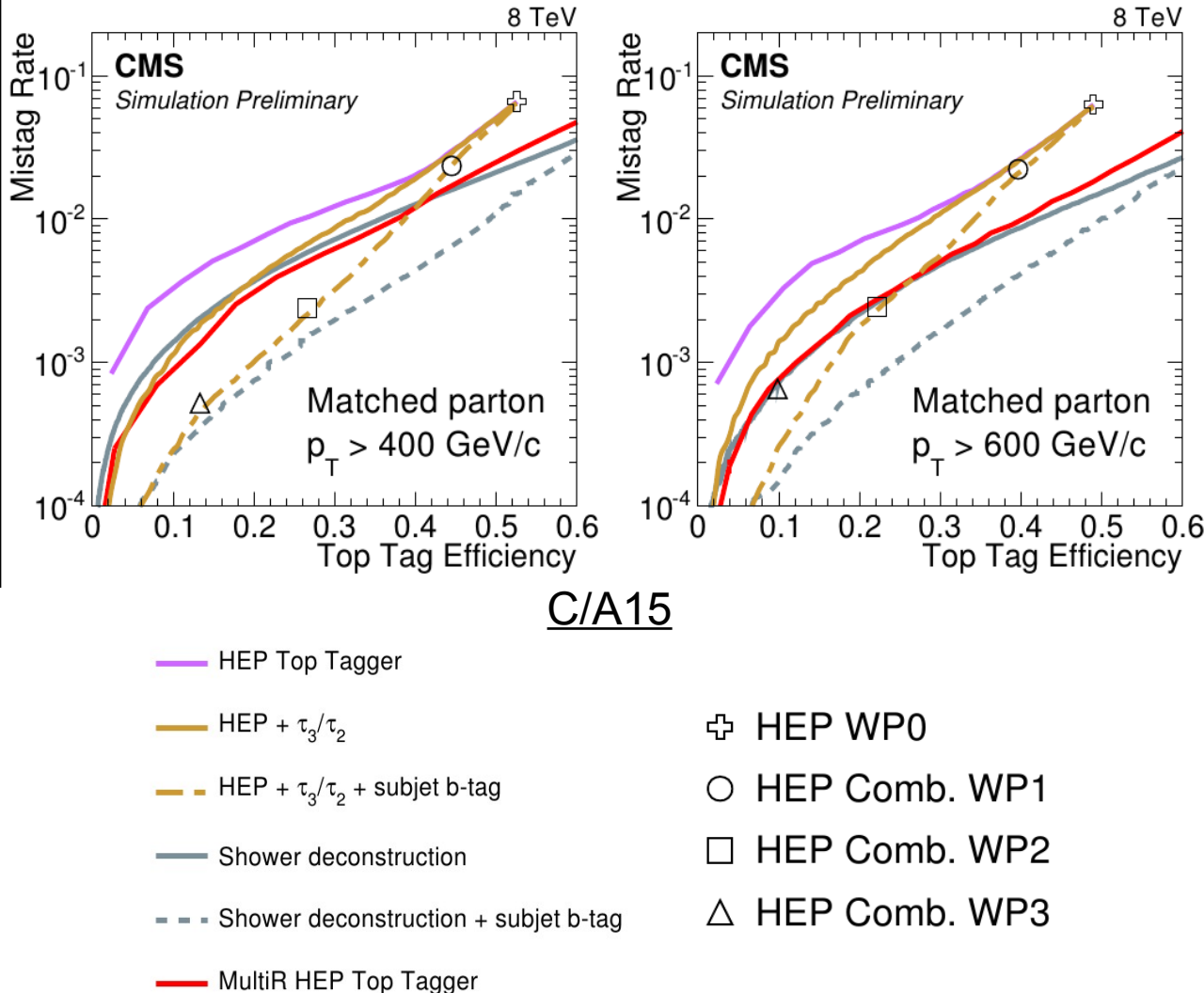
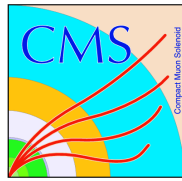
For signal jet matched to hadronically decaying top quarks and anti-top quarks



C/A15

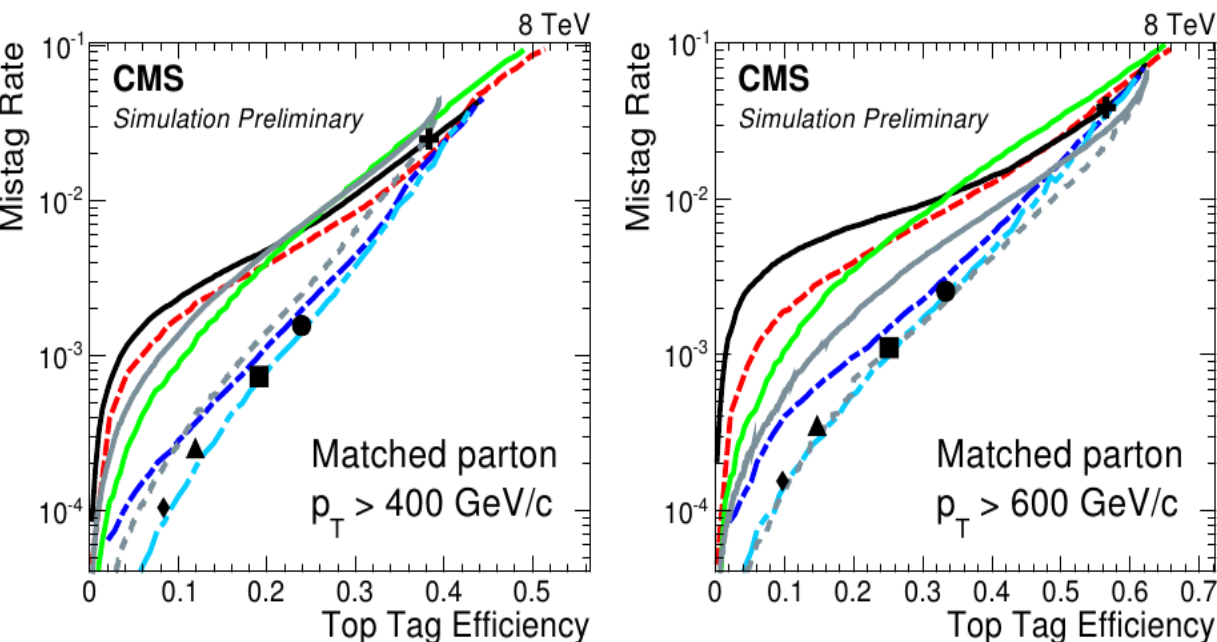
- HEP Top Tagger curve determined by fixing $140 \text{ GeV} < m_{123} < 250 \text{ GeV}$ and varying the width of the W mass selection (f_W)
- MultiR Hep Top Tagger curves are obtained by a parameter scan over three observables ($m_{\text{Jet}}(R=R_{\text{min}})$, $f_W(R=R_{\text{min}})$, ΔR)
- Subjet b-tag curve determined by also varying the subjet CSV discriminant

Performance in Simulation



- Shower deconstruction curve is obtained by scanning χ
- Over the whole p_T range the Shower deconstruction tagger with an additional b-tag performs best
- For $p_T > 800 \text{ GeV}$ the MultiR Tagger and the shower deconstruction tagger show a huge improvement

Performance in Simulation



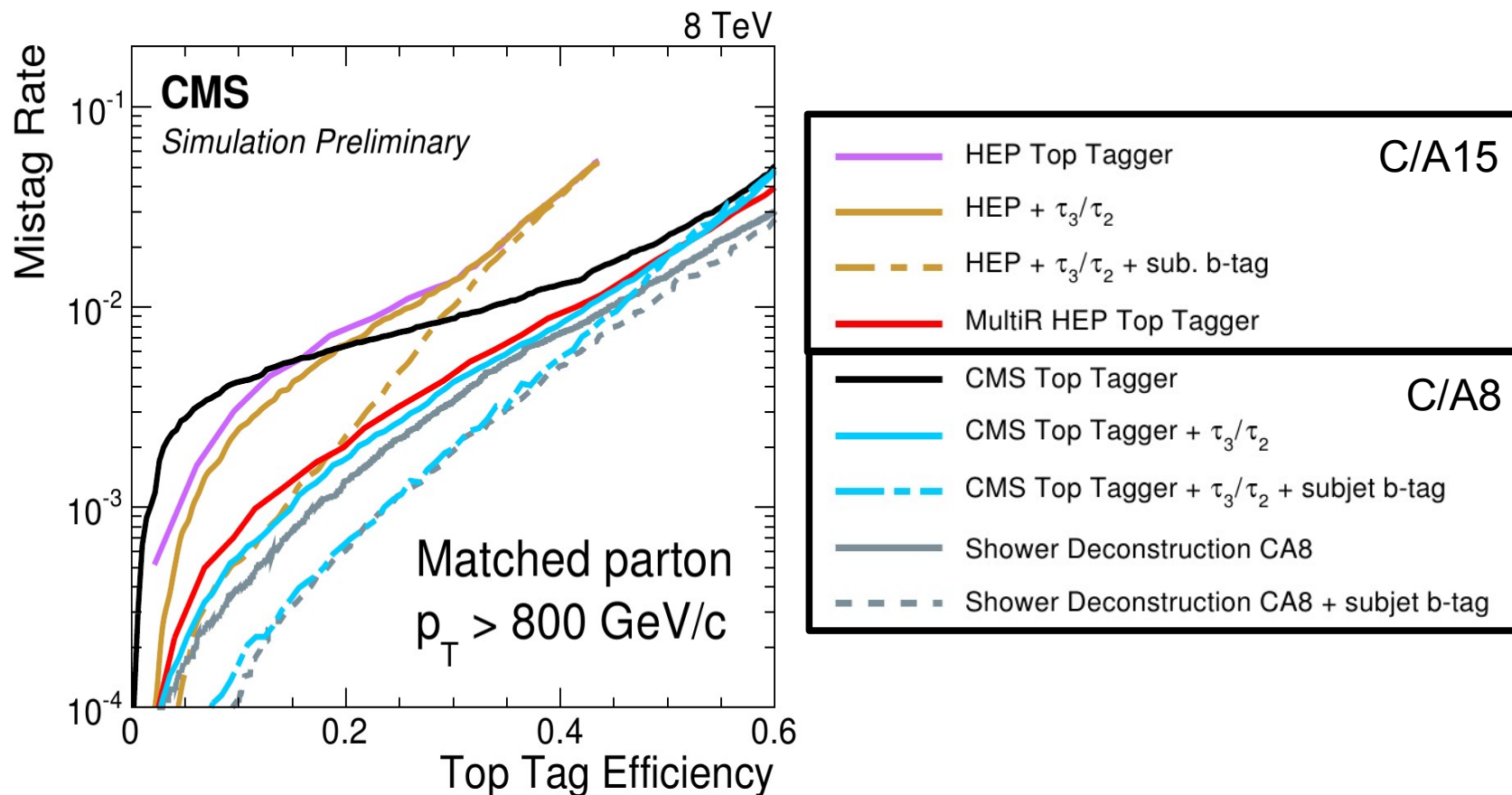
C/A8

- CMS Top Tagger
- - - subjet b-tag
- N-subjettiness ratio τ_3/τ_2
- - - CMS Top Tagger + subjet b-tag
- - - CMS Top Tagger + τ_3/τ_2 + subjet b-tag
- Shower deconstruction
- - - Shower deconstruction + subjet b-tag

- + CMS WP0
- CMS Comb. WP1
- CMS Comb. WP2
- ▲ CMS Comb. WP3
- ◆ CMS Comb. WP4

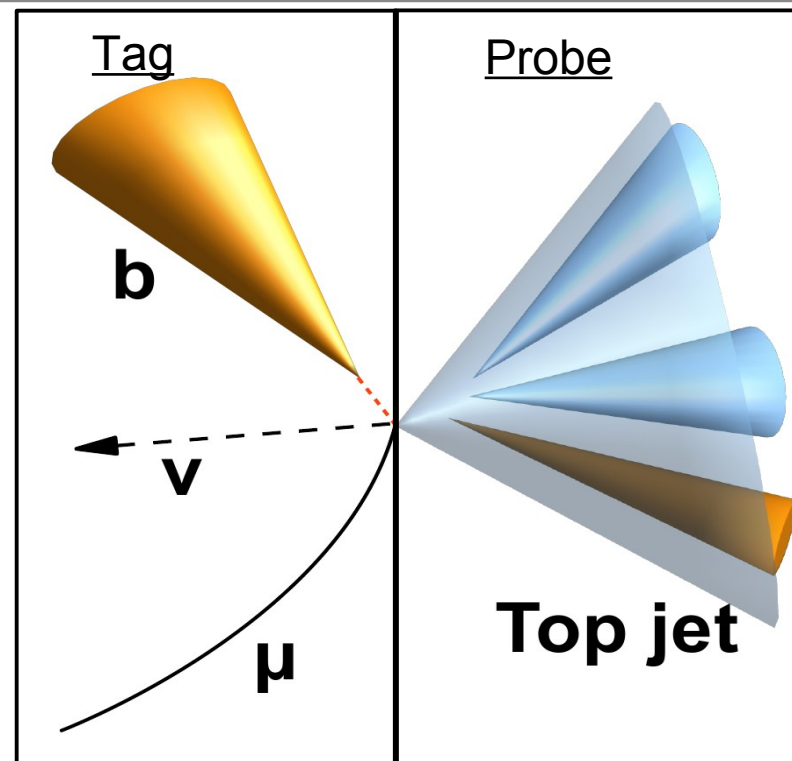
- CMS Top tagger curve determined by fixing $140 \text{ GeV} < m_{\text{Jet}} < 250 \text{ GeV}$ and $N_{\text{subjets}} > 2$, m_{min} is varied
- In whole p_T range the CMS Top tagger + N subjettiness + subjet b-tag is performing the best
- For $p_T > 600 \text{ GeV}$ also the Shower deconstruction tagger is working good

- For $p_T > 800$ GeV comparison between algorithms with different cone sizes possible



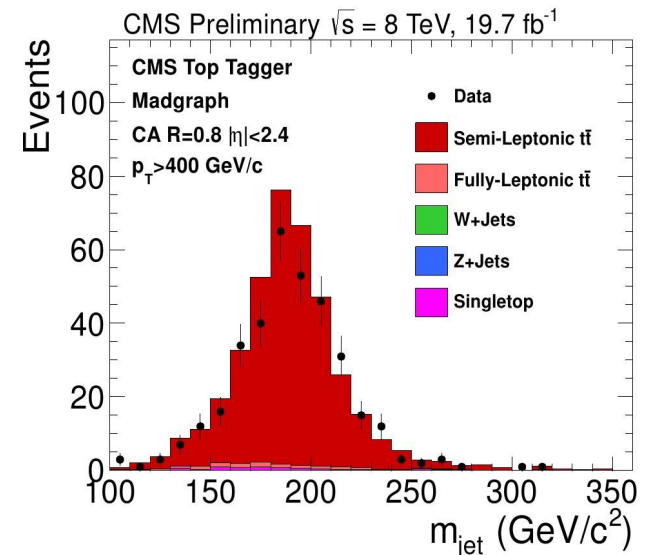
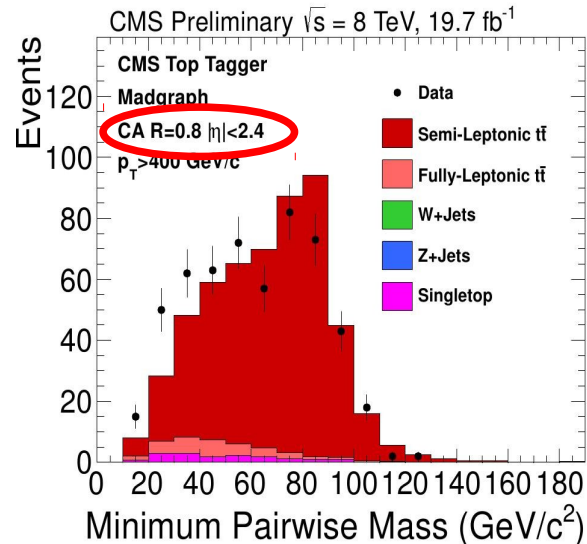
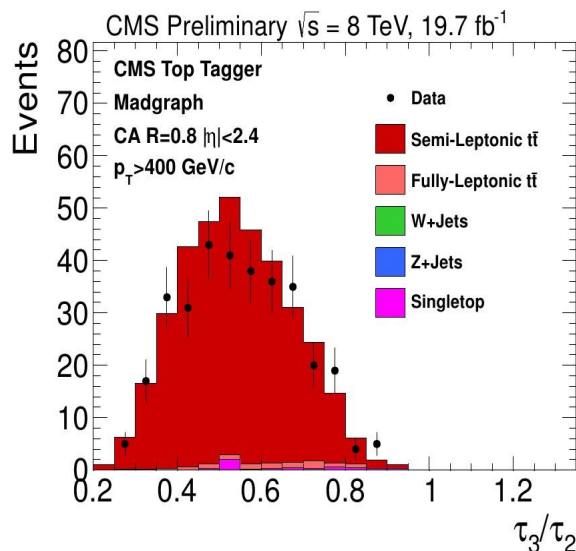
Muon + jets semileptonic ttbar selection

- Exactly one high p_T muon with $p_T > 45$ GeV
- Min one jet tagged with the CSV medium b-tagging algorithm
- B-tagged jet, $p_T > 30$ GeV and $\Delta R_{\text{muon,jet}} < \Pi/2$
- The jet with the highest p_T in the hemisphere $\Delta R_{\text{muon,CA jet}} > \Pi/2$ is a top candidate
- Top candidate for CMS Top Tagger is C/A jet with $R=0.8$, $p_T > 400$ GeV and
- Top candidate for HEP Top Tagger is C/A jet with $R=1.5$, $p_T > 200$ GeV and



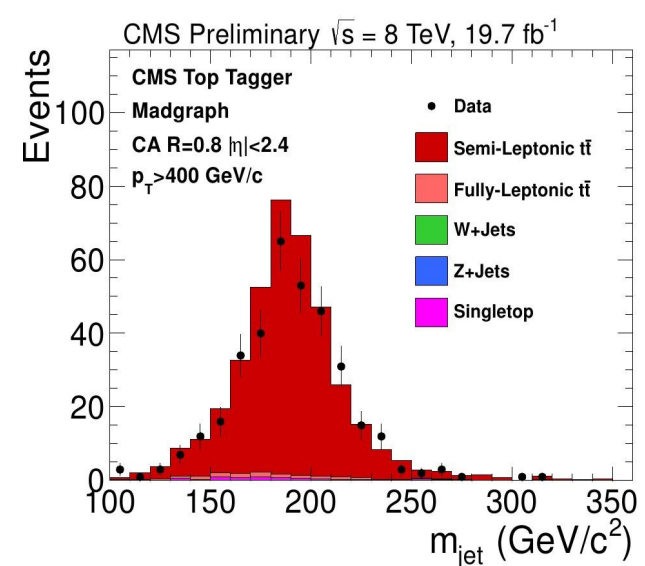
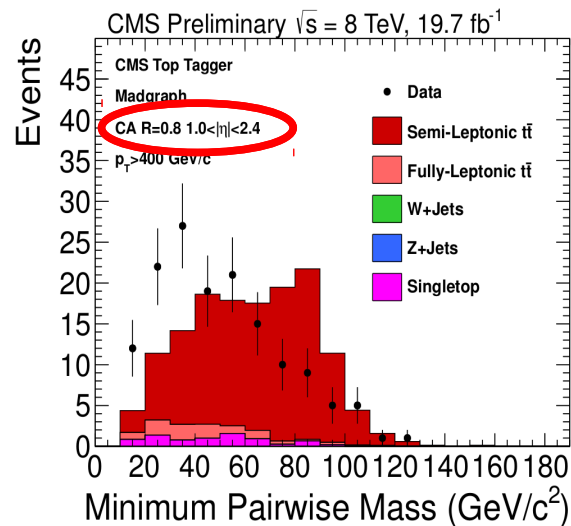
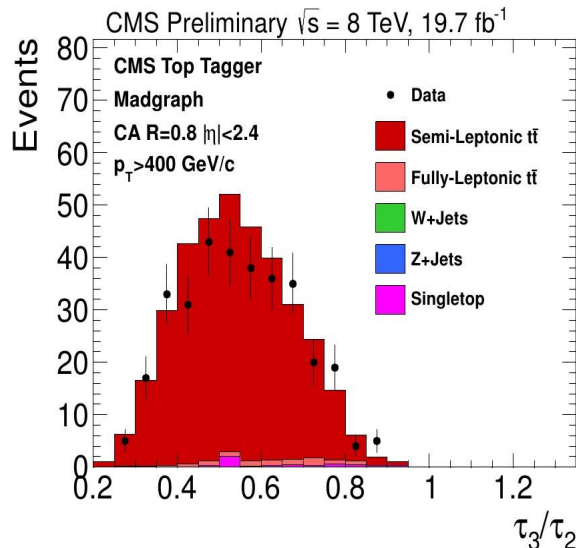
CMS Top Tagger observables:

- M_{\min} not well modelled by simulation
- Effect maybe because of mis-modeling of radiation or merged subjects
- M_{\min} better described in the for the central region $|\eta| < 1.0$
 → pseudorapidity-dependent scale factor
- Other variables well described



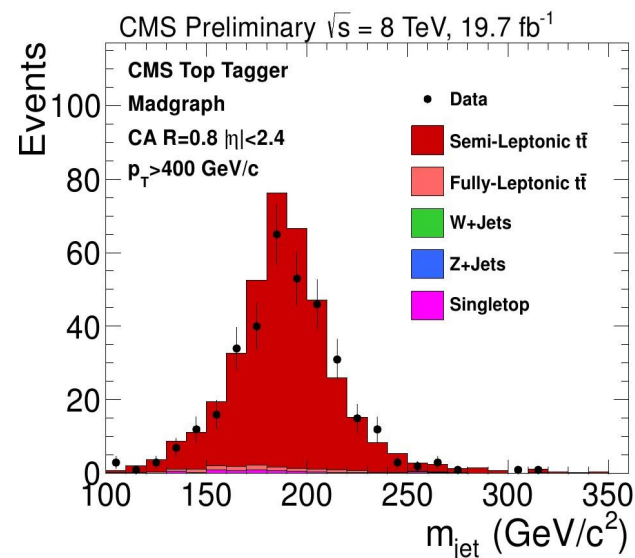
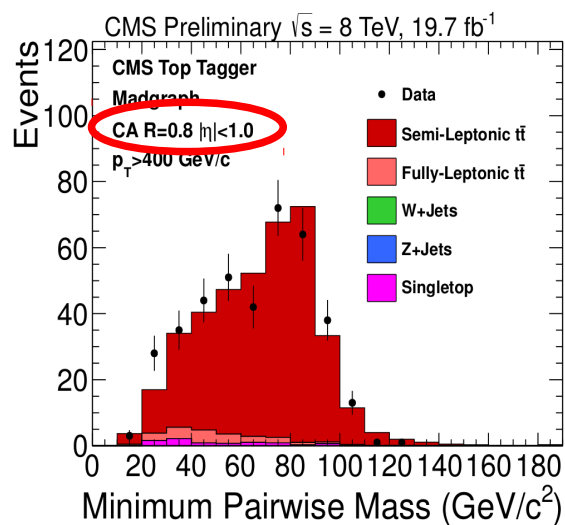
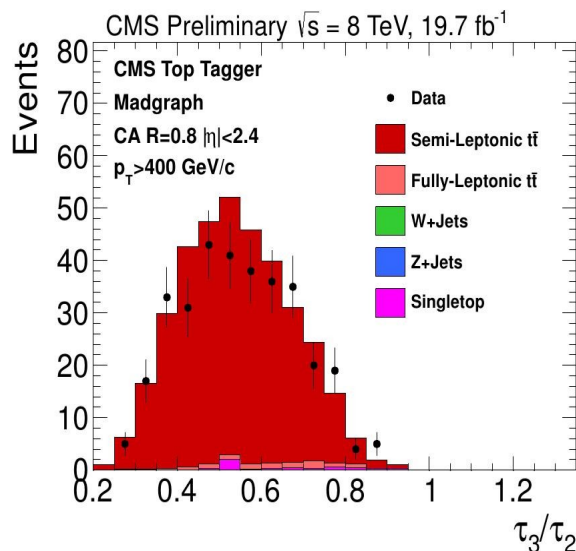
CMS Top Tagger observables:

- M_{\min} not well modelled by simulation
- Effect maybe because of mis-modeling of radiation or merged subjects
- M_{\min} better described in the for the central region $|\eta| < 1.0$
 → pseudorapidity-dependent scale factor
- Other variables well described



CMS Top Tagger observables:

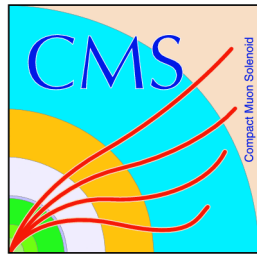
- M_{\min} not well modelled by simulation
- Effect maybe because of mis-modeling of radiation or merged subjects
- M_{\min} better described in the for the central region $|\eta| < 1.0$
 → pseudorapidity-dependent scale factor
- Other variables well described





Universität Hamburg

DER FORSCHUNG | DER LEHRE | DER BILDUNG



V Tagging

CMS PAS JME-14-002

Jet grooming techniques	Parameters
Filtering [1]	3 hardest CA subjects with $R=0.2$
Trimming [2]	$R_{\text{sub}}=0.05$, p_T fraction of mother jet $> 3\%$
Pruning [3]	momentum fraction 0.1, maximal distance 0.5
Soft-Drop [4]	soft threshold fixed to 0.1, $\beta=\{-1,0,2\}$

Variable	Parameter
Gluon/Quark Likelihood [5]	
Subjet Gluon/Quark Likelihood [5]	
Energy Correlation Functions [6]	$\beta = \{0, 0.2, 0.5, 1, 2\}$
N-subjettiness	τ_2/τ_1
Qjet volatility [7]	NTrees=50

[1] J. M. Butterworth, A. R. Davison, M. Rubin, and G. P. Salam: Phys.Rev.Lett. 100 (2008) 242001

[2] D. Krohn, J. Thaler, and L.-T. Wang: JHEP 1002 (2010) 084

[3] S. D. Ellis, C. K. Vermilion, and J. R. Walsh: Phys. Rev. D 80 (2009) 051501

[4] A. Larkoski, S. Marzani, G. Soyez, and J. Thaler: JHEP05(2014)146

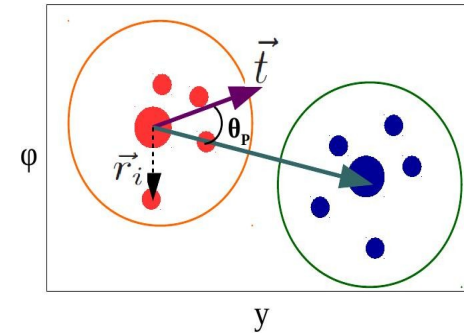
[5] CMS Collaboration: CMS-PAS-JME-13-002

[6] A. Larkoski, G. Salam, and J. Thaler: JHEP06(2013)108

[7] S. D. Ellis et al.: PhysRevLett.108.182003

Jet Pull Angle [J. Gallicchio, M. Schwartz: arXiv:1001.5027v3]:

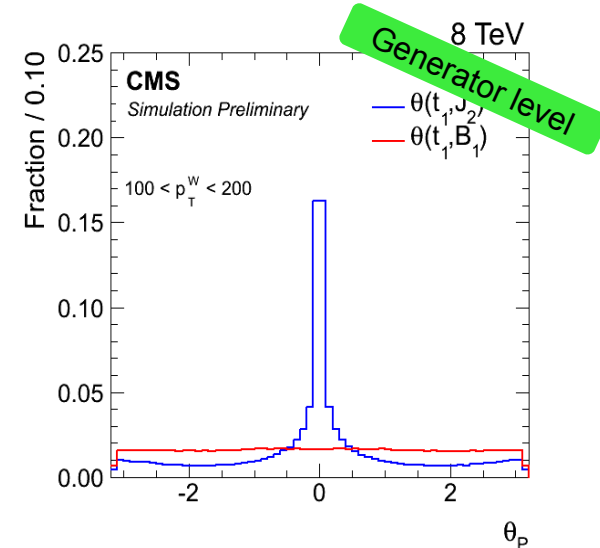
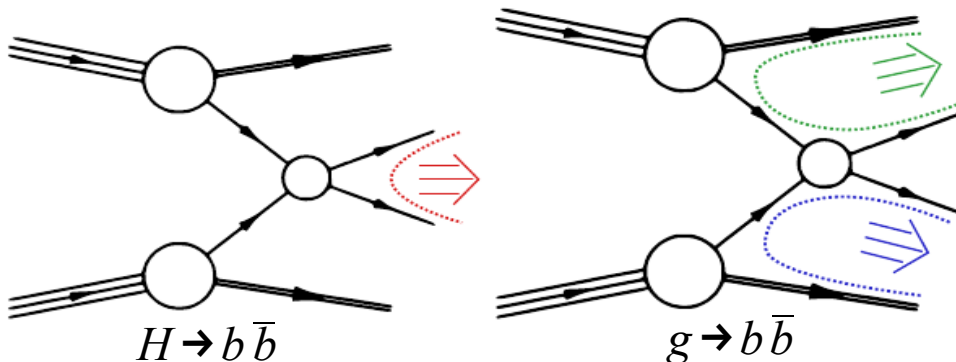
- Compute weighted vector sum of constituent positions relative to the jet axis in y - ϕ space
- The angle between the pull vector and the relative displacement of another jet is the pull angle, θ_p
- θ_p should peak around zero for color connected jet pairs
- pairs are uniformly distributed for unconnected jet pairs



$$\vec{t} = \sum_{i \in \text{jet}} \frac{p_T^i |r_i|}{p_T^{\text{jet}}} \vec{r}_i$$

$$\vec{r}_i = (\Delta y_i, \Delta \phi_i)$$

Jet Pull Magnitude: the magnitude of the jet pull vector for pruned subjets

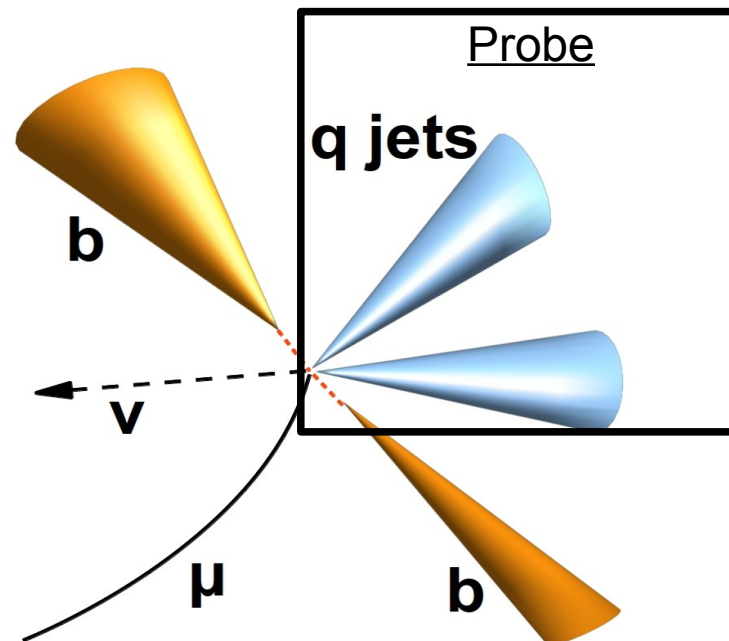


Resolved W/Z selection

Resolved scenario:
electroweak boson $p_T < 160$ GeV

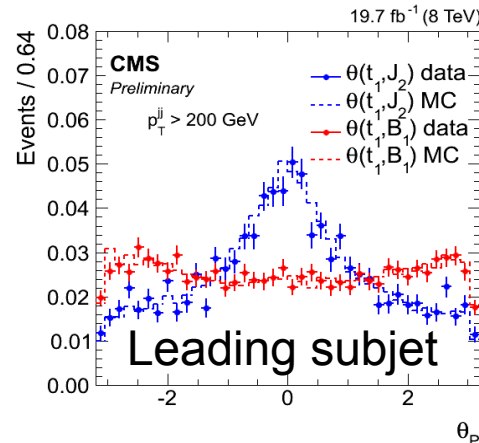
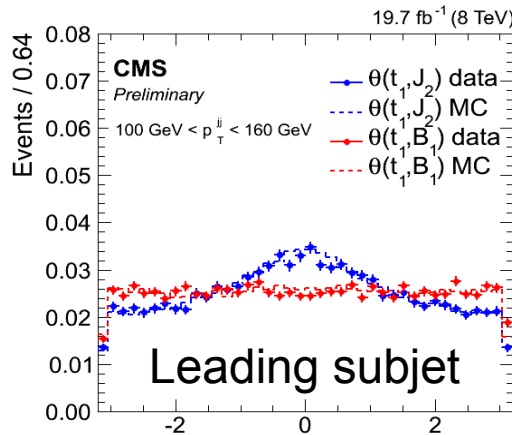
Ttbar selection (signal):

- Min four Anti- k_T jets ($R=0.5$) with $p_T > 30$ GeV,
 $|\eta| < 4.7$
- Exactly one muon $p_T > 30$ GeV
- Min two b-tagged jets
- W candidates:
 - Pairs of dijets (not b-tagged) with a dijet mass between 40 GeV and 130 GeV

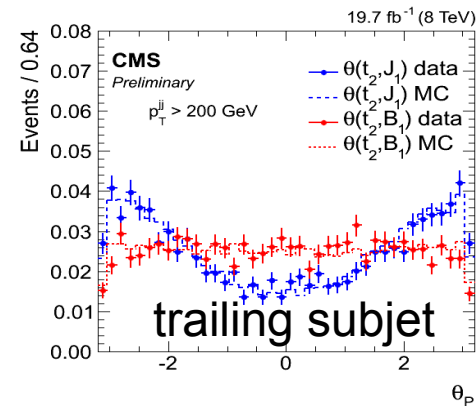
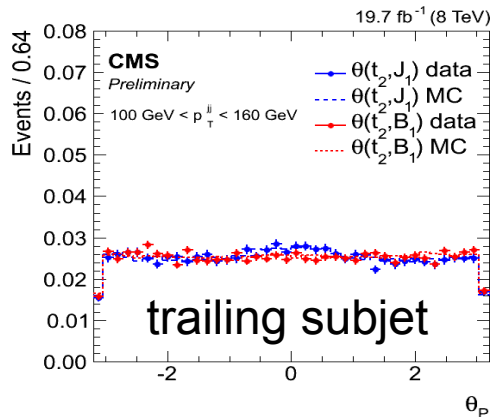


Jet pull angle

- Good data/MC agreement for pull angle computed using leading and subleading jet of the W candidates
- Weak separation power at low dijet p_T

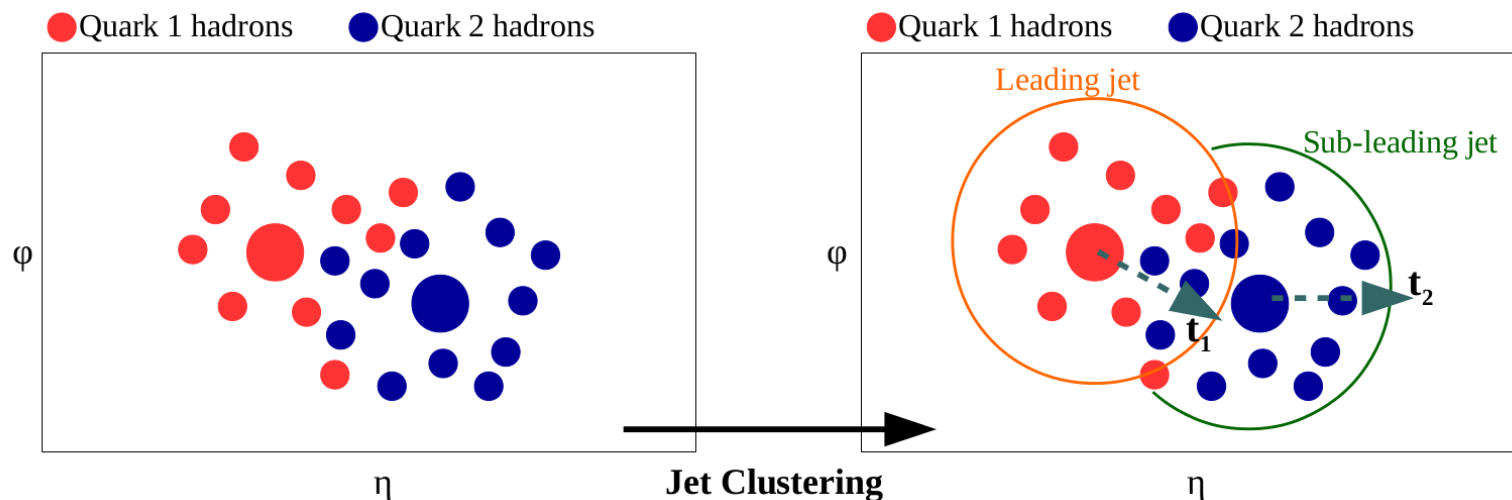


- At high dijet p_T , the pull angle shows opposing behavior between leading and sub-leading jets
→ Consequence of jets overlapping...



Pull Angle: Overlapping Jets

- The asymmetric behavior with reco jets is the effect of partially merged jets from the W and jet clustering of the reconstruction

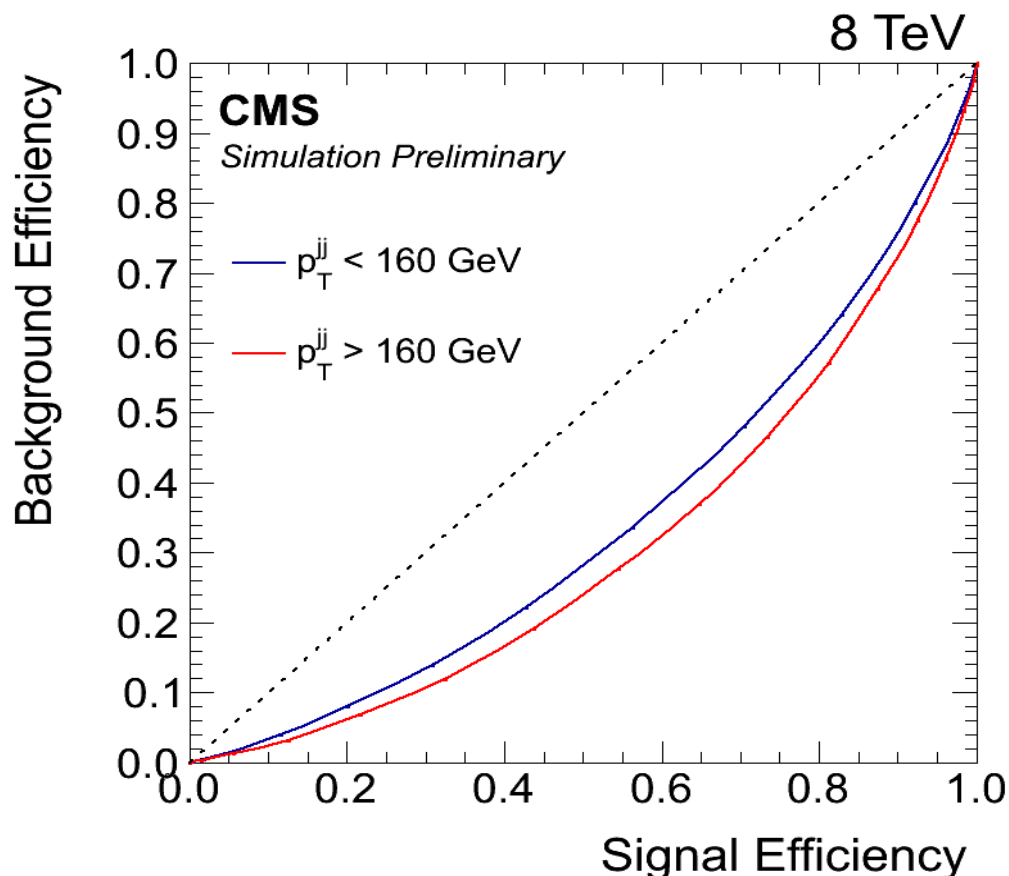


- Leading jet “gobbles up” some hadrons from the other quark
- Consequences:
 - Sub-leading jet pull points away from leading jet, having lost constituents that are “closer” to the leading jet
 - θ_p peak can be enhanced as leading jet absorbs hadrons of other jet

Resolved Jets performance



- QGL, jet pull angle, dijet charge sum used for BDT
- Variables each provide some separation power
- Variables are weakly correlated
- Training was done for two different dijet p_T bins



Unresolved W/Z selection



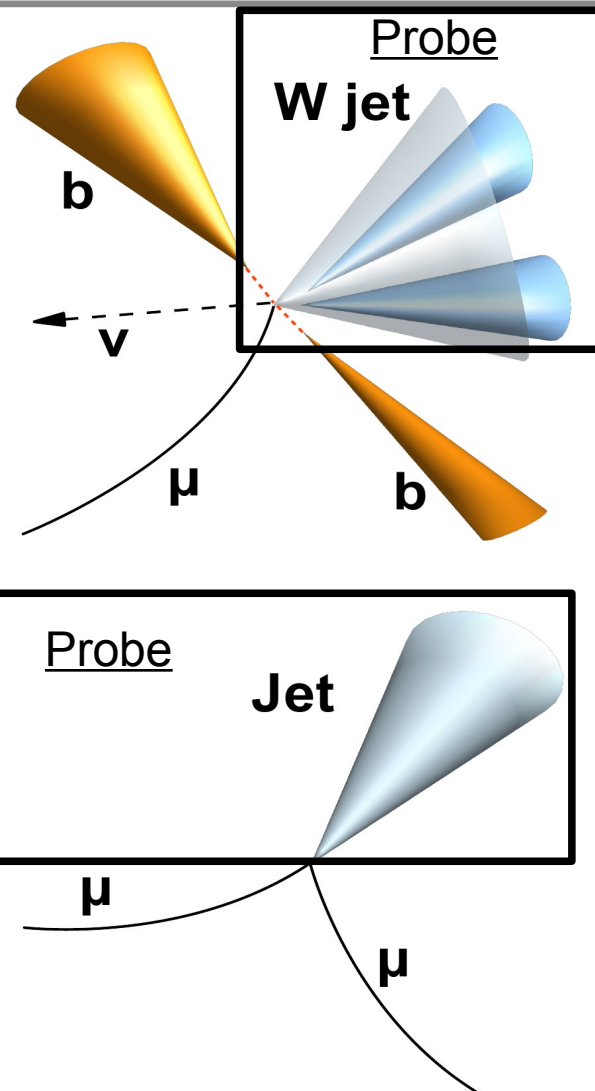
Unresolved scenario:
electroweak boson $p_T > 250 \text{ GeV}$

Ttbar selection (**signal**):

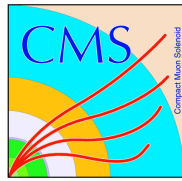
- Anti- k_T jets ($R=0.8$) with $p_T > 250 \text{ GeV}$,
 $|\eta| < 2.5$, $\Delta R(\text{jet}, \text{lepton}) > 0.3$
- Exactly one muon $p_T > 30 \text{ GeV}$
- Min two b-tagged jets (no match with boosted jet)

Z+Jets selection (**background**):

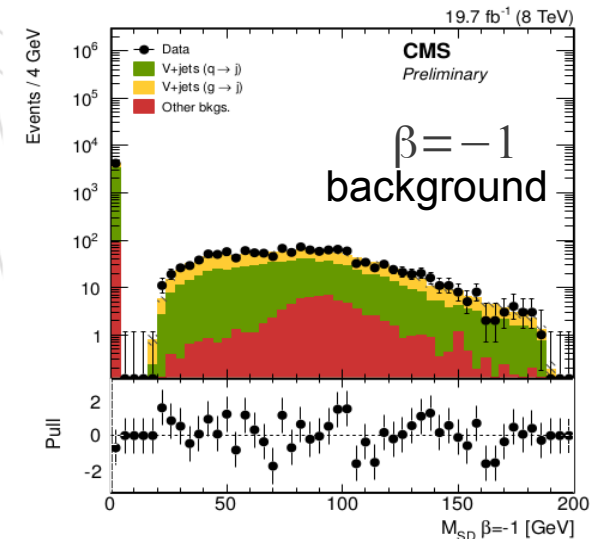
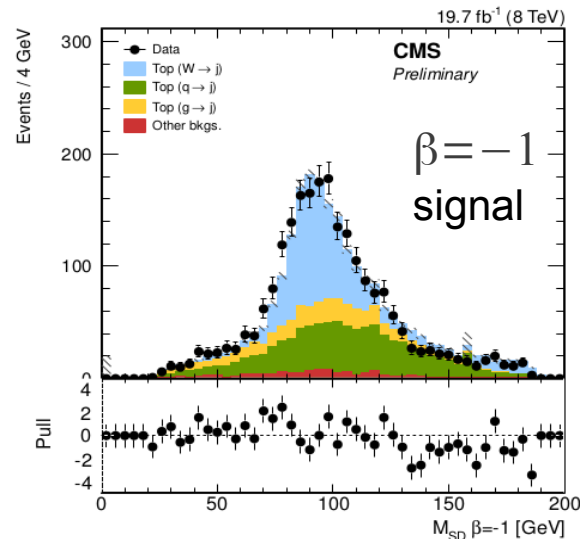
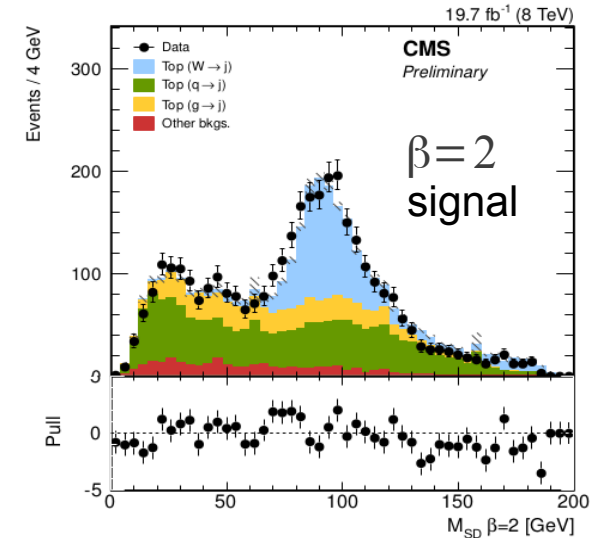
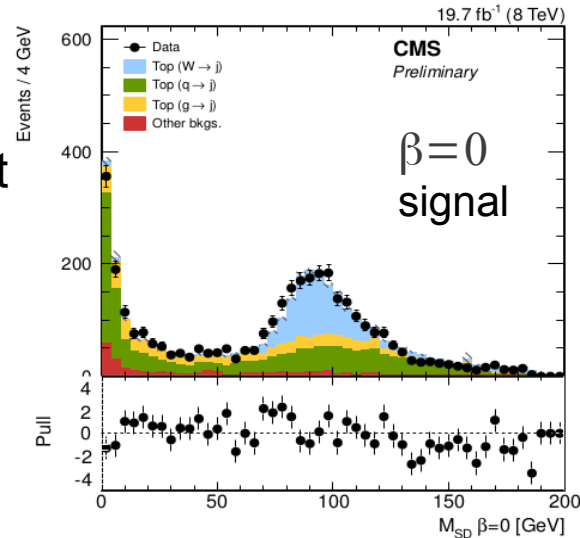
- Anti- k_T jets ($R=0.8$) with $p_T > 250 \text{ GeV}$,
 $|\eta| < 2.5$, $\Delta R(\text{jet}, \text{lepton}) > 0.3$
 - Two opposite sign muons with $p_T > 30 \text{ GeV}$
 - Dimoun mass within 15 GeV of the nominal Z boson mass
 - Dilepton $p_T > 100 \text{ GeV}$
- relatively pure sample of quark jets



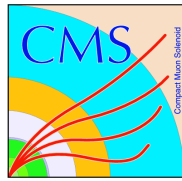
Data/MC comparison: soft drop



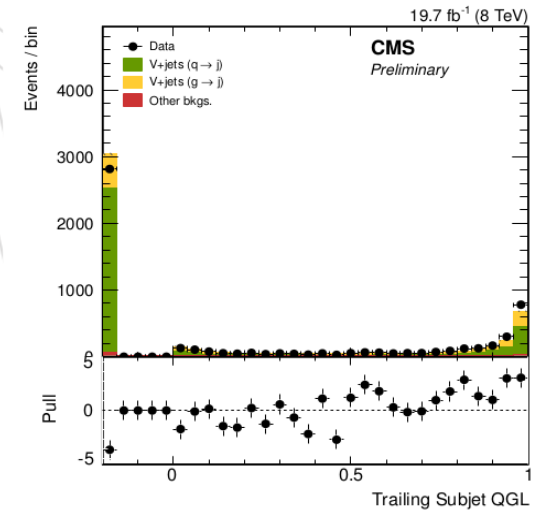
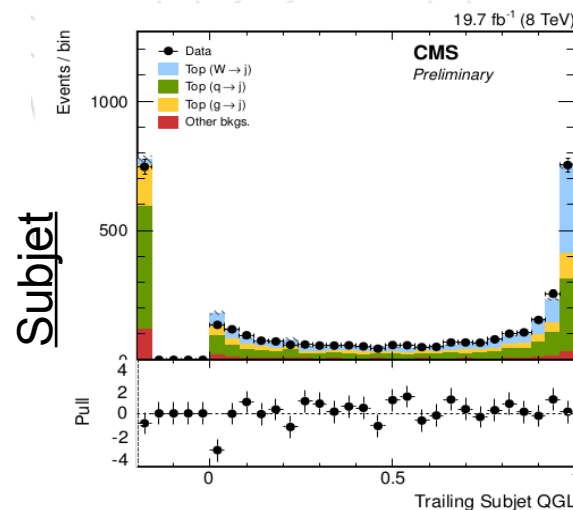
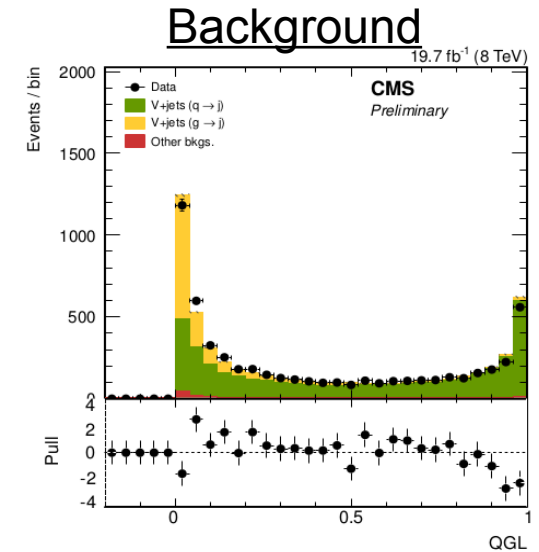
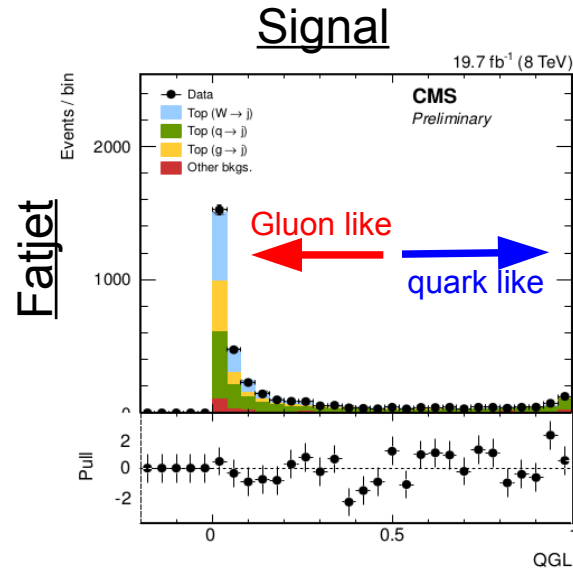
- $z_{\text{cut}}=0.1, R_0=0.8$
- Good data/MC agreement



Data/MC comparison: QGL

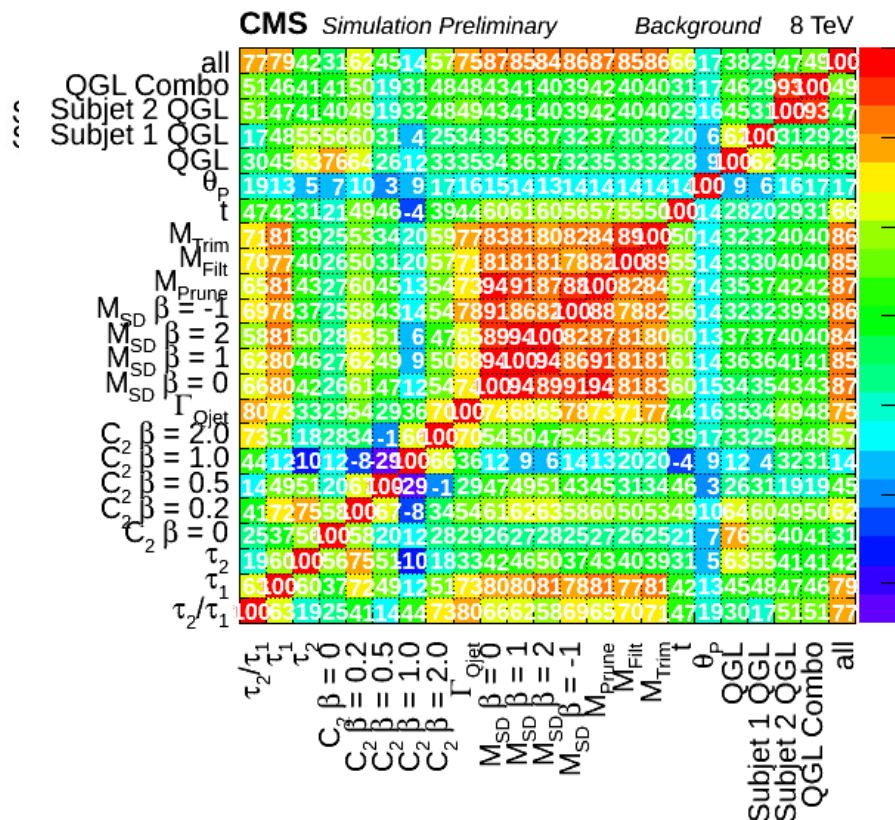


- “Fat” jet appears very gluon-like to QGL
- Subjet QGL recovers expected behavior
- Trailing subjet QGL shows more discriminating power than leading subjet QGL
- QGL combo: defined as a linear combination of the leading subjet QGL with twice the second leading subjet QGL

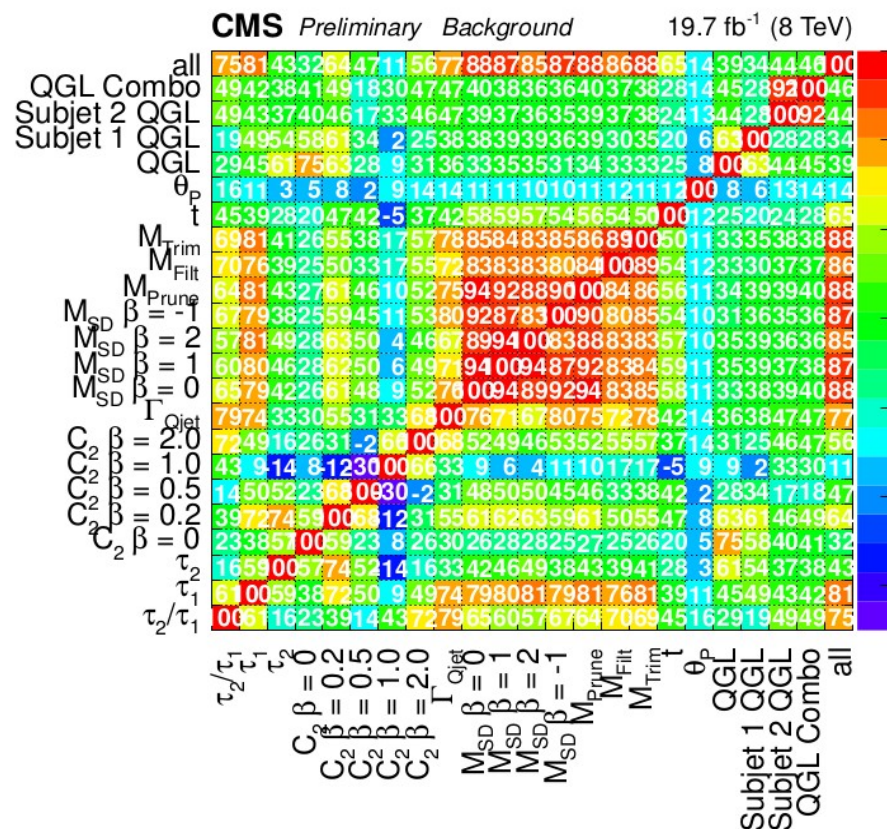


Correlation matrices for BDT single variables for background MC and data (Z+jets selection)

Z+Jets selection simulation



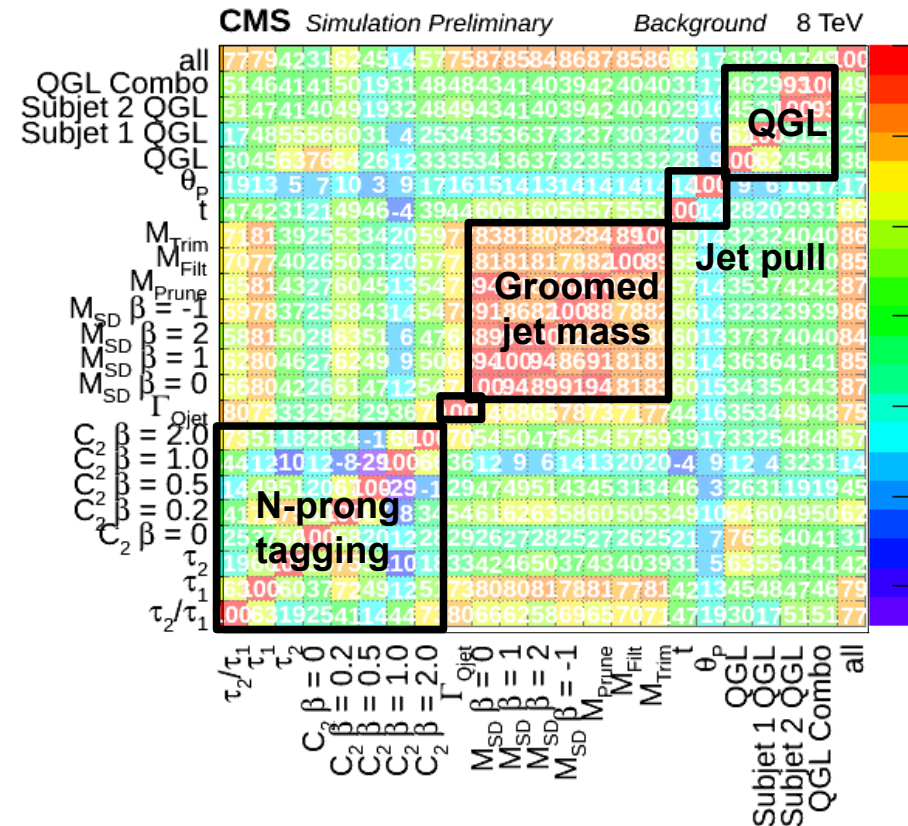
Z+Jets selection data



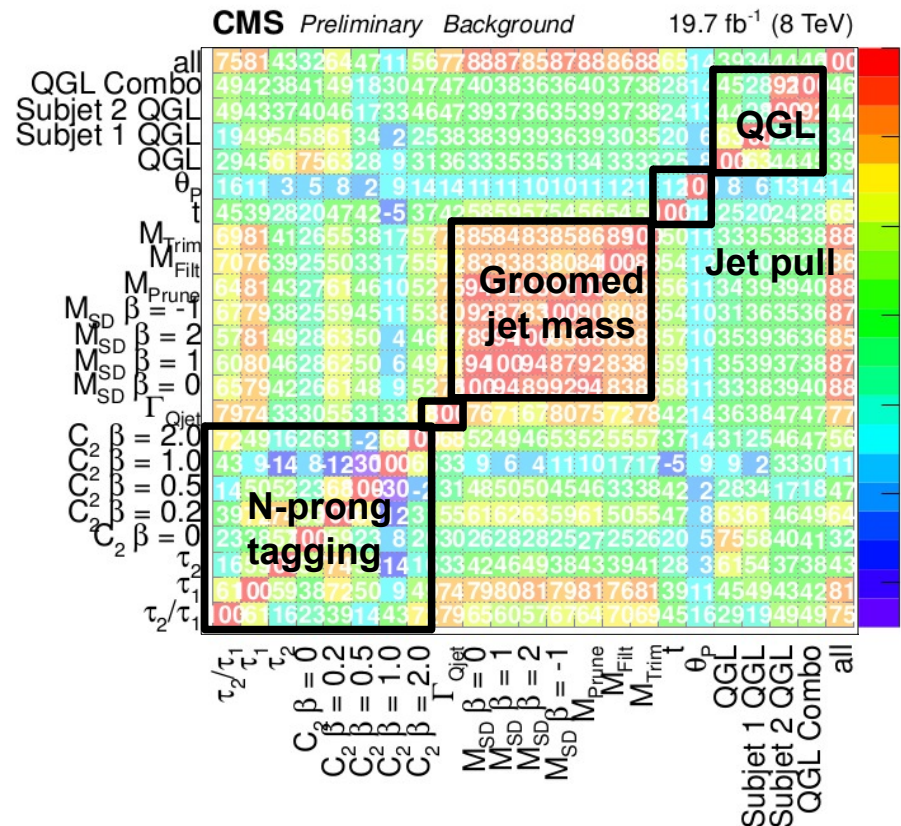
Correlations

- Easier to view correlation matrix in “blocks”
 - Typically stronger correlations within blocks
- Correlations between data and MC look similar

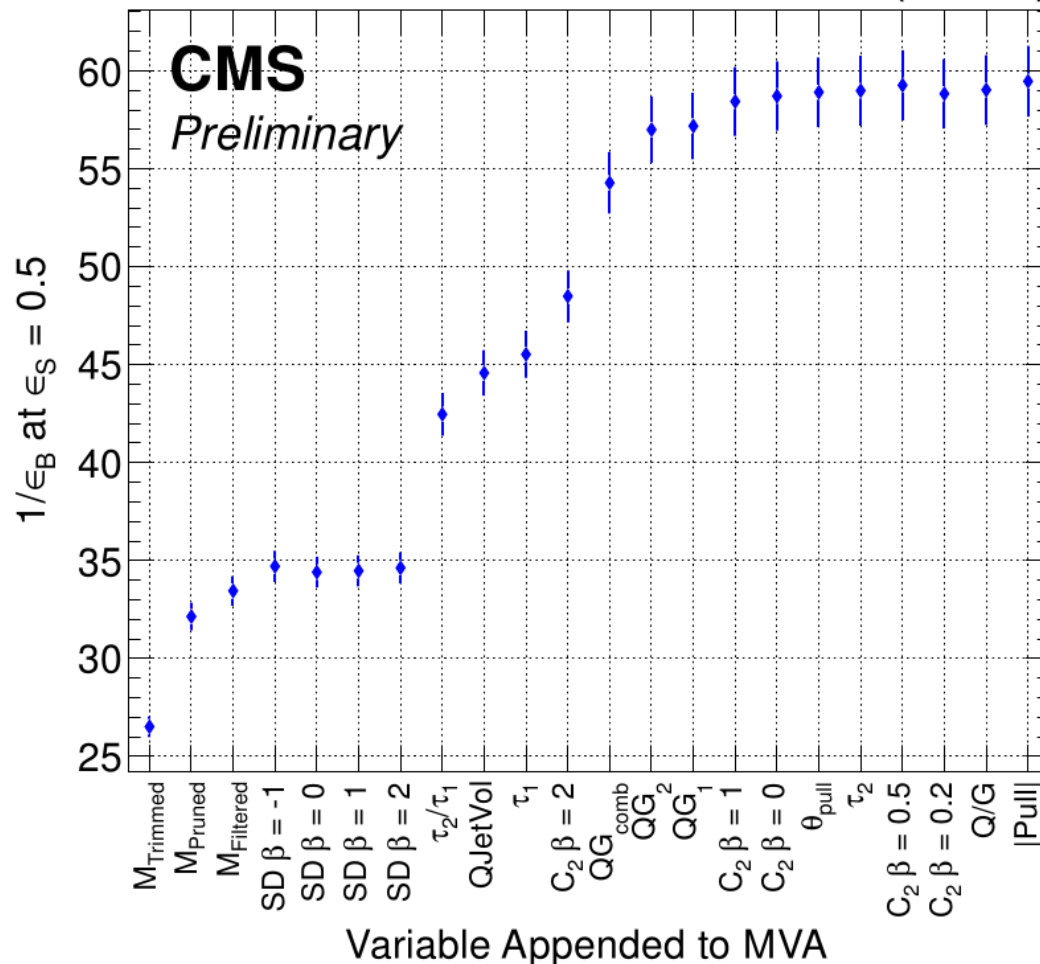
Z+Jets selection simulation



Z+Jets selection data



19.7 fb⁻¹ (8 TeV)



- Multi-dimensional analysis based on Boosted Decision Trees (BDT), using the TMVA framework
- Working point is set to 50% signal efficiency
- Iteratively added one variable on top of the next variable
- Saturation after use of 11 variables

Summary

Top Tagging

- Different top tagger were compared in Simulation
- N subjettiness has a good separation power and can improve existing top tagger
- The MultiR HEP Top tagger is a powerful improvement and makes the HEP Top Tagger usable in higher p_T regions
- Shower deconstruction is a completely other approach for top tagging and has a great performance
- Validation in data is on going

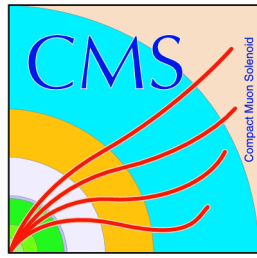
V tagging

- New variables like QGL, pull angle, and pull magnitude are used
- Variables are good described (ttbar selection, Z+Jets selection)
- Variables have a high discriminating power
- New variables have low correlations to any other variables



Universität Hamburg

DER FORSCHUNG | DER LEHRE | DER BILDUNG



Thank you for your attention!



Universität Hamburg

DER FORSCHUNG | DER LEHRE | DER BILDUNG



Top Tagging

CMS PAS JME-13-007
BACKUP

Samples

The following samples were used:

TTbar:

MADGRAPH + PYTHIA 6

POWHEG v1 + PYTHIA 6

MC@NLO + HERWIG

QCD:

PYTHIA 6

MADGRAPH + PYTHIA 6

DiBoson:

PYTHIA 6

CMS detector simulation:

GEANT 4

Scale factors

Cumulative data-simulation scale factor - CMS Tagger, CMS Combined Tagger

$ \eta < 1.0$			
Selection	MADGRAPH	POWHEG	MC@NLO
CMS Tagger WP0	0.985 ± 0.073	1.173 ± 0.092	1.033 ± 0.081
CMS Combined Tagger WP3	0.891 ± 0.118	1.063 ± 0.146	0.933 ± 0.129

$1.0 < \eta < 2.4$			
Selection	MADGRAPH	POWHEG	MC@NLO
CMS Tagger WP0	0.644 ± 0.100	0.704 ± 0.110	0.768 ± 0.118
CMS Combined Tagger WP3	0.685 ± 0.199	0.906 ± 0.277	0.802 ± 0.230

→ scale factors are worse for the high η region

Working points



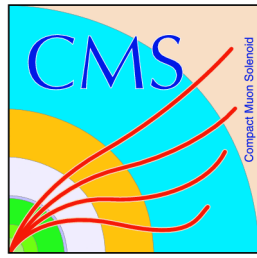
Working point	m_{123} selection	f_W selection	subject b-tag WP	τ_3/τ_2 selection
HEP WP0	140-250 (GeV/ c^2)	0.495	none	none
HEP Combined WP1	140-250 (GeV/ c^2)	0.495	CSV-loose	none
HEP Combined WP2	140-250 (GeV/ c^2)	0.15	CSV-medium	none
HEP Combined WP3	140-250 (GeV/ c^2)	0.15	CSV-medium	< 0.63

Working point	m_{jet} selection	m_{min} selection	subject b-tag WP	τ_3/τ_2 selection
CMS Tagger WP0	140-250 (GeV/ c^2)	> 50 (GeV/ c^2)	none	none
CMS Combined WP1	140-250 (GeV/ c^2)	> 50 (GeV/ c^2)	CSV-loose	< 0.7
CMS Combined WP2	140-250 (GeV/ c^2)	> 50 (GeV/ c^2)	CSV-loose	< 0.6
CMS Combined WP3	140-250 (GeV/ c^2)	> 50 (GeV/ c^2)	CSV-medium	< 0.55
CMS Combined WP4	140-250 (GeV/ c^2)	> 65 (GeV/ c^2)	CSV-medium	< 0.4



Universität Hamburg

DER FORSCHUNG | DER LEHRE | DER BILDUNG



V Tagging

CMS PAS JME-14-002

Jet grooming techniques:

Filtering: three hardest CA subjets with $R=0.2$

Trimming: trimming reclusters the jets' constituents with a radius R_{Sub} and then accepts only the subjets that have $p_{T,\text{sub}} > f_{\text{cut}}$, subjets obtained with k_T clustering, $R_{\text{sub}}=0.05$, p_T fraction of mother jet $> 3\%$

Pruning: technique to remove softest components of the jet, minimal momentum fraction 0.1, maximal distance 0.5

Soft-Drop: declustering fatjet, soft threshold fixed to 0.1, $\beta=\{-1,0,2\}$

$$\frac{\min(p_{T_1}, p_{T_2})}{p_{T_1} + p_{T_2}} > z_{\text{cut}} \left(\frac{\Delta R_{12}}{R_0} \right)^\beta$$

Gluon/Quark Likelihood: capable of distinguishing between jets created by gluons/Quarks

Subjet Gluon/Quark Likelihood: applied on the two leading pruned subjets

Energy Correlation Functions: 3 point correlation function is defined as:

$$C_2^\beta = \frac{\sum_{i,j,k} p_{Ti} p_{Tj} p_{Tk} (R_{ij} R_{ik} R_{jk})^\beta \sum_i p_{Ti}}{(\sum_{i,j} p_{Ti} p_{Tj} (R_{ij})^\beta)^2} \quad \beta = \{0, 0.2, 0.5, 1, 2\}$$

N-subjettiness: τ_2/τ_1

Qjet volatility: Defined as the RMS of the mass distribution of jet trees over the average jet mass, volatility = RMS/ m . Where Ntrees is chosen to be 50.

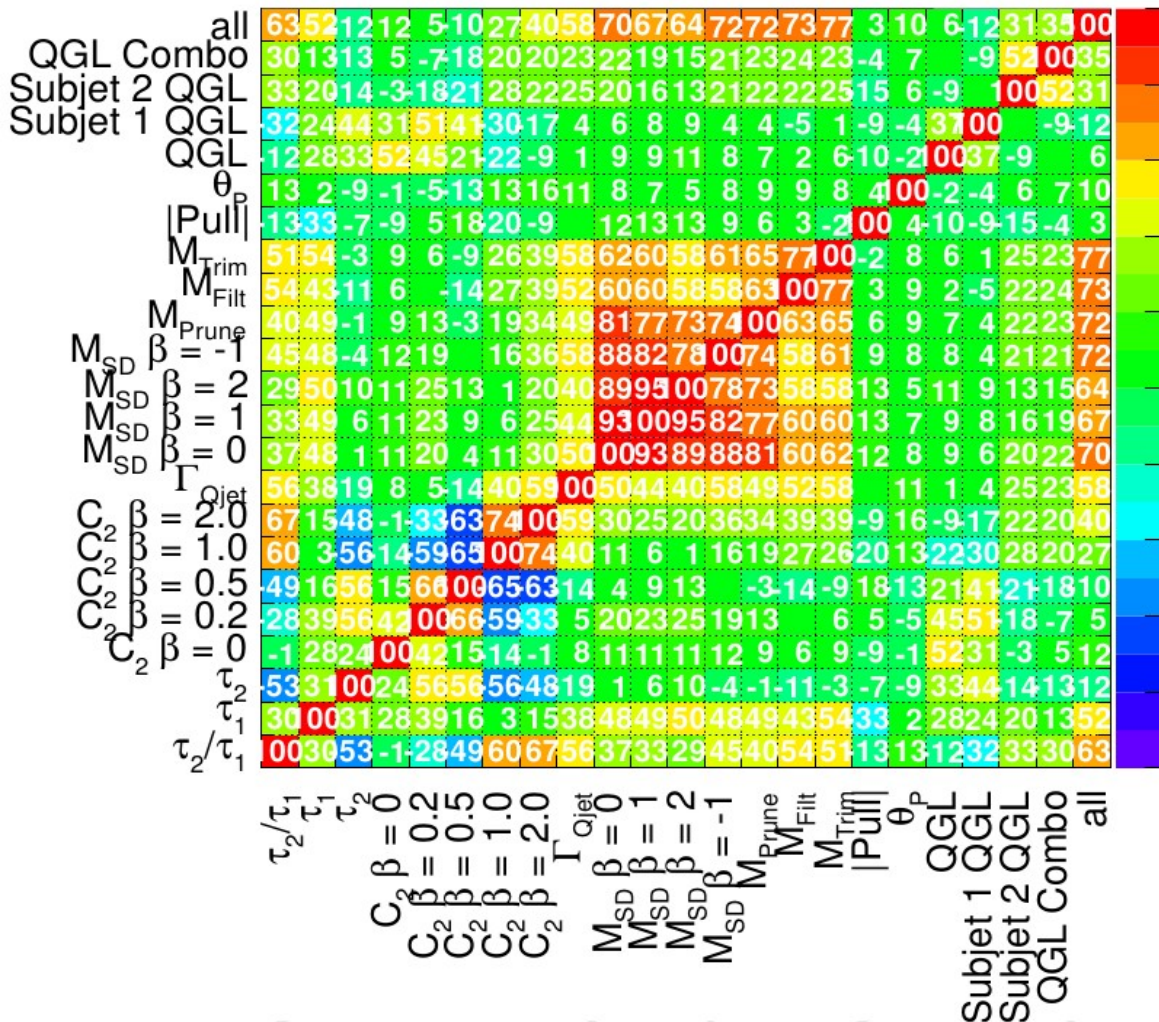
Jet Charge: Jet charge algorithm for boosted W-tagging

$$Q^\kappa = \frac{\sum_i (q_i (p_T^i)^\kappa)}{(p_T^{jet})^\kappa}$$

Correlations

CMS (Preliminary) Signal

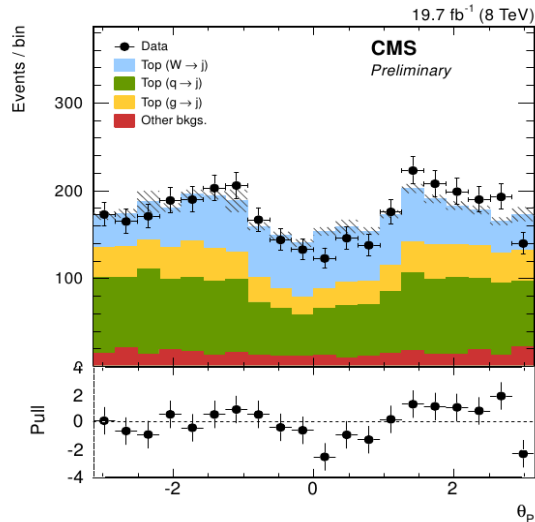
19.7 fb⁻¹ (8TeV)



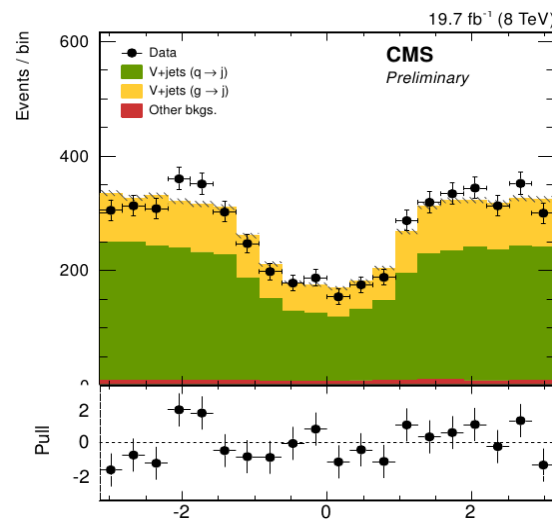
- Mass variables are strongly correlated, trimmed mass the least correlated
- sub-leading subjet QGL, pull angle, and pull magnitude are not correlated
- Correlation with column “all” indicates the most discriminating variables

Data/MC comparison

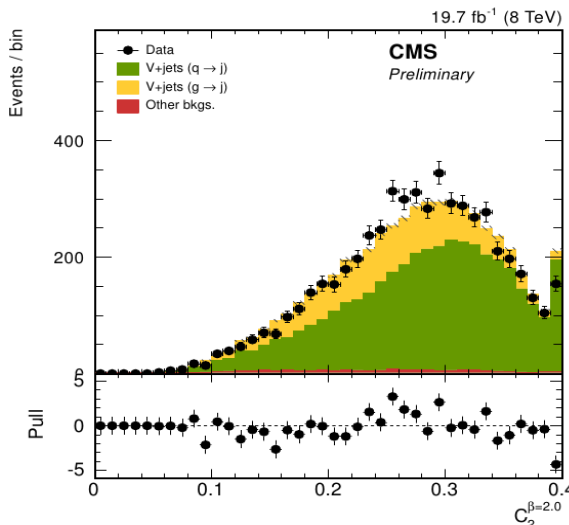
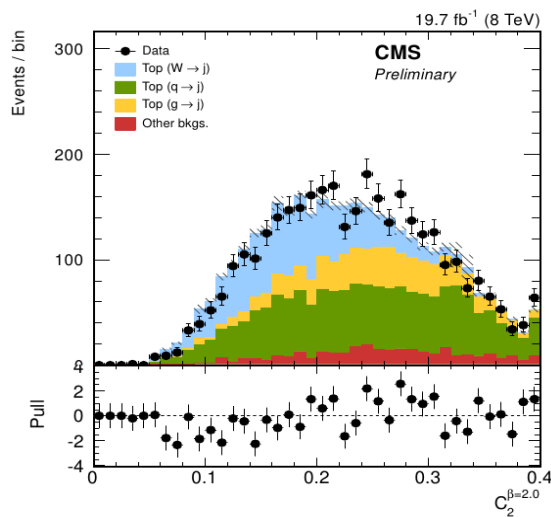
$t\bar{t}$ – Selection



Z+Jets – Selection

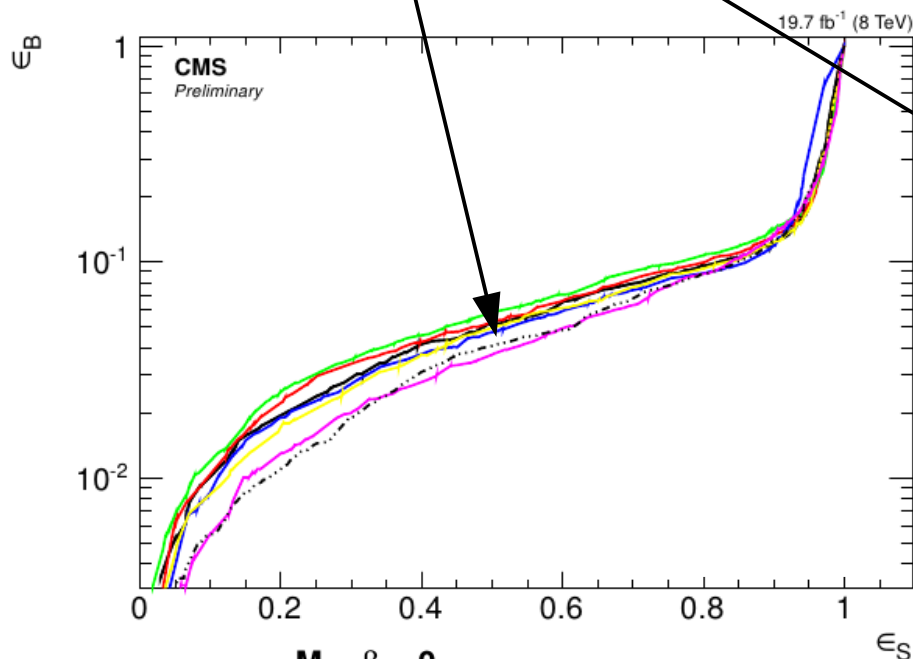


Jet source determined
 by calculating ΔR , to
 the closest generator
 level parton, $\Delta R < 0.7$

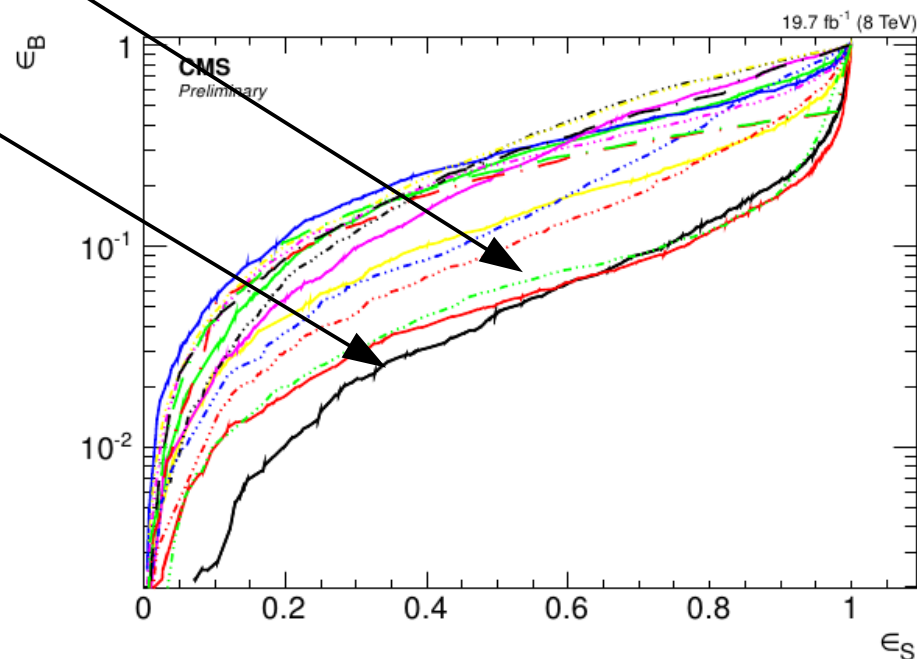


ROC curves for all single variables

Best variables: M_{trim} , N-subjettiness, Qjet volatility



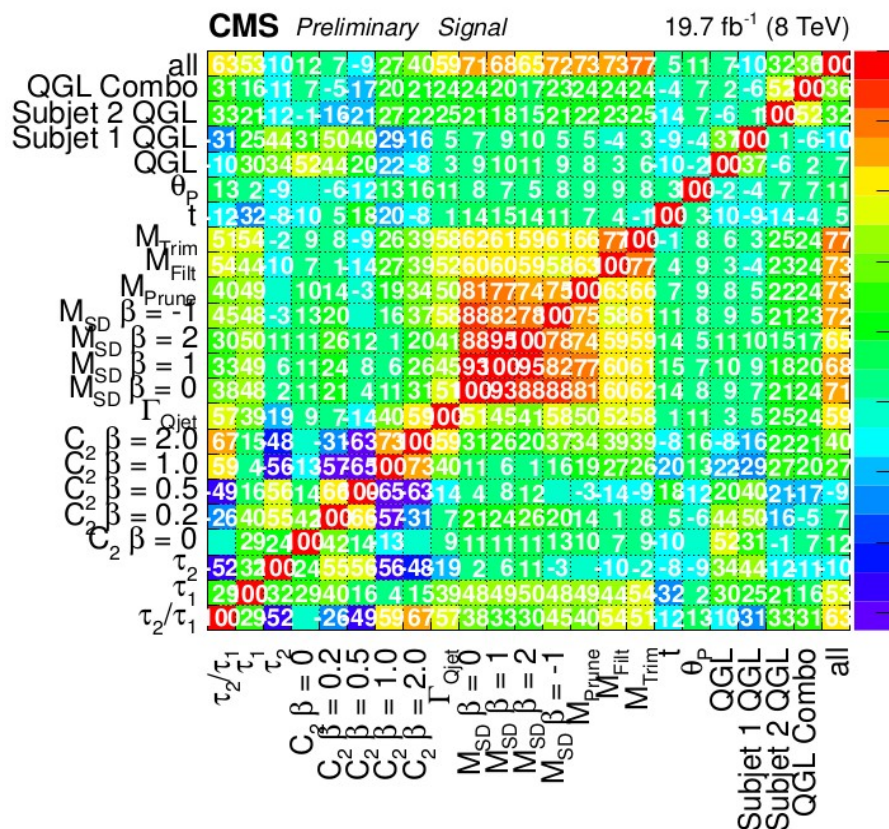
- $M_{SD} \beta = 0$
- $M_{SD} \beta = 1$
- $M_{SD} \beta = 2$
- $M_{SD} \beta = -1$
- M_{Prune}
- M_{Filt}
- - - M_{Trim}



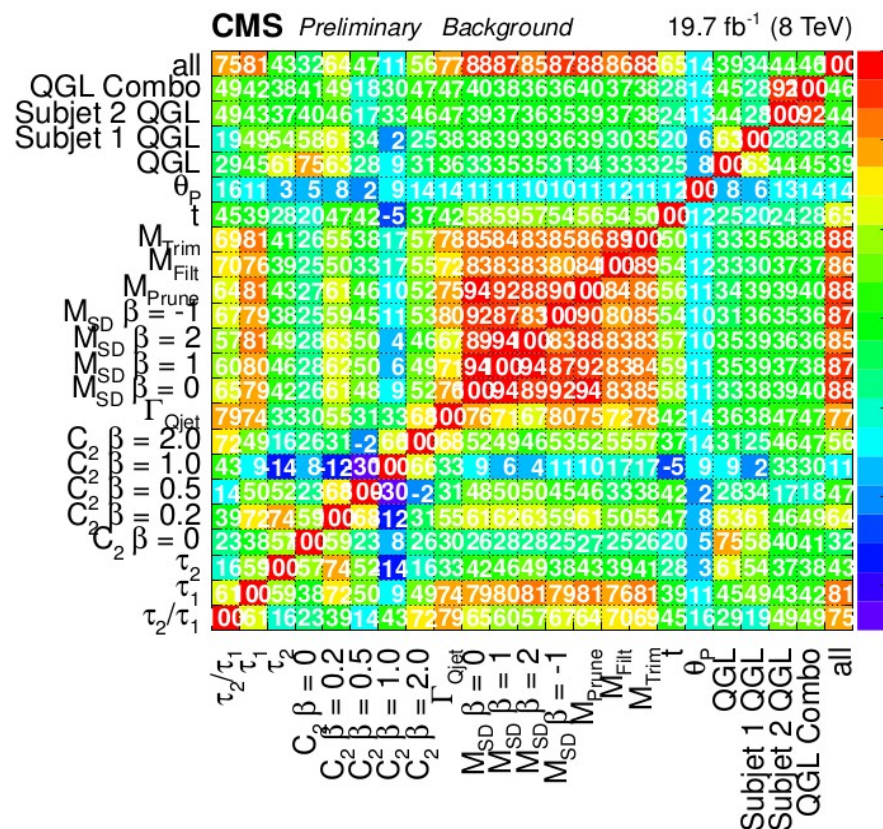
- | | |
|---------------------|-------------------|
| — τ_2/τ_1 | — Γ_{Qjet} |
| — τ_1 | — $ \text{Pull} $ |
| — τ_2 | — θ_p |
| — $C_2 \beta = 0$ | — QGL |
| — $C_2 \beta = 0.2$ | — Subjet 1 QGL |
| — $C_2 \beta = 0.5$ | — Subjet 2 QGL |
| — $C_2 \beta = 1.0$ | — QGL Combo |
| — $C_2 \beta = 2.0$ | |

Correlation matrices for signal (ttbar selection) and background (Z+jets selection)

ttbar selection simulation

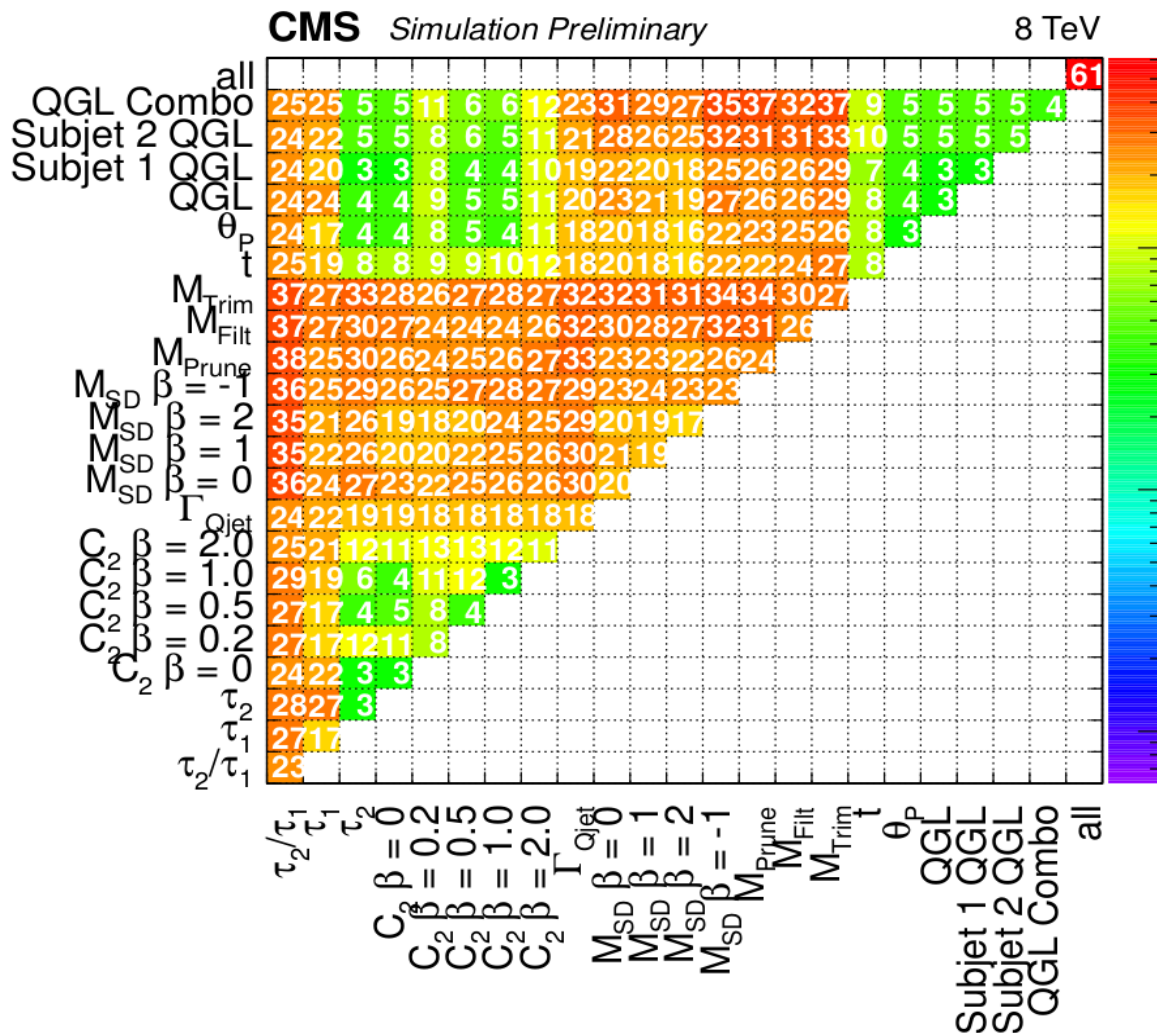


Z+Jets selection data



Z score for variable pairs

- BDT trained with pair of variables
- Shown is the Z score, which is defined as $1/\epsilon_{mis}$
- Efficiency working point is set to 50%
- Signal: MC
- Background: MC



- CMS** Preliminary 19.7 fb⁻¹ (8 TeV)



- BDT trained with triplets of variables
- Shown is the Z score, which is defined as $1/\epsilon_{mis}$
- Efficiency working point is set to 50%

