# Identification of b jets in boosted event topologies with CMS

Dinko Ferenček
*Ruđer Bošković Institute, Zagreb, Croatia*
on behalf of the CMS Collaboration

BOOST 2016
July 19, 2016
Zürich

# Introduction/motivation

- Many searches for new physics as well as SM measurements rely on accurate identification of jets originating from b quarks (*b tagging*)
  - SM top quark, SM Higgs boson, SUSY and many other BSM scenarios

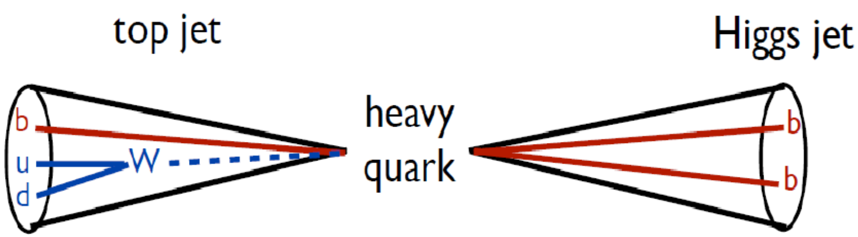
# Introduction/motivation

- Many searches for new physics as well as SM measurements rely on accurate identification of jets originating from b quarks (*b tagging*)
  - SM top quark, SM Higgs boson, SUSY and many other BSM scenarios
- What about boosted event topologies?

# Introduction/motivation

- Many searches for new physics as well as SM measurements rely on accurate identification of jets originating from b quarks (*b tagging*)
  - SM top quark, SM Higgs boson, SUSY and many other BSM scenarios
- What about boosted event topologies?

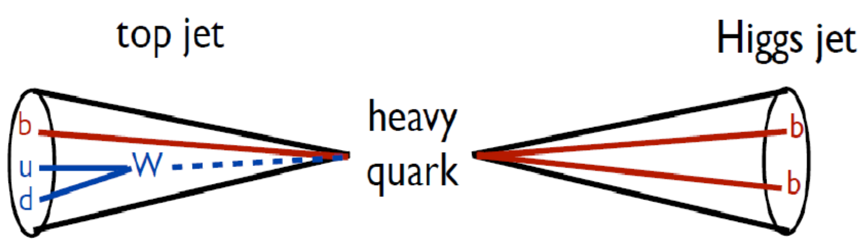
$Br(t \rightarrow bW) \approx 100\%$ $Br(H(125) \rightarrow b\bar{b}) \approx 58\%$



# Introduction/motivation

- Many searches for new physics as well as SM measurements rely on accurate identification of jets originating from b quarks (*b tagging*)
  - SM top quark, SM Higgs boson, SUSY and many other BSM scenarios
- What about boosted event topologies?

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

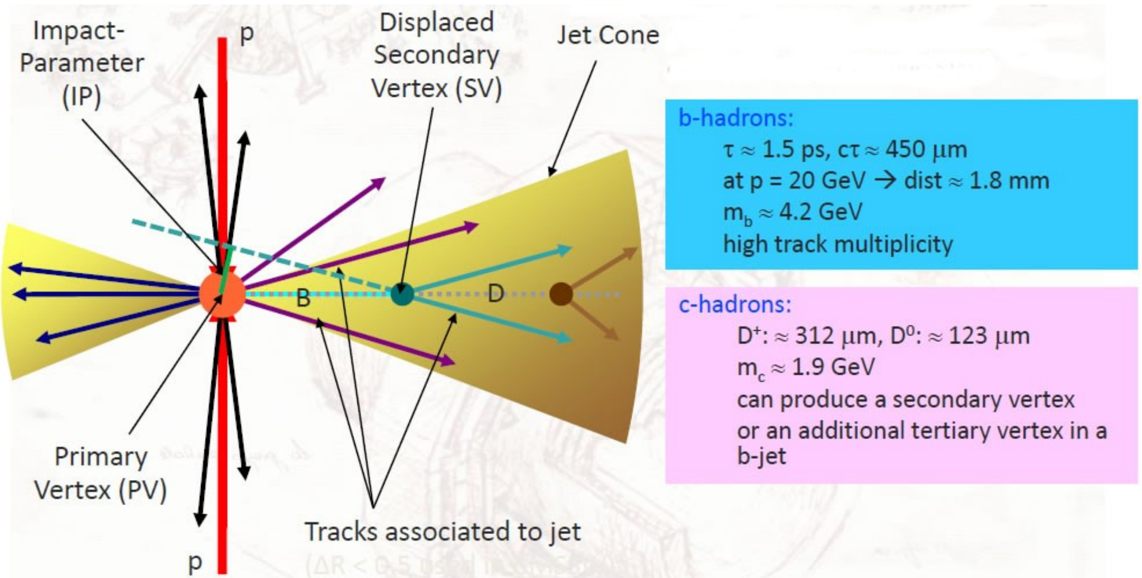



- *b tagging, using information largely complementary to the jet substructure, can greatly improve the sensitivity of tagging algorithms for boosted objects*

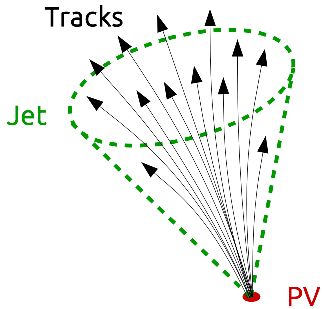
# b jets and b tagging

- b tagging tries to "determine" whether a jet originated from the hadronization of a b quark by looking for signatures of b hadrons inside the jet

- Exploits distinctive properties of b hadrons:
  - Long lifetime
  - Large mass
  - Decays with high track multiplicities (~5 on average)
  - Relatively large semileptonic branching fraction (≈20% for electron or muons with cascade decays included)
  - Hard fragmentation function



# b tagging in CMS

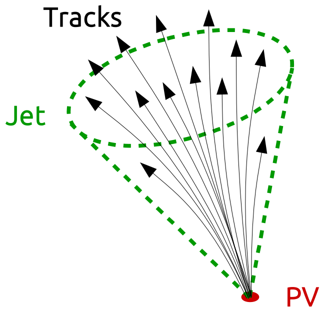



**Jet-track association**

Jet-track association (JTA):

- Cone-based JTA, e.g. ΔR(track,jet)<0.3

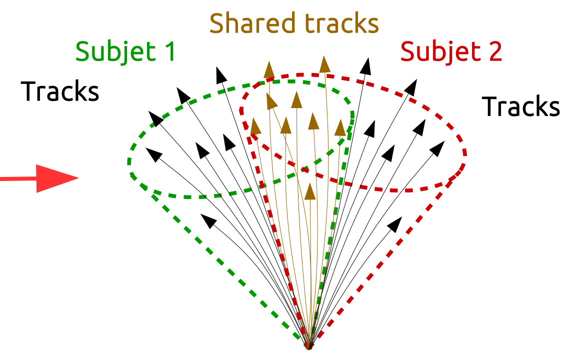
# b tagging in CMS



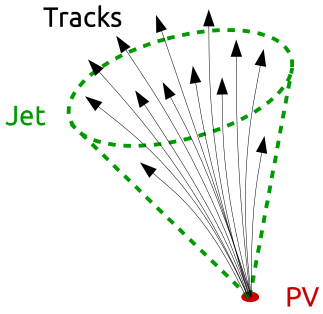

**Jet-track association**

## Jet-track association (JTA):

- Cone-based JTA, e.g. ΔR(track,jet)<0.3



# b tagging in CMS

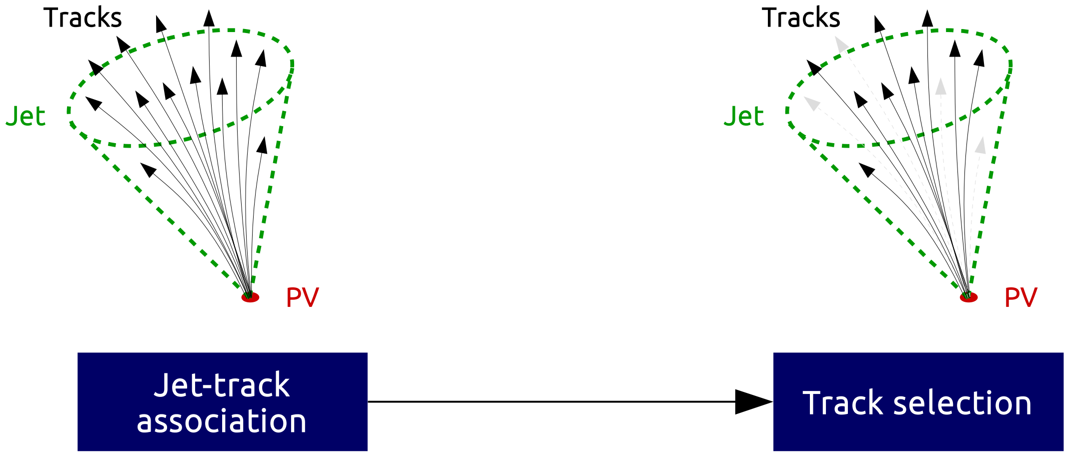



Jet-track association

Jet-track association (JTA):

- Cone-based JTA, e.g. ΔR(track,jet)<0.3
- Explicit JTA based on tracks linked to charged constituents of PF jets (default for subjets)

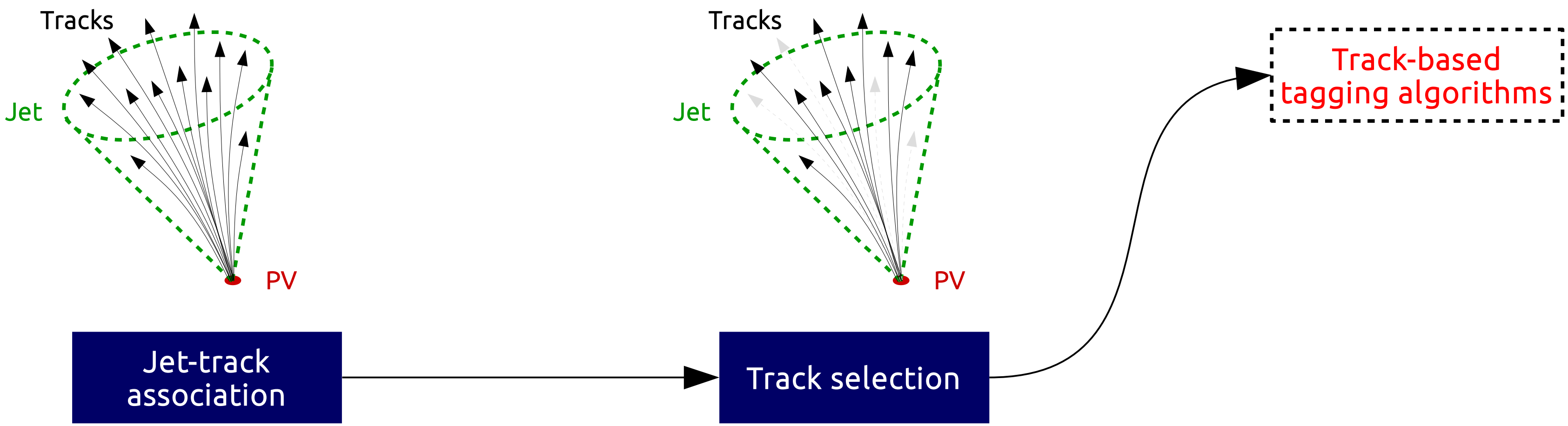
# b tagging in CMS



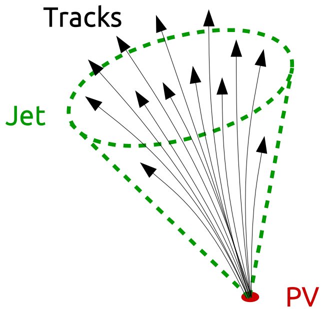

Track selection:

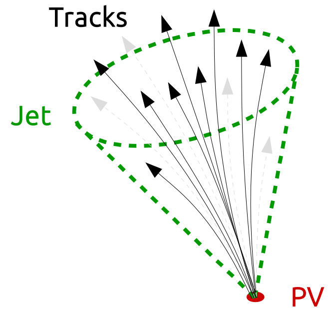
- Consists of applying a set of track quality criteria

# b tagging in CMS



# b tagging in CMS



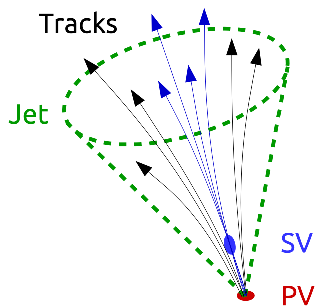




Jet-track association → Track selection → Track-based tagging algorithms

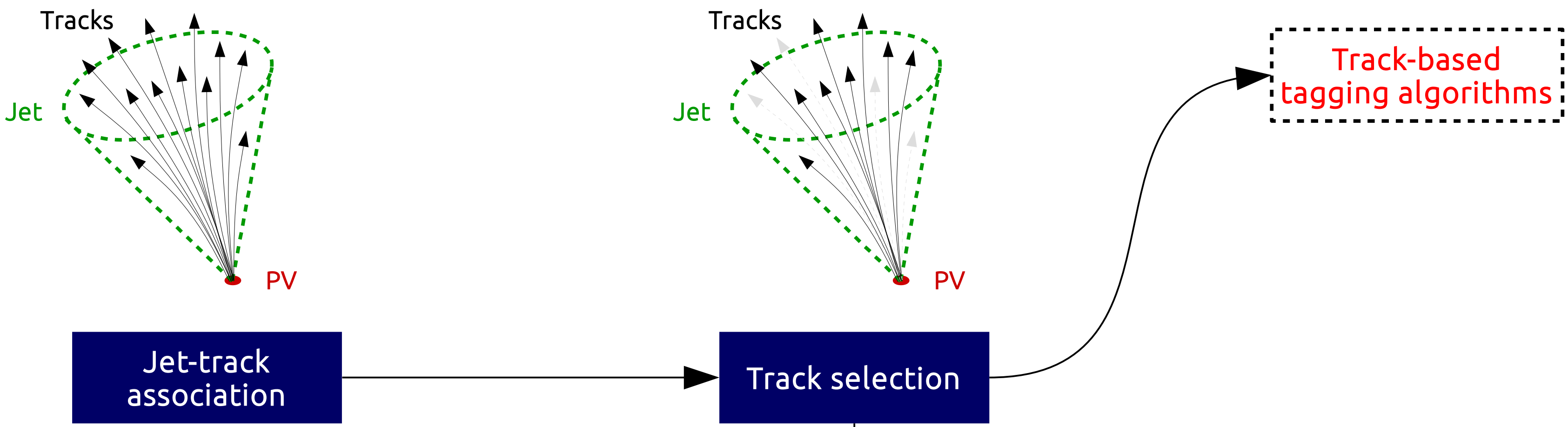
Track selection → Secondary vertex reconstruction

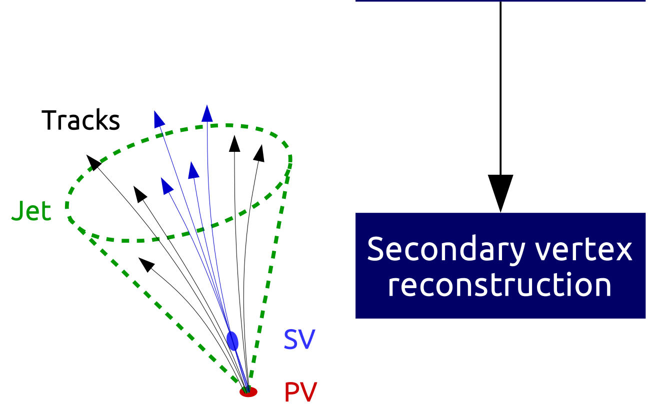
SV reconstruction:

- Adaptive vertex reconstruction (**AVR**) from tracks associated to the jet



# b tagging in CMS

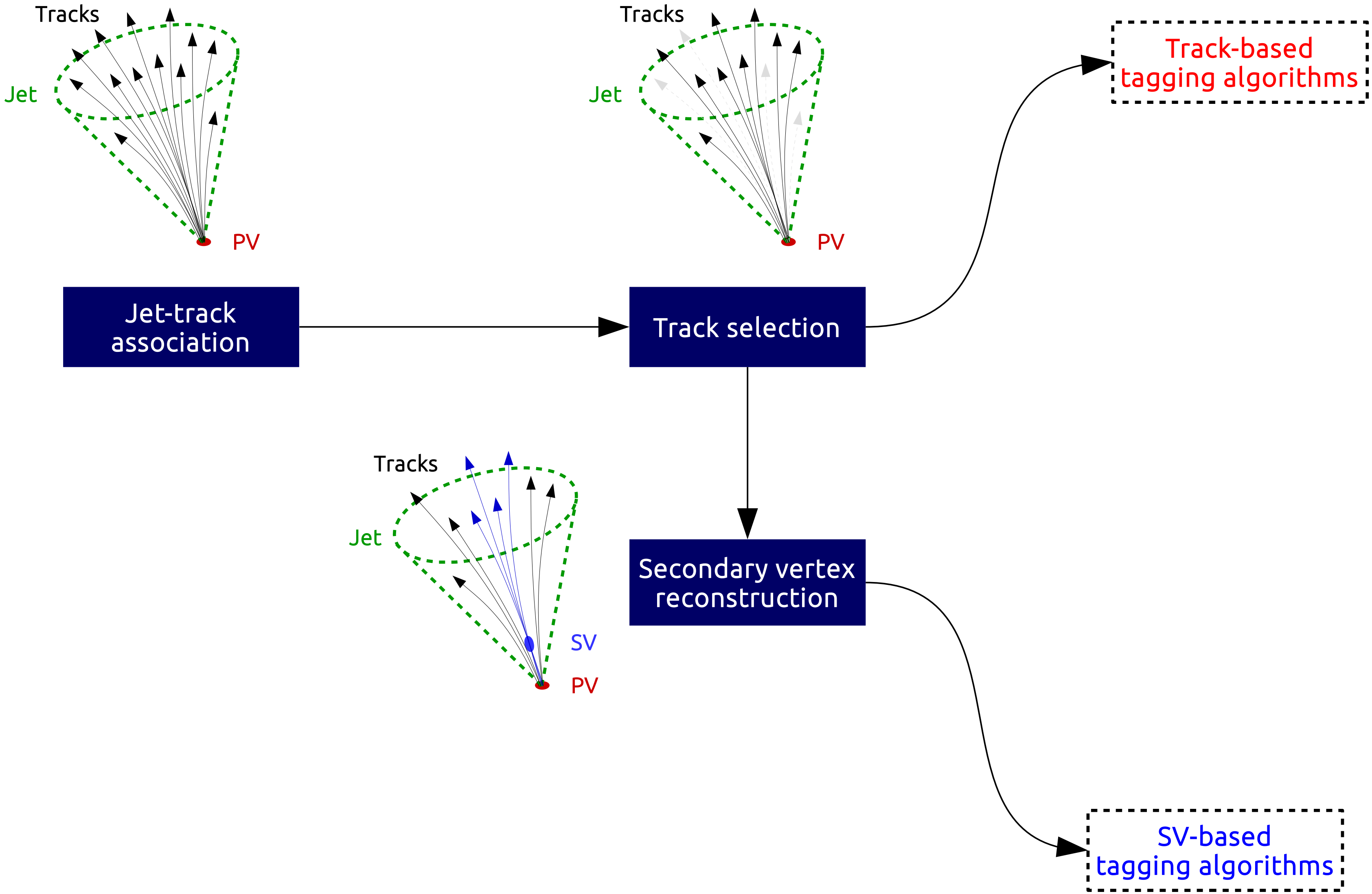


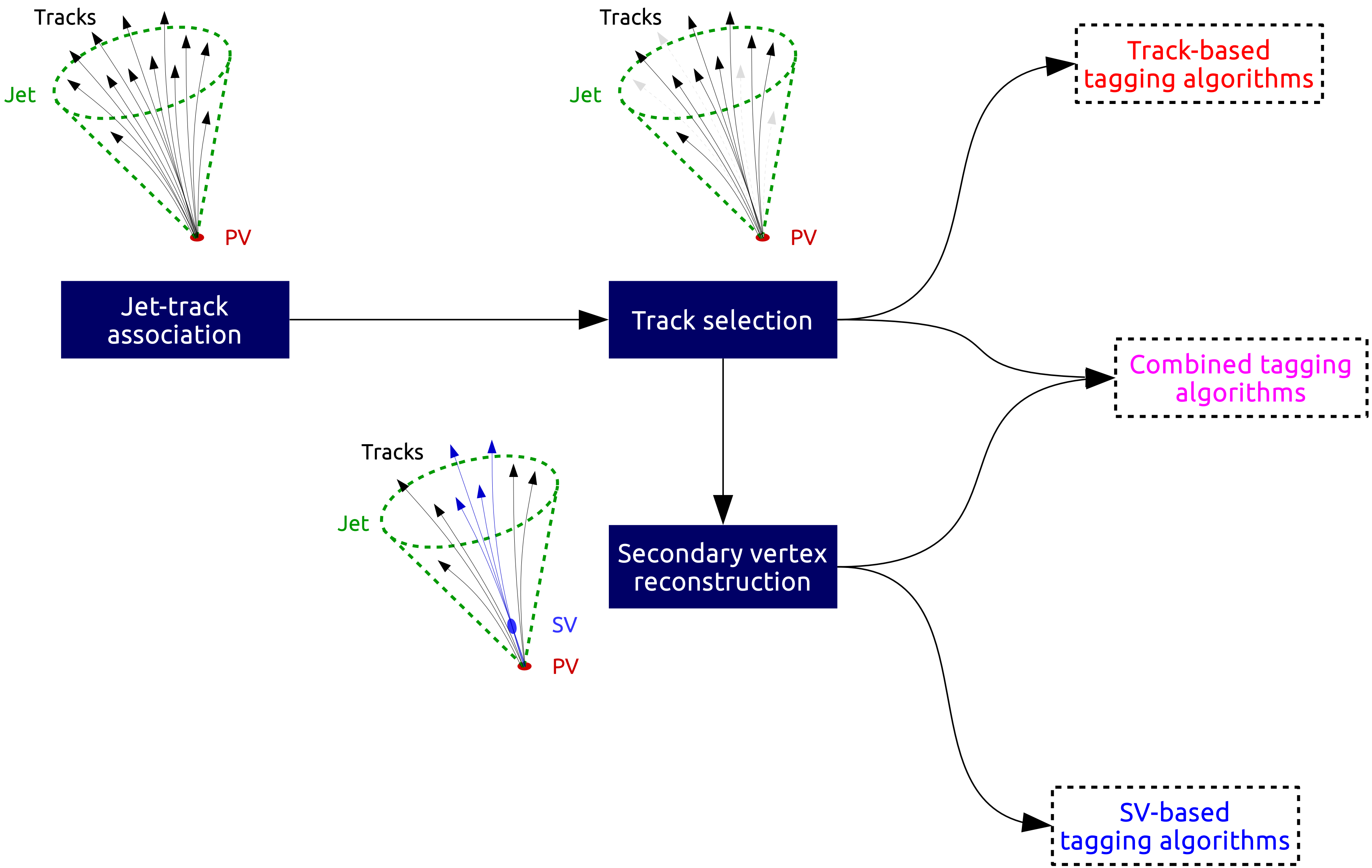

SV reconstruction:

- Adaptive vertex reconstruction (**AVR**) from tracks associated to the jet
- Inclusive Vertex Finder (**IVF**) secondary vertices reconstructed from all tracks independently from jets (current default)

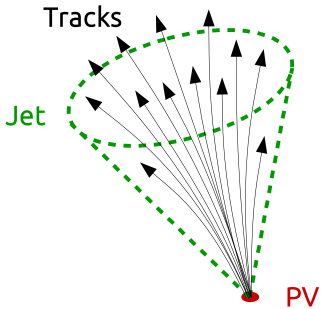
# b tagging in CMS



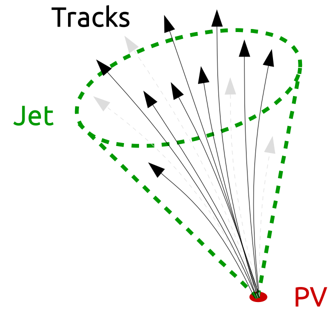
# b tagging in CMS



# b tagging in CMS



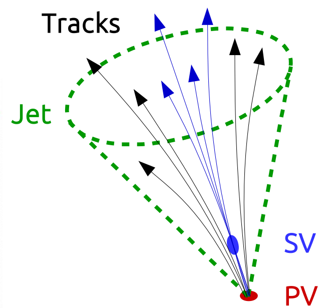

Jet-track association → Track selection




Track selection → Track-based tagging algorithms

Track selection → Combined tagging algorithms

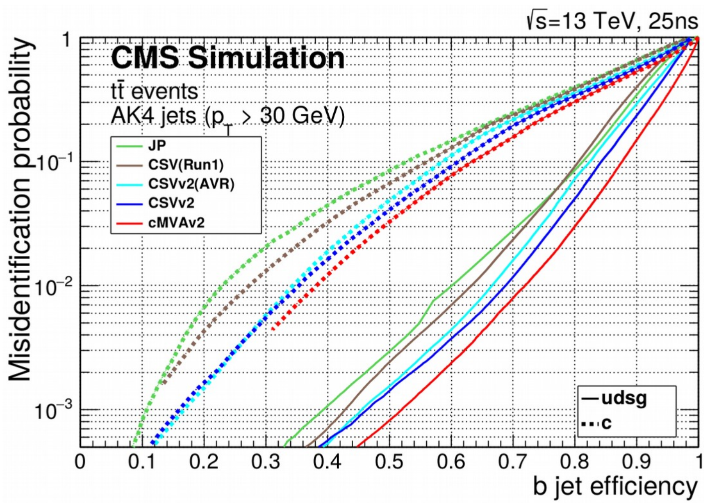
Track selection → Secondary vertex reconstruction




Secondary vertex reconstruction → Combined tagging algorithms

Secondary vertex reconstruction → SV-based tagging algorithms

e.g. **CSVv2** = **v2** of the **C**ombined **S**econdary **V**ertex algorithm



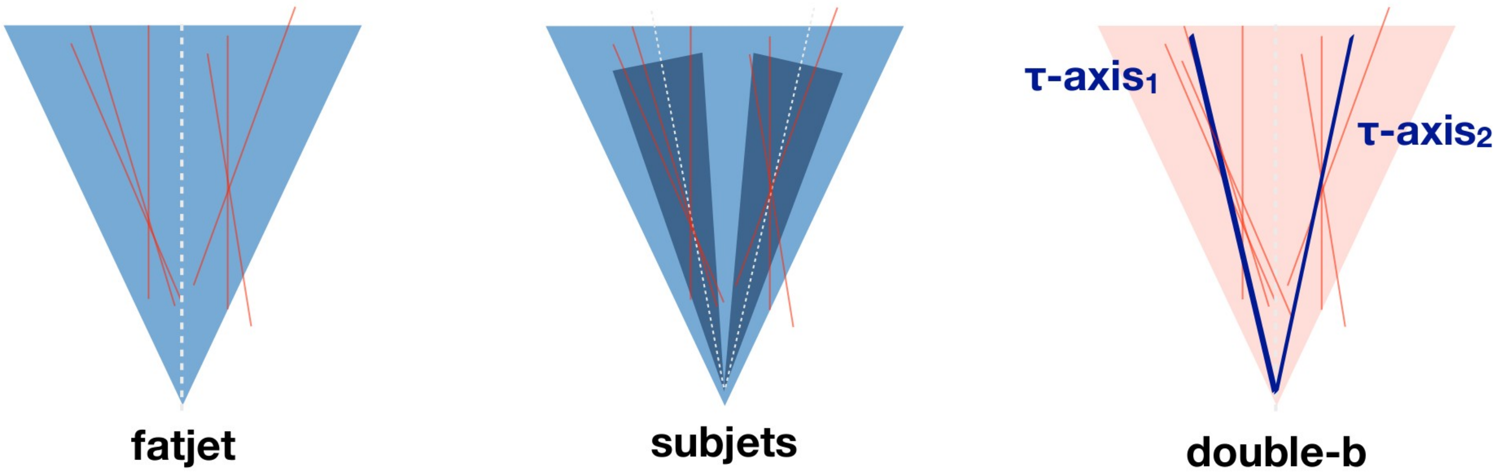

**Tagger operating points:**

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

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

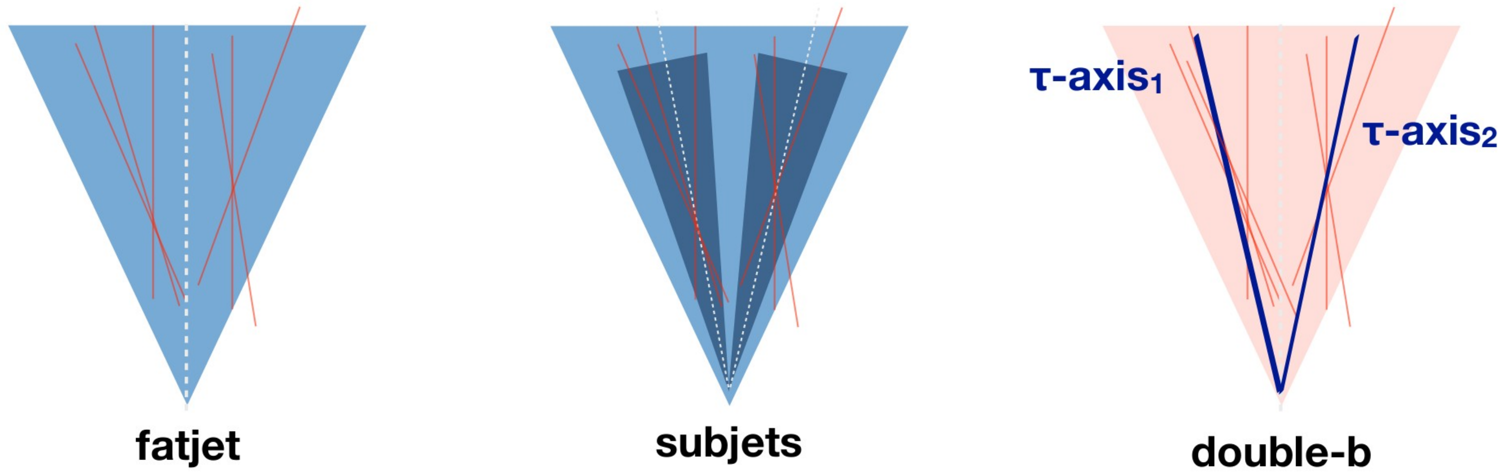
T = tight (≈0.1% light-flavor mistag rate)

# Boosted b tagging in CMS



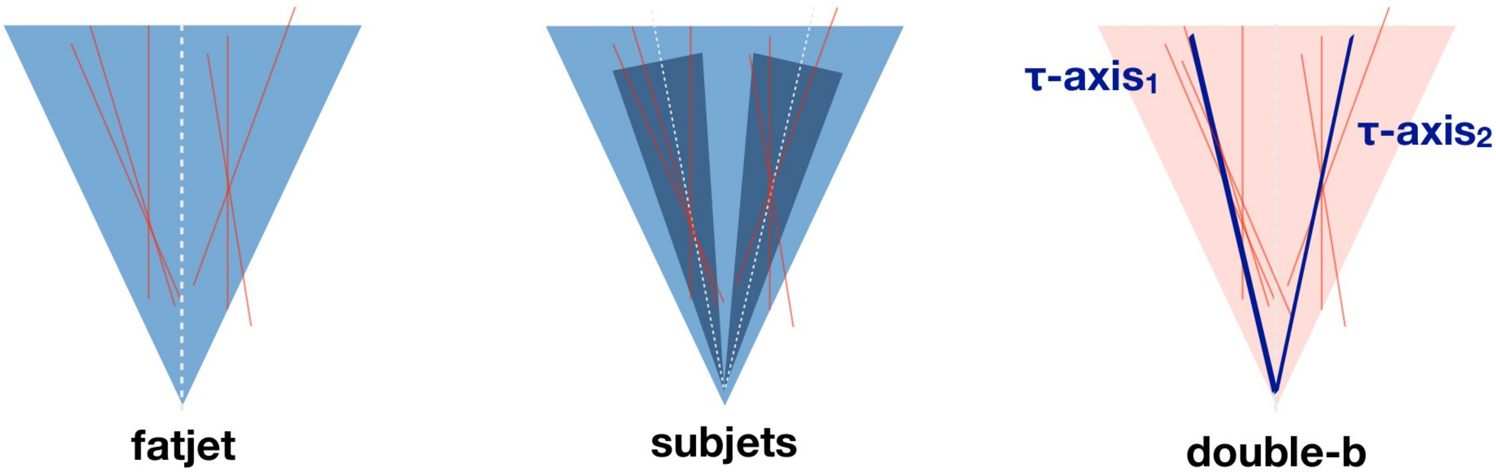

- 3 b tagging scenarios considered:
  - **Fat jet b tagging:** CSVv2 w/o dedicated retraining and w/ relaxed track and SV association criteria applied to a fat jet (anti-$k_T$ R=0.8)

# Boosted b tagging in CMS



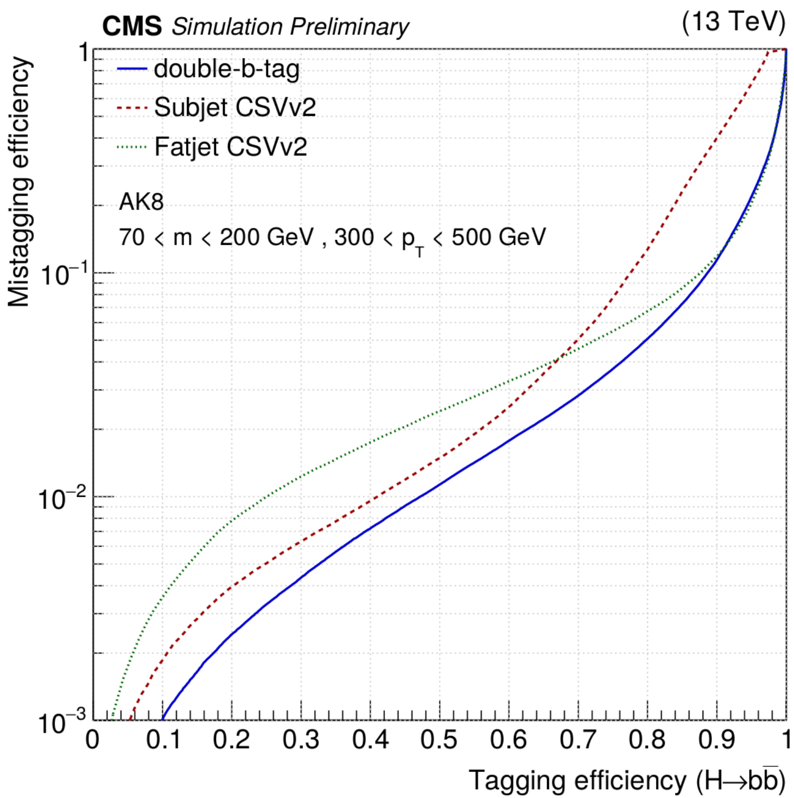

- 3 b tagging scenarios considered:
  - **Fat jet b tagging:** CSVv2 w/o dedicated retraining and w/ relaxed track and SV association criteria applied to a fat jet (anti-$k_T$ R=0.8)
  - **Subjet b tagging:** Standard CSVv2 w/ explicit JTA and ghost-associated SVs applied to subjets (soft drop, pruned,...) of a fat jet

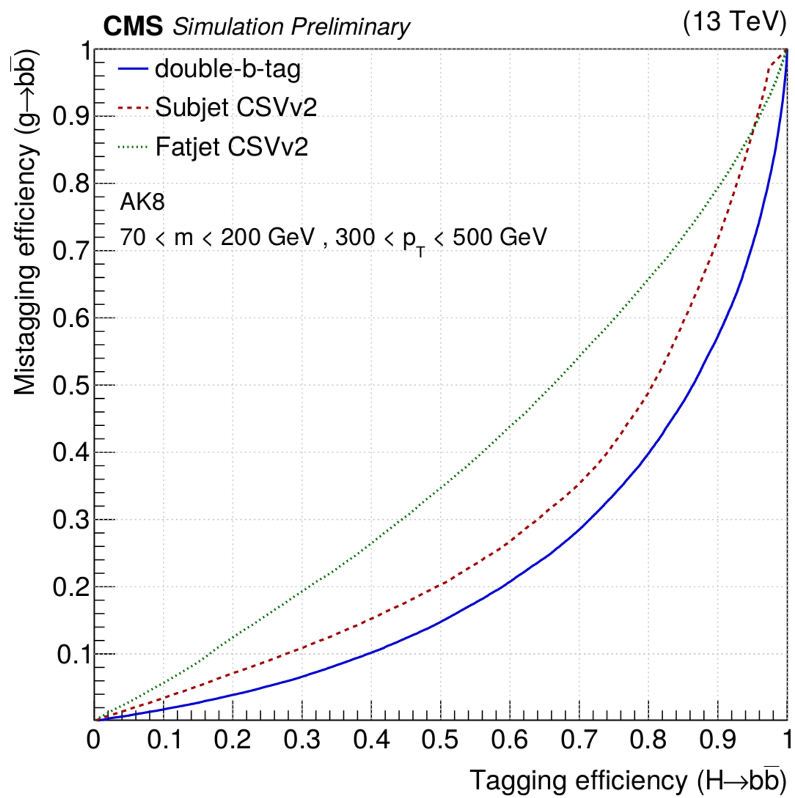
# Boosted b tagging in CMS




- 3 b tagging scenarios considered:
  - **Fat jet b tagging:** CSVv2 w/o dedicated retraining and w/ relaxed track and SV association criteria applied to a fat jet (anti-$k_T$ R=0.8)
  - **Subjet b tagging:** Standard CSVv2 w/ explicit JTA and ghost-associated SVs applied to subjets (soft drop, pruned,...) of a fat jet
  - **Double-b tagger:** Dedicated b tagging algorithm targeting boosted resonances decaying into a pair of b quarks (e.g. $H \rightarrow b\bar{b}$)

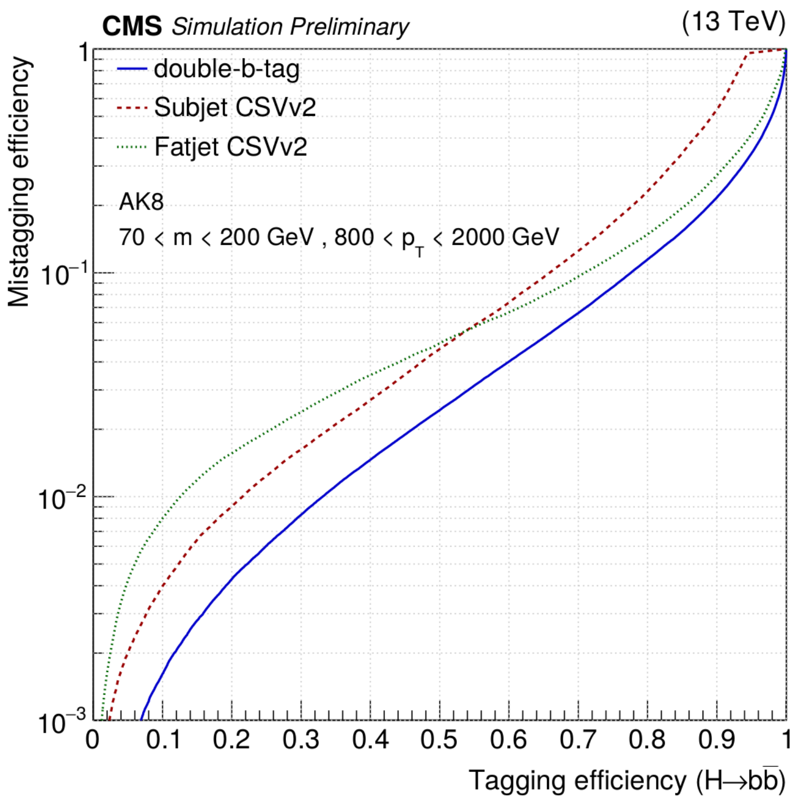
# Boosted b tagging performance

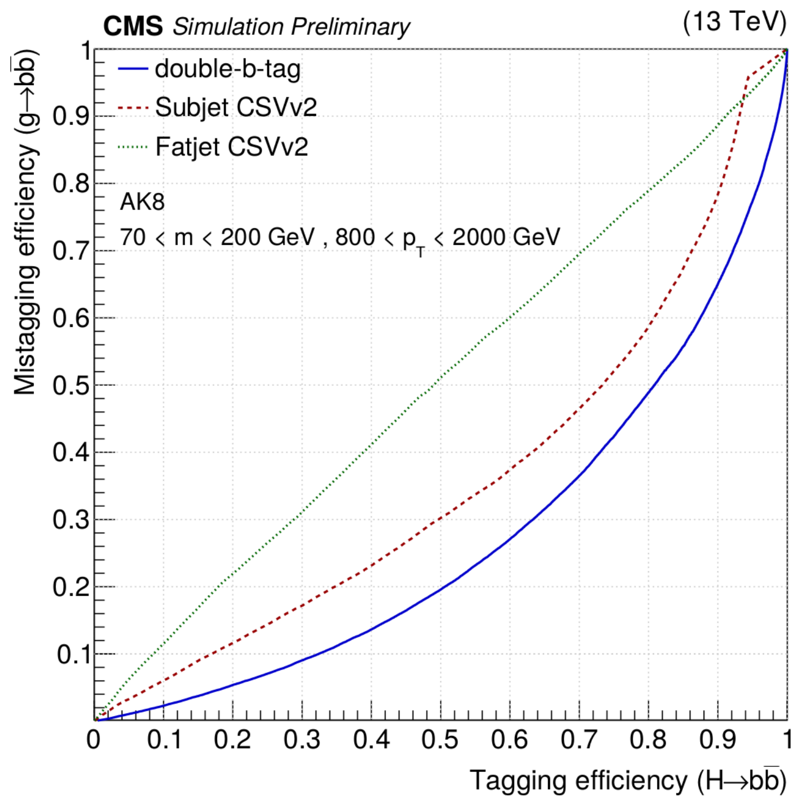



- Both subjets required to pass the same CSVv2 cut
- Fat jet b tagging not optimal for this topology → Poor discrimination against $g \rightarrow b\bar{b}$
- Fat jet and subjet b tagging complementary → Prompted development of a dedicated double-b tagger

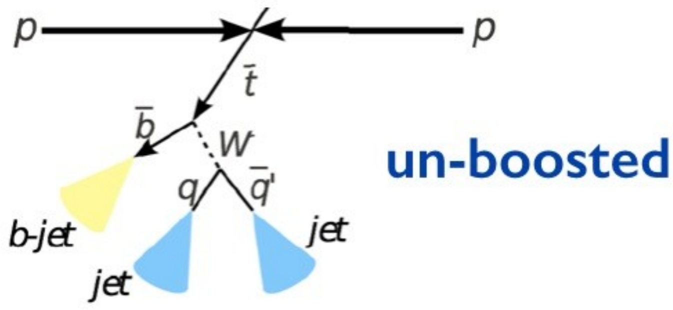
# Boosted b tagging performance (cont'd)

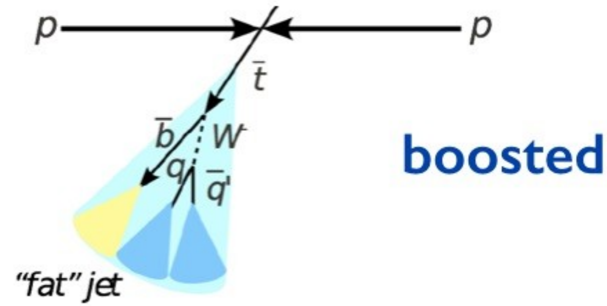


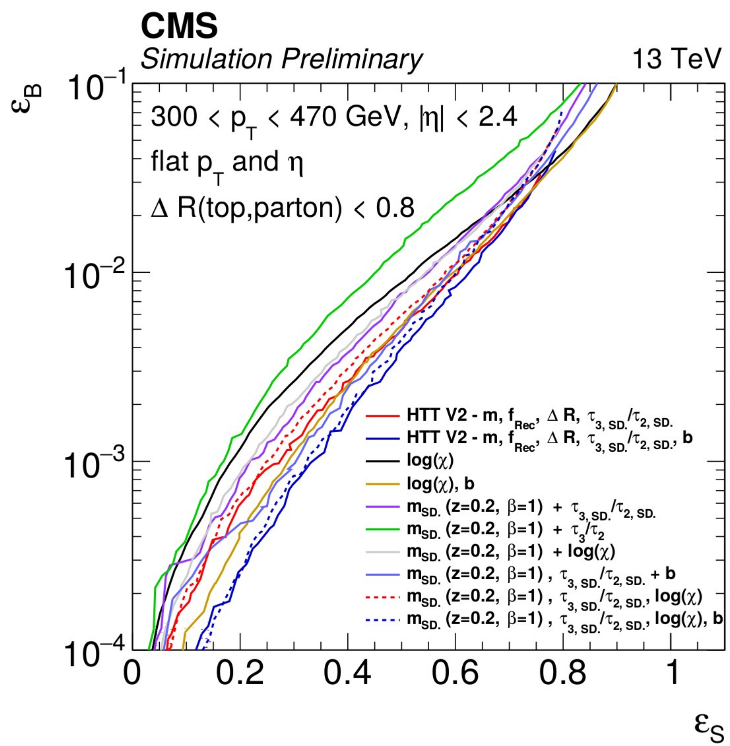

- Double-b tagger performance improvement wrt other approaches increases at high $p_T$

# Boosted b tagging performance (cont'd)




- Subjet b tagging an integral part of the top tagging algorithms commissioned in CMS [*]
  - Helps to improve the tagging performance




[*] http://cds.cern.ch/record/2126325/files/JME-15-002-pas.pdf

# Performance measurements in data

- Why performance measurements in data?

# Performance measurements in data

- Why performance measurements in data?
- Simulation does not perfectly reproduce b tagging performance in data (imperfect physics modeling, detector simulation,...) $\rightarrow$ Need to correct simulation by introducing and applying scale factors

$$\mathrm{SF} = \frac{\varepsilon_{\mathrm{DATA}}}{\varepsilon_{\mathrm{MC}}}$$

# Performance measurements in data

- Why performance measurements in data?
- Simulation does not perfectly reproduce b tagging performance in data (imperfect physics modeling, detector simulation,...) $\rightarrow$ Need to correct simulation by introducing and applying scale factors

$$\mathrm{SF} = \frac{\varepsilon_{\mathrm{DATA}}}{\varepsilon_{\mathrm{MC}}}$$

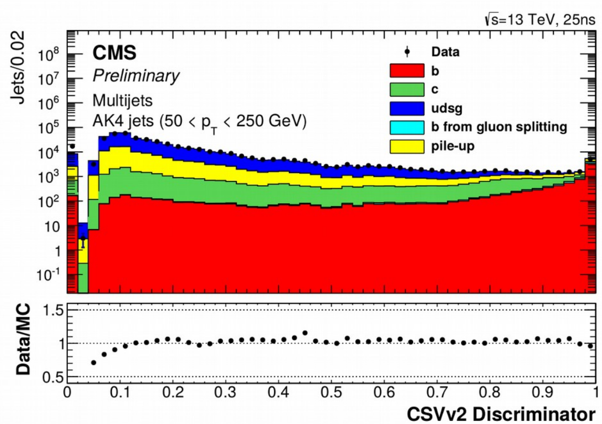
- Entire program of performance measurements using inclusive and muon-enriched multijet and $t\bar{t}$-enriched event samples defined

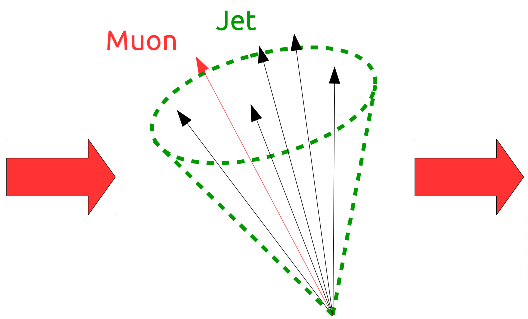
# Performance measurements in data

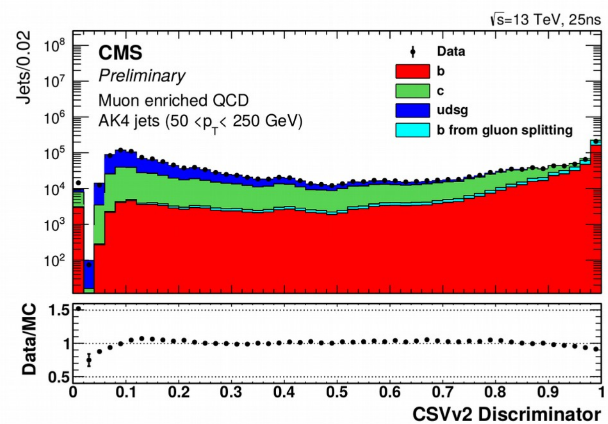
- Why performance measurements in data?
- Simulation does not perfectly reproduce b tagging performance in data (imperfect physics modeling, detector simulation,...) → Need to correct simulation by introducing and applying scale factors

$$\mathrm{SF} = \frac{\varepsilon_{\mathrm{DATA}}}{\varepsilon_{\mathrm{MC}}}$$

- Entire program of performance measurements using inclusive and muon-enriched multijet and $t\bar{t}$-enriched event samples defined
- Muon-enriched multijet sample obtained by muon-tagging jets



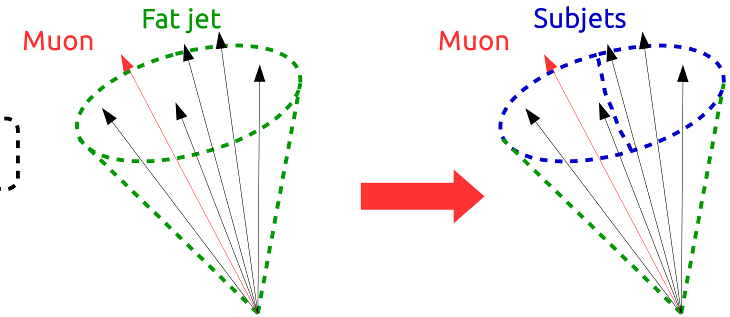



# Performance measurements in data (cont'd)

- **Subjet CSVv2:** Efficiency measured using muon-tagged subjets and mistag rate using subjets from an inclusive multijet sample

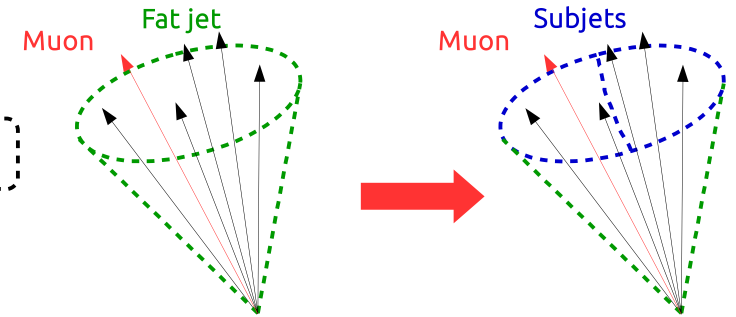
Muon-tagged subjets



# Performance measurements in data (cont'd)

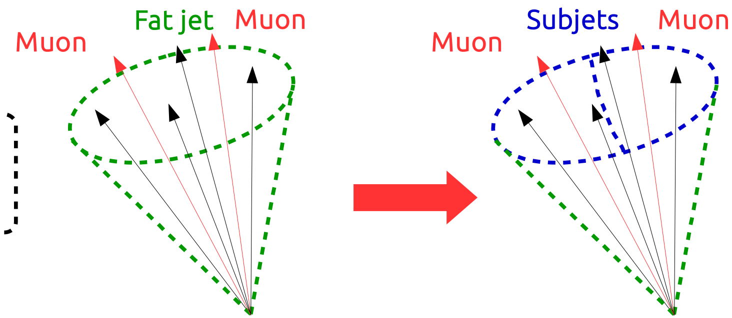
- **Subjet CSVv2:** Efficiency measured using muon-tagged subjets and mistag rate using subjets from an inclusive multijet sample

Muon-tagged subjets




- **Double-b tagger:** Efficiency measured using double-muon-tagged fat jets and mistag rate from a $t\bar{t}$-enriched sample

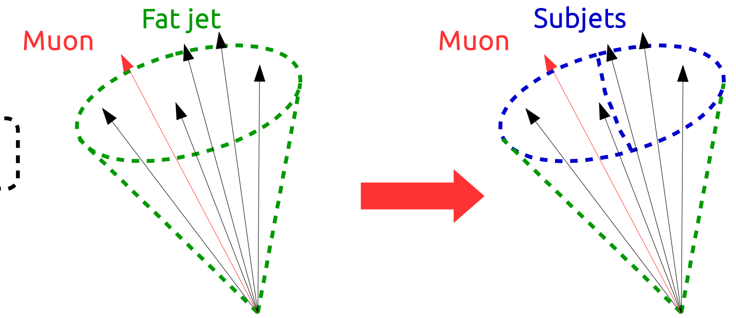
Double-muon-tagged fat jets



# Performance measurements in data (cont'd)

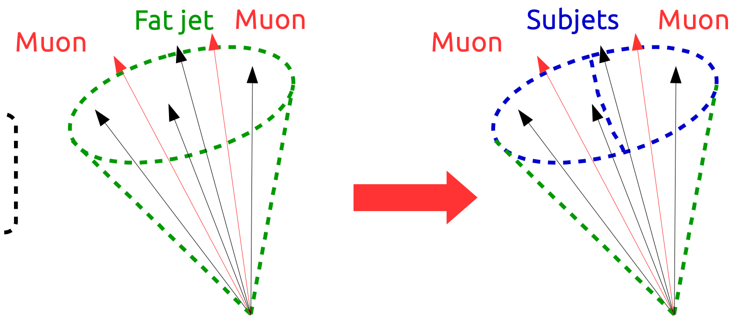
- **Subjet CSVv2:** Efficiency measured using muon-tagged subjets and mistag rate using subjets from an inclusive multijet sample

Muon-tagged subjets



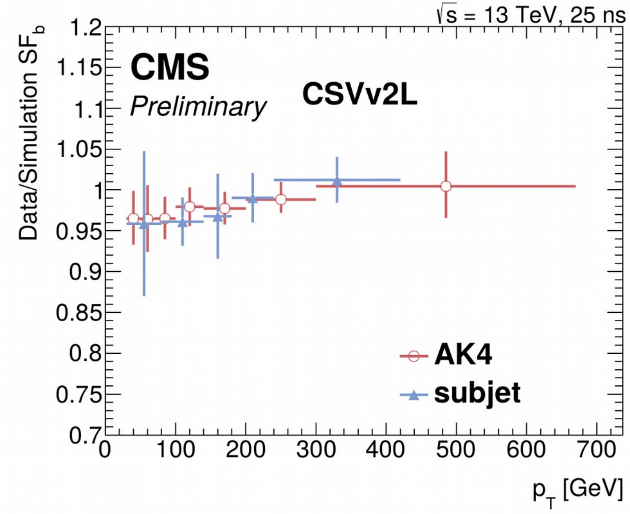

- **Double-b tagger:** Efficiency measured using double-muon-tagged fat jets and mistag rate from a $t\bar{t}$-enriched sample

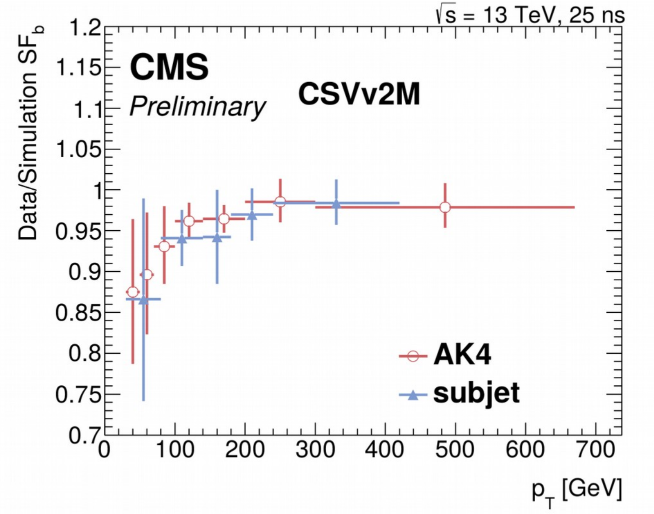
Double-muon-tagged fat jets




Enriched in gluon splitting jets with signal-like topology → Serves as proxy for H/Z → $b\bar{b}$ which have low cross section and are difficult to select with high purity

# Measurements for subjet CSVv2

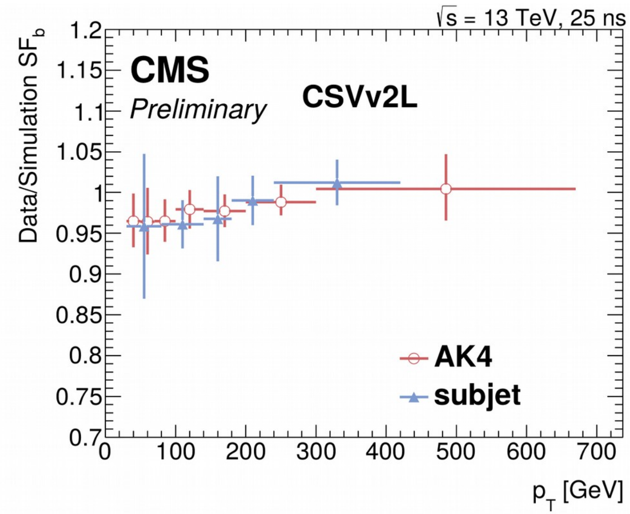


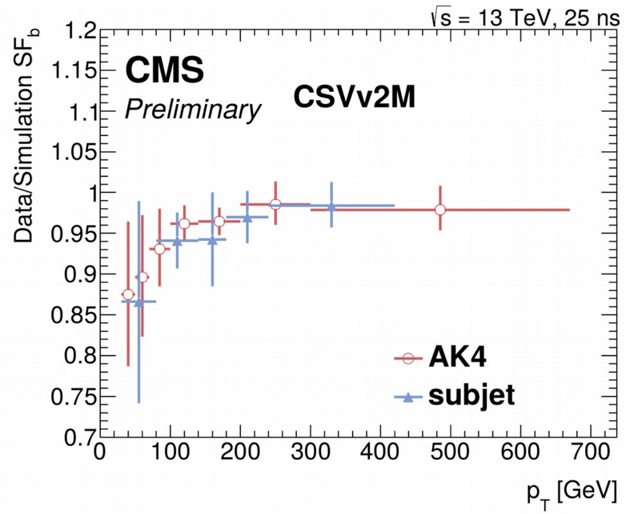




Data from 2015

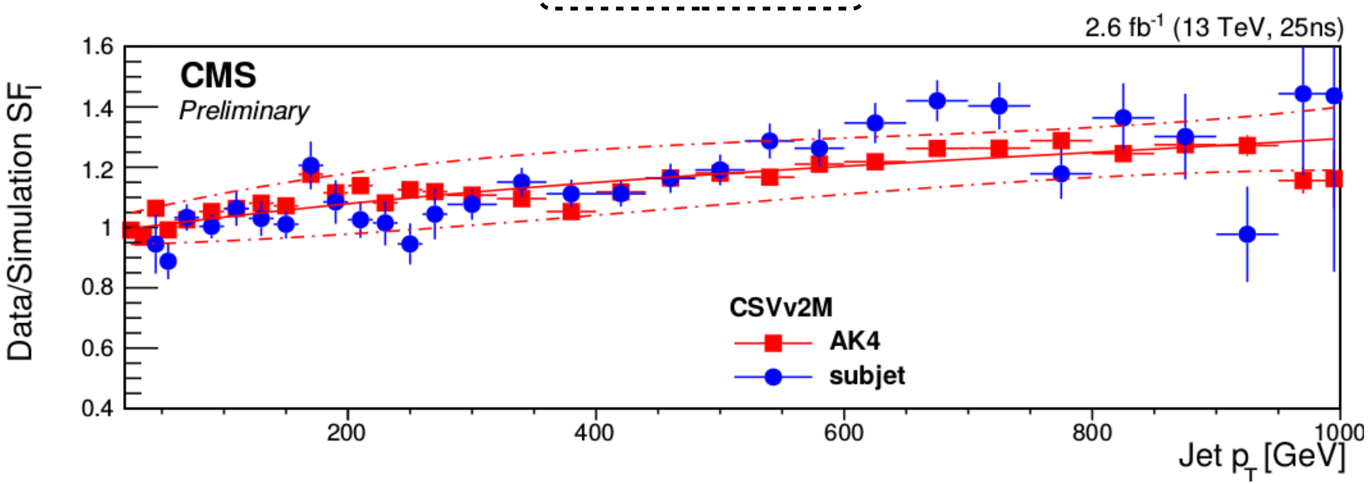
- Efficiency SFs measured using template fits (LT method) to JP tagger distributions for muon-tagged subjets

# Measurements for subjet CSVv2



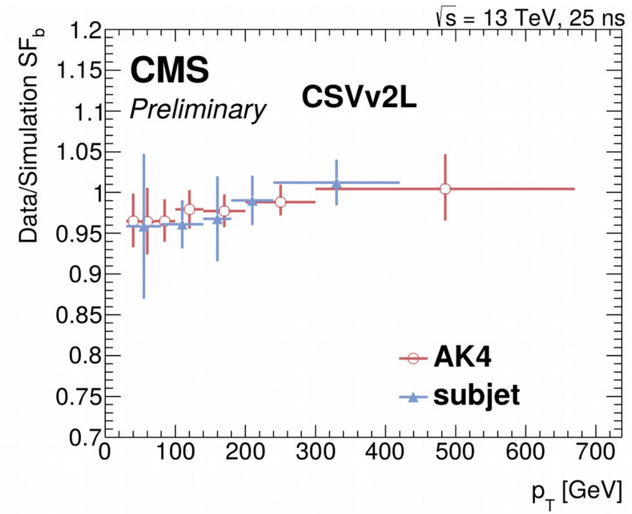



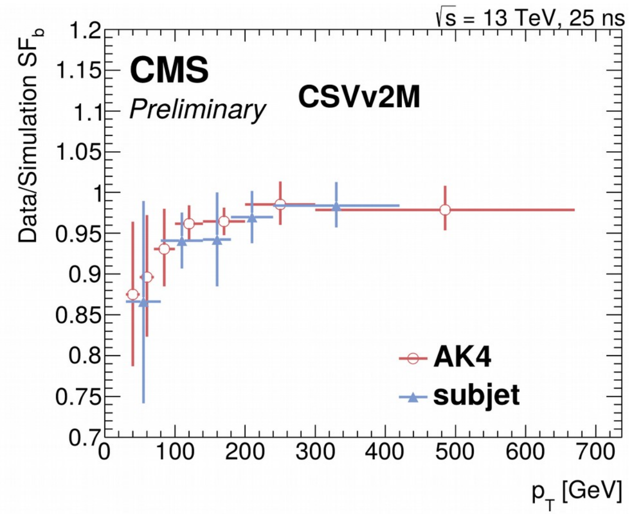

Data from 2015



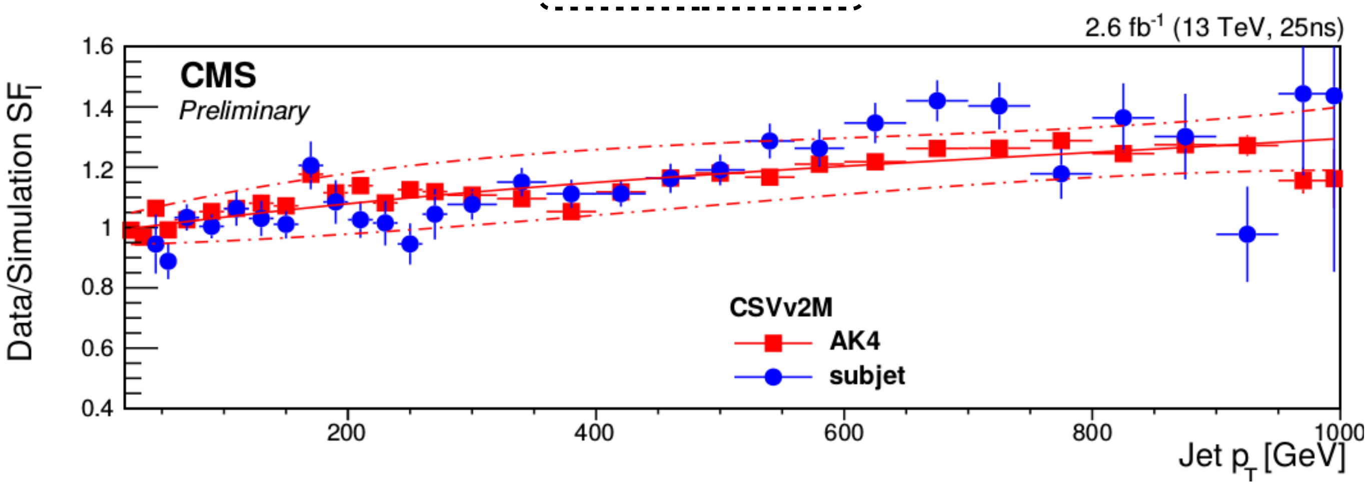

- Efficiency SFs measured using template fits (LT method) to JP tagger distributions for muon-tagged subjets
- Mistag rate SFs measured using negative tag method applied to subjets of inclusive fat jets

# Measurements for subjet CSVv2



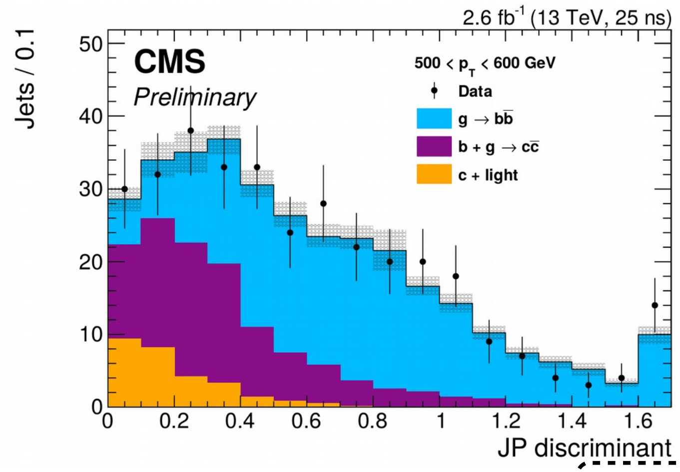

Data from 2015



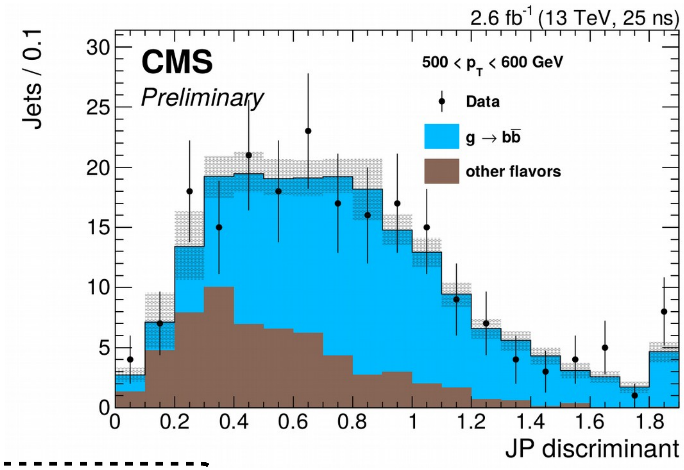

- Efficiency SFs measured using template fits (LT method) to JP tagger distributions for muon-tagged subjets
- Mistag rate SFs measured using negative tag method applied to subjets of inclusive fat jets
- Good agreement with SFs for the “standard” anti-$k_T$ R=0.4 jets

# Measurements for double-b tagger

Double-muon tagged fat jets




Left + passing loose double-b tag

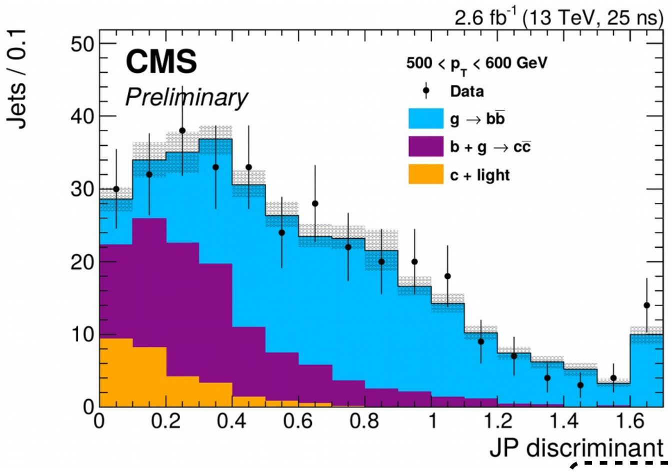



Data from 2015

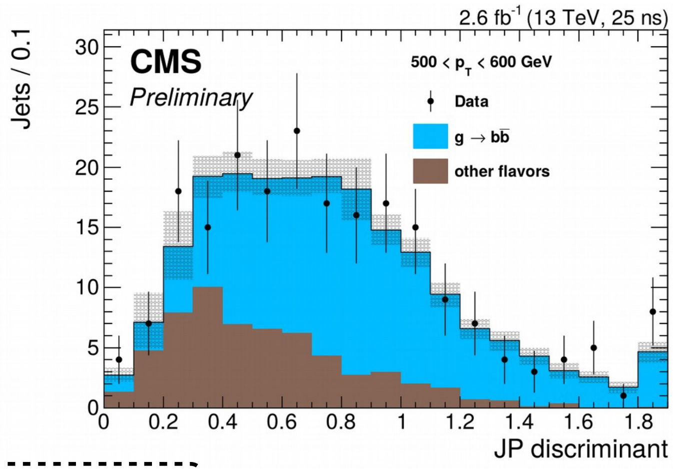
- Efficiency SFs measured using template fits (LT method) to JP tagger distributions for double-muon-tagged fat jets

# Measurements for double-b tagger

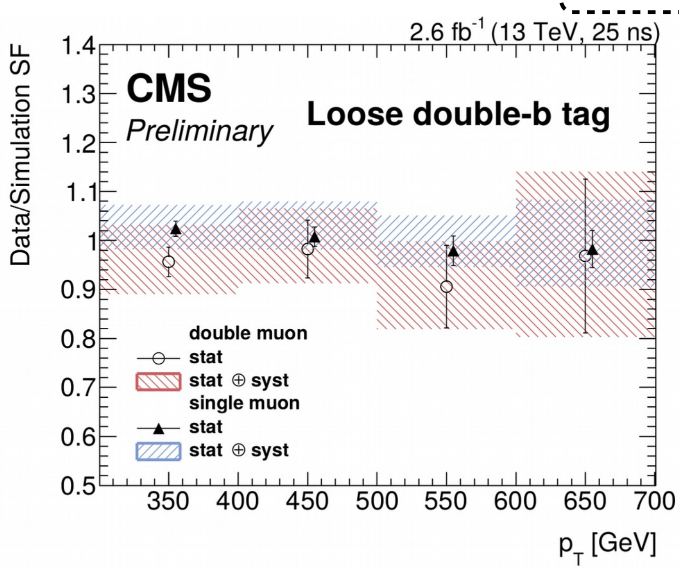
Double-muon tagged fat jets

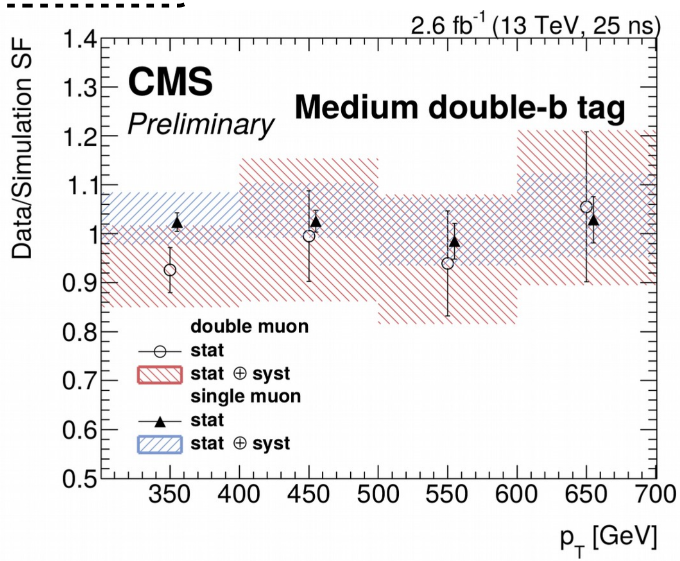



Left + passing loose double-b tag




Data from 2015

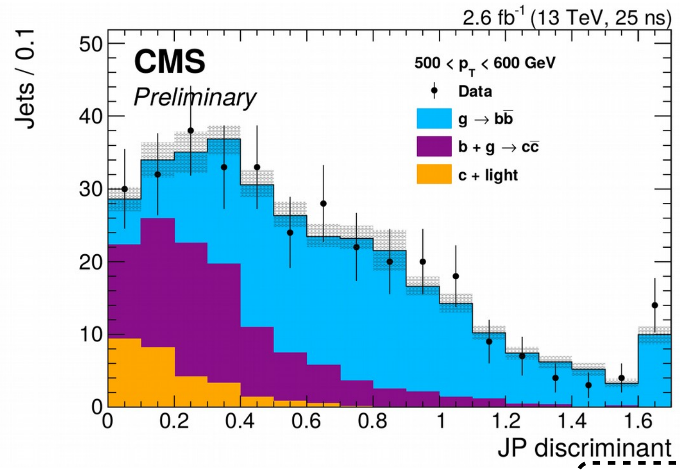


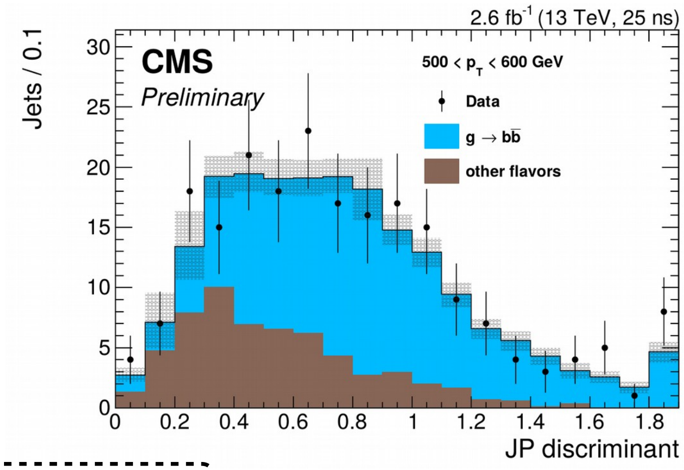

- Efficiency SFs measured using template fits (LT method) to JP tagger distributions for double-muon-tagged fat jets
- SFs consistent with unity but with sizable uncertainties due to limited sample size

# Measurements for double-b tagger

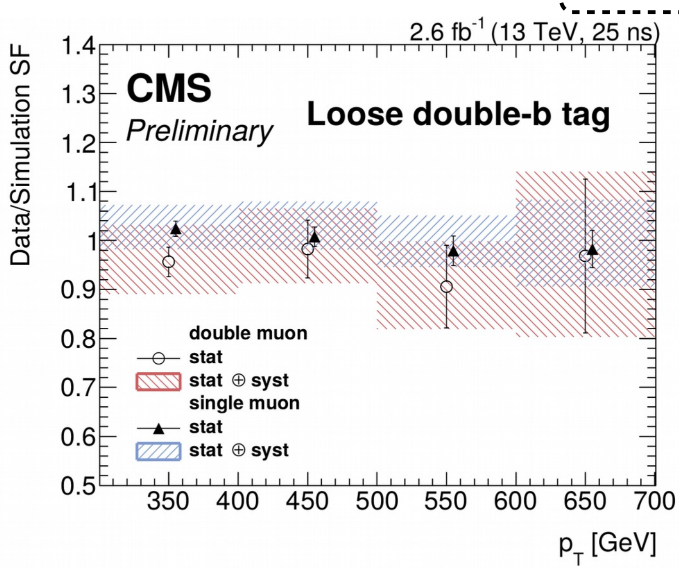
Double-muon tagged fat jets

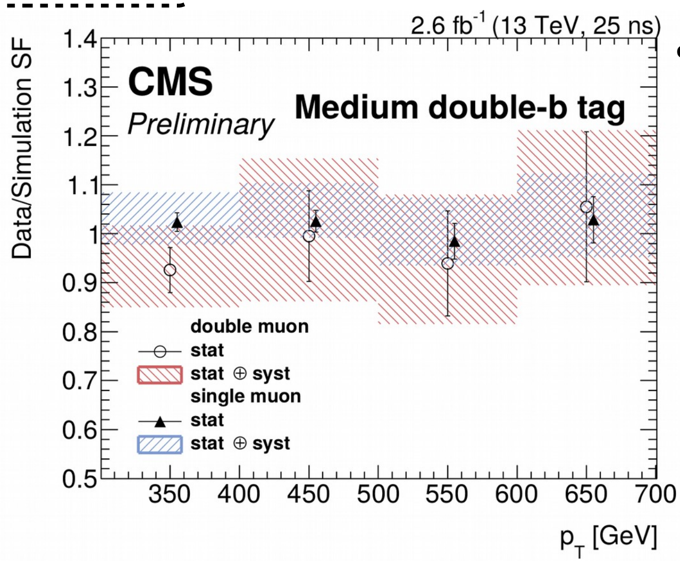



Left + passing loose double-b tag




Data from 2015



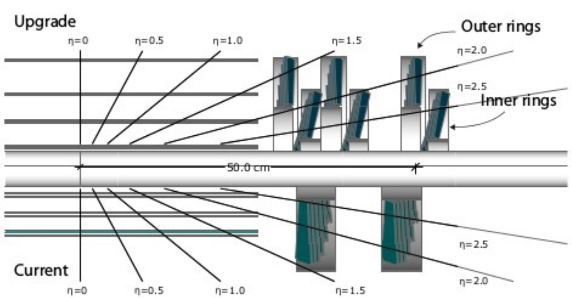

- Efficiency SFs measured using template fits (LT method) to JP tagger distributions for double-muon-tagged fat jets
- SFs consistent with unity but with sizable uncertainties due to limited sample size
- SFs for boosted top quarks faking boosted Higgs bosons measured in a sample enriched in semileptonic $t\bar{t}$ events (results in the backup slides)

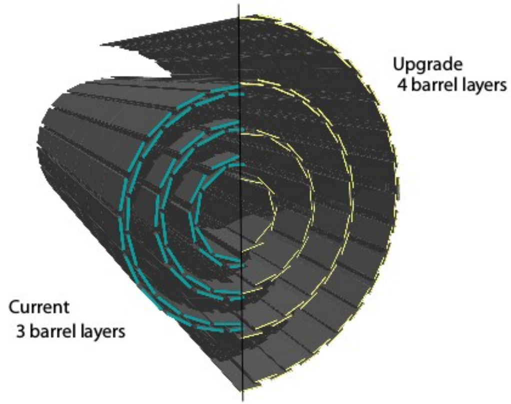
# Summary and outlook

- CMS has developed a powerful set of tools for boosted b tagging
- Subjet b tagging successfully commissioned during Run 1 with subjet scale factors now included in the standard list of deliverables of the b tagging group
- Dedicated double-b tagger algorithm developed and now commissioned for the first time using 13 TeV collision data recorded in 2015
- Growing base of analyses using boosted b tagging (primarily searches)
- More precise performance measurements expected using 2016 data
- *Further developments and improvements possible, especially in light of the* ***Phase I upgrade*** *of the CMS pixel detector*

# Summary and outlook

- CMS has developed a powerful set of tools for boosted b tagging
- Subjet b tagging successfully commissioned during Run 1 with subjet scale factors now included in the standard list of deliverables of the b tagging group
- Dedicated double-b tagger algorithm developed and now commissioned for the first time using 13 TeV collision data recorded in 2015
- Growing base of analyses using boosted b tagging (primarily searches)
- More precise performance measurements expected using 2016 data
- *Further developments and improvements possible, especially in light of the **Phase I upgrade** of the CMS pixel detector*



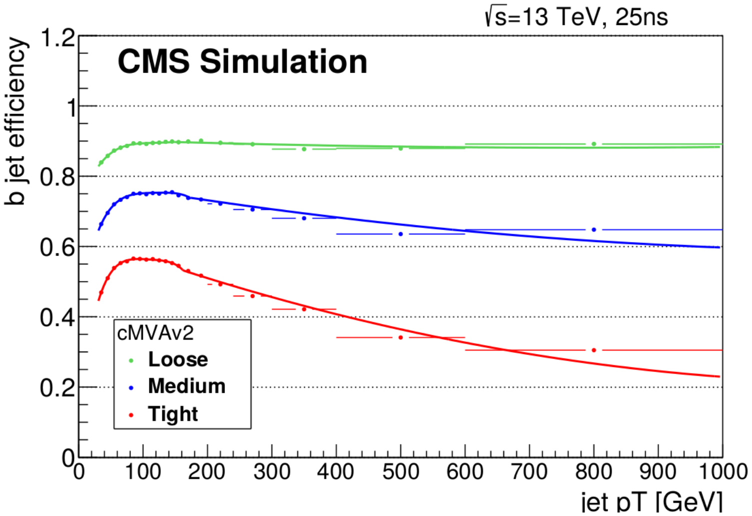
# Summary and outlook

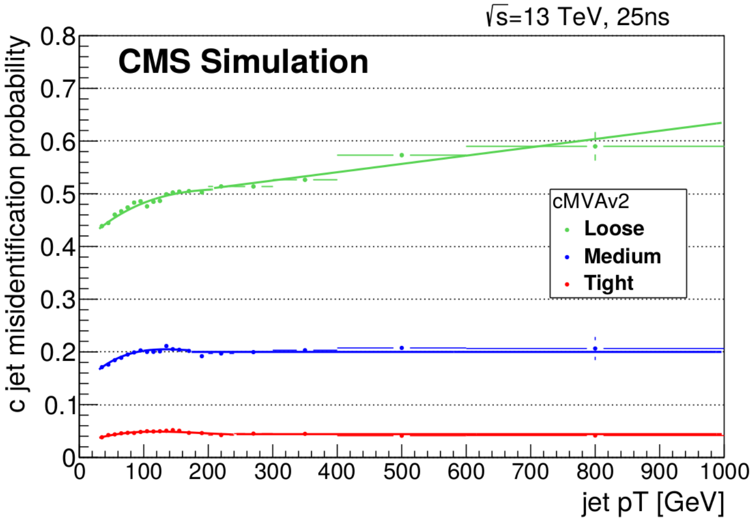
- CMS has developed a powerful set of tools for boosted b tagging
- Subjet b tagging successfully commissioned during Run 1 with subjet scale factors now included in the standard list of deliverables of the b tagging group
- Dedicated double-b tagger algorithm developed and now commissioned for the first time using 13 TeV collision data recorded in 2015
- Growing base of analyses using boosted b tagging (primarily searches)
- More precise performance measurements expected using 2016 data
- *Further developments and improvements possible, especially in light of the **Phase I upgrade** of the CMS pixel detector*

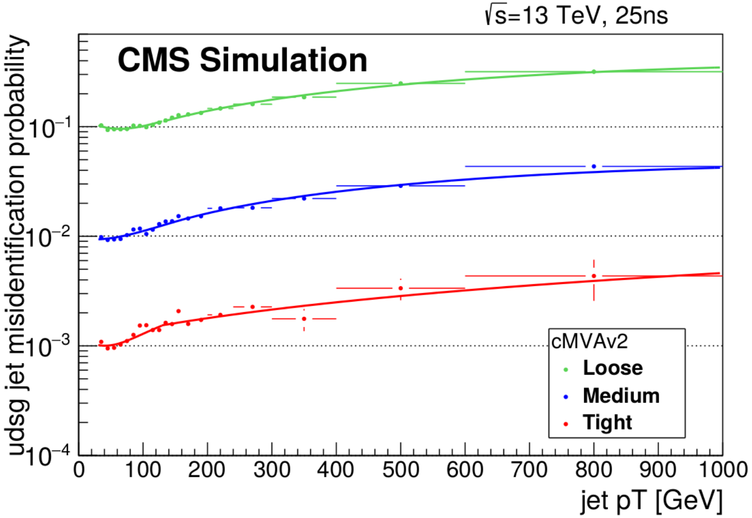
- More information about the current b tagging performance and results, including selection details, can be found in
  - http://cds.cern.ch/record/2138504/files/BTV-15-001-pas.pdf
  - http://cds.cern.ch/record/2195743/files/BTV-15-002-pas.pdf

# Backup Slides

# cMVAv2 $p_T$ dependence




More details about the cMVAv2 algorithm in http://cds.cern.ch/record/2138504/files/BTV-15-001-pas.pdf

# CSV algorithm

## Run 1 CSV algorithm:

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

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

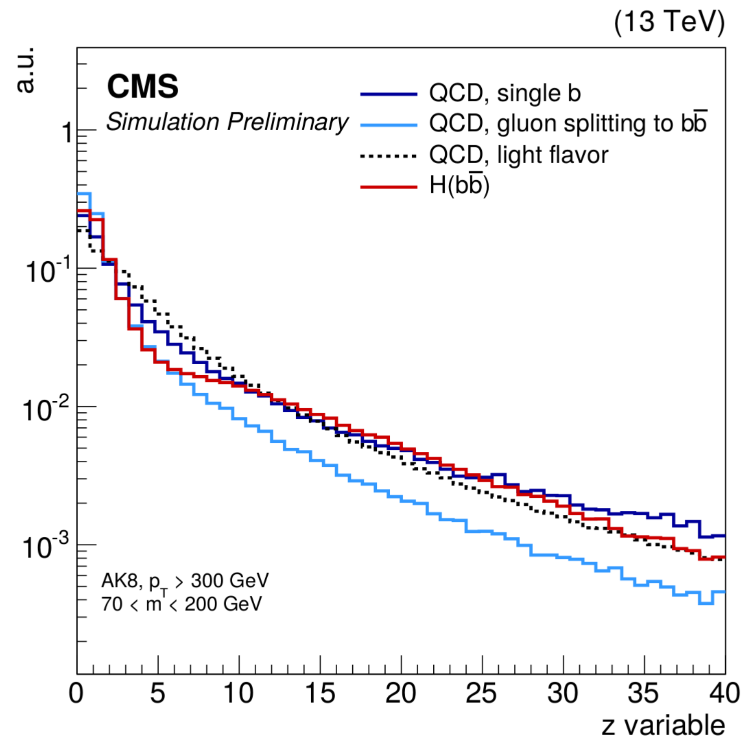
## Improved CSVv2 algorithm:

- 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 | ✔ | ✘ | ✘ |

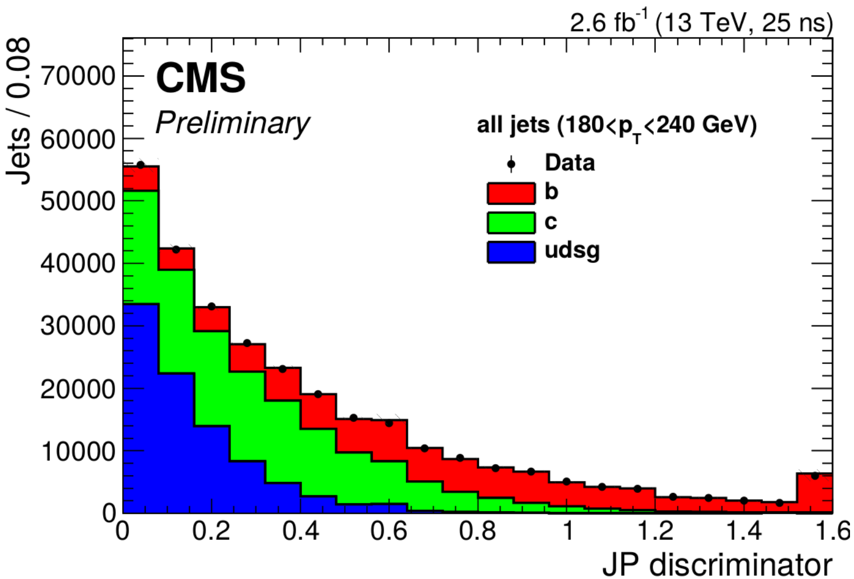
# Double-b tagger input variables

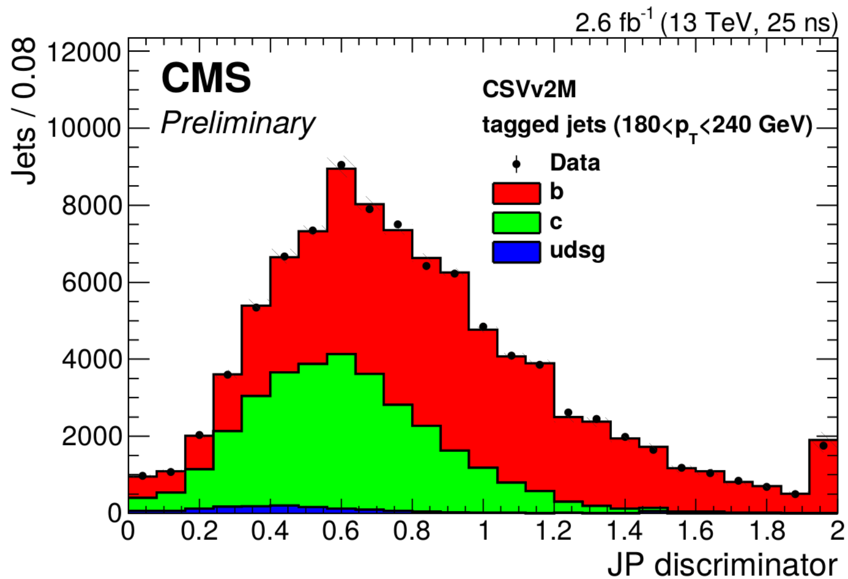
| | Variable | | Variable |
|---|---|---|---|
| 1. | trackSip3dSig_0 | 15. | tau1_trackEtaRel_0 |
| 2. | trackSip3dSig_1 | 16. | tau1_trackEtaRel_1 |
| 3. | trackSip3dSig_2 | 17. | tau1_trackEtaRel_2 |
| 4. | trackSip3dSig_3 | 18. | tau2_trackEtaRel_0 |
| 5. | trackSip2dSigAboveCharm_0 | 19. | tau2_trackEtaRel_1 |
| 6. | trackSip2dSigAboveBottom_0 | 20. | tau2_trackEtaRel_2 |
| 7. | trackSip2dSigAboveBottom_1 | 21 | tau1_vertexMass |
| 8. | jetNTracks | 22. | tau2_vertexMass |
| 9. | jetNSecondaryVertices | 23. | tau1_vertexEnergyRatio |
| 10. | tau1_trackSip3dSig_0 | 24. | tau2_vertexEnergyRatio |
| 11. | tau1_trackSip3dSig_1 | 25. | tau1_flightDistance2dSig |
| 12. | tau2_trackSip3dSig_0 | 26. | tau2_flightDistance2dSig |
| 13. | tau2_trackSip3dSig_1 | 27. | tau1_vertexDeltaR |
| 14. | z_ratio | | |




$$z = \Delta R(\mathrm{SV}_0, \mathrm{SV}_1) \cdot \frac{p_{\mathrm{T},\mathrm{SV}_1}}{m(\mathrm{SV}_0, \mathrm{SV}_1)}$$

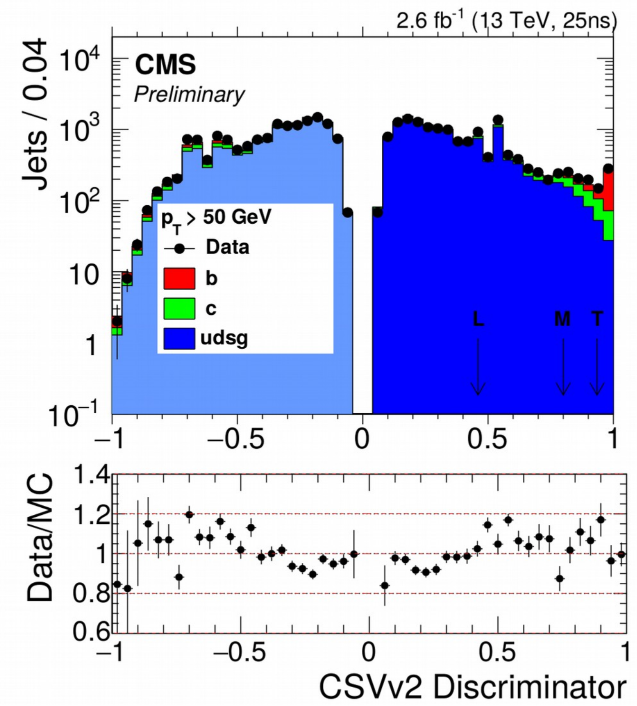
- JP template fits before and after tagging




$$\varepsilon_{\mathrm{b}}^{\mathrm{tag}} = \frac{C_{\mathrm{b}} \cdot f_{\mathrm{b}}^{\mathrm{tag}} \cdot N_{\mathrm{data}}^{\mathrm{tag}}}{f_{\mathrm{b}}^{\mathrm{before\ tag}} \cdot N_{\mathrm{data}}^{\mathrm{before\ tag}}}$$

# Negative tag method

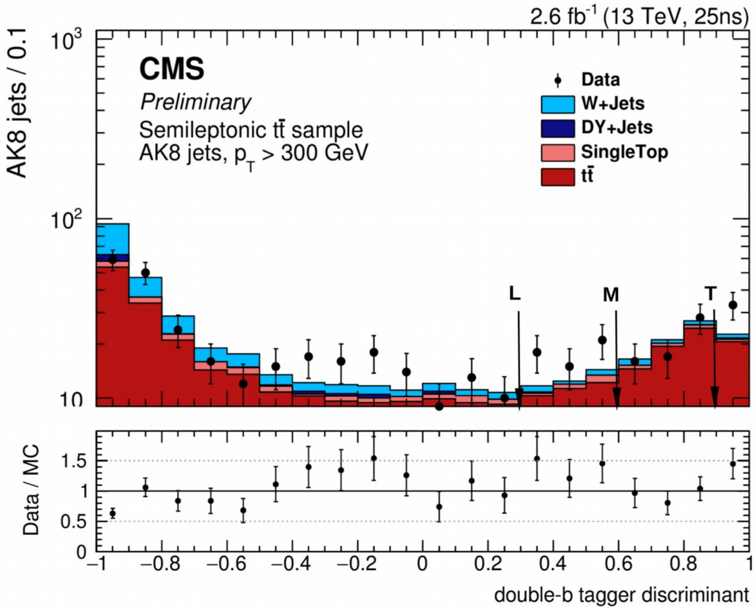



$$\varepsilon_{\mathrm{data}}^{\mathrm{misid}} = \varepsilon_{\mathrm{data}}^{-} \cdot R_{\mathrm{light}}$$

$$R_{\mathrm{light}} = \frac{\varepsilon_{\mathrm{MC}}^{\mathrm{misid}}}{\varepsilon_{\mathrm{MC}}^{-}}$$

# Measurements for double-b tagger

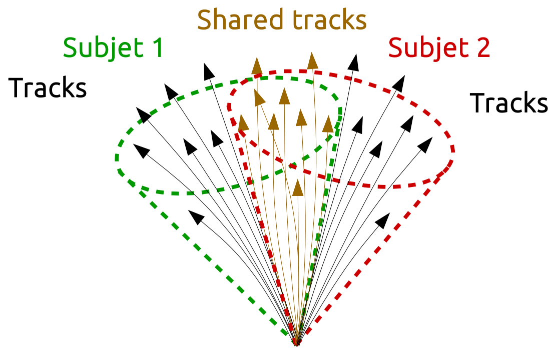
- Scale factors for boosted top quarks faking boosted Higgs bosons



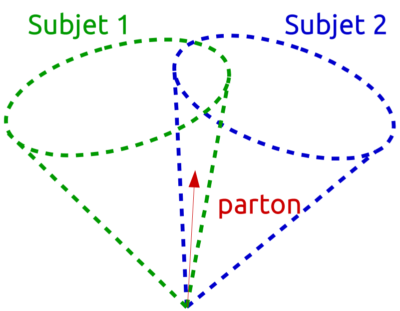

| $p_T$ bin (GeV) | 300 - 400 | 400 - 550 | 550 - 700 | inclusive (300 - 700 GeV) |
|---|---|---|---|---|
| loose double-b | | | | |
| $\epsilon$ (Data) | 0.40±0.04 | 0.40±0.05 | 0.47±0.09 | 0.41±0.03 |
| $\epsilon$ (MC) | 0.33±0.01 | 0.36±0.0.01 | 0.34±0.01 | 0.34±0.01 |
| SF | 1.24±0.13 | 1.12±0.13 | 1.40±0.32 | 1.20±0.09 (stat.) ±0.05 (syst.) |
| medium double-b | | | | |
| $\epsilon$ (Data) | 0.26±0.04 | 0.25±0.04 | 0.25±0.03 | 0.26±0.03 |
| $\epsilon$ (MC) | 0.23±0.01 | 0.25±0.01 | 0.22±0.01 | 0.24±0.01 |
| SF | 1.14±0.16 | 1.01±0.17 | 1.13±0.39 | 1.09±0.11 (stat.) ±0.05 (syst.) |
| $p_T$ bin (GeV) | 300 - 400 | 400 - 500 | | inclusive (300 - 500 GeV) |
| tight double-b | | | | |
| $\epsilon$ (Data) | 0.10±0.02 | 0.08±0.02 | | 0.06±0.01 |
| $\epsilon$ (MC) | 0.07±0.01 | 0.06±0.01 | | 0.09±0.02 |
| SF | 1.54±0.36 | 1.41±0.39 | | 1.49±0.27 (stat.) ±0.05 (syst.) |

# Limitations of the Run 1 setup

- Run 1 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
  - However, 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



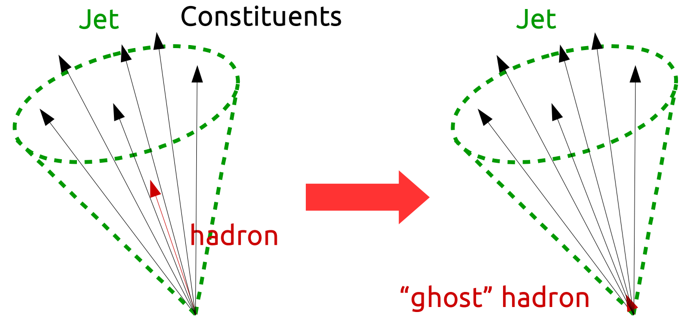

- Jet flavor assignment:
  - Also based on a fixed-size cone ($\Delta R<0.3$)
  - Can lead to subjet flavor ambiguities



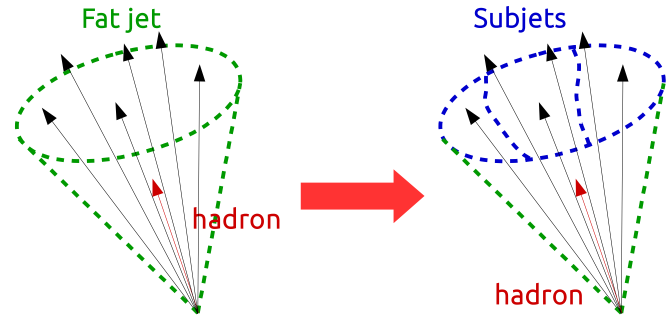

- Secondary vertex reconstruction:
  - Using tracks associated to jets (not optimal when the fraction of shared tracks becomes significant)

# Run 2 developments

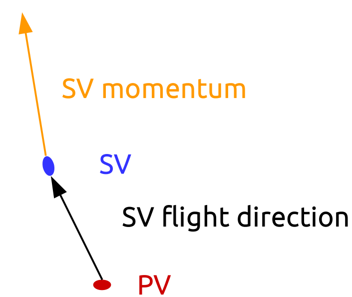
- Improved (sub)jet flavor definition [1]
  - Using b and c hadrons instead of b and c quarks
  - Based on ghost association of hadrons/partons instead of ΔR matching → Subjet flavor ambiguities eliminated




- 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 significantly reduces possible track sharing
  - Ghost association used to assign SVs to jets and then to the closest (sub)jets based on rapidity-based ΔR

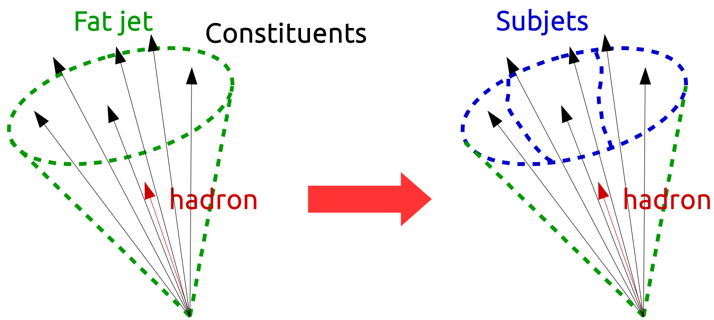



- Improved CSV algorithm (CSVv2)

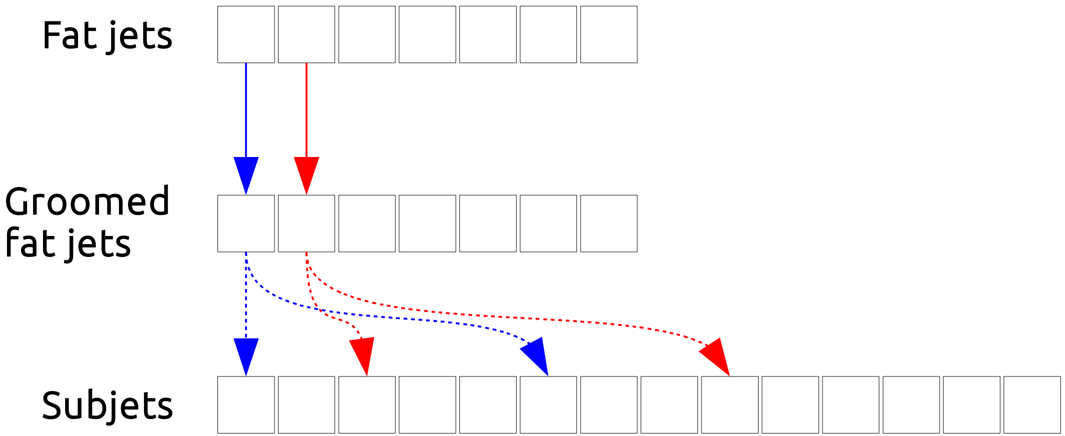
[1] https://twiki.cern.ch/twiki/bin/view/CMSPublic/SWGuideBTagMCTools

# Subjet flavor

- Subjet flavor definition:
  - “Ghost” hadrons/partons clustered inside a fat jet are 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)

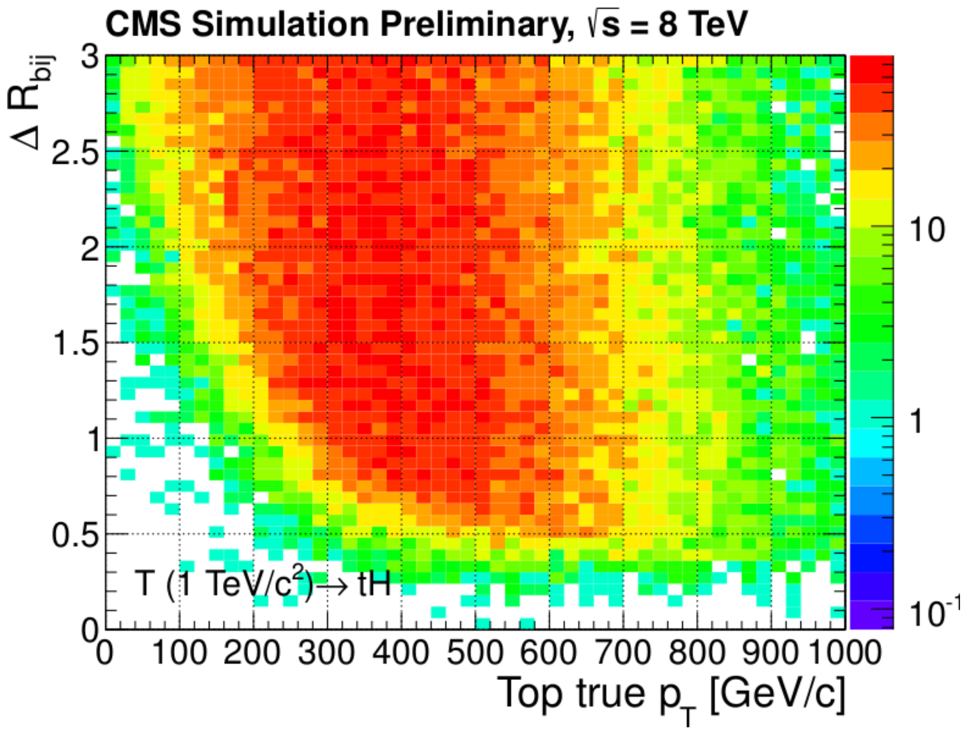


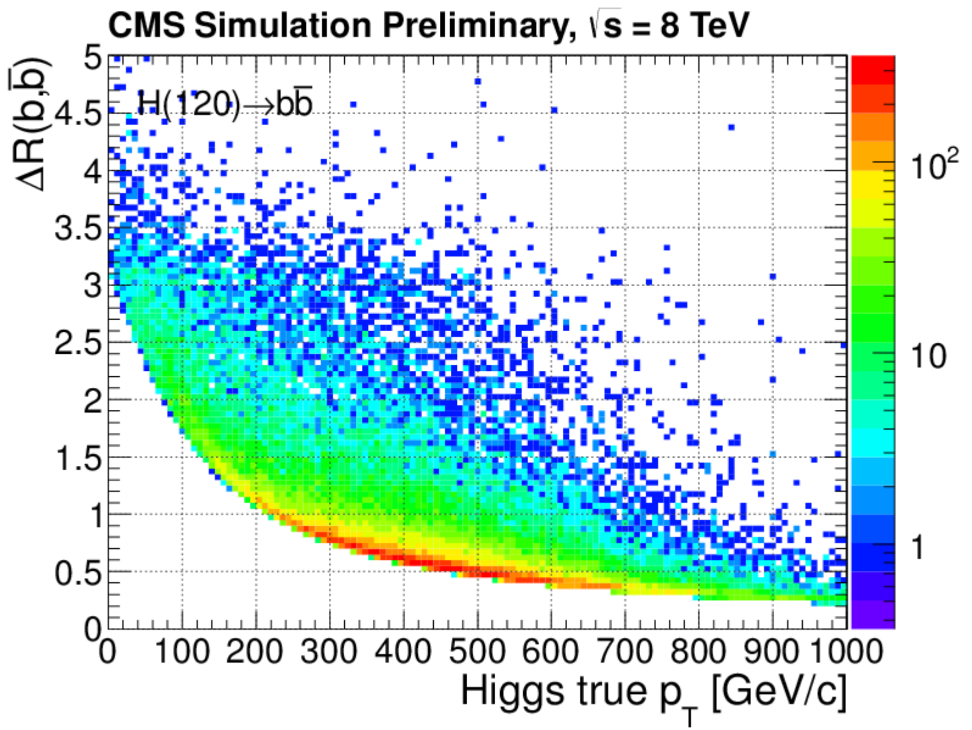
# Inclusive Vertex Finder SV reconstruction

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

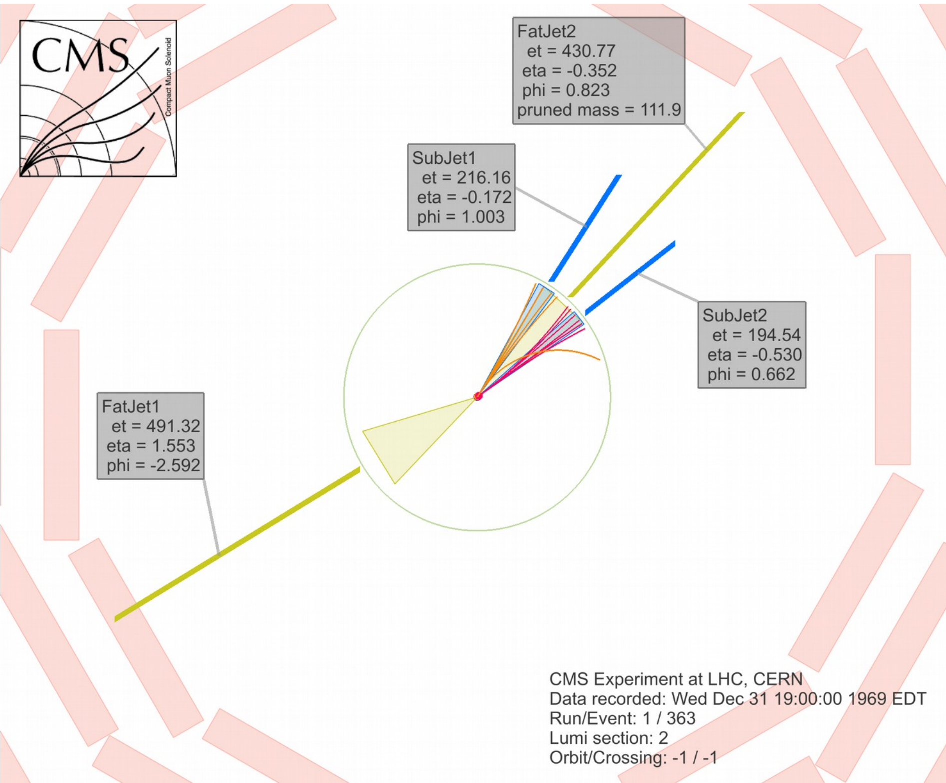
# Boosted hadronic top quarks and H→b$\bar{b}$

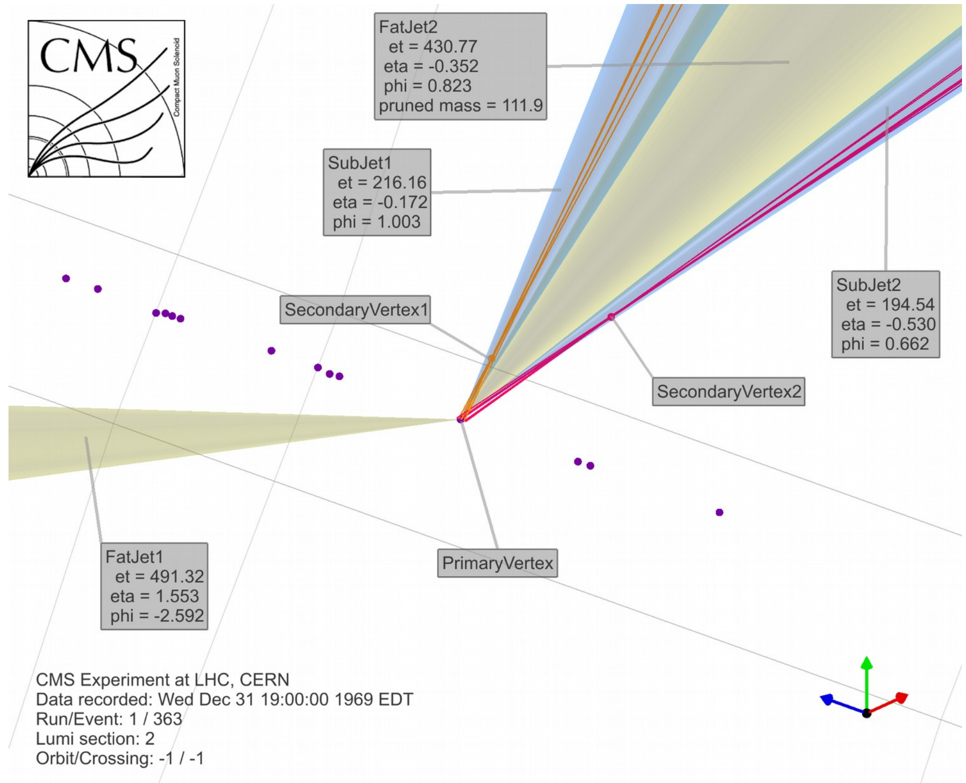


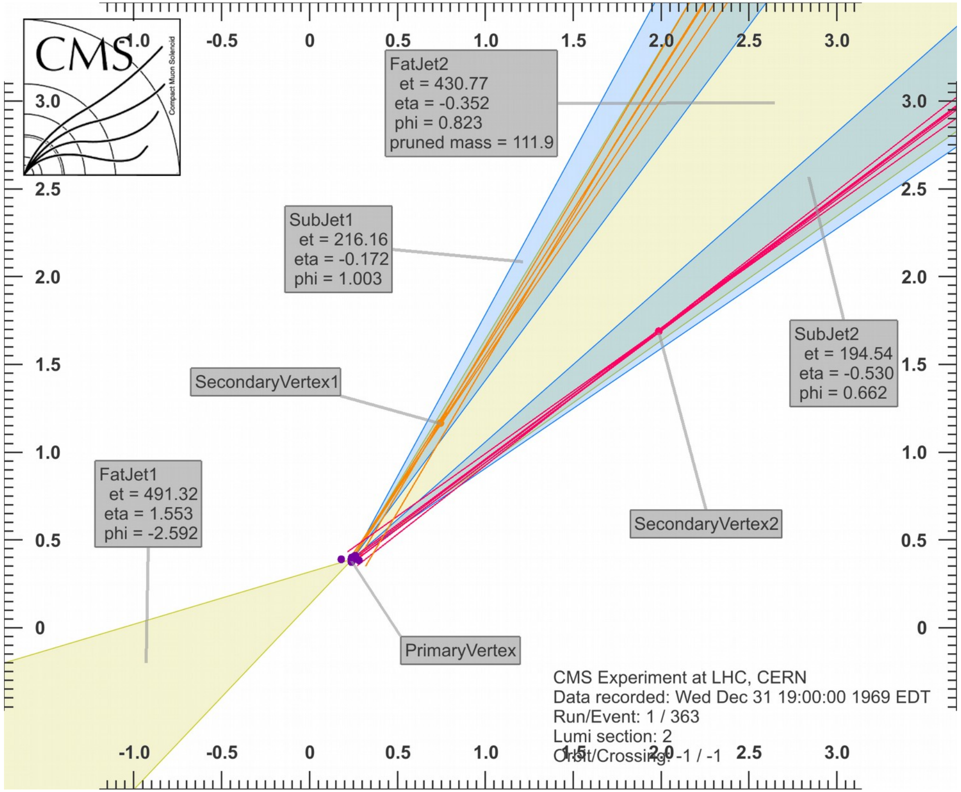

$$\Delta R(\mathrm{b},\bar{\mathrm{b}}) \gtrsim \frac{2m}{p_{\mathrm{T}}}$$

# Boosted H→b$\bar{b}$ (simulation)

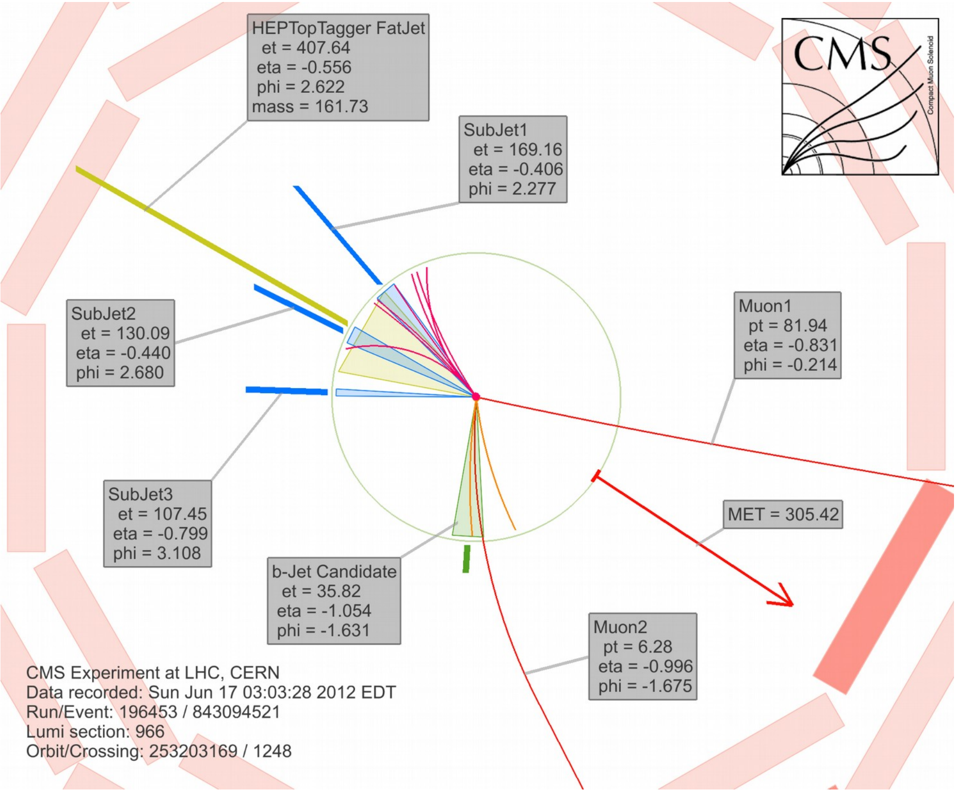


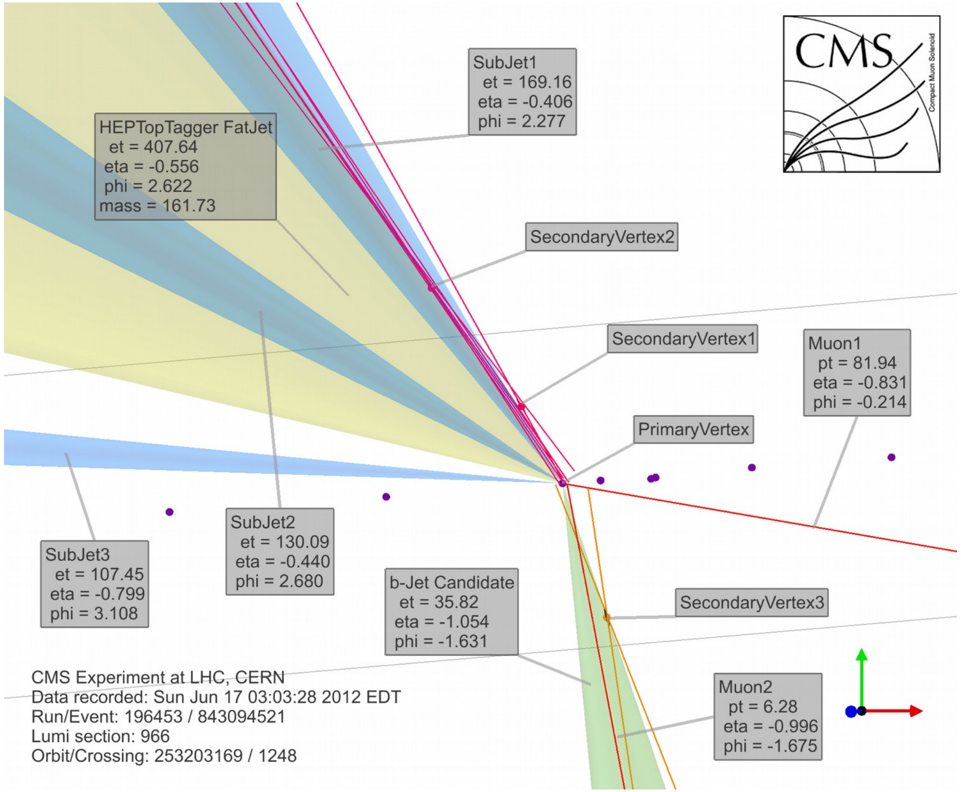



# Boosted H→b$\bar{b}$ (simulation)



# Boosted top candidate



# Boosted top candidate