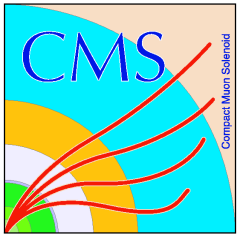
CMS b and quark/gluon tagging



Dinko Ferenček
Rutgers, The State University of New Jersey
(On behalf of the CMS Collaboration)

BOOST2014
August 18–22, 2014
University College London

Outline

Introduction/motivation

b tagging in CMS

With emphasis on boosted topologies

Quark/gluon tagging in CMS

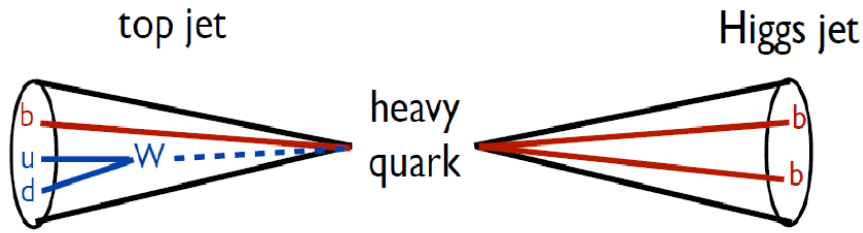
Summary and outlook

Introduction/motivation

- Most physics analyses look for final states that are flavor-specific
  - b jet discrimination: SM top quark, SM Higgs boson, SUSY,...
  - Quark/gluon jet discrimination: forward jets in VBF Higgs production, "tagging" of Z bosons in $H \rightarrow ZZ \rightarrow 2l2q$, characterization of monojets in dark matter searches,...

- What about boosted hadronically decaying resonances?
  - Classic examples of b-enriched final states are $t \rightarrow bW$ and $H \rightarrow b\bar{b}$
  - Classic example of a quark-enriched final state is $W \rightarrow q\bar{q}'$

$Br(t \rightarrow bW) \approx 100\%$ $Br(H(125) \rightarrow b\bar{b}) \approx 58\%$

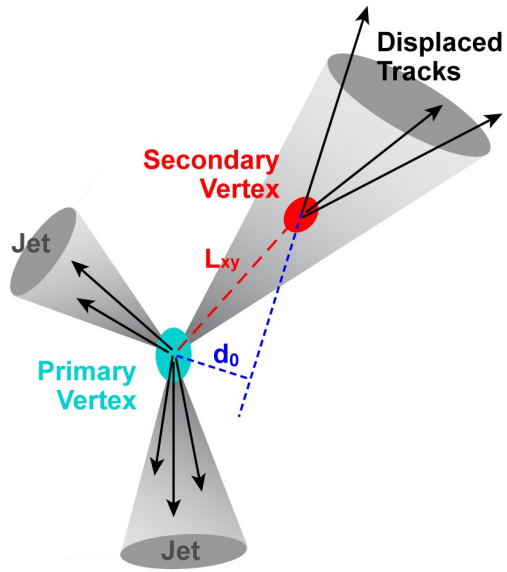


- *b and quark/gluon tagging, being largely complementary to the more traditional boosted tagging algorithms, can lead to significant improvements in the sensitivity of tagging algorithms for boosted objects*

b tagging in CMS

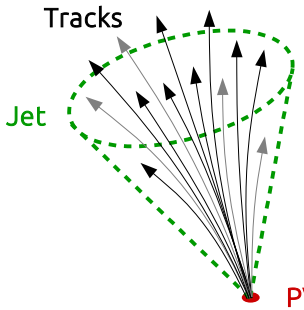
b jets and b tagging

- b tagging is a reconstruction technique that tries to “determine” whether a jet contains a b hadron
- Exploits distinct properties of b hadrons:
  - Long lifetime ($\tau \approx 1.5$ ps, $c\tau \sim 500$ μm, $\beta\gamma c\tau \approx 5$ mm @ 50 GeV; for comparison, primary vertex position resolution ~few tens of μm)
  - Large mass (~5 GeV)
  - Decays with high track multiplicities (~5 on average)
  - Relatively large semileptonic branching fraction (for electrons and muons, ≈20% each with cascade decays included)
  - Hard fragmentation function (a large fraction of the original b quark momentum carried by the b hadron)

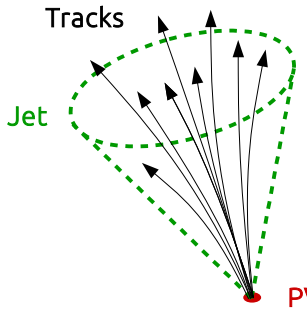


- Relies on the track reconstruction and can be based on:
  - Displaced track information
  - Secondary vertex information
  - Soft lepton information
  - Some combination of the above
- Several b-tagging algorithms available in CMS
  - Each producing a single discriminator value per tagged jet; the more positive the value the more b-like the jet is

b tagging in CMS



Fixed-size association cone ΔR(track,jet)<0.3



Jet-track association

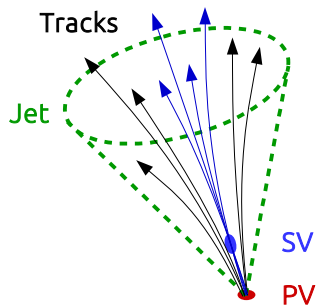
Track selection

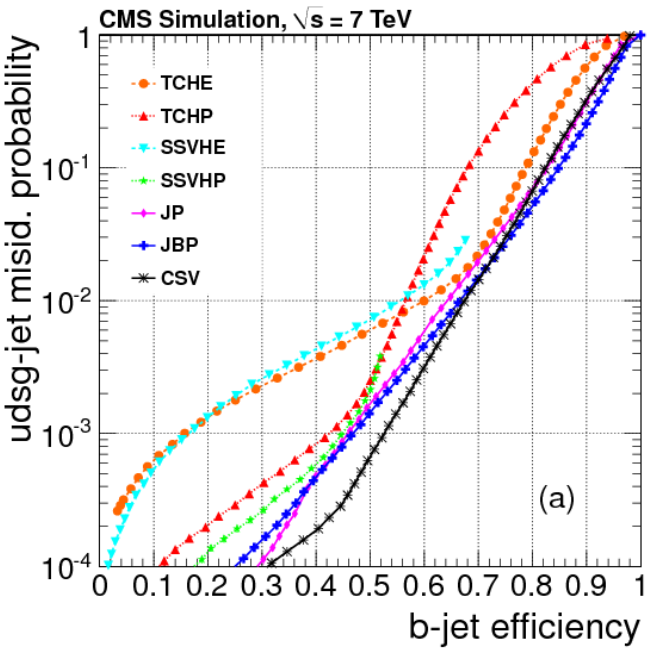
Track-based tagging algorithms

Combined tagging algorithms

Secondary vertex reconstruction

SV-based tagging algorithms





Tagger operating points:

L = loose (≈10% light-flavor mistag rate)

M = medium (≈1% light-flavor mistag rate)

T = tight (≈0.1% light-flavor mistag rate)

Boosted b tagging in Run I

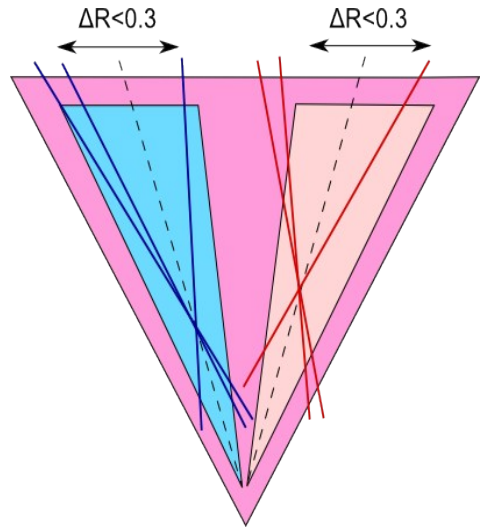
- Using Combined Secondary Vertex (CSV) algorithm
- b-tagging scenarios considered

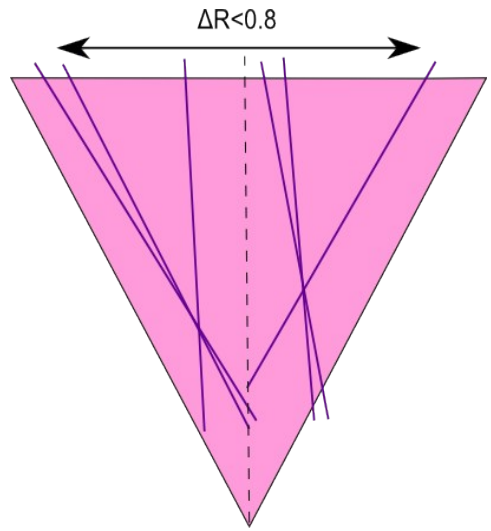
Subjet b tagging:

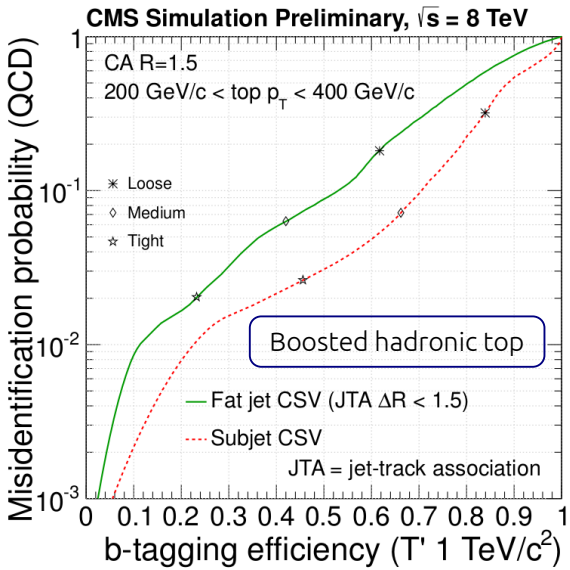
- Standard CSV applied to subjets of fat jets (2 (≥1) subjet tags for boosted Higgs (top) candidates)
- Standard jet-track association ΔR<0.3

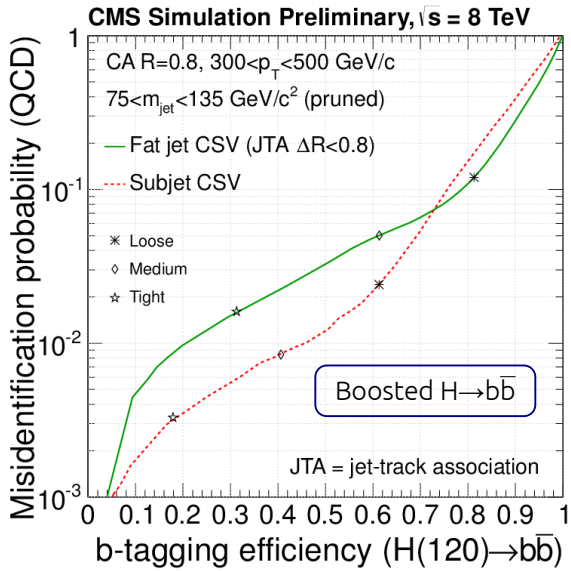
Fat jet b tagging:

- Standard CSV applied to Higgs/top candidate fat jets
- Extended jet-track association ΔR<R_{jet} (0.8 or 1.5)



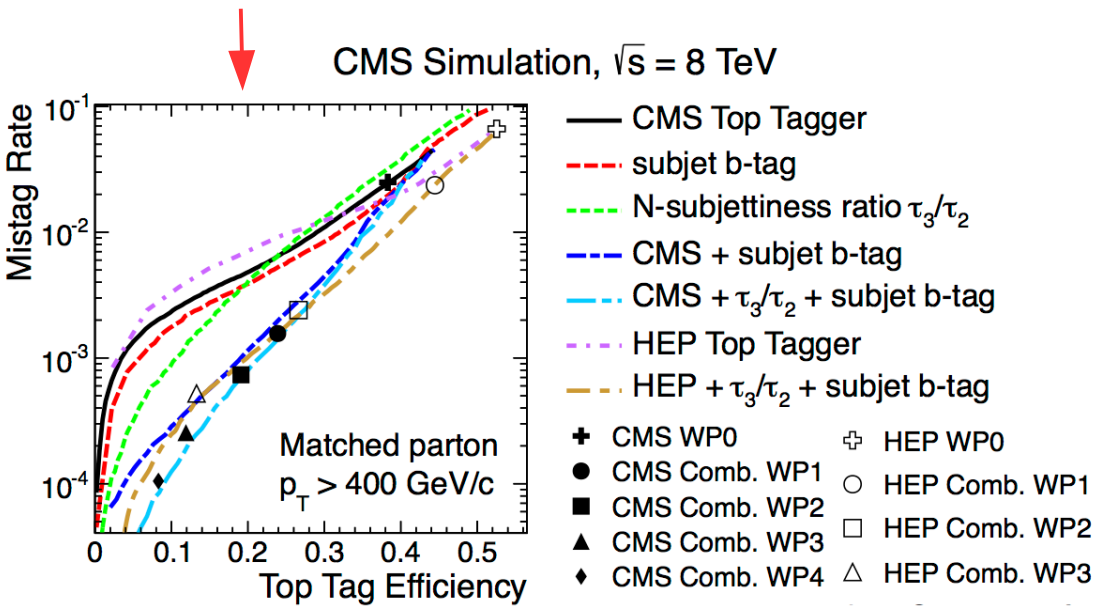


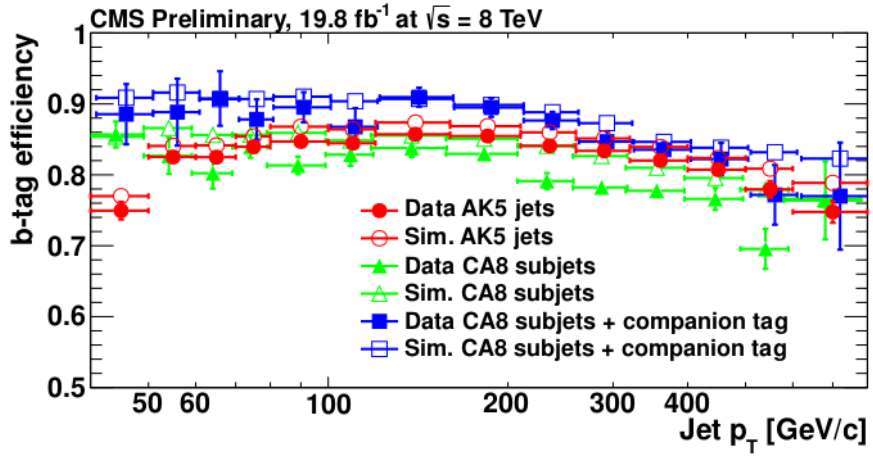


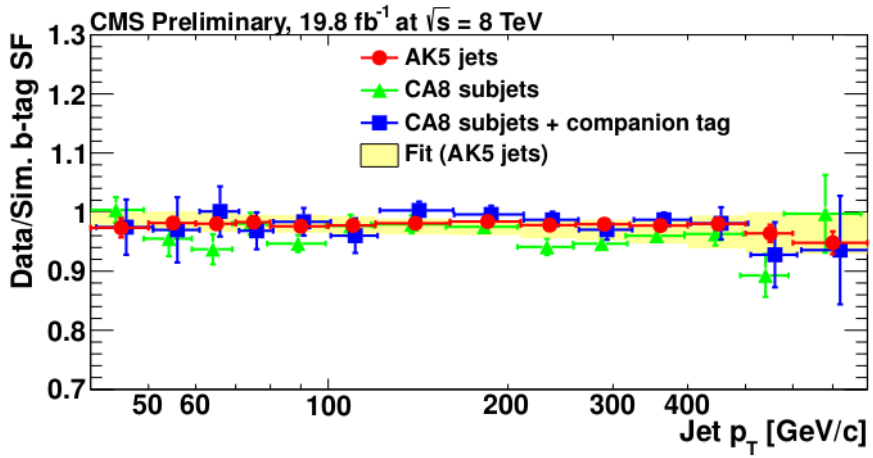


Boosted b tagging in Run I (cont'd)

- b tagging for boosted topologies successfully commissioned using 8 TeV collision data [1,2] and first CMS analyses exploiting subjet b tagging now public
- Subjet b tagging also an integral part of boosted top tagging algorithms commissioned in CMS [3]



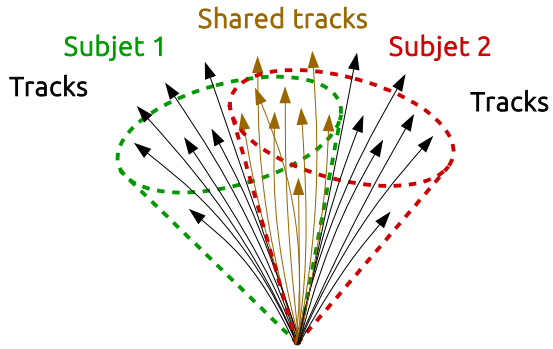


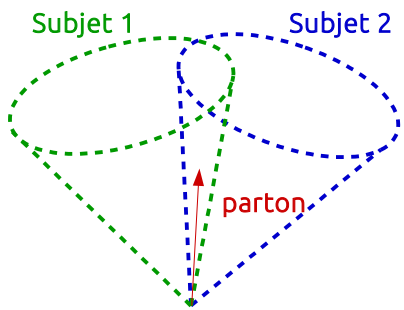


[1] http://cds.cern.ch/record/1581306/files/BTV-13-001-pas.pdf
[2] https://twiki.cern.ch/twiki/bin/view/CMSPublic/PhysicsResultsBTV13001
[3] http://cds.cern.ch/record/1647419/files/JME-13-007-pas.pdf

Limitations of Run I setup

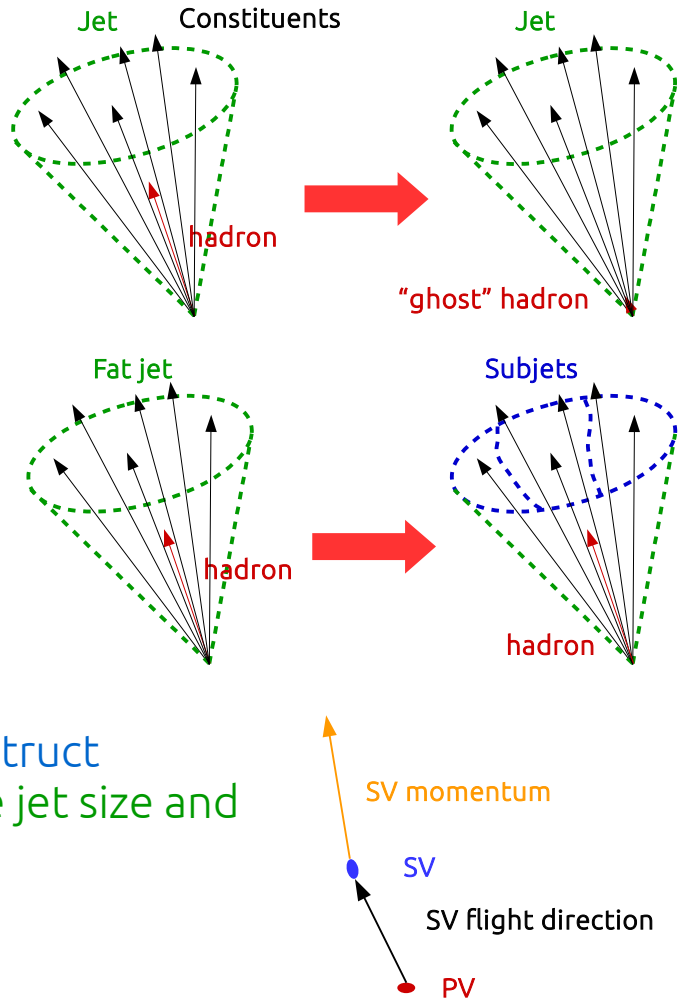
- Current boosted b-tagging setup based on the software framework and tagging algorithms designed for R=0.5 jets
  - Facilitated commissioning studies and early adoption in physics analyses
  - Certain aspects suboptimal for boosted topologies
- Jet-track association:
  - Based on a fixed-size cone
  - Can lead to double-counting of tracks at high p_T and subjet tag correlations (problematic for the application of data/MC scale factors)
  - Default cone size also not optimal for fat jet b tagging
- Jet flavor assignment:
  - Also based on a fixed-size cone ($\Delta R<0.3$)
  - Can lead to subjet flavor ambiguities
- Secondary vertex reconstruction:
  - Using tracks associated to jets (not optimal when the fraction of shared tracks becomes significant)





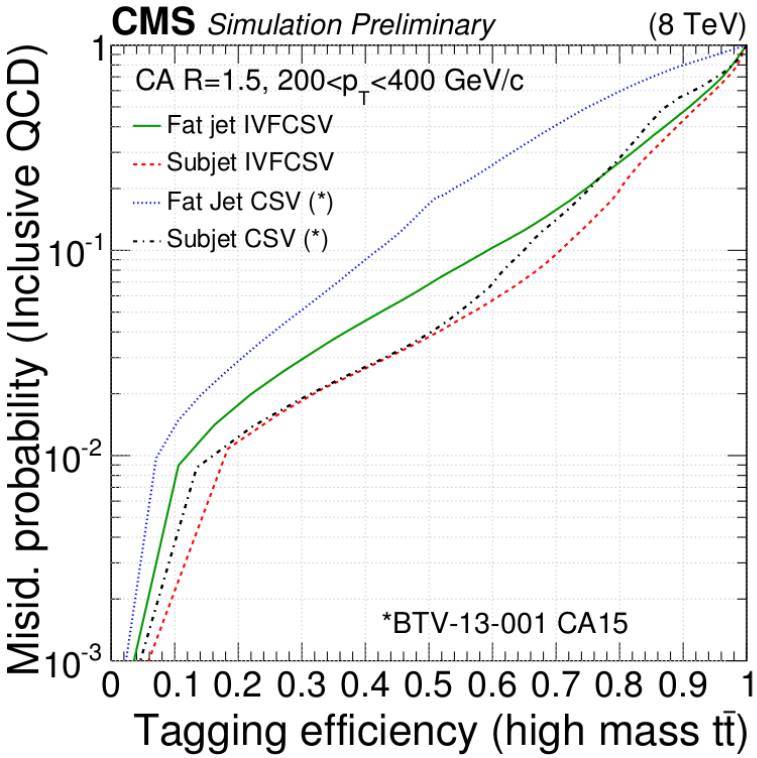
Run II developments

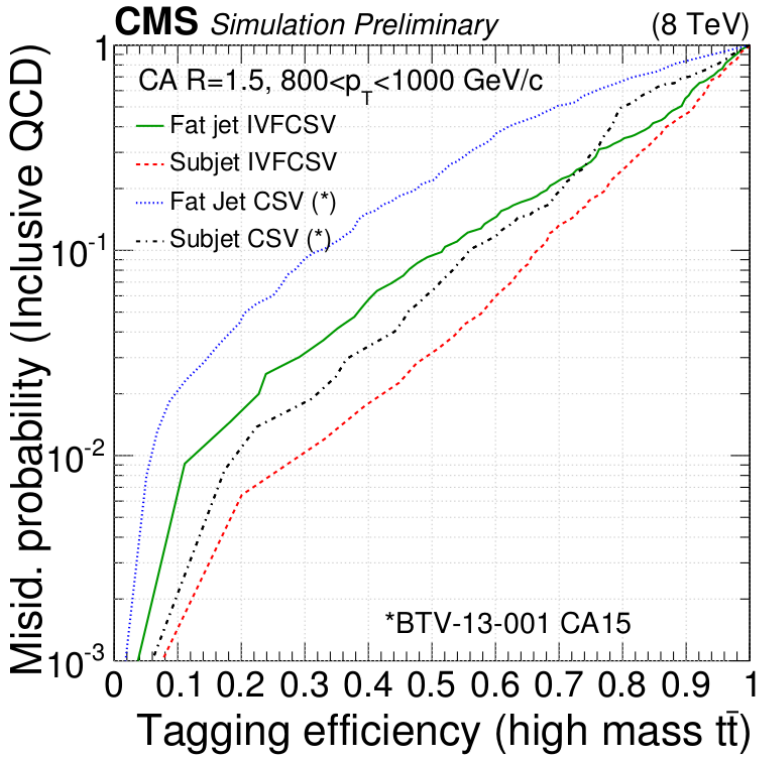
- Improved (sub)jet flavor definition [1]
  - Using b and c hadrons instead of b and c quarks
  - Based on clustering "ghost" hadrons/partons instead of ΔR matching → Subjet flavor ambiguities eliminated
- Explicit jet-track association
  - Uses tracks linked to charged constituents of particle-flow jets
  - Eliminates the problem of shared tracks
- Inclusive Vertex Finder (IVF) secondary vertices
  - Does not require jets and instead uses all tracks to reconstruct secondary vertices → By construction independent of the jet size and does not introduce track sharing
  - Jet clustering used to assign SVs to (sub)jets
- Improved CSV algorithm



[1] https://twiki.cern.ch/twiki/bin/view/CMSPublic/SWGuideBTagMCTools

Boosted hadronic top (Inclusive QCD as background)

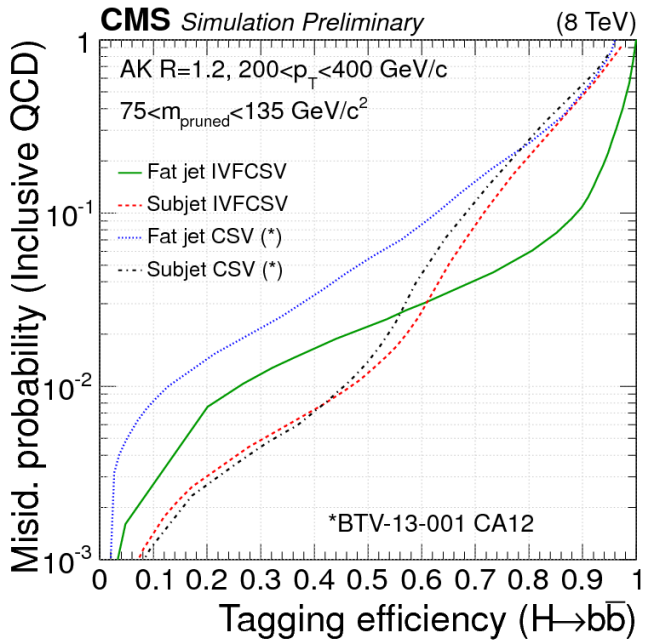


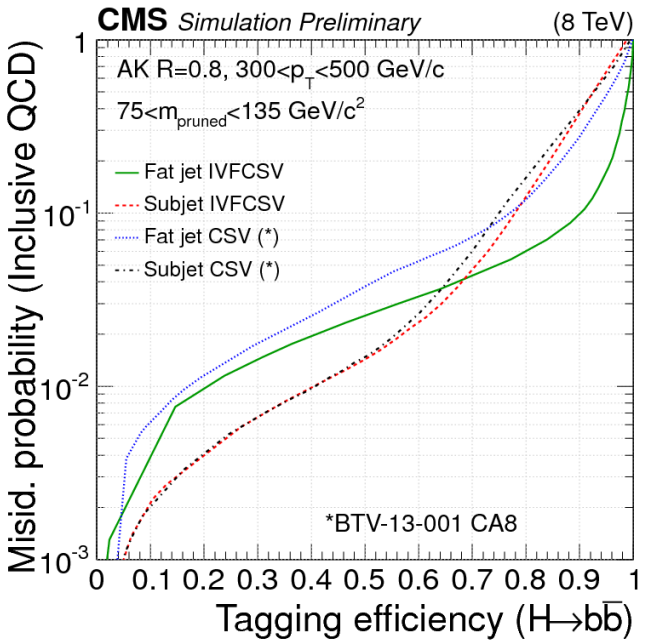


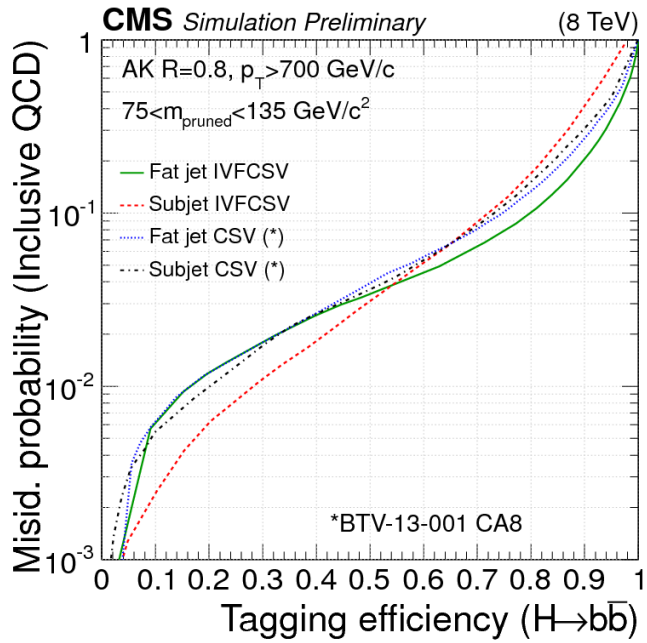
Subjet tagging efficiency refers to tagging ≥1 subjets

- CA R=1.5 fat jets and HEPTopTagger subjets
- Improved CSV algorithm based on IVF vertices performs better than the older generation CSV algorithm
- Subjet b tagging outperforms fat jet b tagging in the entire p_T range considered

Boosted H→$b\bar{b}$ (Inclusive QCD as background)



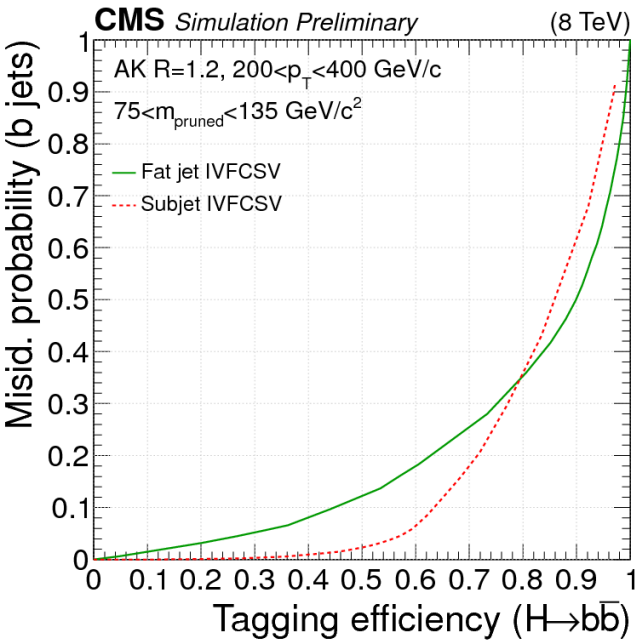


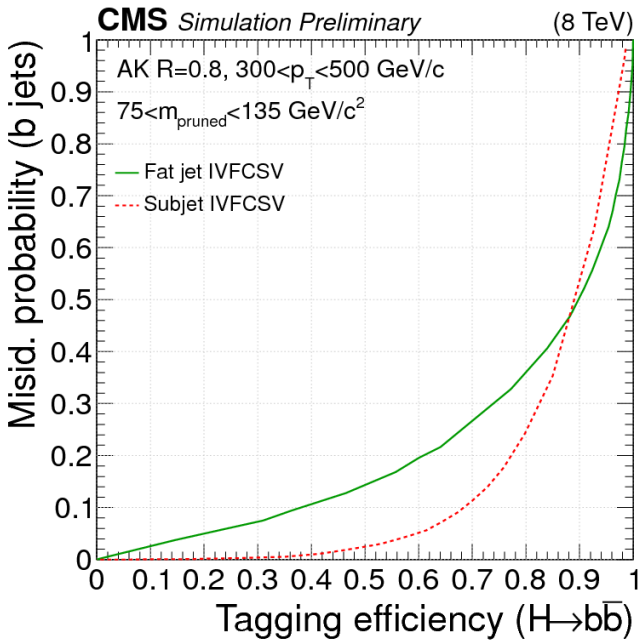


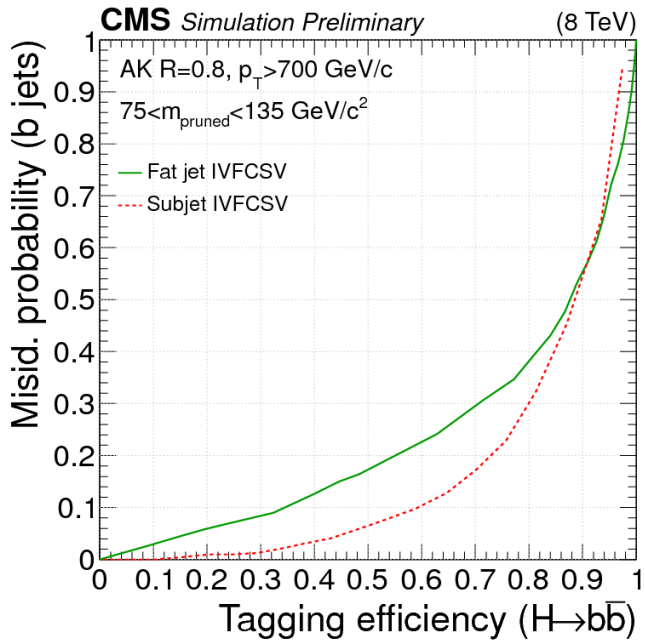
Subjet tagging efficiency refers to tagging both subjets

- AK R=0.8 or 1.2 (depending on the p_T range) fat jets and pruned subjets (z_{cut}=0.1 and R_{cut}=0.5)
- Improved CSV algorithm based on IVF vertices performs better than the older generation CSV
  - Older CSV algorithm applied to CA jets but the choice of the clustering algorithm found to have negligible impact on the b-tagging performance
- Subjet and fat jet b tagging curves cross each other with fat jet b tagging performing better at high tagging efficiencies

Boosted H→$b\bar{b}$ (b jets as background)

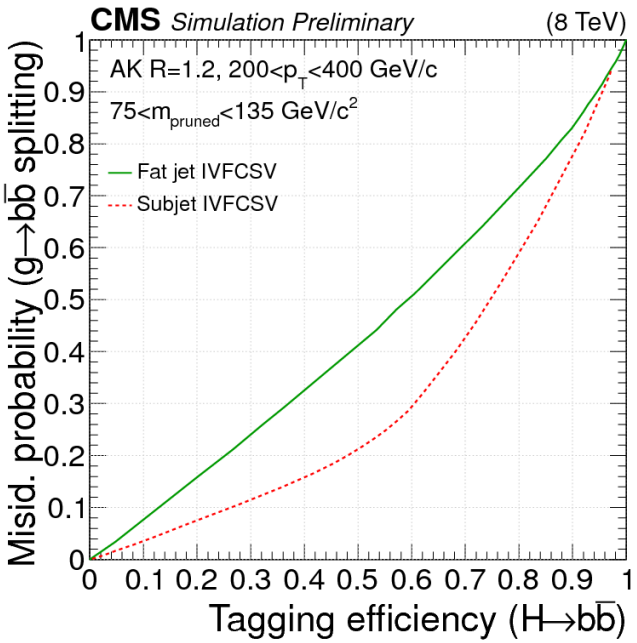


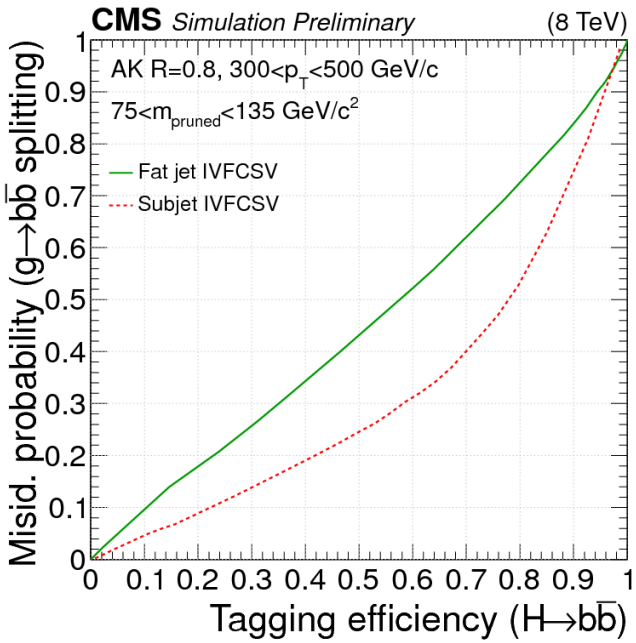


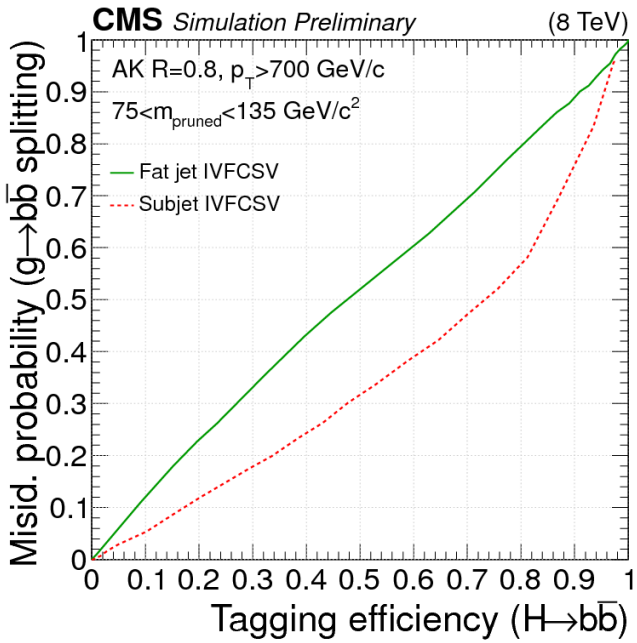


- Subjet b tagging generaly outperforms fat jet b tagging except at high tagging efficiencies for lower p_T

Boosted H→$b\bar{b}$ (Gluon splitting as background)

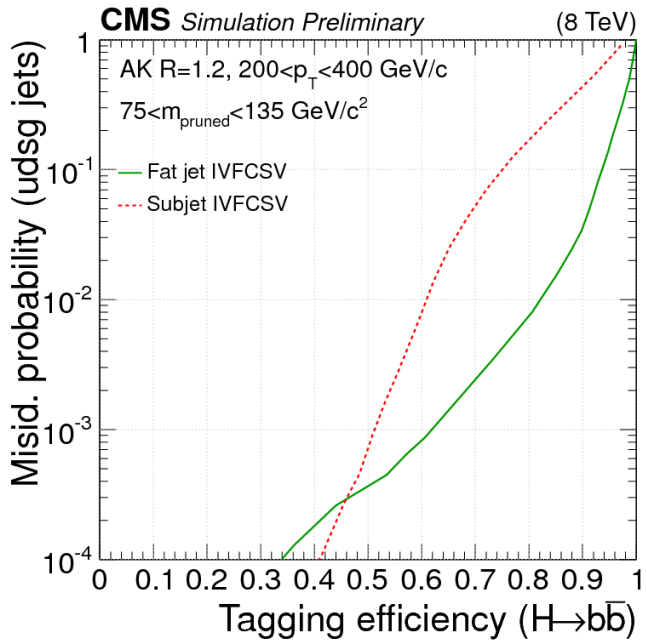


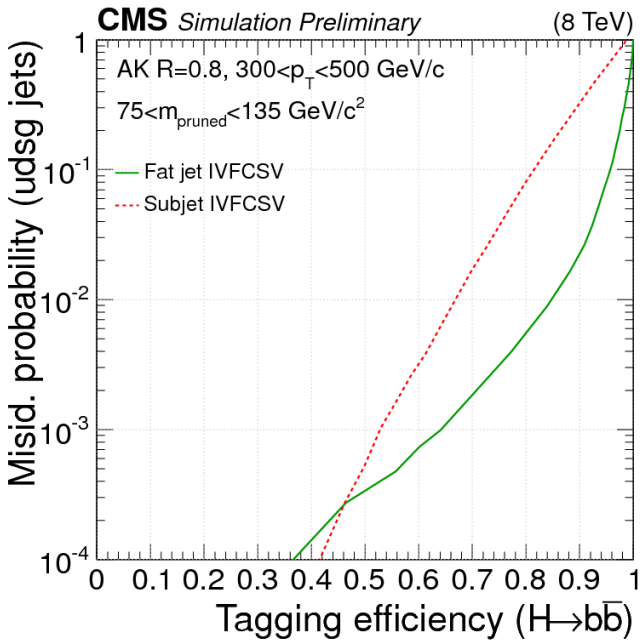


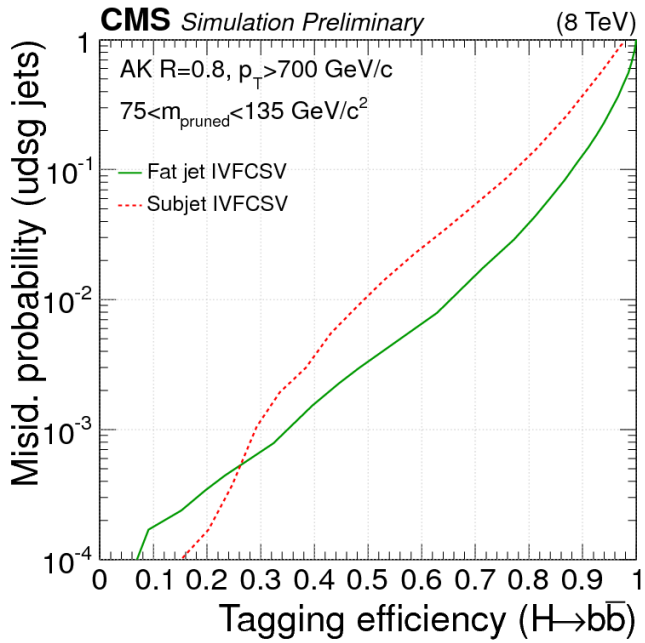


- Subjet b tagging outperforms fat jet b tagging in the entire p_T range considered

Boosted H→$b\bar{b}$ (udsg jets as background)

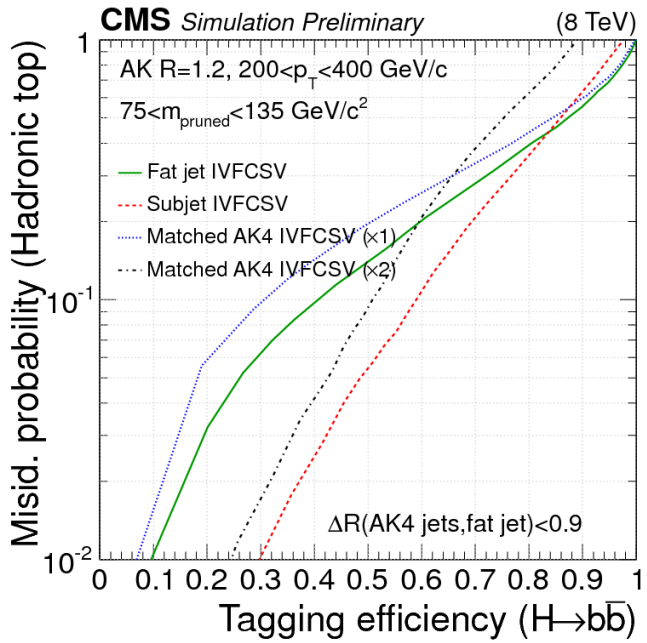


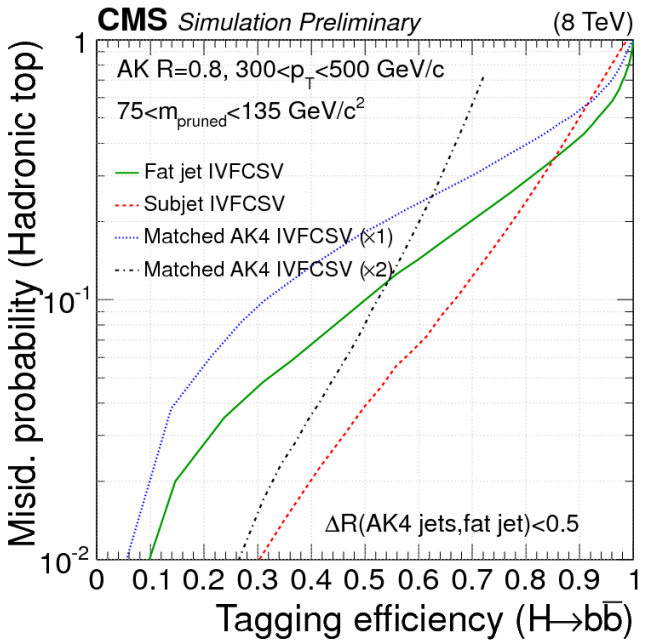


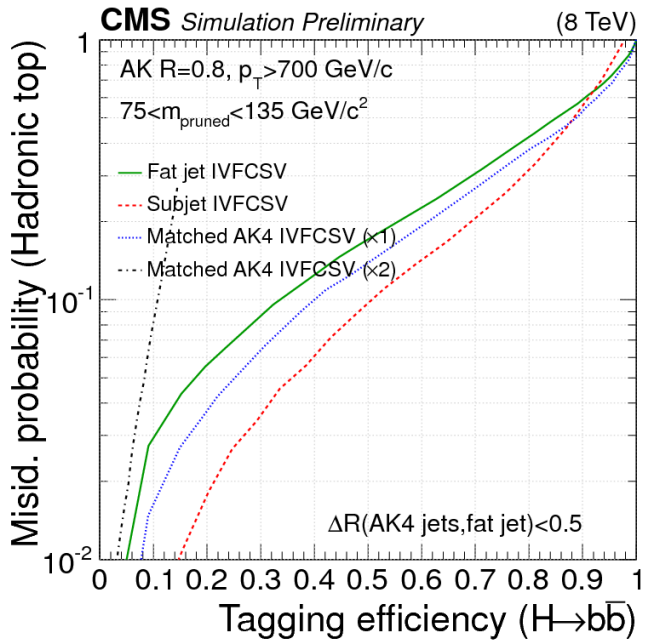


- Fat jet b tagging generally outperforms subjet b tagging in the entire p_T range considered, expect at low tagging efficiencies

Boosted H→$b\bar{b}$ (Hadronic top as background)







- Subjet b tagging outperforms both fat jet b tagging and matched AK4 jets in the entire p_T range considered

Quark/gluon tagging in CMS

http://cds.cern.ch/record/1599732/files/JME-13-002-pas.pdf
https://twiki.cern.ch/twiki/bin/view/CMSPublic/PhysicsResultsJME13002

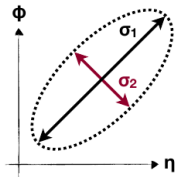
Quark vs gluon jets

Compared to gluon jets, quark jets have:

- Fewer constituents
- Narrower shape
- Harder fragmentation function and less symmetric energy sharing among constituents

Multiplicity variables
Charged multiplicity
Neutral multiplicity
Total multiplicity
Width variables
Major axis of η-φ shape (σ_1)
Minor axis of η-φ shape (σ_2)
σ
Energy sharing variables
Pull
R
$p_T D$

From particle flow



obtained by diagonalizing

$$\frac{1}{\sum_i p_{\mathrm{T},i}^2} \sum_i p_{\mathrm{T},i}^2 \begin{pmatrix} (\Delta\phi_i)^2 & (\Delta\phi_i \Delta\eta_i) \\ (\Delta\eta_i \Delta\phi_i) & (\Delta\eta_i)^2 \end{pmatrix}$$

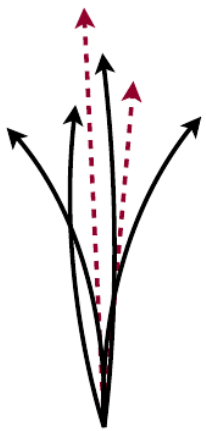
$$\sigma = \sqrt{\sigma_1^2 + \sigma_2^2}$$

$$|\vec{t}| = \left| \frac{\sum_i p_{\mathrm{T},i}^2 \, |r_i| \, \vec{r}_i}{\sum_i p_{\mathrm{T},i}^2} \right| \qquad \vec{r}_i = (\Delta\eta_i, \Delta\phi_i)$$

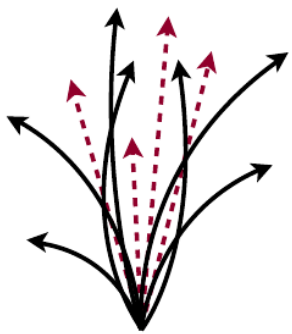
$$R = \frac{\max\left(p_{\mathrm{T},i}\right)}{\sum_i p_{\mathrm{T},i}}$$

$$p_{\mathrm{T}} D = \frac{\sqrt{\sum_i p_{\mathrm{T},i}^2}}{\sum_i p_{\mathrm{T},i}}$$

Quark jets:

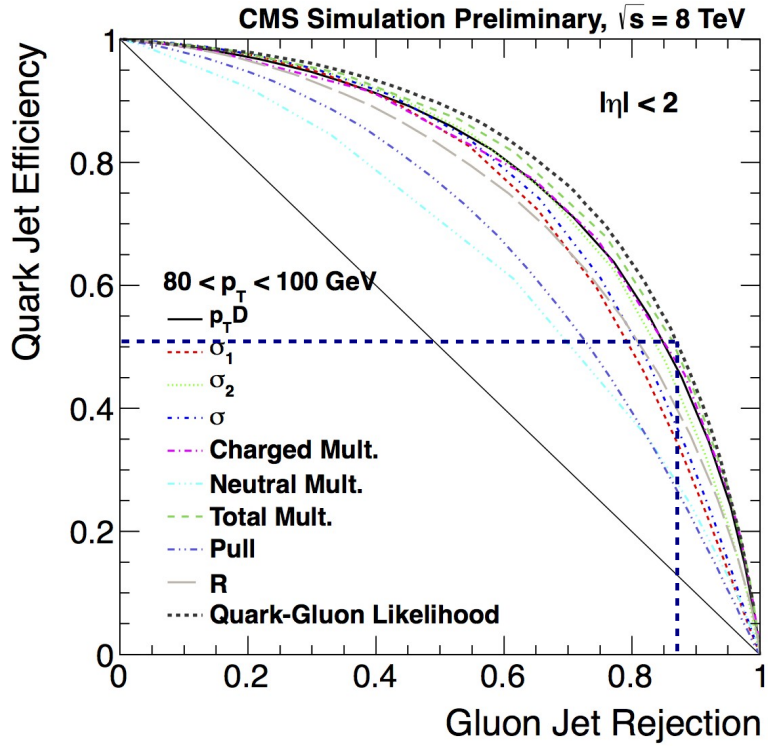


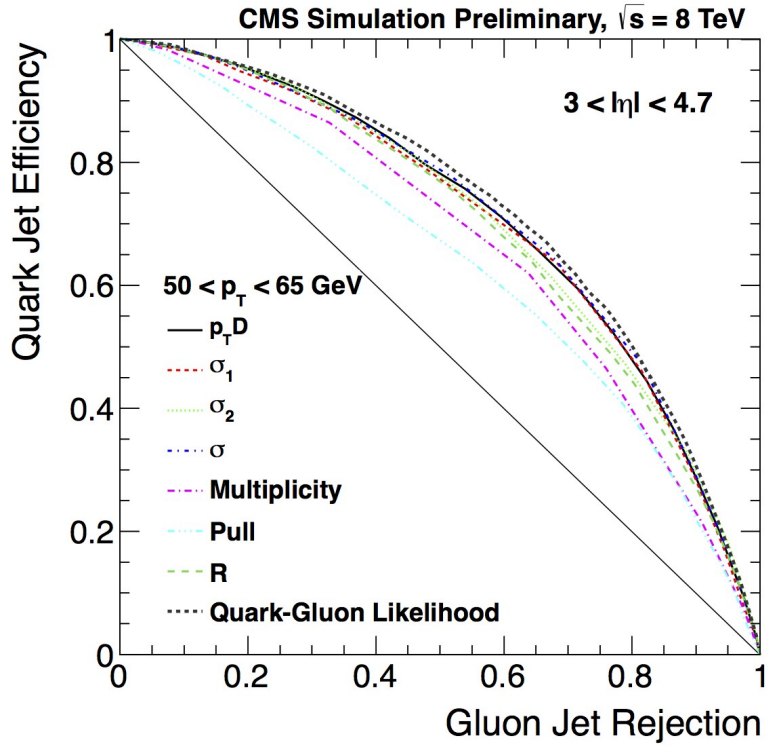
Gluon jets:



Discriminating power of quark/gluon variables

Pythia6



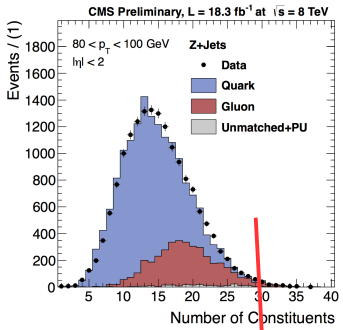


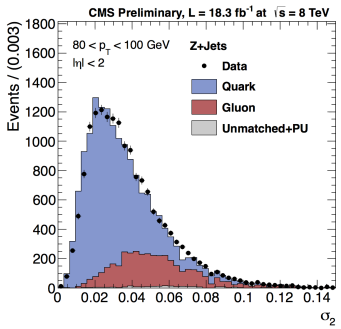
≈87% background rejection for 50% signal efficiency (for $80<p_T<100$ GeV, $|\eta|<2$)

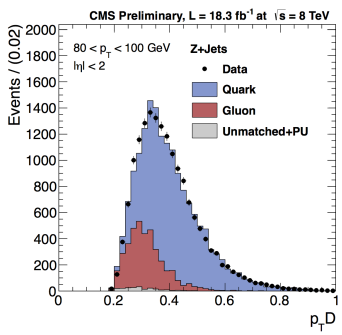
Quark-gluon discriminator

Likelihood-based discriminator obtained by combining 3 variables

- Total multiplicity
- Minor axis (σ_2)
- $p_T D$





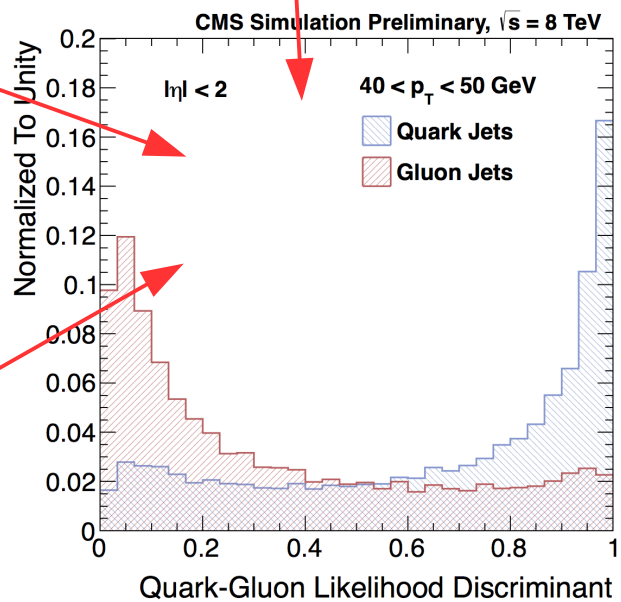


Discriminant defined for jets with p_T>30 GeV and |η|<5

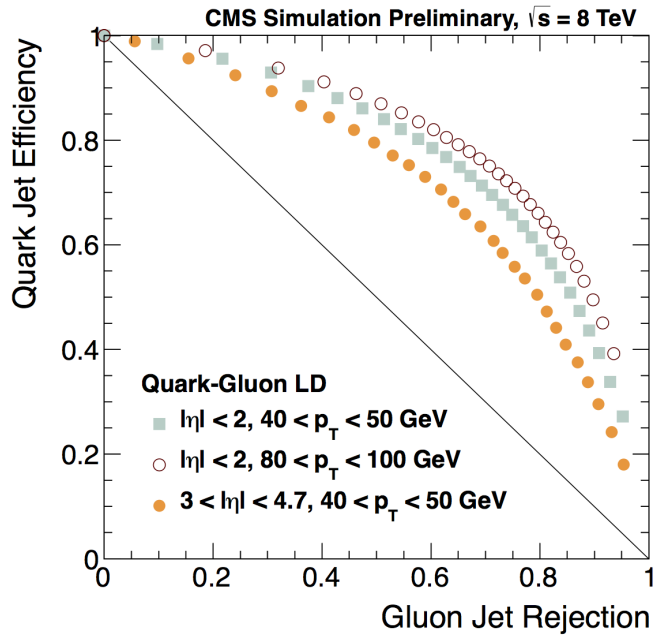
For better discrimination and resilience to pileup

- Using charged hadron subtraction
- Neutral constituents with p_T>1 GeV

PDFs binned in jet p_T and η and pileup transverse momentum density ρ

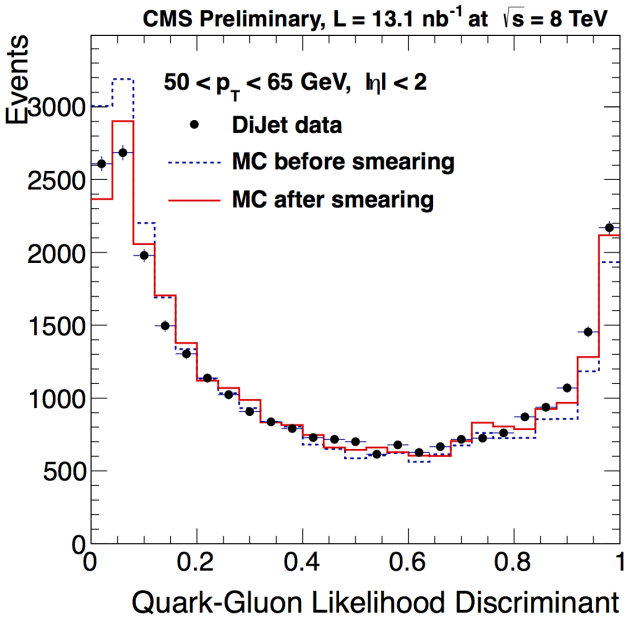


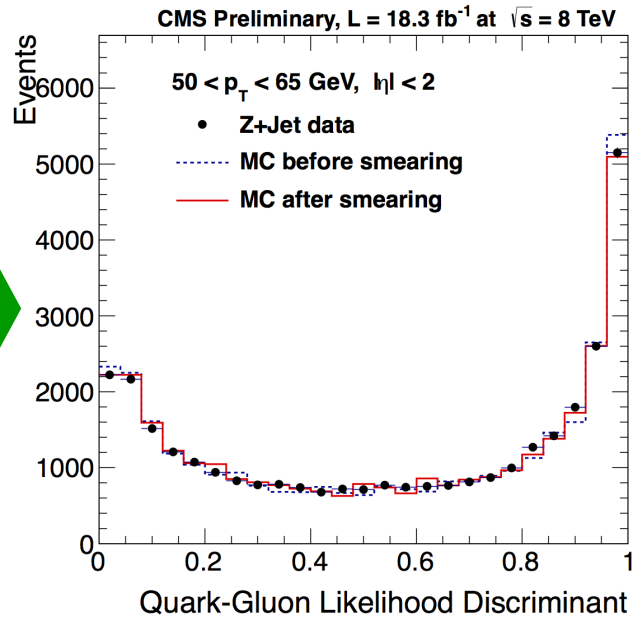
Pythia6

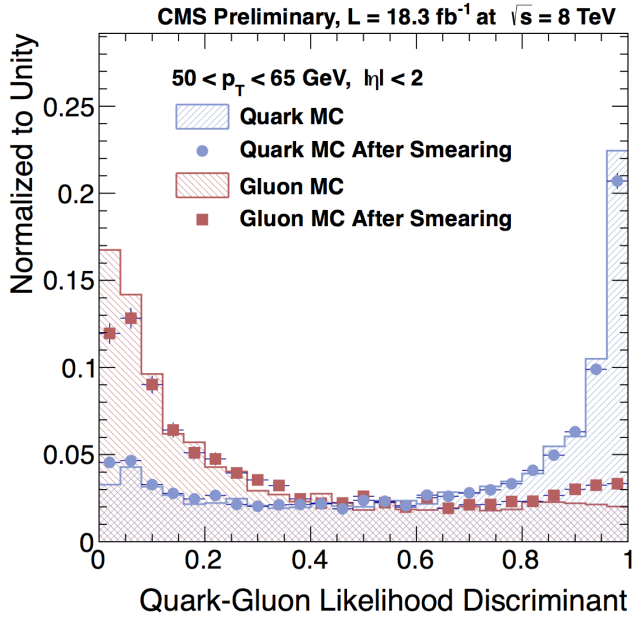


Data validation

- Discrepancy observed in gluon-enriched dijet control sample → Need to smear MC distribution to better match data
- Reshaping corrections derived in a dijet control sample (>60% gluon jets) and validated in a Z+jets control sample (>70% quark jets)
- Smearing function: $g(x, a, b) = \tanh\left[a \operatorname{arctanh}(2x - 1) + b\right]/2 + \frac{1}{2}$
  - Remaps q/g discriminant distributions on jet-by-jet basis (separately for quark and gluon jets)

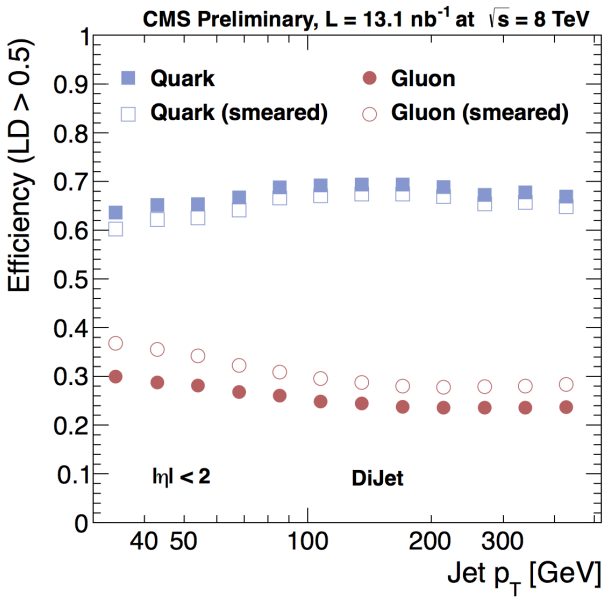




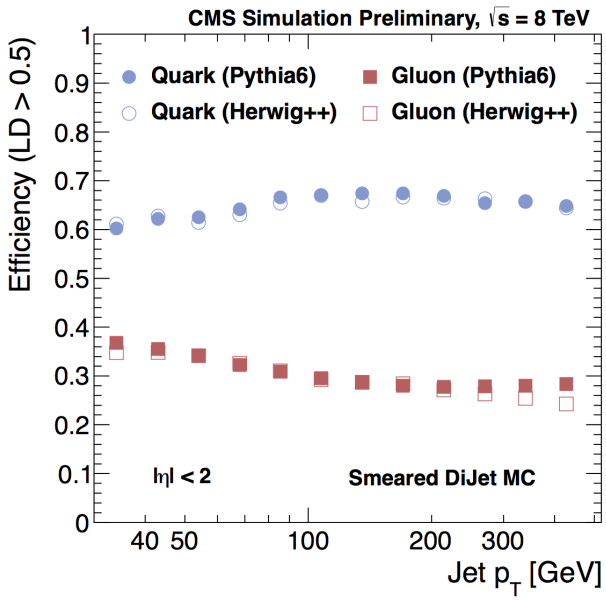


Nature vs Pythia6 and Herwig++

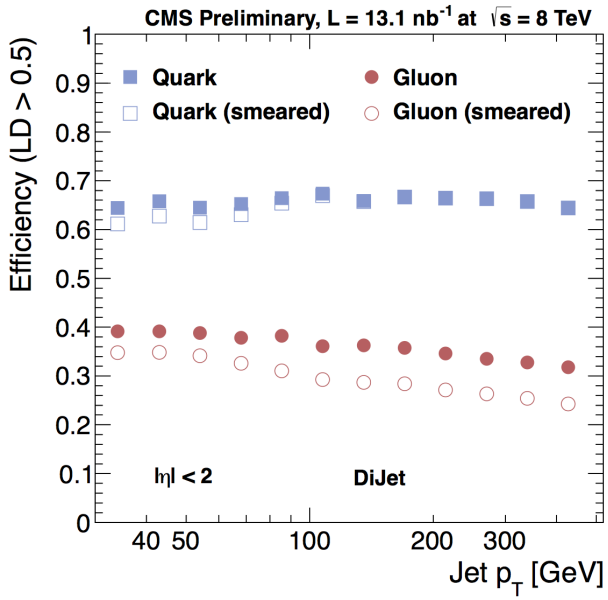
Pythia6



Smeared MC



Herwig++

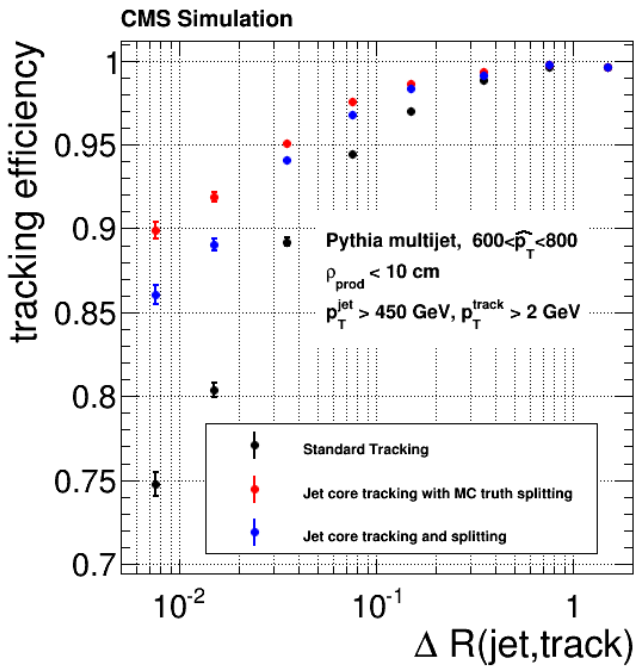


- Nature lies in-between Pythia6 and Herwig++ predictions

Summary and outlook

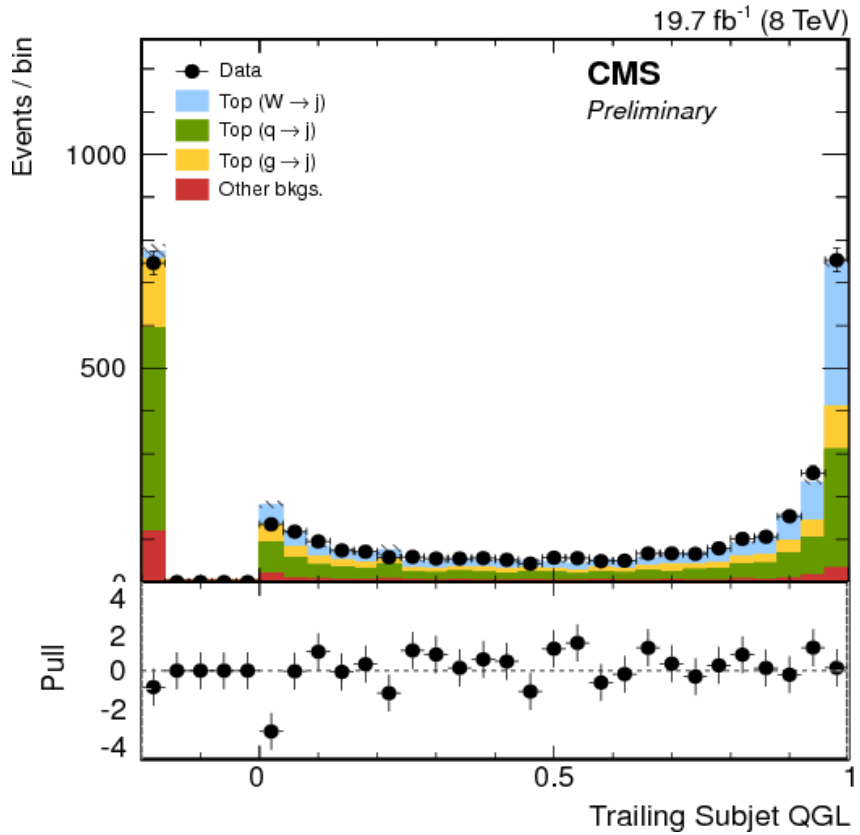
Summary and outlook

- b tagging in boosted topologies successfully commissioned during Run I
  - Recommended approach is to use subjet b tagging and at very high boosts switch to standard b tagging applied to fat jets
- First analyses using subjet b tagging now public (see CMS talks on boosted Higgs bosons and top quarks in physics analyses). More analyses in the pipeline
- Several new developments addressing some of the shortcomings of the Run I setup have been presented
  - Overall see improved performance
  - Significantly improved fat jet b tagging which in the case of $H \rightarrow b\bar{b}$ jets outperforms subjet b tagging at high tagging efficiencies
  - Further improvements still possible
- Improved track reconstruction in the core of high-p_T jets expected to further improve the performance of b tagging in boosted topologies



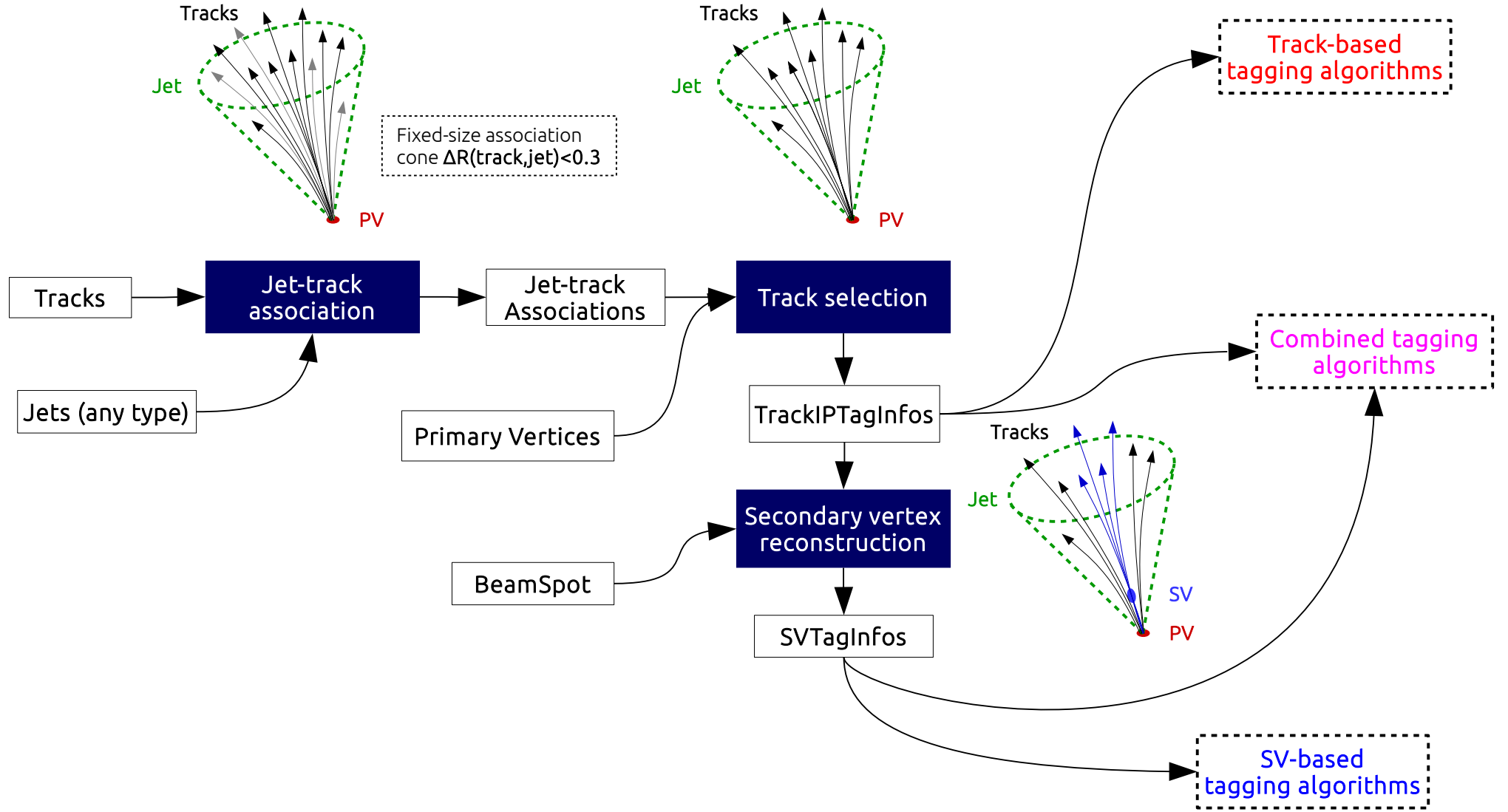
Summary and outlook (cont'd)

- Quark/gluon tagging also successfully commissioned during Run I
  - Available for p_T>30 GeV and full η coverage (|η|<5)
- Quark/gluon tagger based on 3 variables combined into a likelihood-based discriminator
  - Constituent multiplicity
  - Jet width
  - Fragmentation function
- Discriminant reshaping corrections derived for improved data/MC agreement
- New developments involve the use of subjet q/g tagging (see CMS talk on W tagging)
- *CMS has developed a powerful set of tools for discrimination of different (sub)jet flavors which forms a strong foundation for future developments*



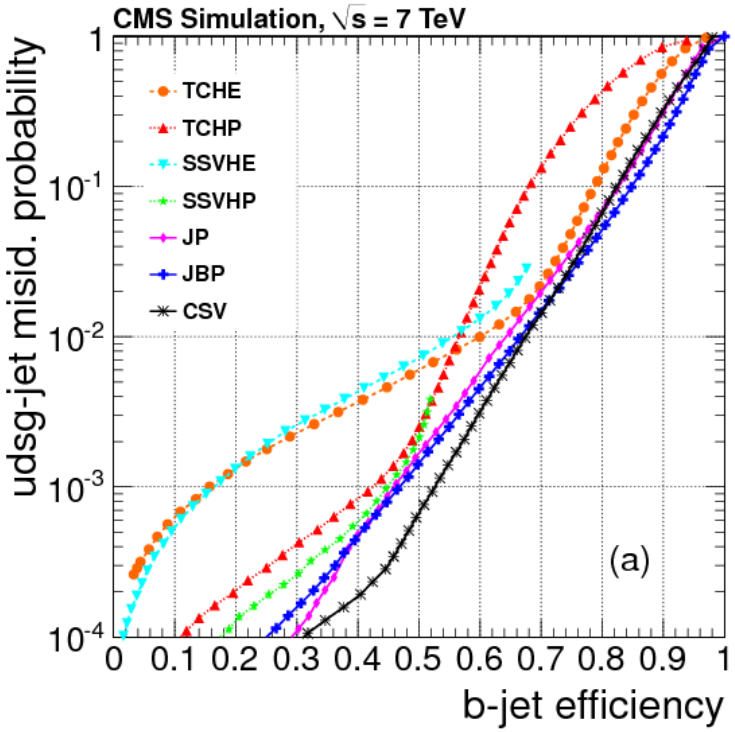
Backup Slides

b tagging in CMS



CMS b-tagging algorithms

Tagging Algorithm	Operating points	Supported at 7 TeV	Supported at 8 TeV
Track Counting High Efficiency	TCHEL	✔	✘
	TCHEM	✔	✘
	TCHET	✘	✘
Track Counting High Purity	TCHPL	✘	✘
	TCHPM	✔	✘
	TCHPT	✔	✔
Jet Probability	JPL	✔	✔
	JPM	✔	✔
	JPT	✔	✔
Jet B Probability	JBPL	✔	✘
	JBPM	✔	✘
	JBPT	✔	✘
Simple Secondary Vertex High Efficiency	SSVHEM	✔	✘
	SSVHET	✘	✘
Simple Secondary Vertex High Purity	SSVHPT	✔	✘
Combined Secondary Vertex	CSVL	✔	✔
	CSVM	✔	✔
	CSVT	✔	✔



From JINST 8 (2013) P04013

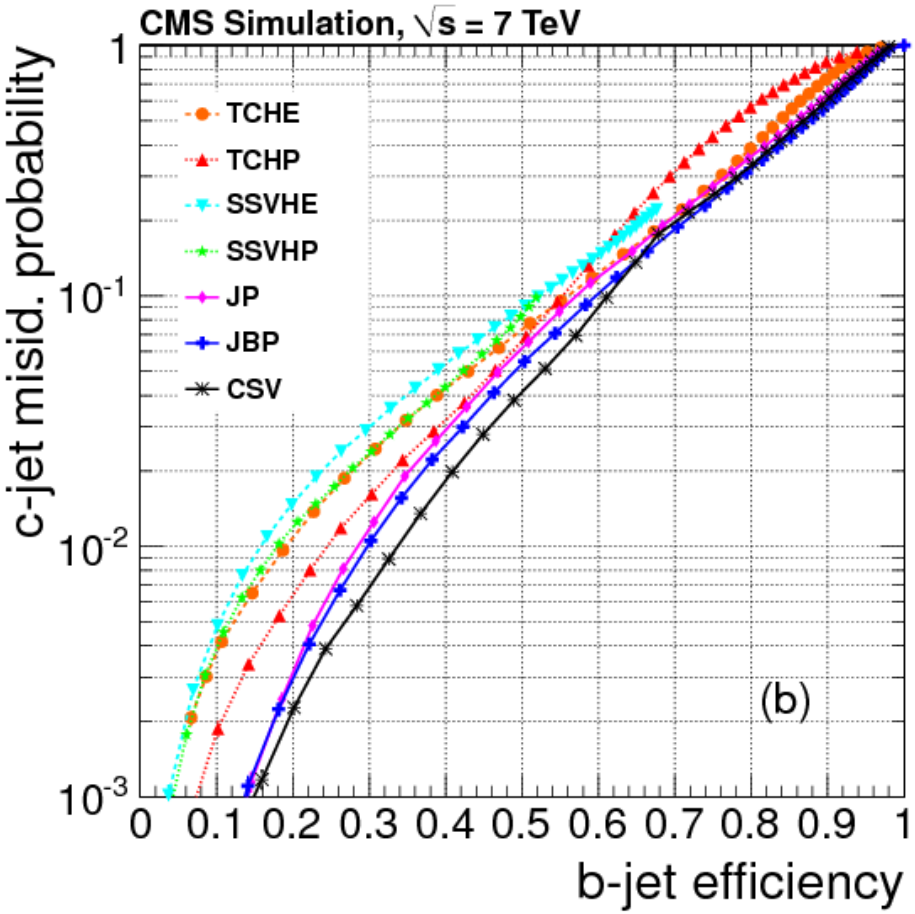
Tagger operating points:

L = loose (≈10% light-flavor mistag rate)

M = medium (≈1% light-flavor mistag rate)

T = tight (≈0.1% light-flavor mistag rate)

CMS b-tagging algorithms (cont'd)



From JINST 8 (2013) P04013

CSV algorithm

Older generation CSV:

- Likelihood-ratio-based discriminator
- Based on the variables listed below

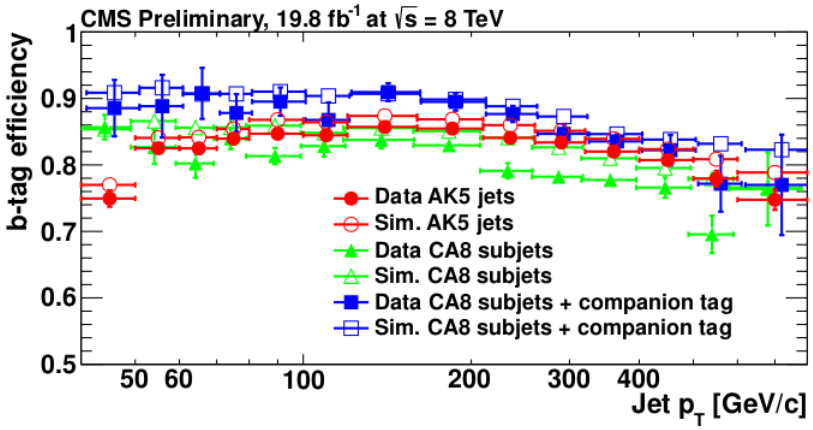
Variable	Vertex category		
	RecoVertex	PseudoVertex	NoVertex
trackSip3dSig	✔	✔	✔
trackSip2dSigAboveCharm	✔	✔	✘
trackEtaRel	✔	✔	✘
vertexMass	✔	✔	✘
vertexNTracks	✔	✔	✘
vertexEnergyRatio	✔	✔	✘
flightDistance2dSig	✔	✘	✘

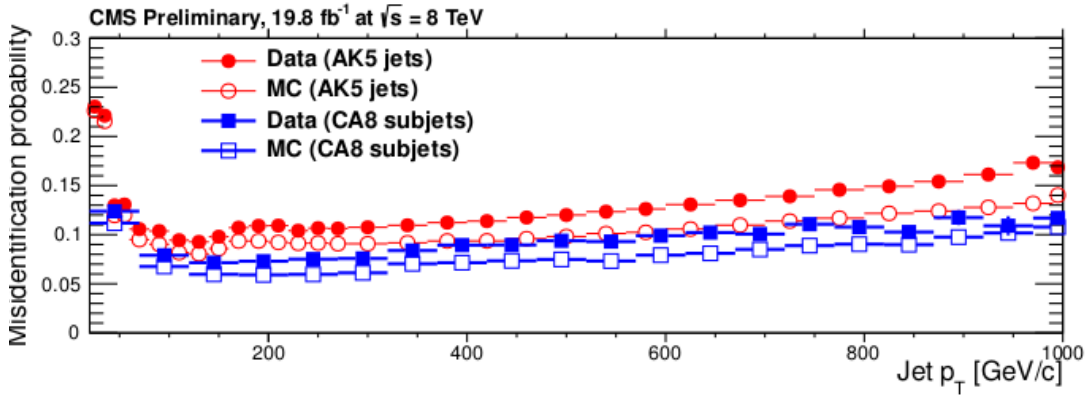
Improved CSV:

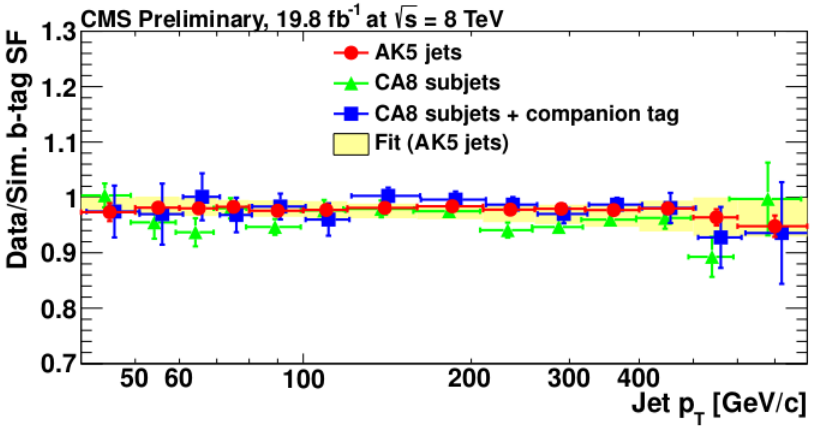
- MLP-based discriminator
- Based on the variables listed below

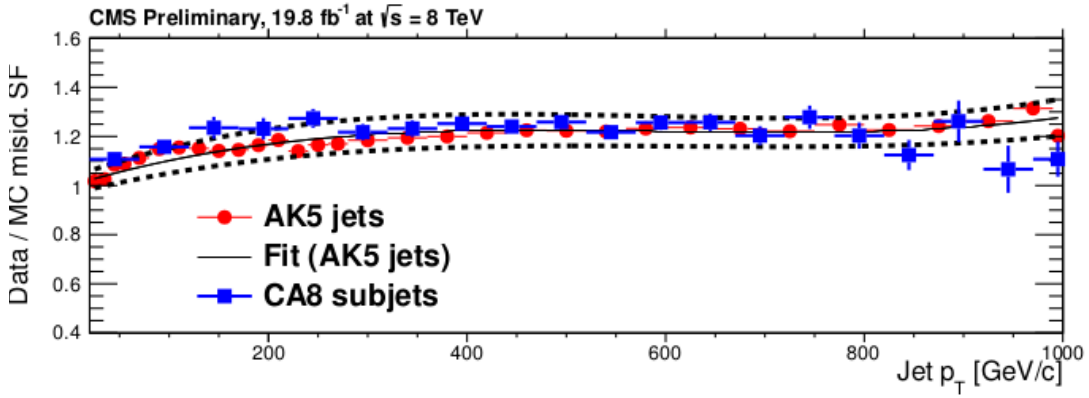
Variable	Vertex category		
	RecoVertex	PseudoVertex	NoVertex
trackSip3dSig	✔	✔	✔
trackSip2dSigAboveCharm	✔	✔	✔
jetNTracks	✔	✔	✔
trackEtaRel	✔	✔	✘
vertexMass	✔	✔	✘
vertexNTracks	✔	✔	✘
vertexEnergyRatio	✔	✔	✘
vertexJetDeltaR	✔	✔	✘
flightDistance2dSig	✔	✘	✘
jetNSecondaryVertices	✔	✘	✘

Validation of “boosted” b tagging in data









More details in http://cds.cern.ch/record/1581306/files/BTV-13-001-pas.pdf

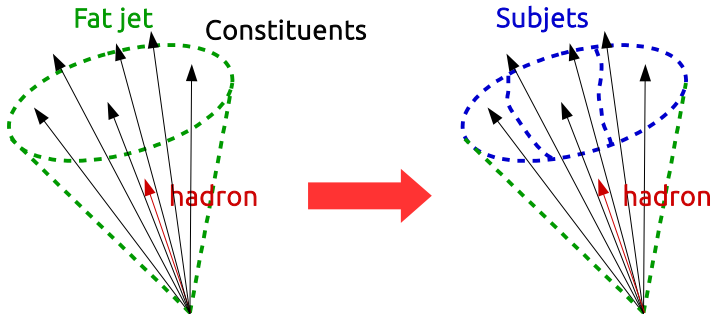
Inclusive Vertex Finder SV reconstruction

1. Coarse track pre-clustering around displaced seed tracks
   - Based on track distances and angles
2. Vertex reconstruction/fitting from the track clusters obtained in step 1 (using "adaptive vertex fit")
3. Vertex merging
   - Check vertices for shared tracks
   - Remove vertex if shared fraction >0.7 and distance significance <2
4. Track-vertex arbitration
   - Trade off tracks between PV and SV based on their compatibility with vertices
   - Refit vertices with new track selection
5. Vertex merging
   - Same as step 3 with max. shared fraction of 0.2 and min. distance significance of 10

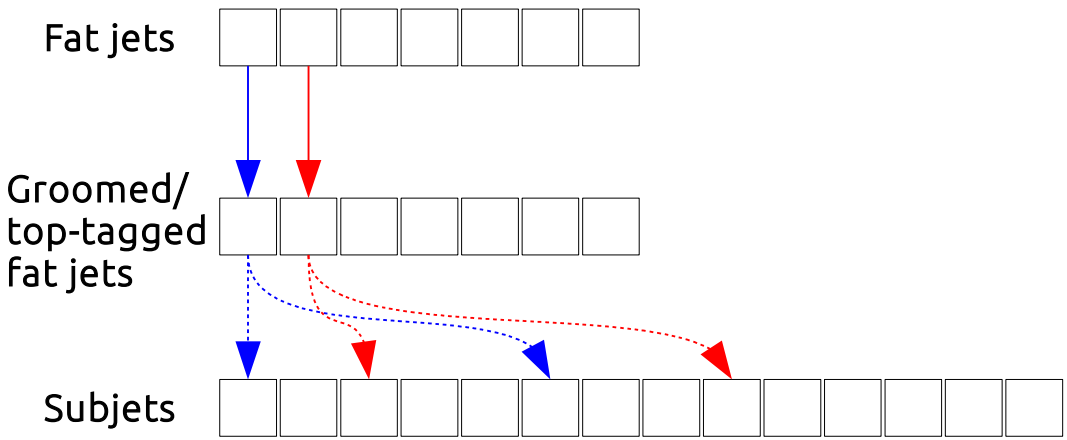
Algorithm employed in JHEP 03 (2011) 136

Subjet flavor

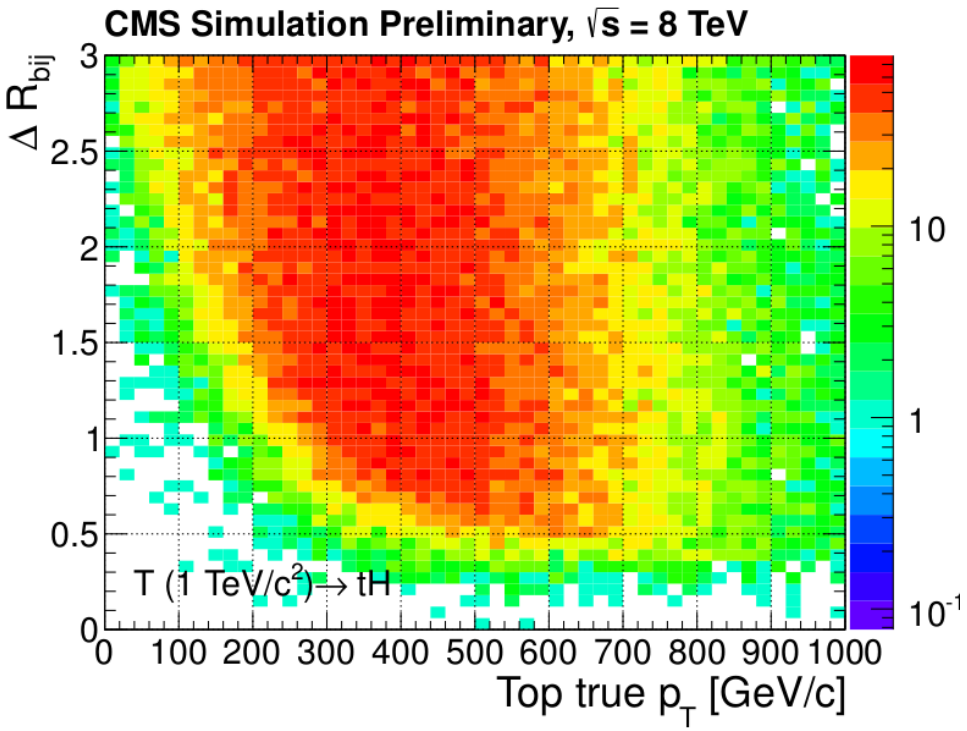
- Subjet flavor definition:
  - "Ghost" hadrons/partons clustered inside a fat jet later assigned to the closest subjet in rapidity-based ΔR

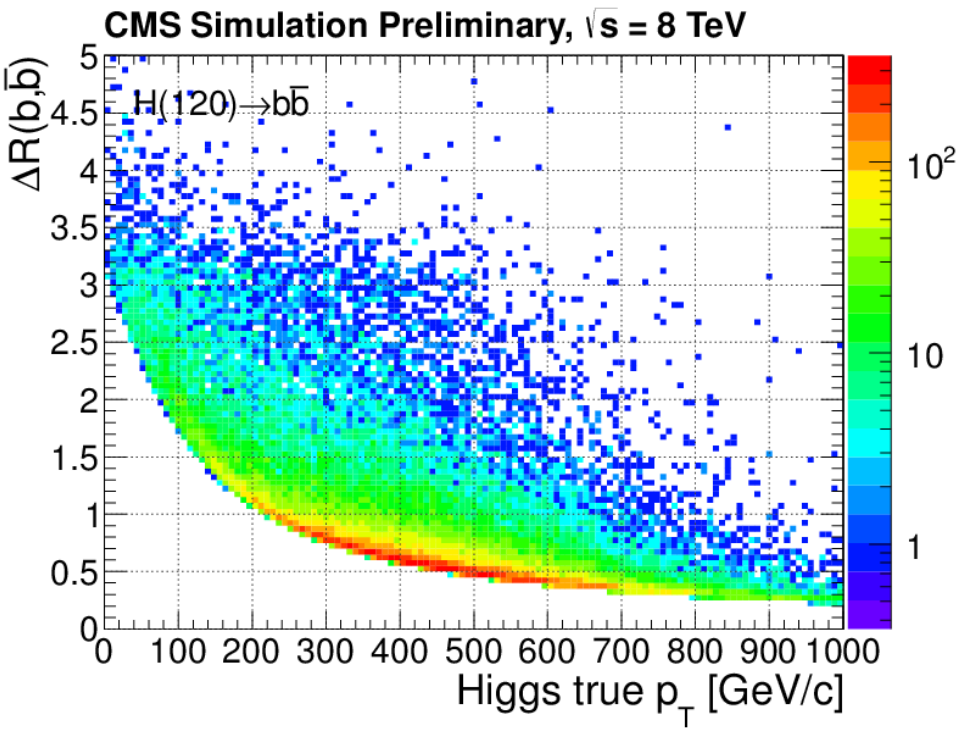


→ In order to assign subjet flavor, need external fat jet collections (to avoid flavor inconsistencies between subjets and fat jets)



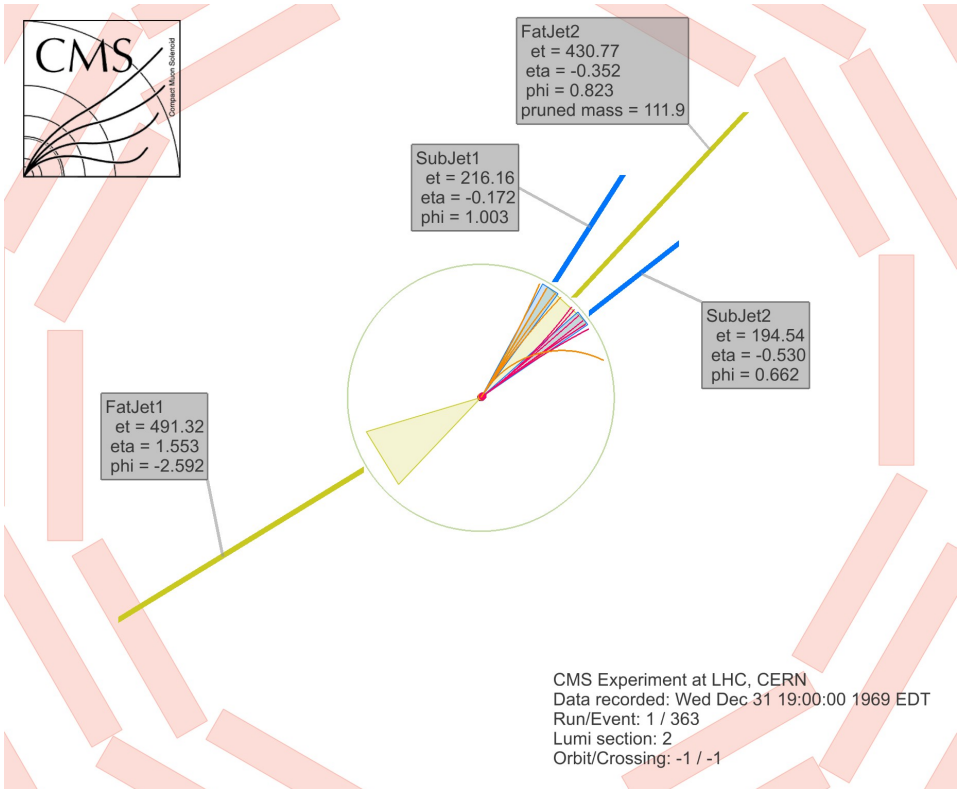
Boosted H→$b\bar{b}$ and hadronic top quarks

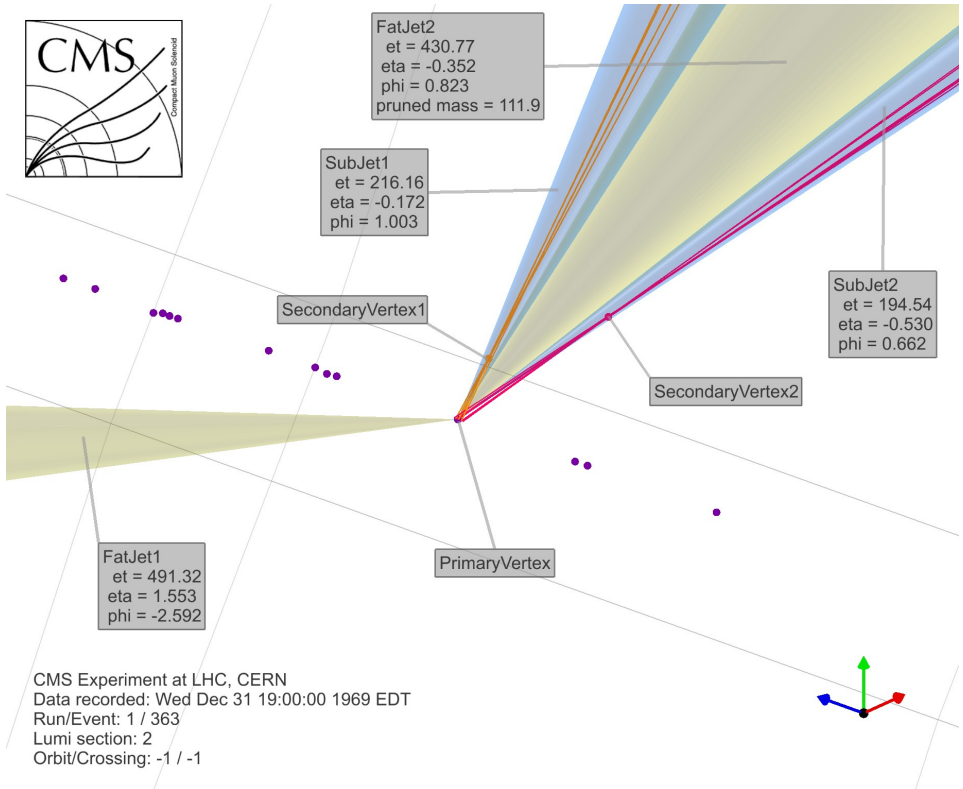




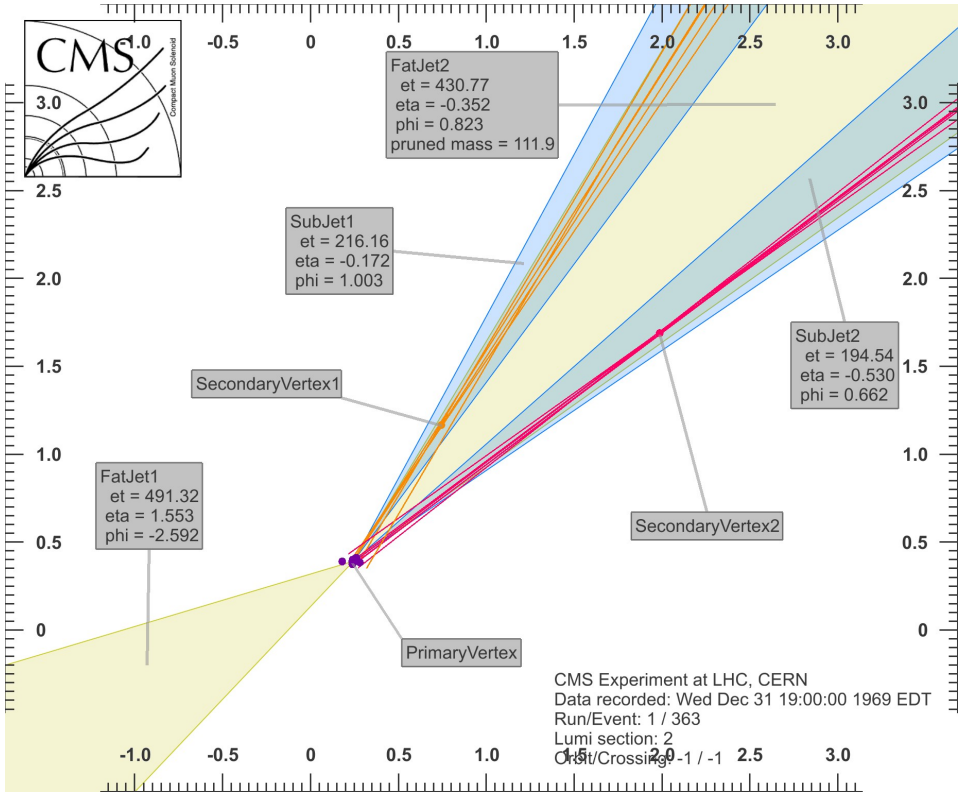
$$\Delta R(\mathrm{b},\bar{\mathrm{b}}) \gtrsim \frac{2m}{p_{\mathrm{T}}}$$

Boosted H→$b\bar{b}$ (simulation)

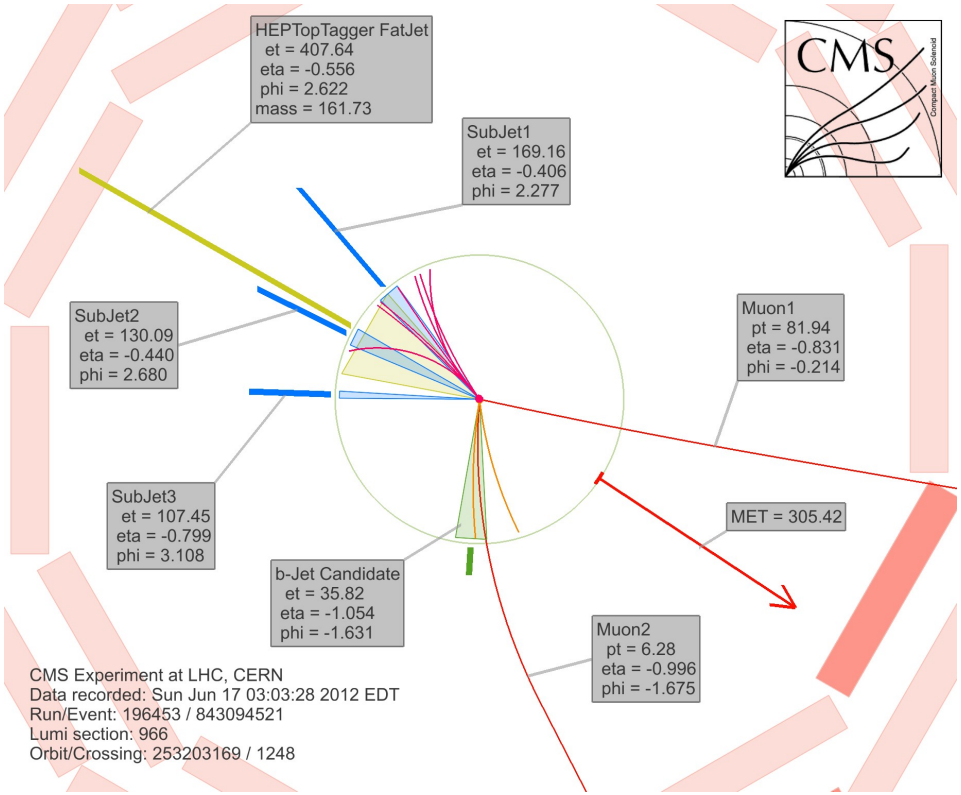


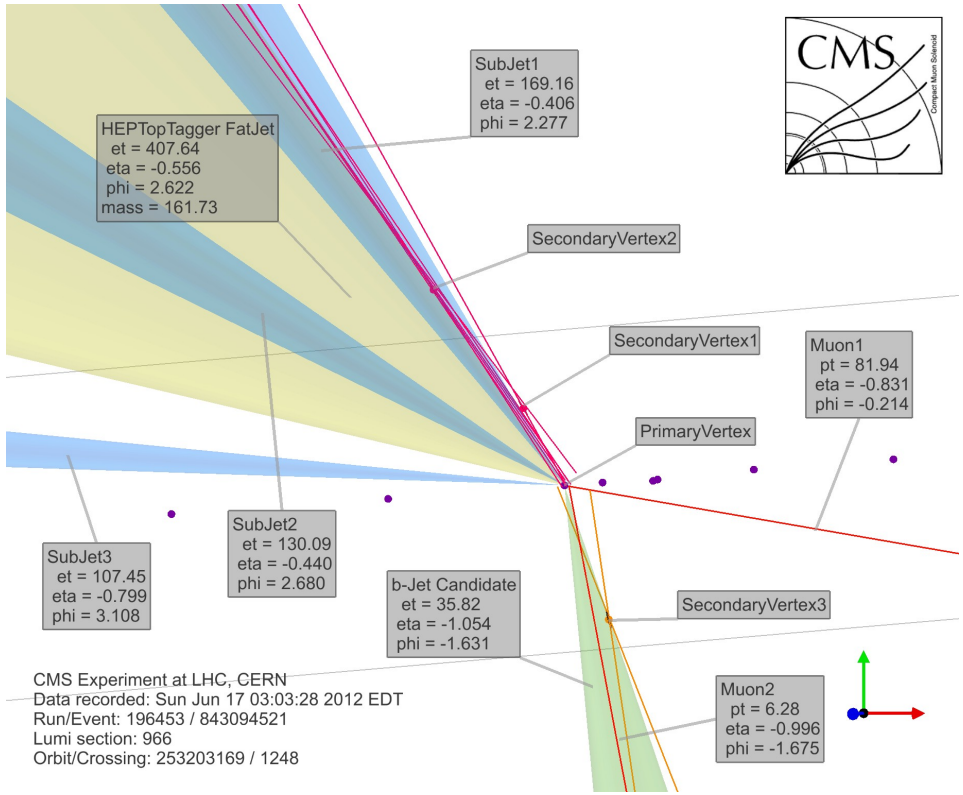


Boosted H→$b\bar{b}$ (simulation)

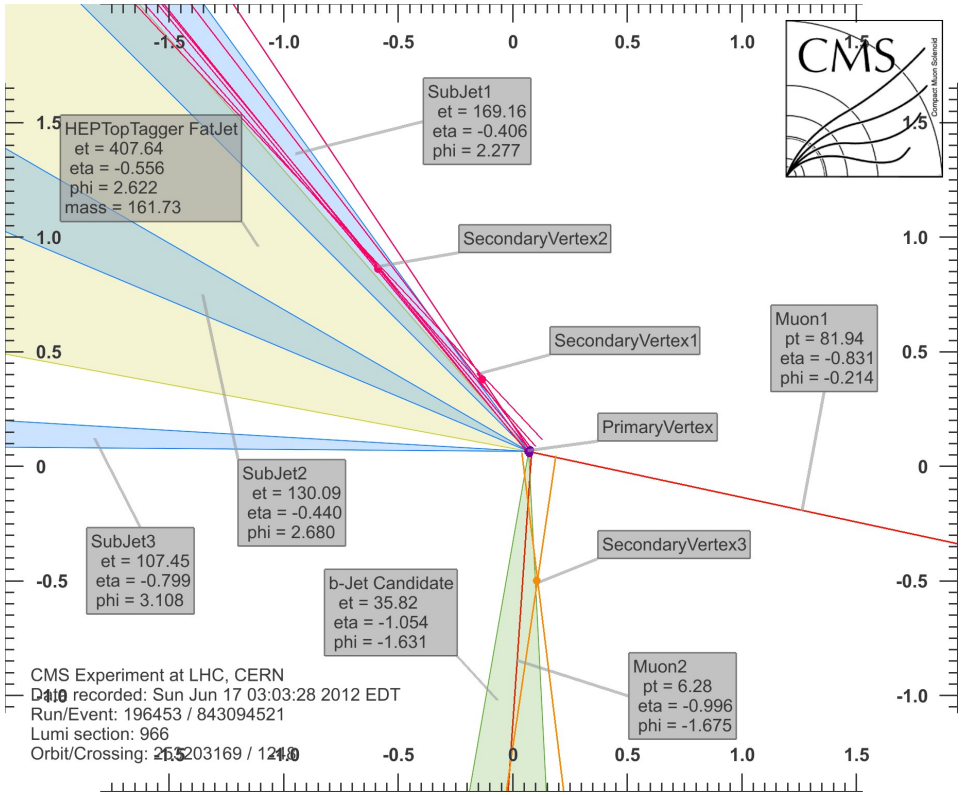


Boosted top candidate





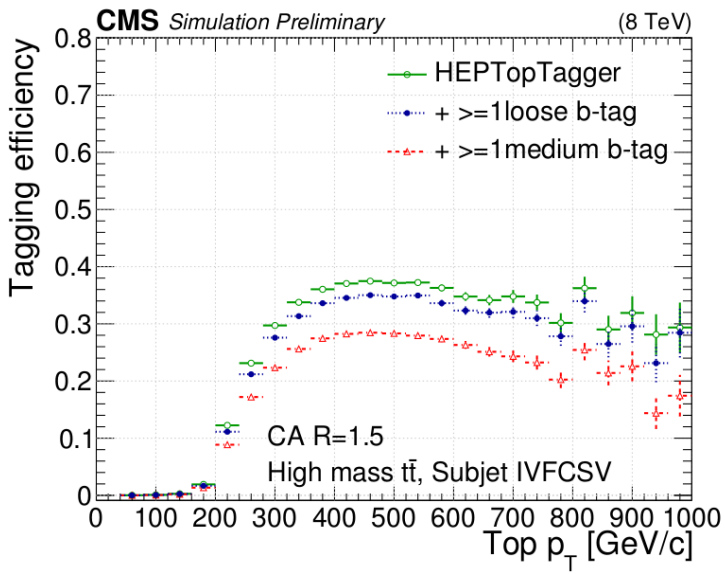
Boosted top candidate

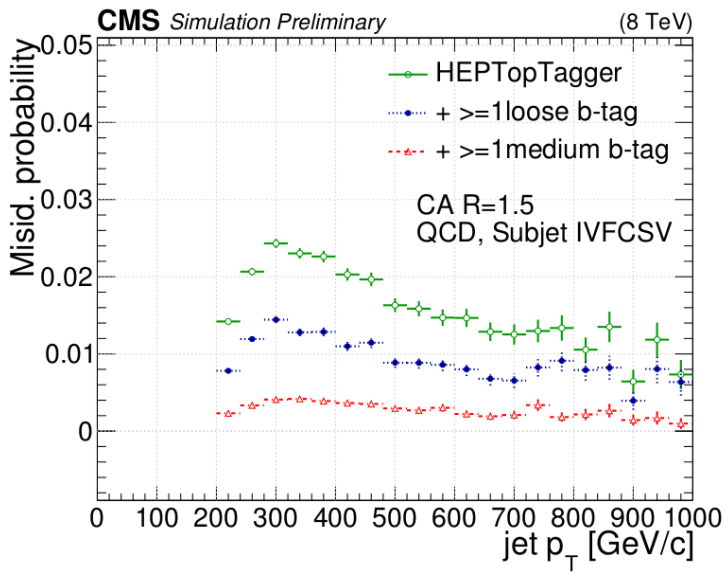


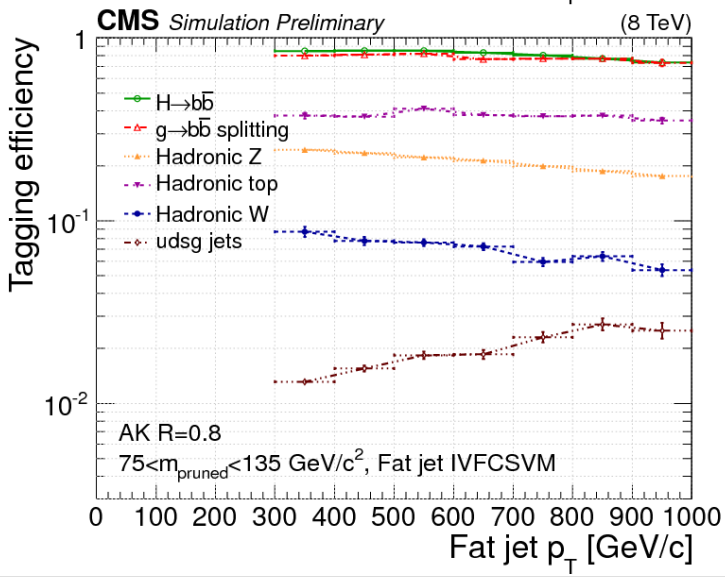
p_T dependence of boosted b tagging

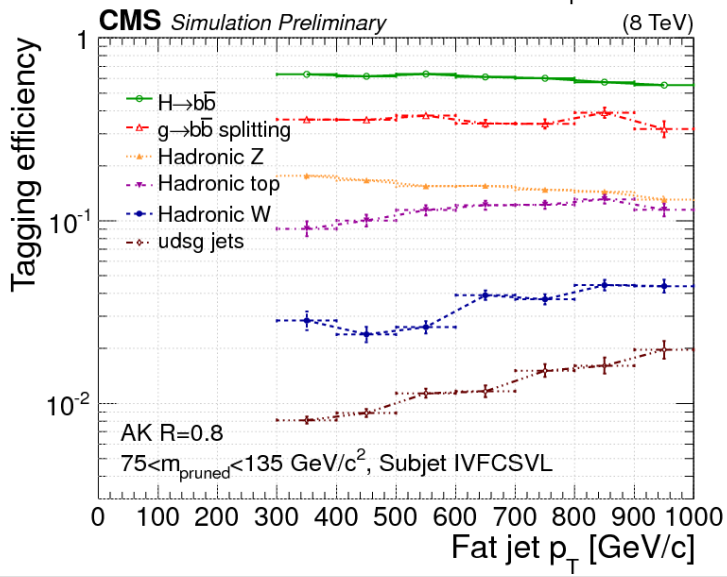
Boosted hadronic top

Boosted H→bb̄



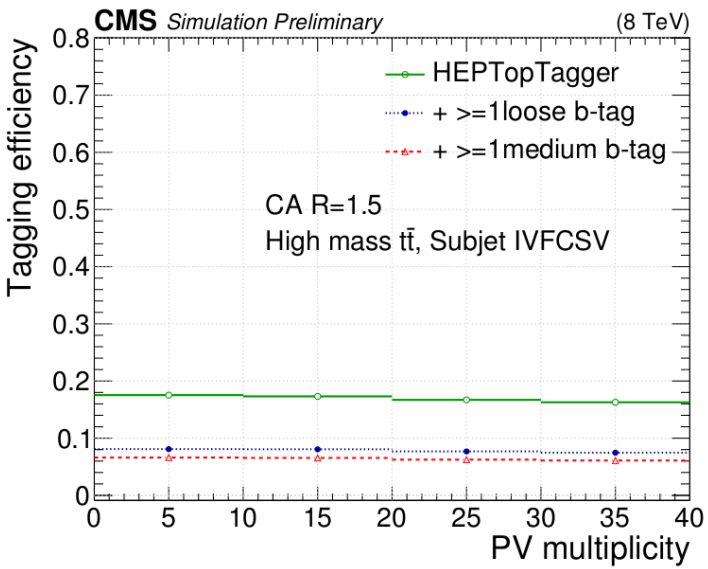


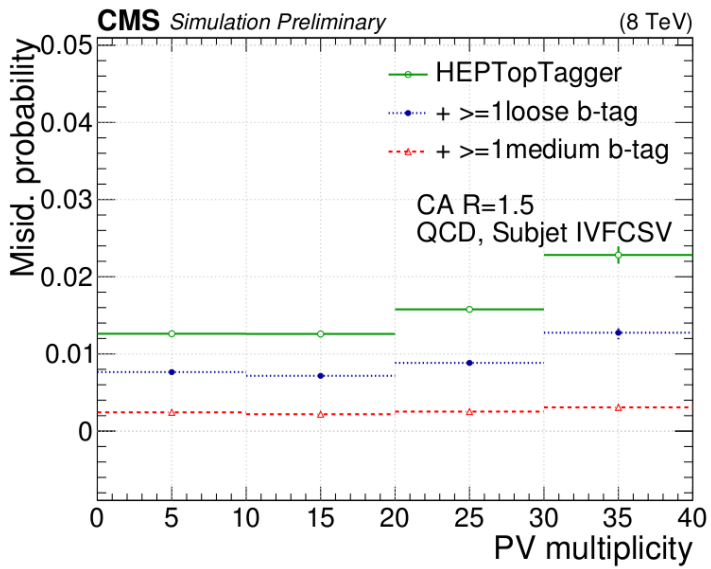




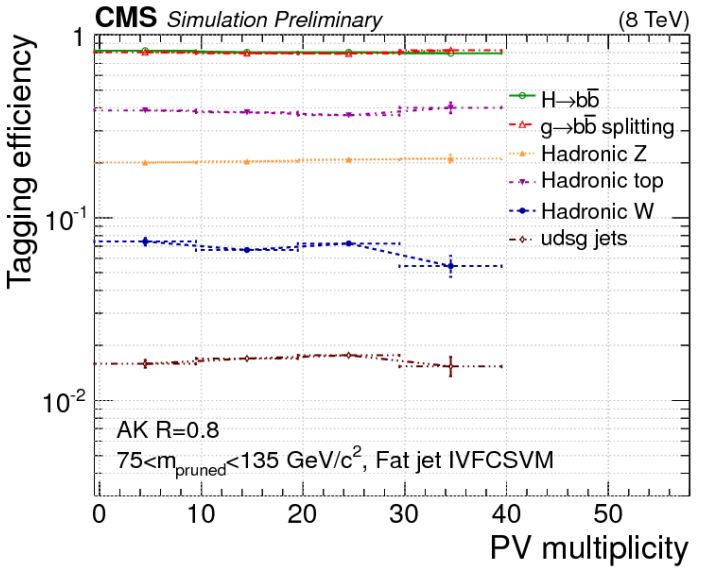
Pileup dependence of boosted b tagging

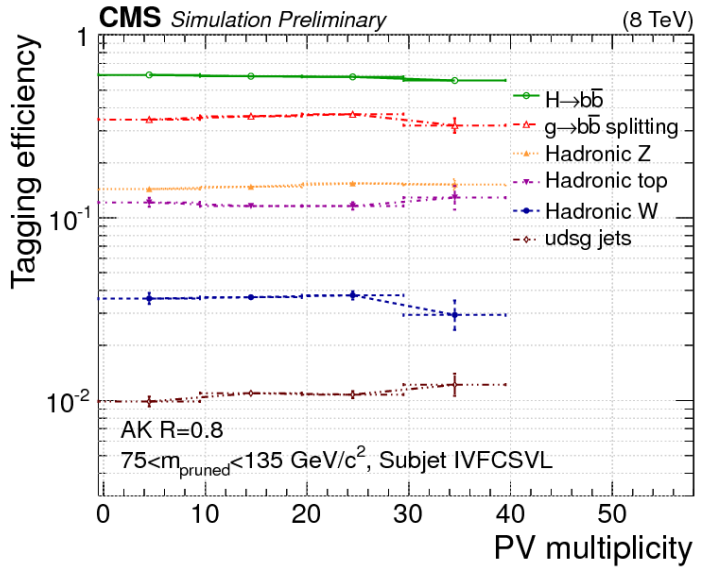
Boosted hadronic top



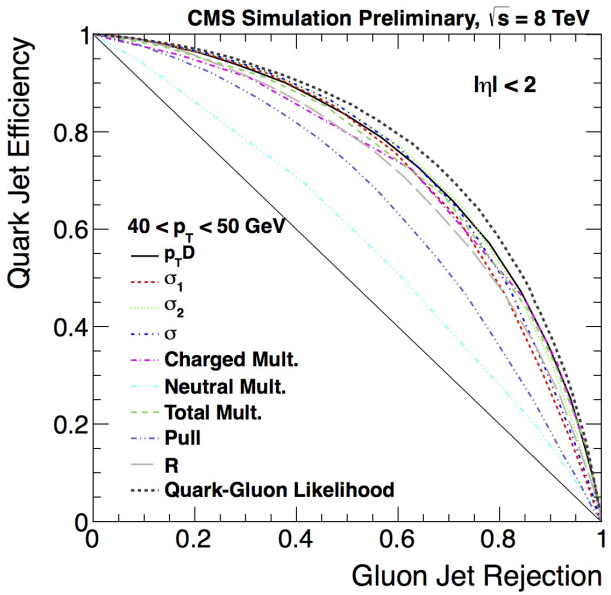


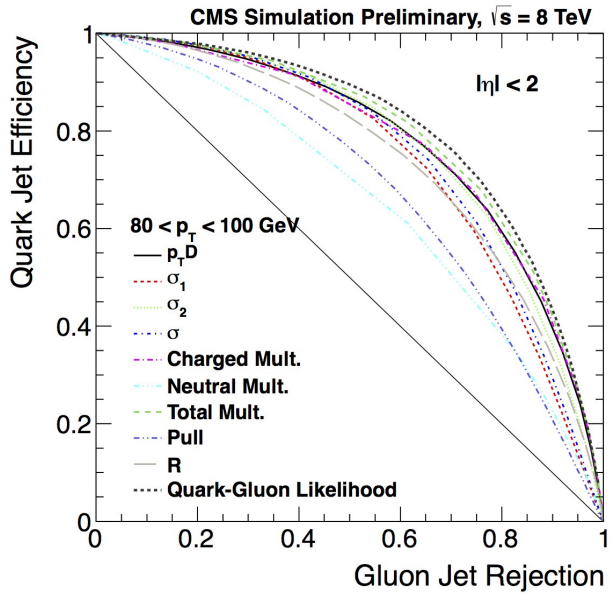
Boosted H→bb̄

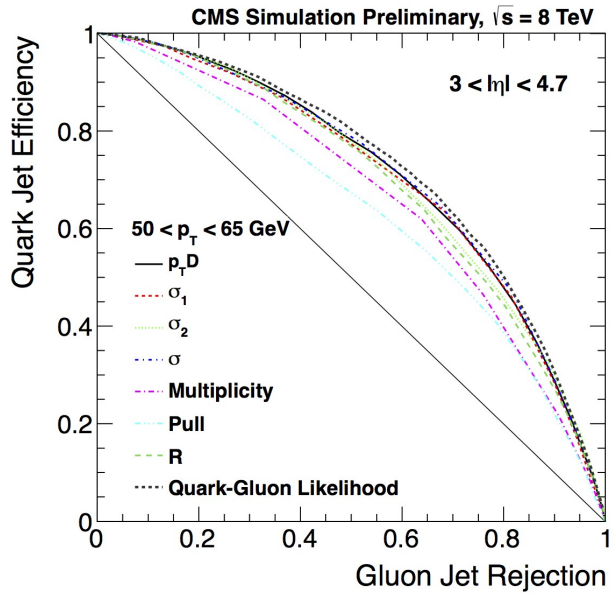




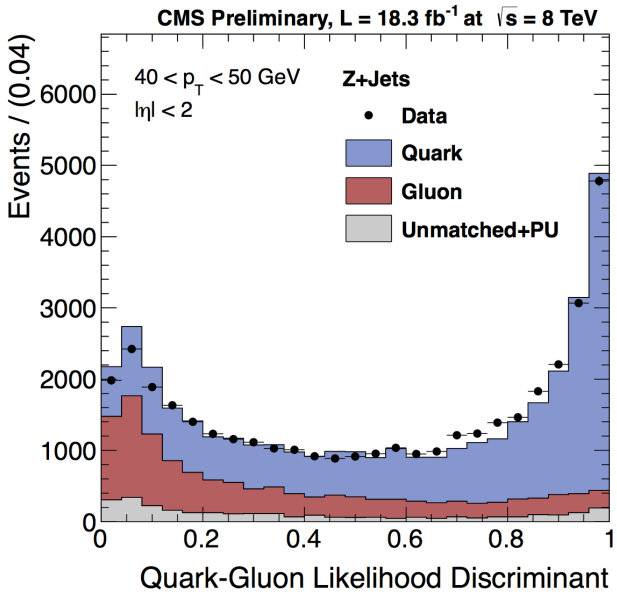
Discriminating power of quark/gluon variables

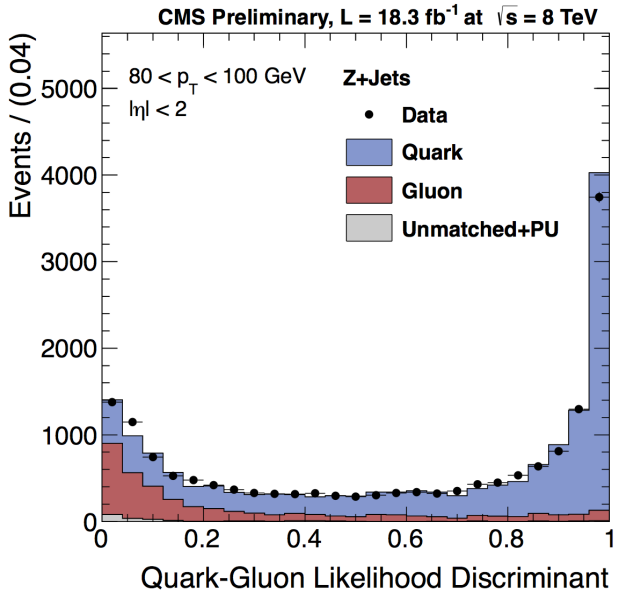


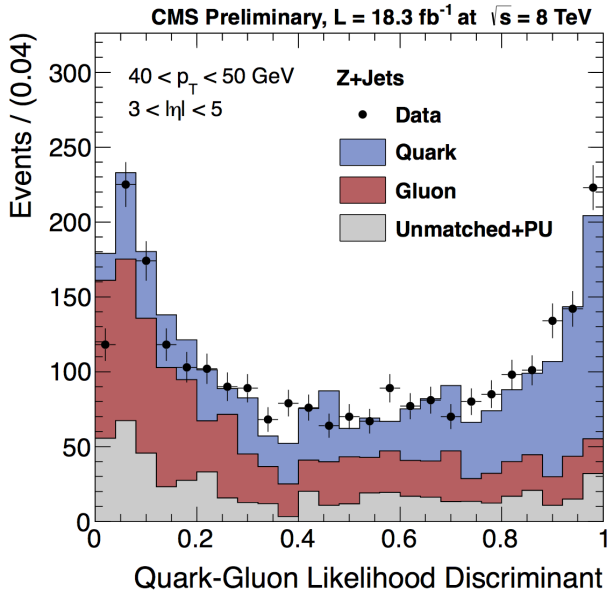




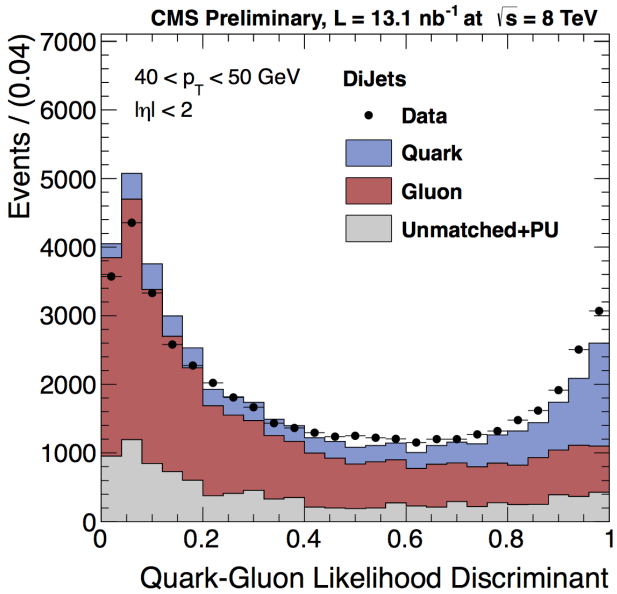
Data validation of quark/gluon tagging

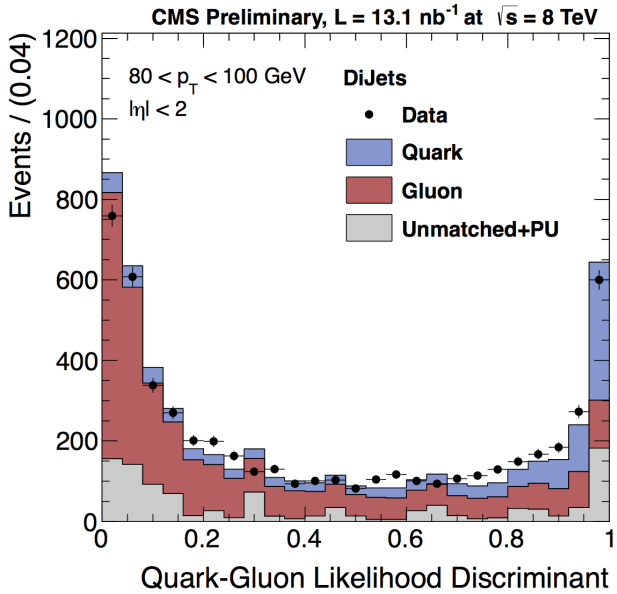


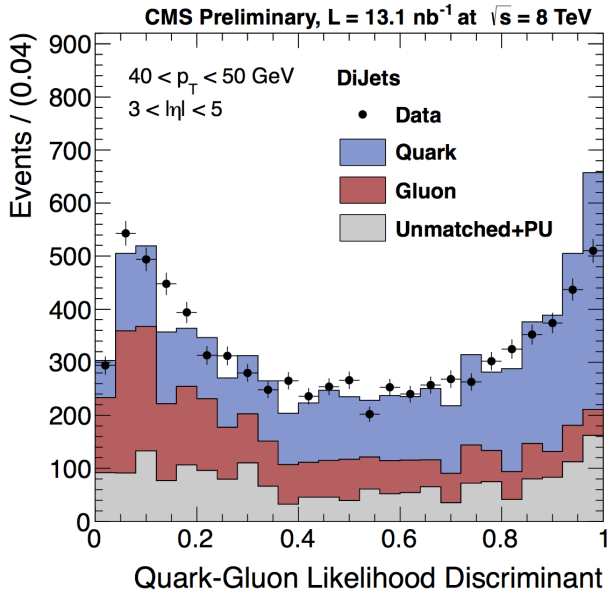




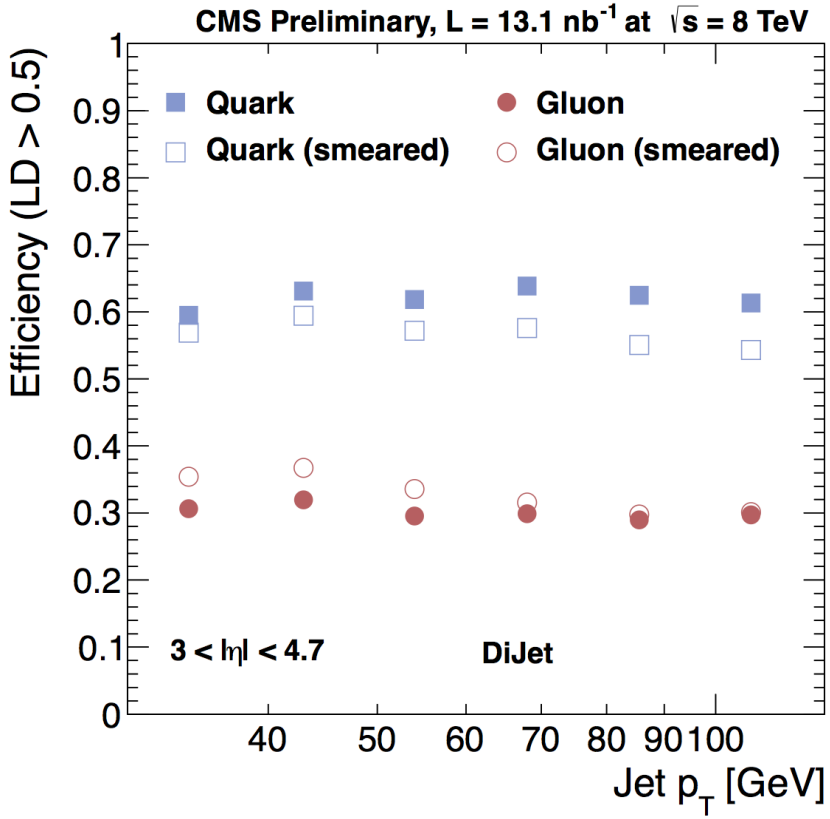
Data validation of quark/gluon tagging (cont'd)



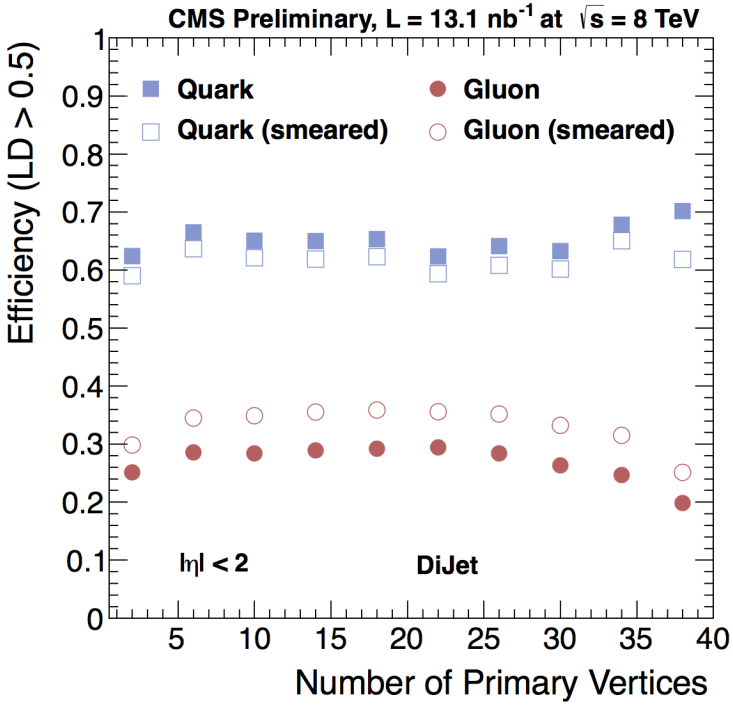


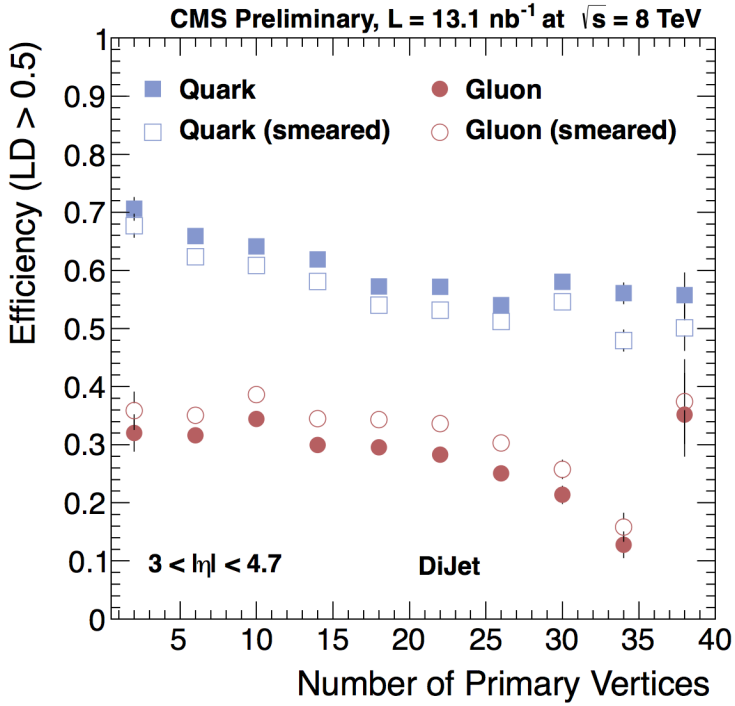


Quark/gluon tagging in forward region



PU dependence of quark/gluon tagging





High-p_T jet tracking