# B-Jet Identification in Boosted Topologies at CMS
# BTV-13-001

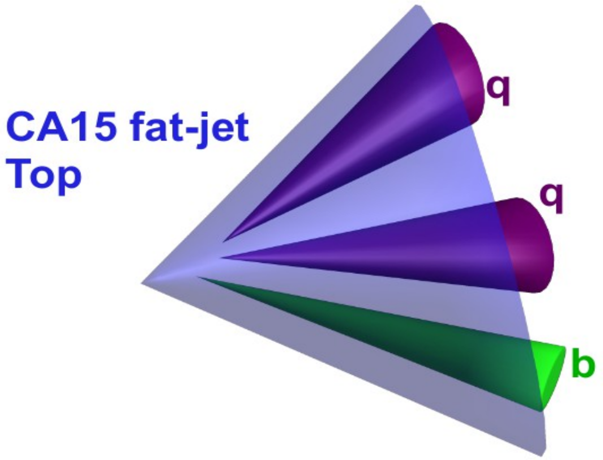
**Ivan Marchesini (University of Hamburg)**
***on behalf of the CMS Collaboration***

# B-Tagging in Boosted Topologies

- B-tagging at CMS traditionally developed on **isolated AK5 jets**, mostly suitable for the **non-boosted regime**.
- The work here presented is the first study at CMS dedicated to b-tagging in the boosted regime. Two topologies considered:

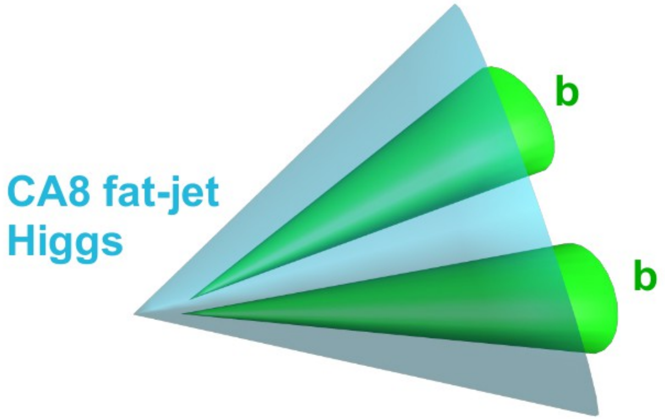
**Boosted top, hadronic decay:**

**→b-jet clustered in large fat-jet, together with W decay products**
**→top decay selected using HEPTopTagger, based on CA15 jet collection**
**→studies based on CA8 cone size and CMSTopTagger underway**



**Boosted Higgs→b$\bar{b}$:**

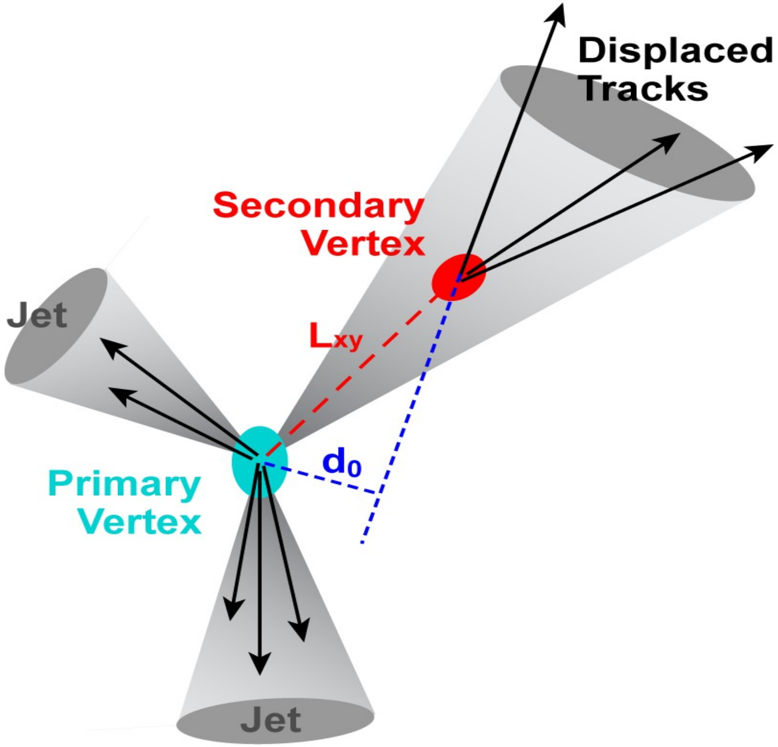
**→2 b-jets clustered together in large fat-jet**
**→studies based on CA8 collection**



# B-Quark Signatures

- B-quarks hadronize in B-hadrons, forming jets.

- Sizable lifetime B-hadron:
  - → **secondary vertex**;
  - → **tracks with large impact parameter**.



- Large mass, ~5 GeV: decay products have **large $p_{Trel}$**, transverse momentum relative to jet-axis.

- B-quark fragmentation function: **high $p_T$** of the b-hadron relatively to the jet $p_T$.

- The B-decay produces often leptons: soft muon or electron within jet.

# B-Tagging at CMS

**JTA**

→**jet-tracks association**: static **cone** **$\Delta R(\text{tracks,jet}) < 0.3$**

**OBSERVABLES**

→apply tight **selection on tracks**, mainly for pile-up rejection

→determine b-tagging observables

**DISCRIMINATORS**

→calculate b-tagging **discriminators**

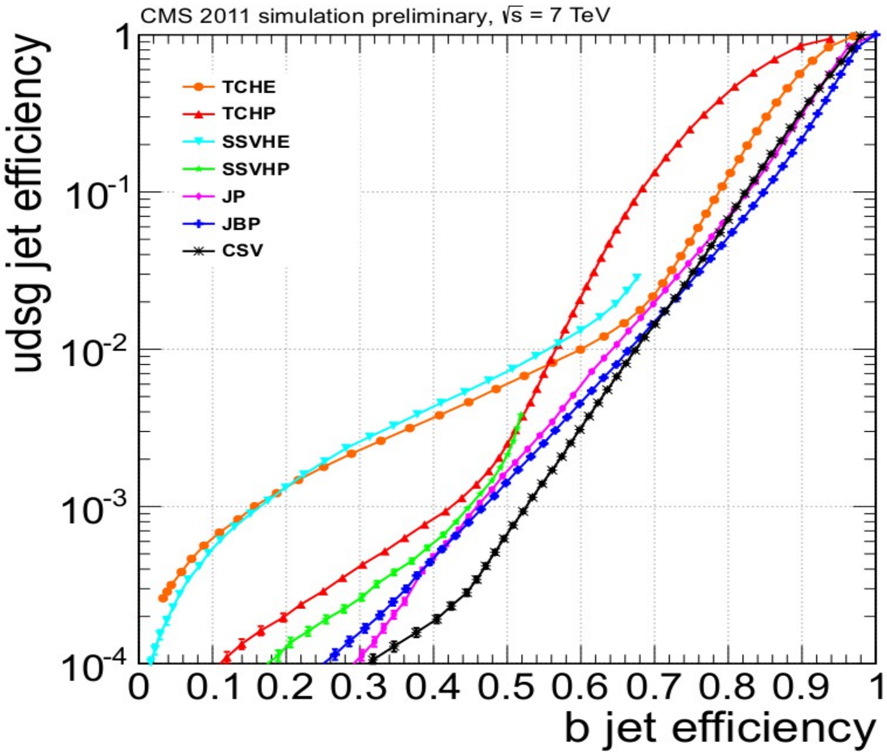
→several **operating points** defined for taggers, selecting different regions of purity/efficiency:

| | |
|---|---|
| • loose **L**; | 10% |
| • medium **M**; | 1% |
| • tight **T**; | 0.1% |

misidentification from light quarks/gluons

# B-Tagging Algorithms

- Boosted studies based on the **Combined Secondary Vertex** **CSV** tagger:
  - likelihood ratio combination of **secondary vertex + single track information**;
  - currently the best tagger in CMS, improvements ongoing.



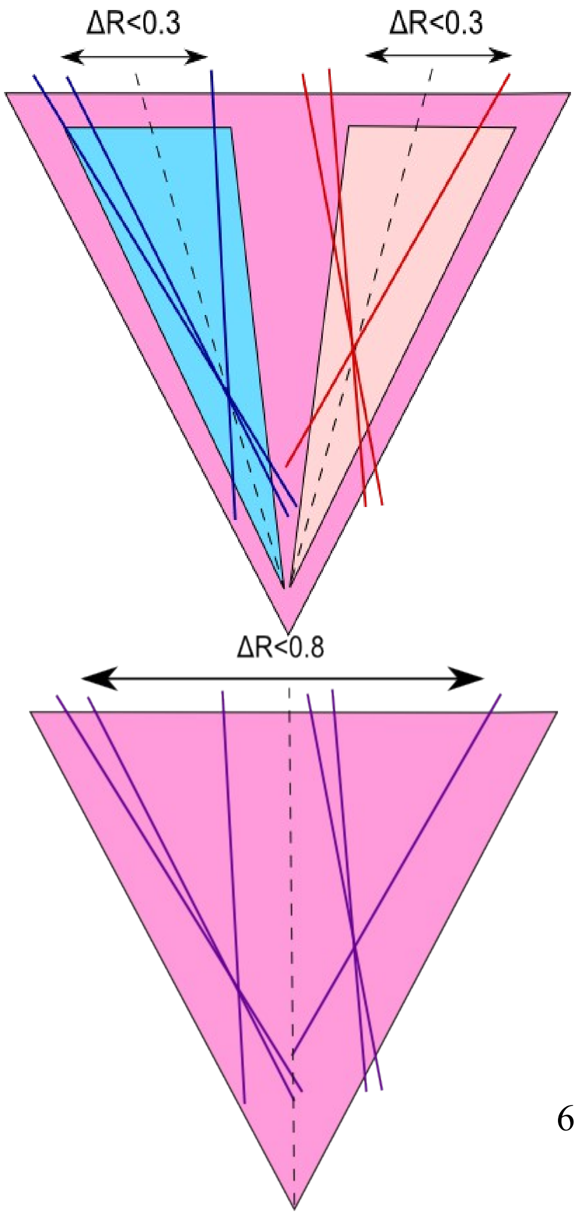
- For performance measurements used also **Jet-Probability** **JP** tagger:
  - likelihood estimate of the probability that the jet-tracks come from the PV, based on the IP significance of all jet-tracks;
  - **calibrated on data** from tracks with negative IP.

# Boosted B-Tagging Scenarios

Scenarios considered for boosted topologies:

- **subjet CSV**:
  - standard CSV b-tagger applied to subjets of the fat-jet (**2 b-tags for Higgs-tagging, ≥1 for top-tagging**);
  - standard track selection, ΔR<0.3.

- **fat-jet CSV**:
  - standard CSV b-tagger applied to the Higgs/top candidate fat-jet;
  - **extended track selection**, **ΔR<0.8 or 1.5** according to jet size.



# Monte Carlo Studies

# Higgs Channel

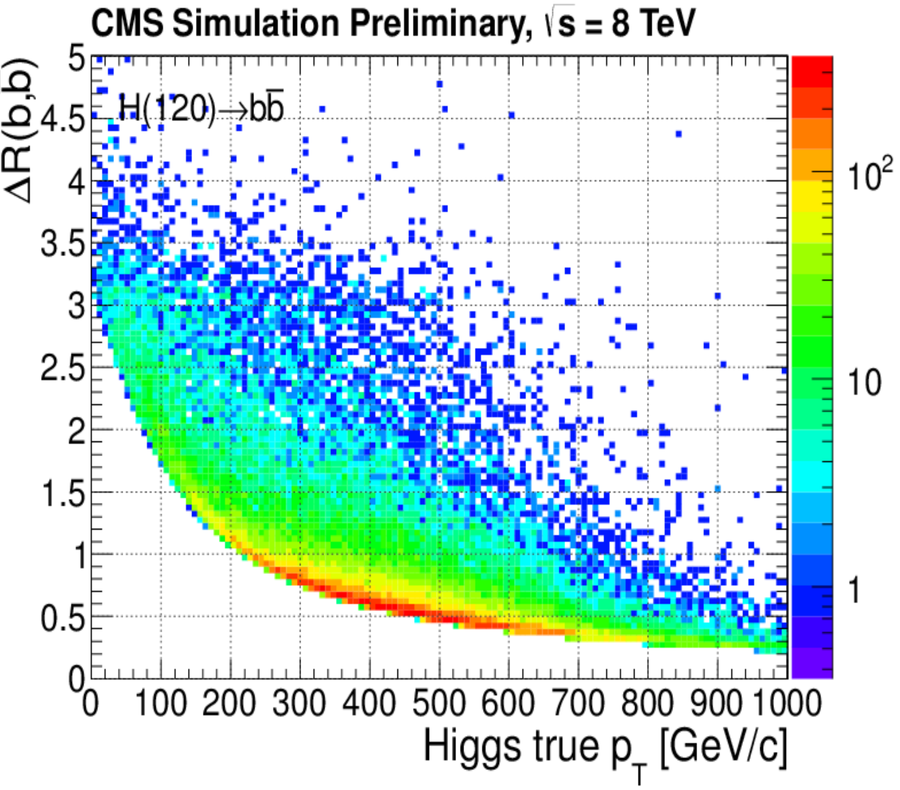
- Based on **CA8 jet collection**: boosted regime for $p_T$ > 300 GeV.

- Signal: **B'→bH** pair production. B-tagging studied on H→bb.

- Inclusive **mistag** from **QCD** and mistags from hadronically-decaying **W/Z/top**.

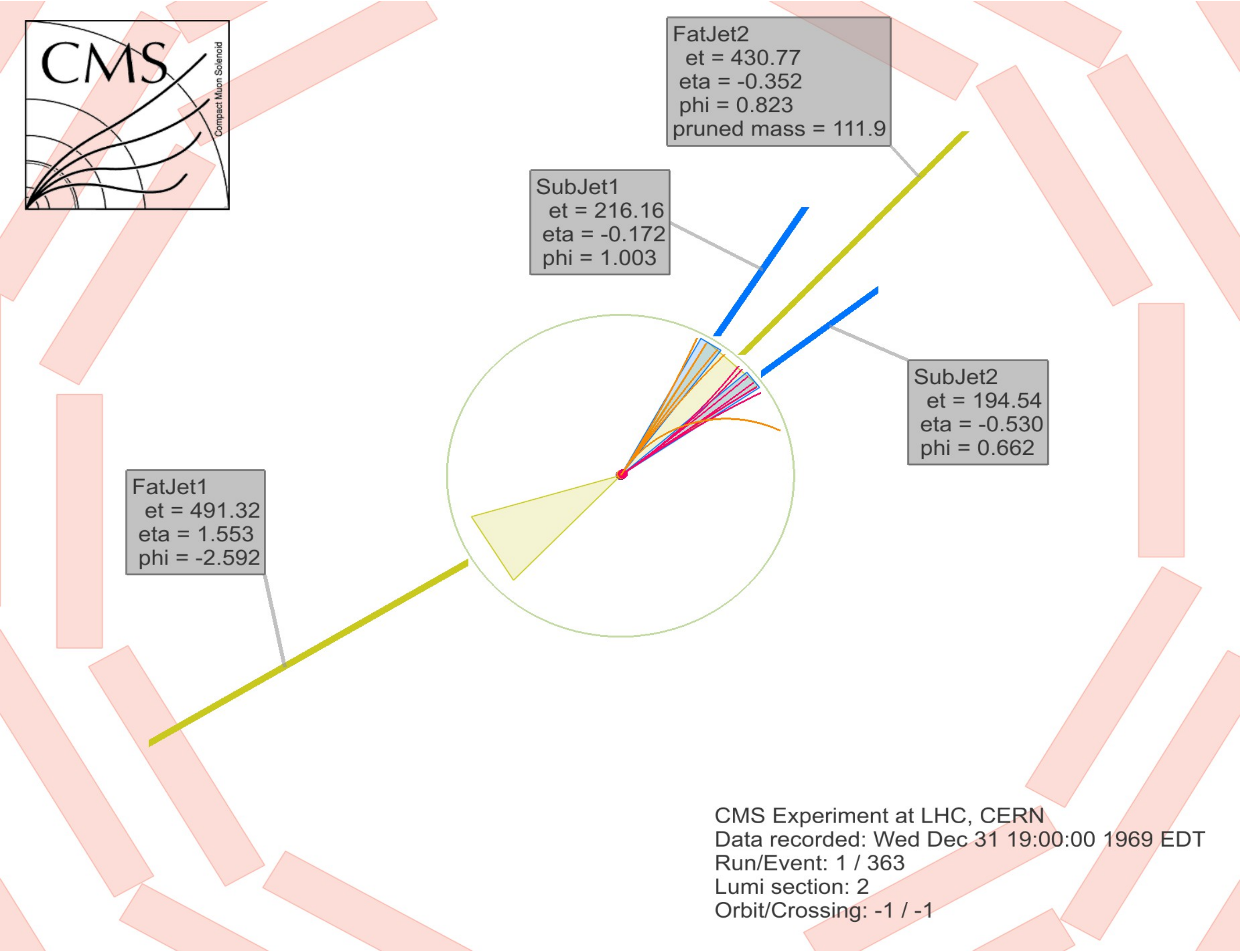
- Subjet b-tagging based on **pruned subjets**:

➔cut on **pruned jet mass** can be combined with b-tagging requirement (see next slides).



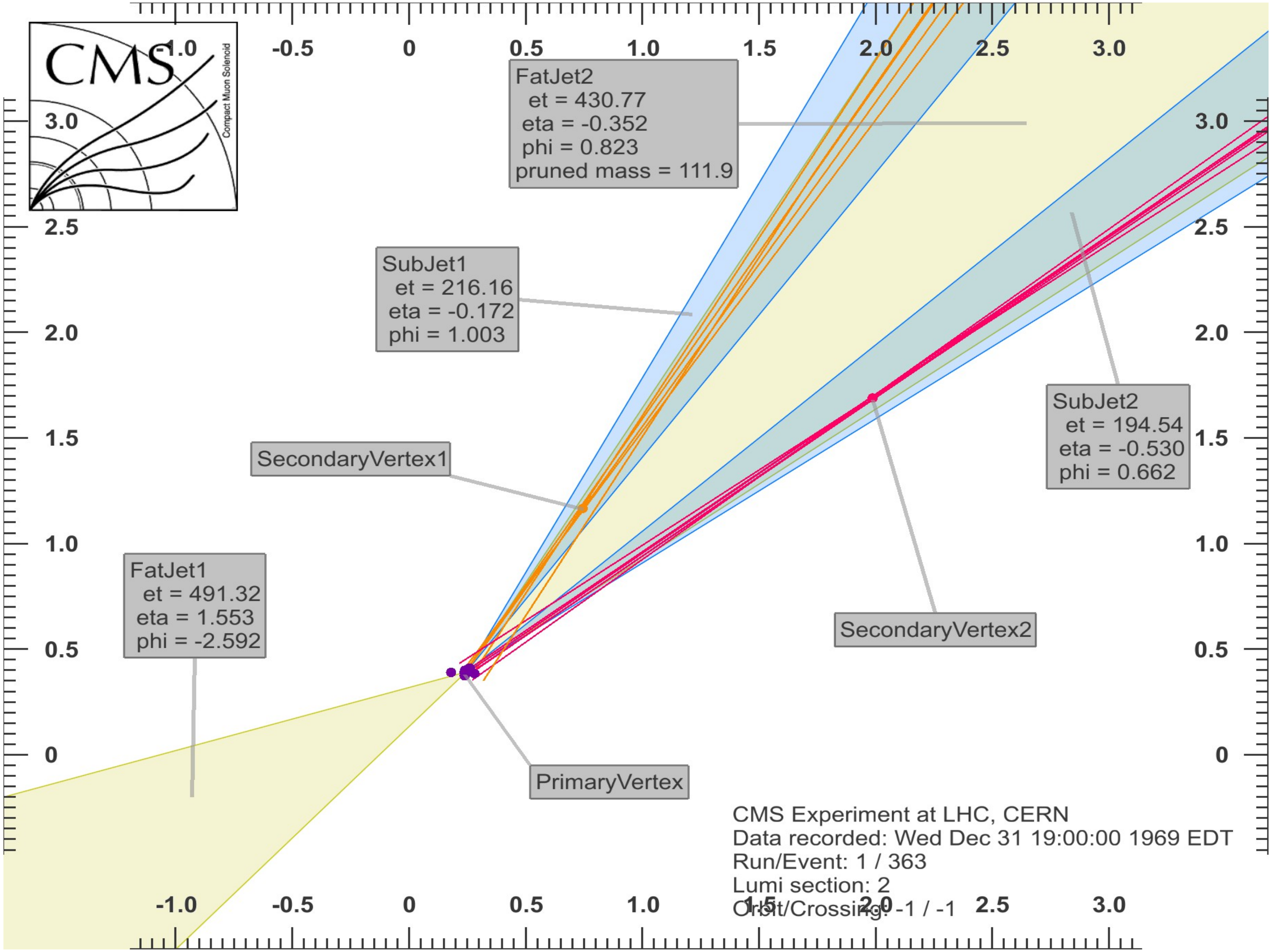
# Event Display (MC)

**Radion (1.5 TeV)→HH→bbbb**



# Event Display (MC): Zoom

**Radion (1.5 TeV)→HH→bbbb**

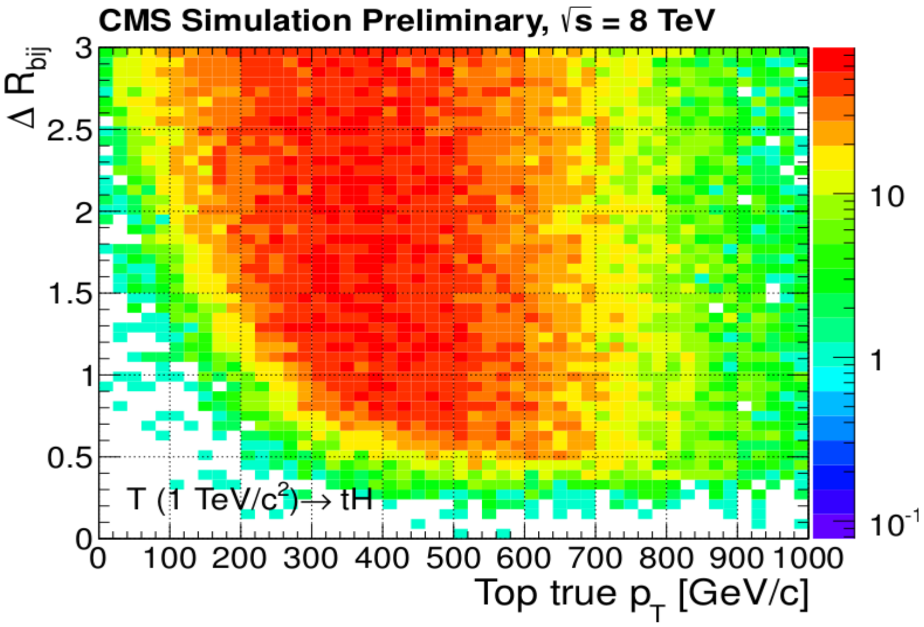


# Top Channel

- Based on **CA15** collection, default for **HEPTopTagger**.

- Large cone-size allows to **reach lower $p_T$'s** (~200GeV) without switching from merged-top to un-merged top selection.

- Signal: **T'→tH** pair production. Consistency of the results checked also on SM ttbar production.



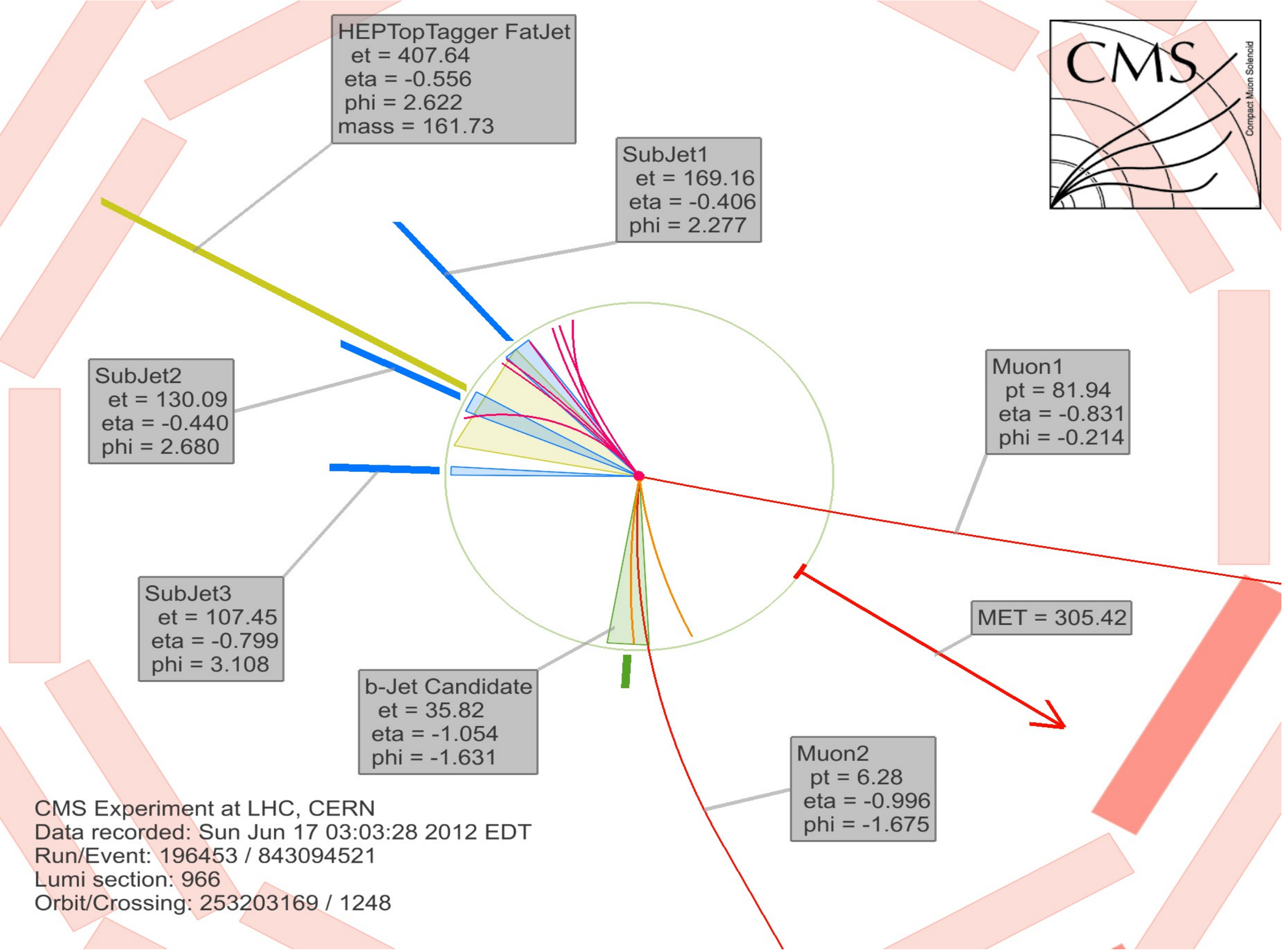
**spread between top decay products (T'→tH)**

- Inclusive **mistag** from **QCD**.

- HEPTopTagger forces **3 filtered subjets**: used for subjet b-tagging.

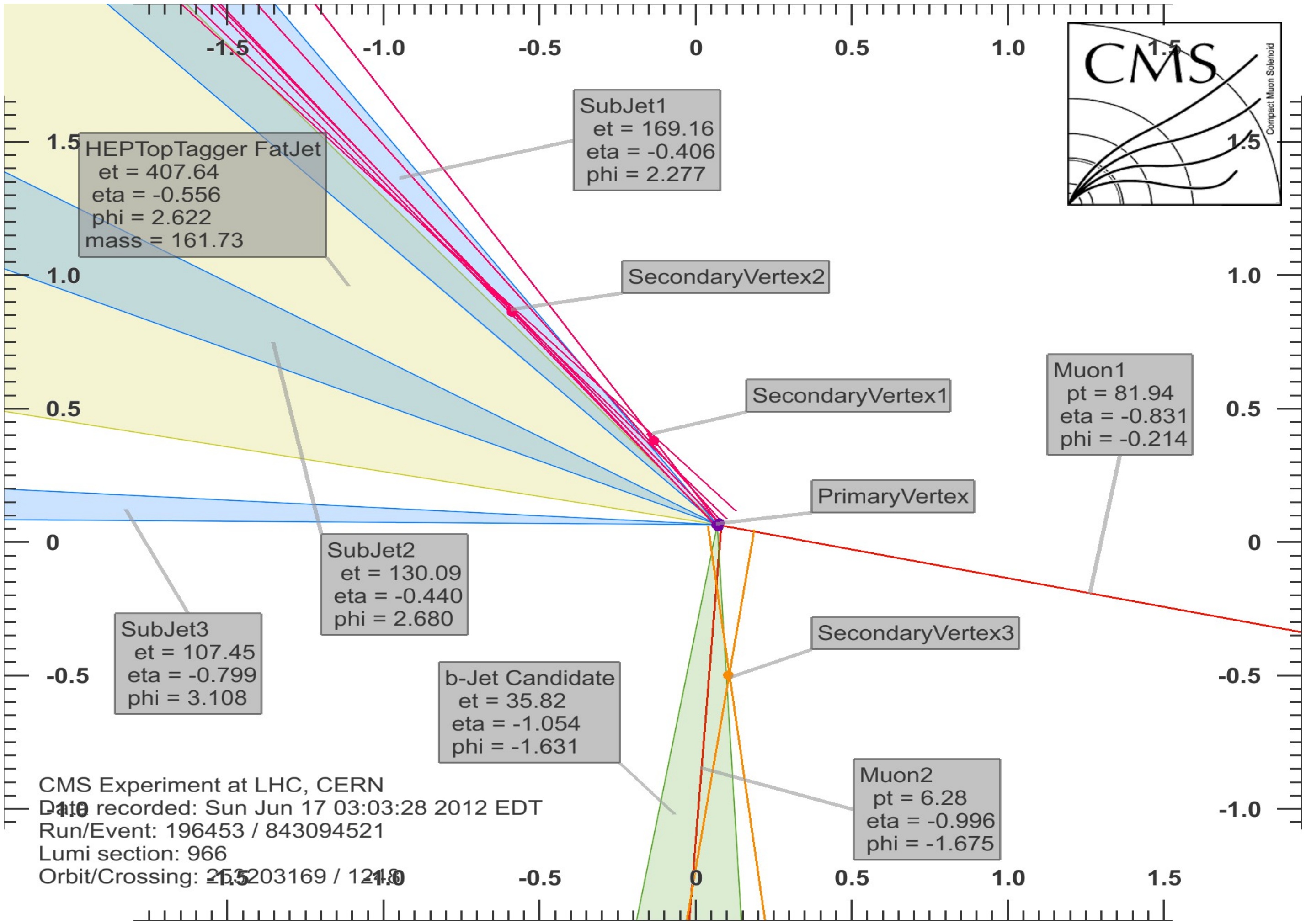
# Event Display (Data)

**Semileptonic ttbar**



# Event Display (Data): zoom

**Semileptonic ttbar**



CMS Experiment at LHC, CERN
Data recorded: Sun Jun 17 03:03:28 2012 EDT
Run/Event: 196453 / 843094521
Lumi section: 966
Orbit/Crossing: 253203169 / 1218

# B-Tagging Performance

Higgs channel

**Subjet b-tagging** performs better

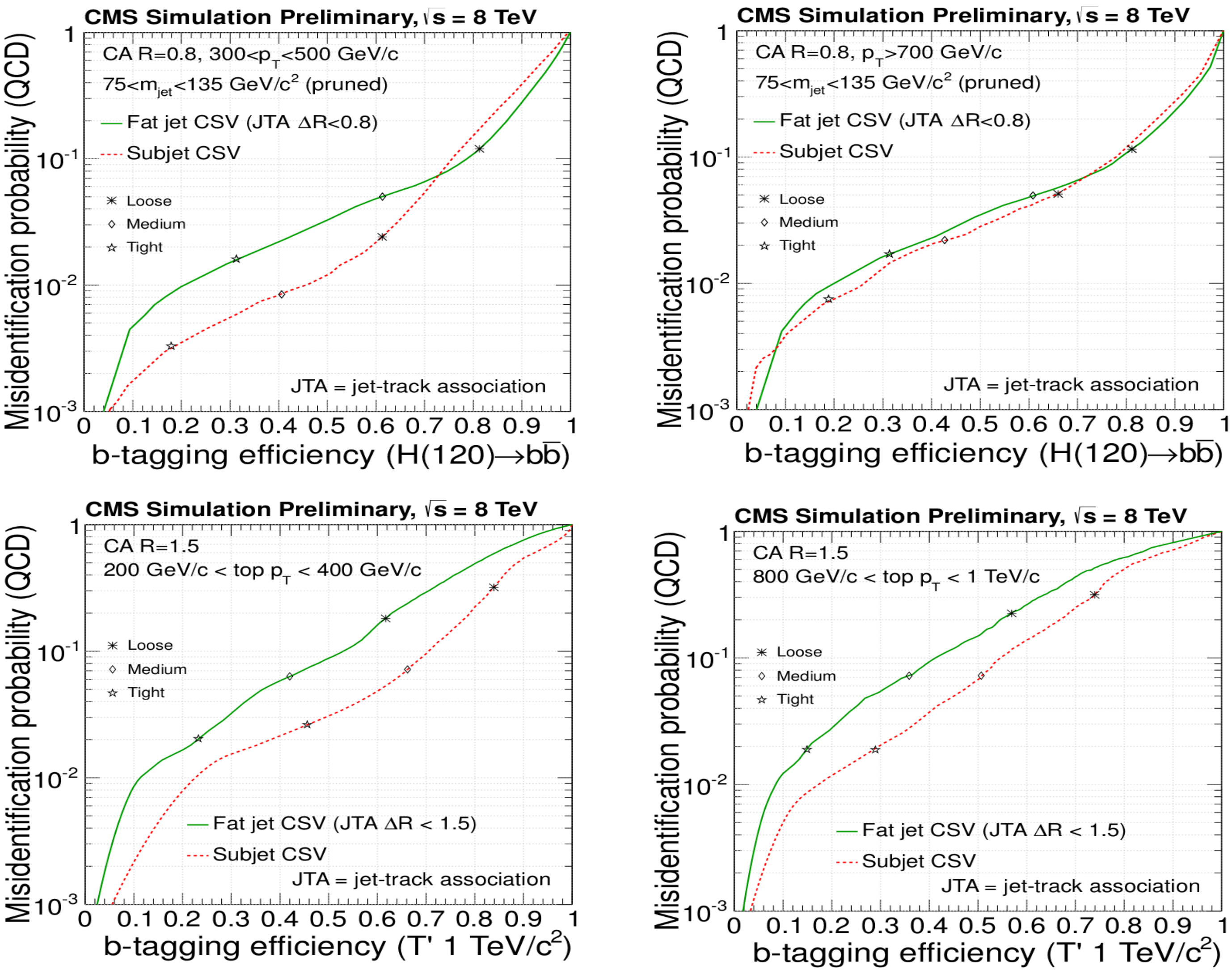
**Fat-jet b-tagging** suitable at very high $p_T$

Top channel

Overall **subjet b-tagging** performs better

medium boost regime

large boost regime



# Tagging Performance

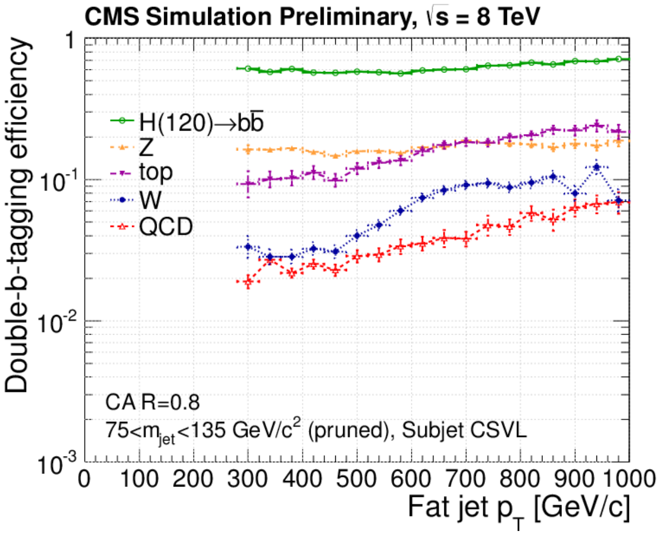
## Higgs channel

Higgs-tagging
=
double b-tagging
+
75 <$m_{jet}$< 135 GeV

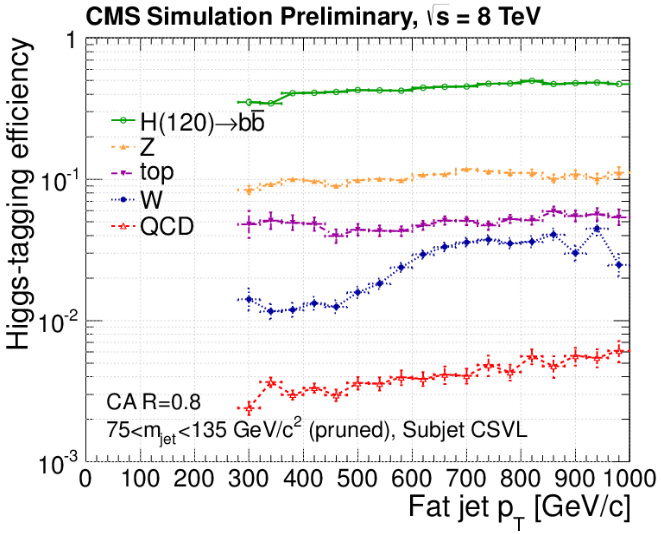
## Top channel

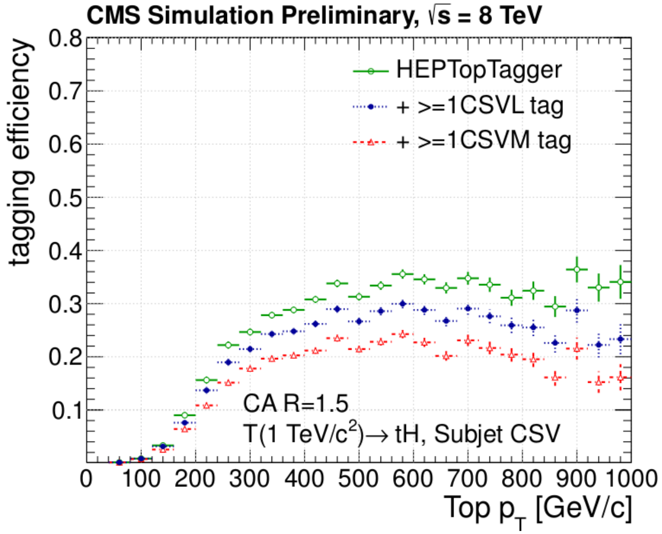
QCD mistag rate reduced up to a factor 10 with minor loss of efficiency

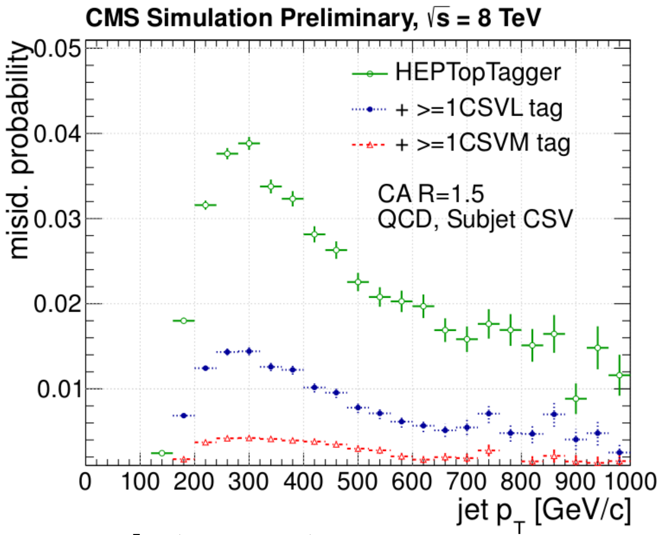
### double b-tagging



### Higgs tagging







### tagging efficiency

### mistag rate

# Validation on Data

# Validation

- Slight discrepancies in b-tagging performance between data and Monte Carlo → corrected applying to simulated events **Scale Factors**:
  - **SF$_b$**: b-tagging;
  - **SF$_{light}$**: misidentification from light flavors;
  - **SF$_c$** :  misidentification from charm.

- Can we apply the standard Scale Factors, **measured on isolated b-jets**, in boosted event topologies? Validation in two steps:

  - **STEP 1**. Check agreement between data and Monte Carlo in the boosted topologies, for relevant b-tagging observables.

  - **STEP 2**. Dedicated measurement of the scale factors in boosted topologies and comparison with standard ones.

# Validation Sample: Higgs Channel

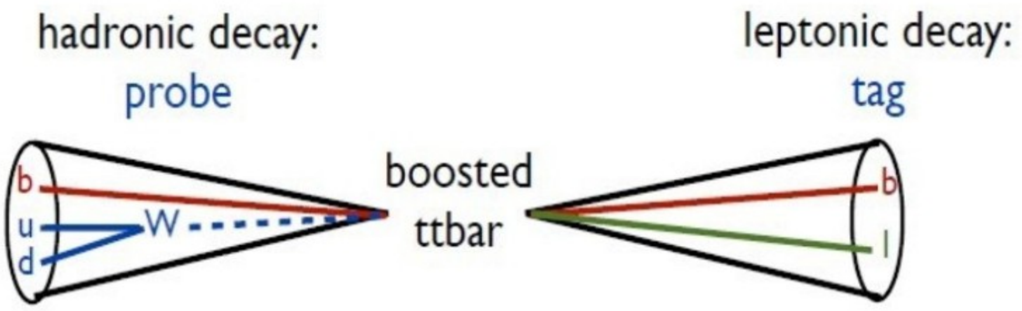
- Challenging definition of the control sample. Similar topology: **gluon splitting jets**, two closeby b's clustered in the same fat-jet.

- Event selection:
  - 1 CA8 jet, $p_T$>400 GeV, |η|<2.4;
  - ΔR(subjets)>$m_{jet}/p_T$: remove infrared unsafe configurations;
  - MC samples: inclusive and muon-enriched QCD, tt, Z→qq.

- **Muon-tag** to b-enrich subjets sample: require muon with $p_T$>5GeV within subjet cone.

- Sample of CA8 fat-jets enriched in gluon splitting, requiring **both subjets to be muon-tagged**: **Higgs-like sample**.

# Validation Sample: Top Channel



- ttbar semi-leptonic decays.

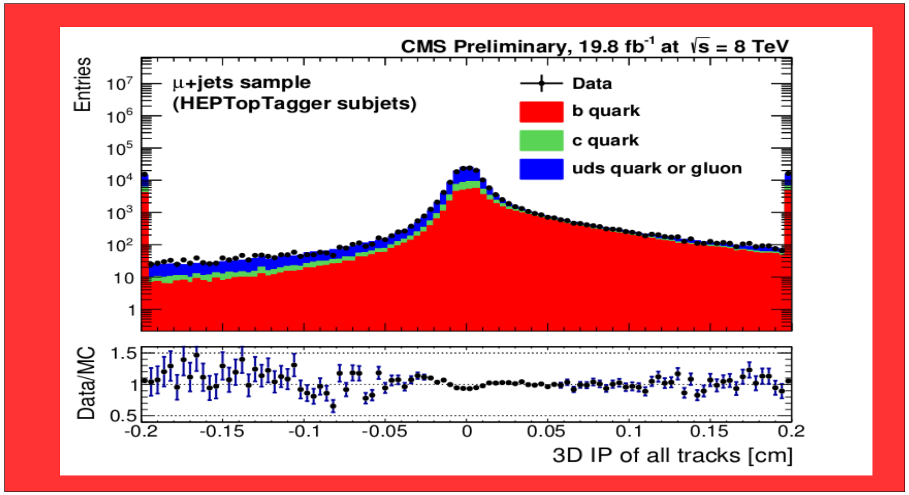
- Leptonic decay:
  - → isolated muon;
  - → 1 standard b-tag.

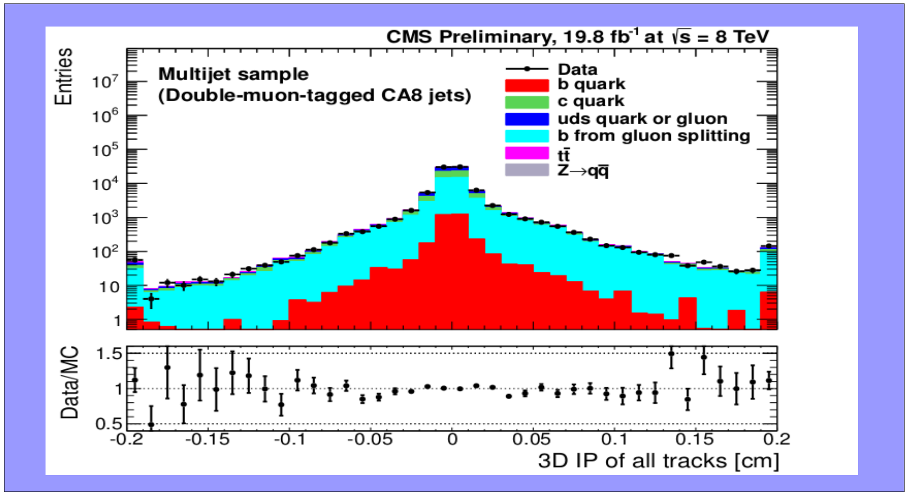
- Hadronic decay selected using HEPTopTagger.

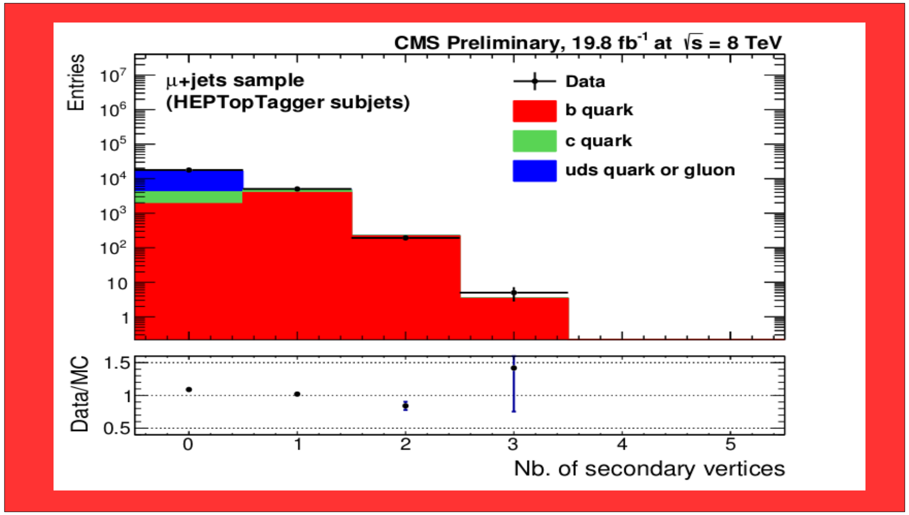
- MC samples: ttbar + all SM backgrounds (single-top, Z/W+jets).

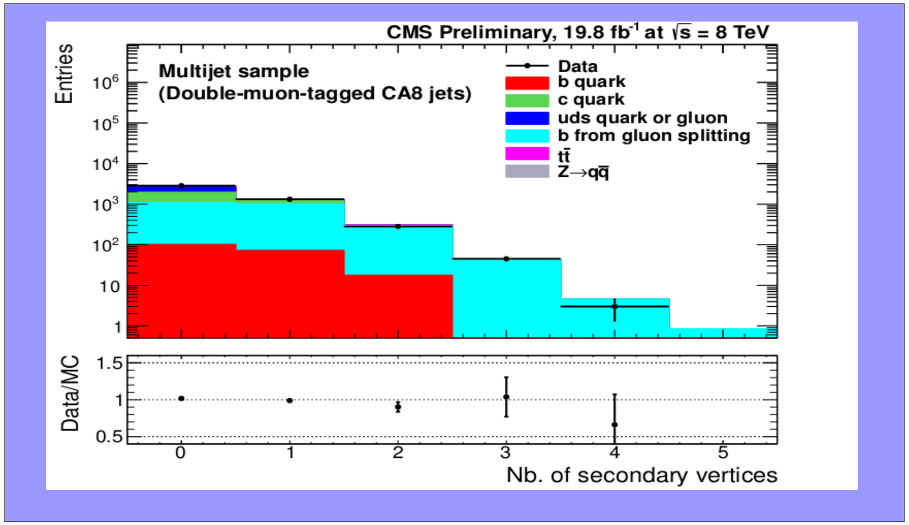
# B-Tagging Observables

- Checking data/Monte Carlo agreement for b-tagging quantities.





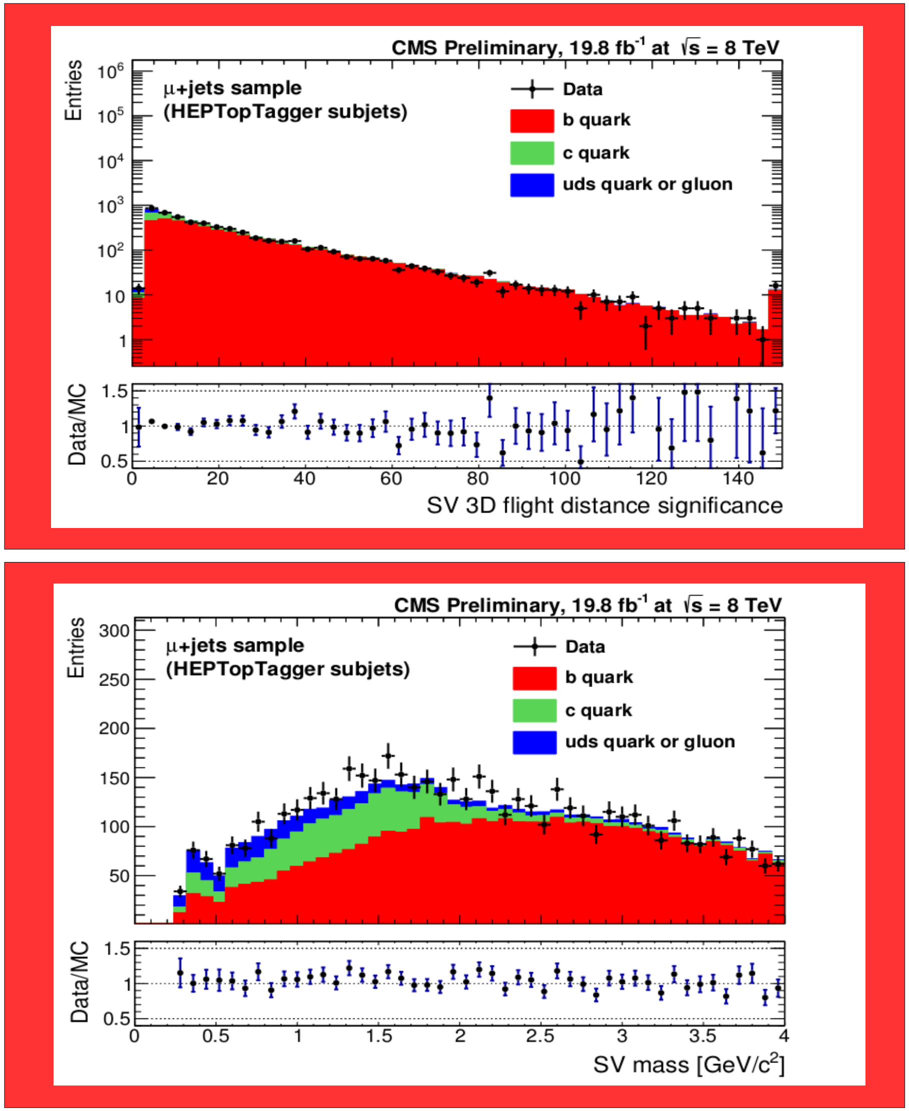




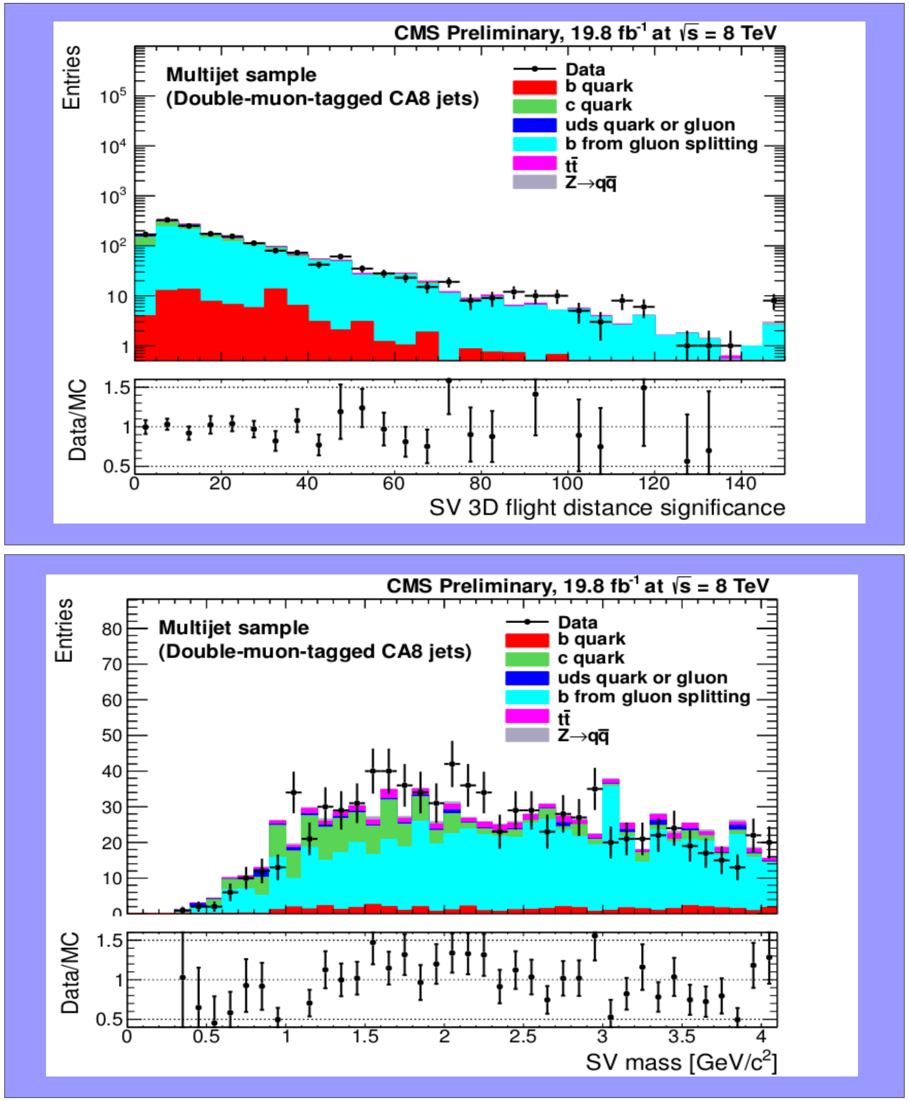
**Top channel**

**Gluon splitting**

# B-Tagging Observables

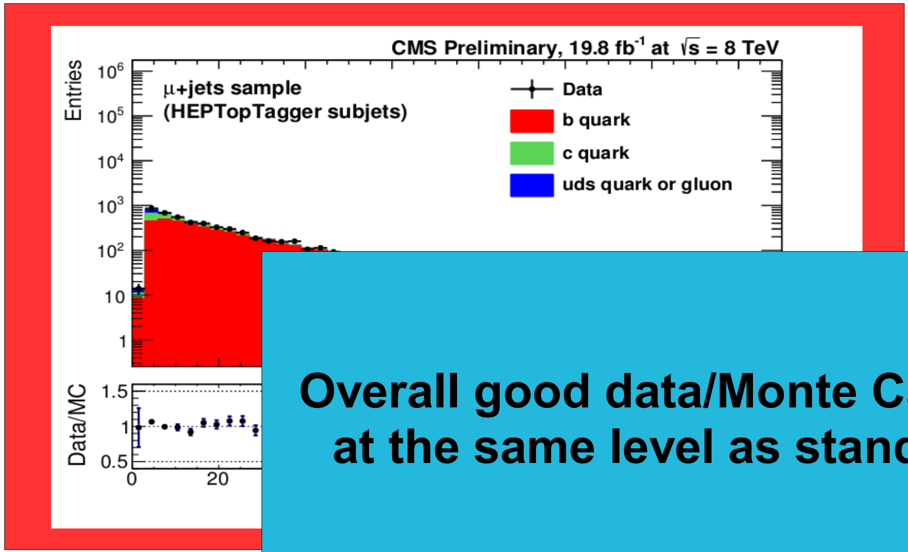


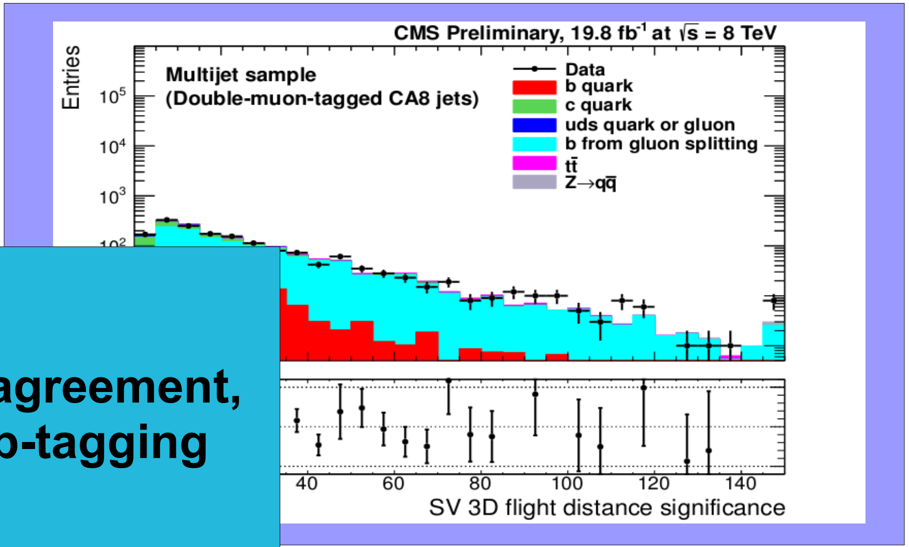
**Top channel**

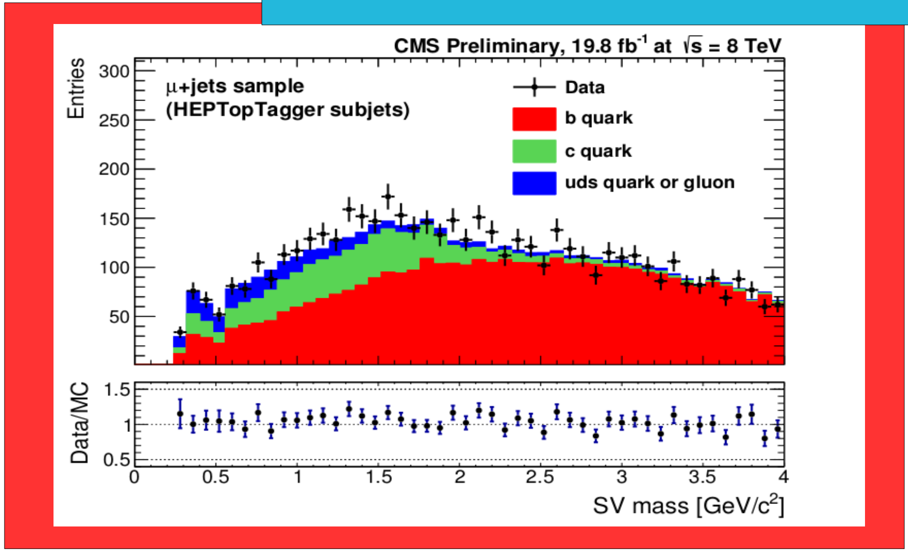


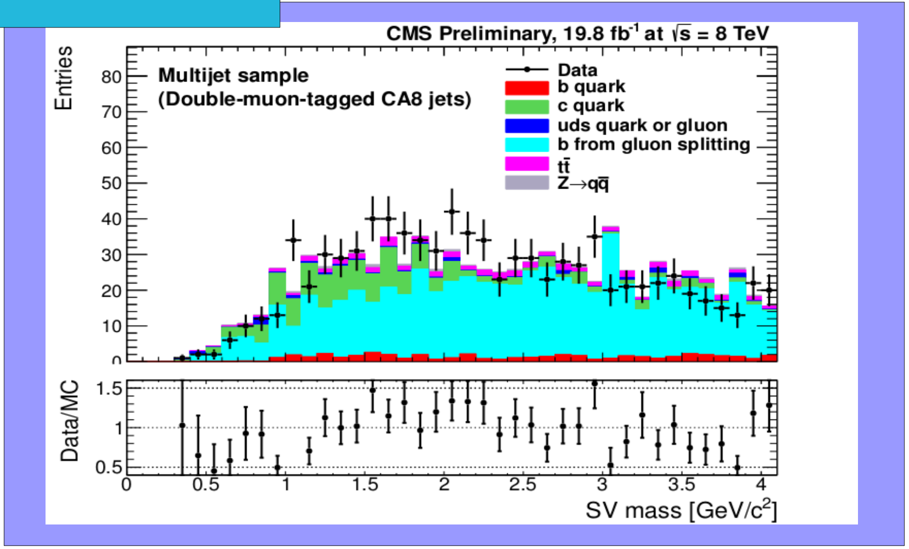
**Gluon splitting**

# B-Tagging Observables









**Overall good data/Monte Carlo agreement, at the same level as standard b-tagging**

**Top channel**

**Gluon splitting**

# Scale Factor Measurement: Higgs

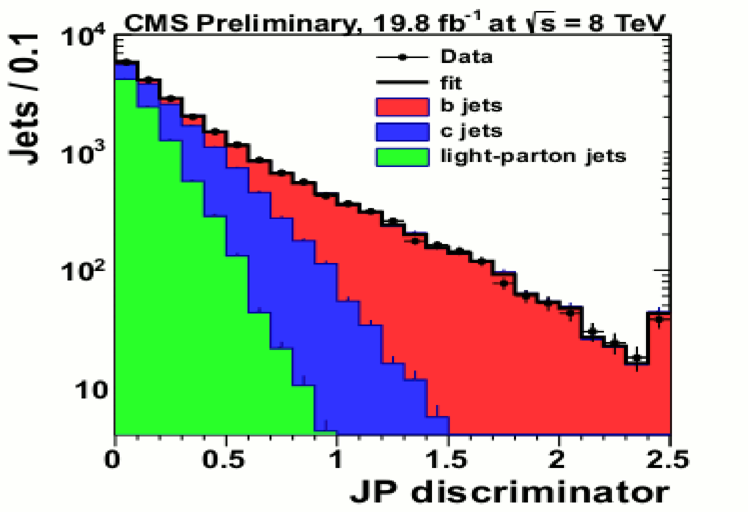
# Lifetime Tagger Method

- Method based on **Jet-Probability b-tagger**. Advantage:
  - JP discriminant can be defined for most jets (>90%);
  - calibrated on data.

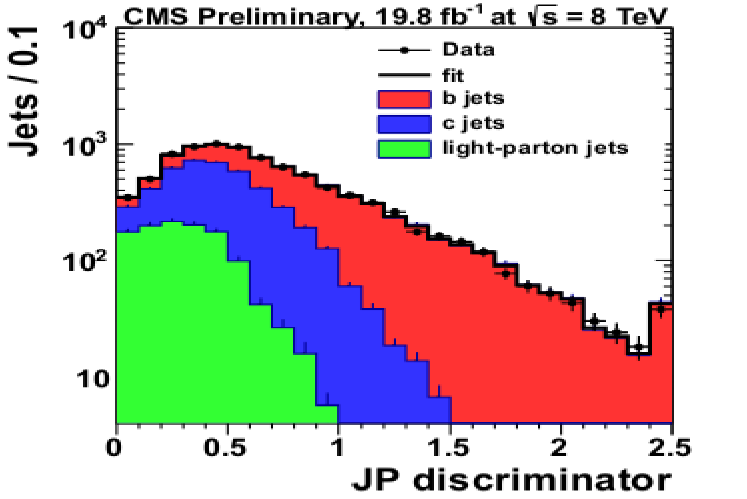
- **Template fit to JP discriminant**, before and after applying CSV. Discriminant shape from MC, while **relative flavor fractions are free parameters**.

- Tagging efficiency in data given by ($C_b$ is fraction of jets for which JP computable):

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



before b-tagging



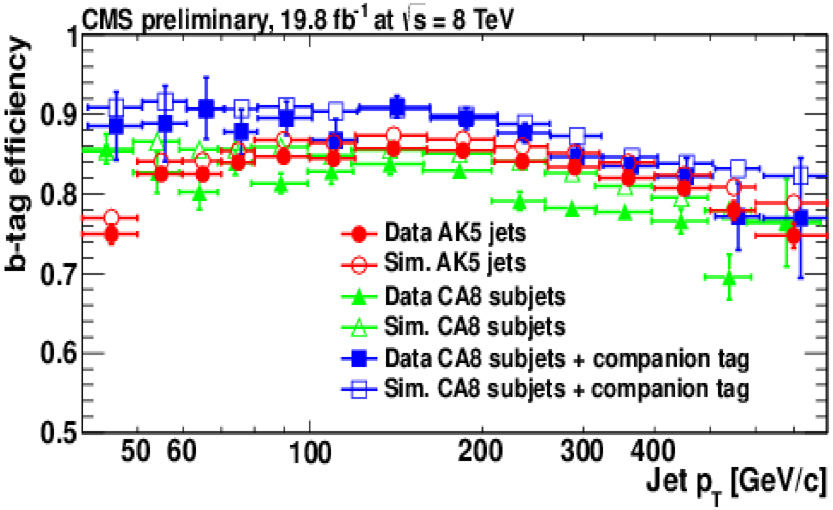
after b-tagging

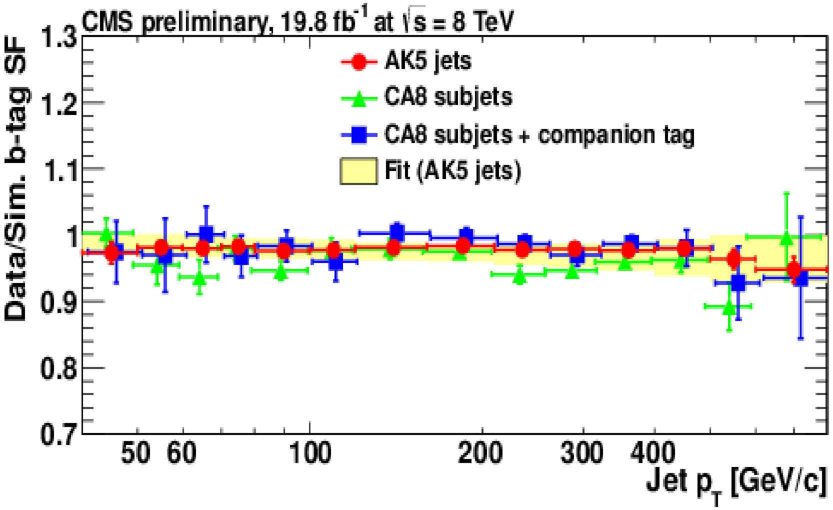
# B-tagging Scale Factor

- LT method applied to individual **muon-tagged subjets of CA8 fat jets** (**w/** and **w/o** the **companion subjet b-tagged**).

- Very good agreement with the **standard scale factors**.

- Results for the loose operating point of CSV.

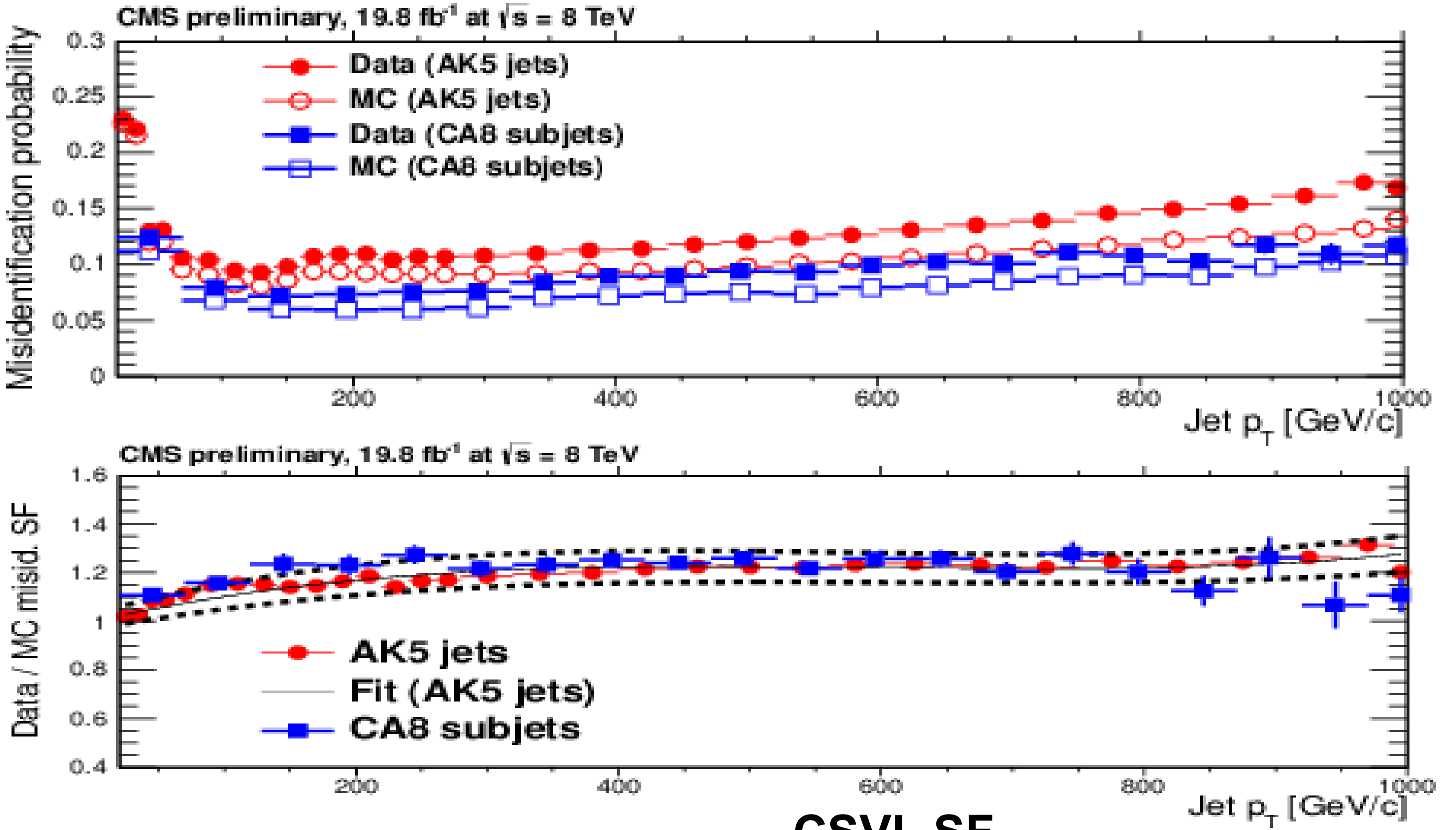




**CSVL $SF_b$**

# Mistag Scale Factor

- Measurement of **mistag rate SF$_{light}$ for CA8 subjets** based on **negative taggers**, which use tracks with negative impact parameter.
- Very good agreement with the **standard scale factors**.



**CSVL SF$_{light}$**

# Scale Factor Measurement: Top

# Flavor Tag Consistency Method

▶ Method based on distribution of **number of b-tags for the 3 subjets of CA15 HEPTopTagged fat-jet**: expected distribution fitted to data, with scale factors as free parameters.

▶ Expected number *n* of tags for ttbar signal can be expressed as:

$$\langle N_n \rangle = \mathcal{L} \cdot \sigma_{t\bar{t}} \cdot \varepsilon \cdot \sum_{i,j,k} F_{ijk} \sum_{i'+j'+k'=n}^{i'\leq i, j'\leq j, k'\leq k} [C_i^{i'} \varepsilon_b^{i'} (1-\varepsilon_b)^{(i-i')} C_j^{j'} \varepsilon_c^{j'} (1-\varepsilon_c)^{(j-j')} C_k^{k'} \varepsilon_l^{k'} (1-\varepsilon_l)^{(k-k')}]$$

→ $\boldsymbol{\varepsilon_b, \varepsilon_c, \varepsilon_l}$ are the tagging efficiencies;

→ $\boldsymbol{C^a_b}$ are the binomial coefficients;

→ ***Fijk*** are the fractions of events with *i* b-subjets, *j* c-subjets and *k* light-subjets: **taken from MC**.

→ **backgrounds included in the fit**.

# Fit Modalities

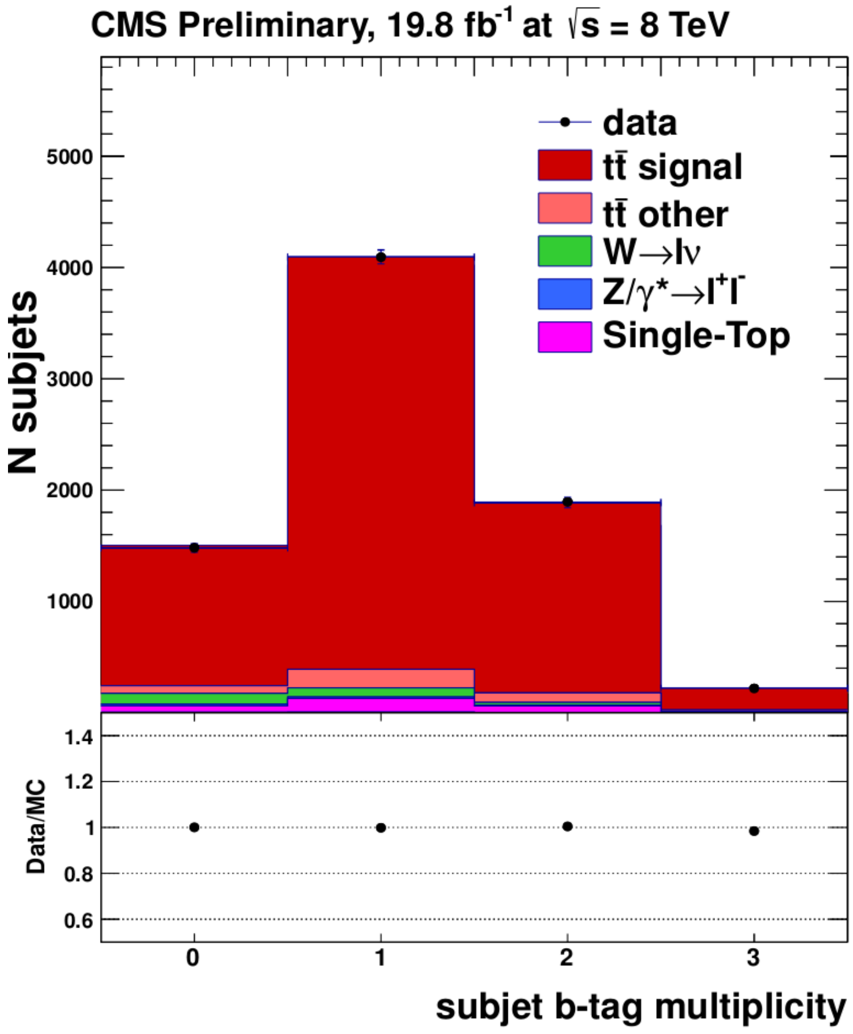
- **2 parameters** fit:

→ $\sigma_{tt}$, $SF_b$ are free parameters. Fixed $SF_c$ = $SF_b$ and fixed $SF_{light}$ to $SF_{light}$ for standard b-tagging on AK5 jets.

- **3 parameters** fit:

→$\sigma_{tt}$, $SF_b$ and $SF_{light}$ are free parameters. Fixed $SF_c$ = $SF_b$.

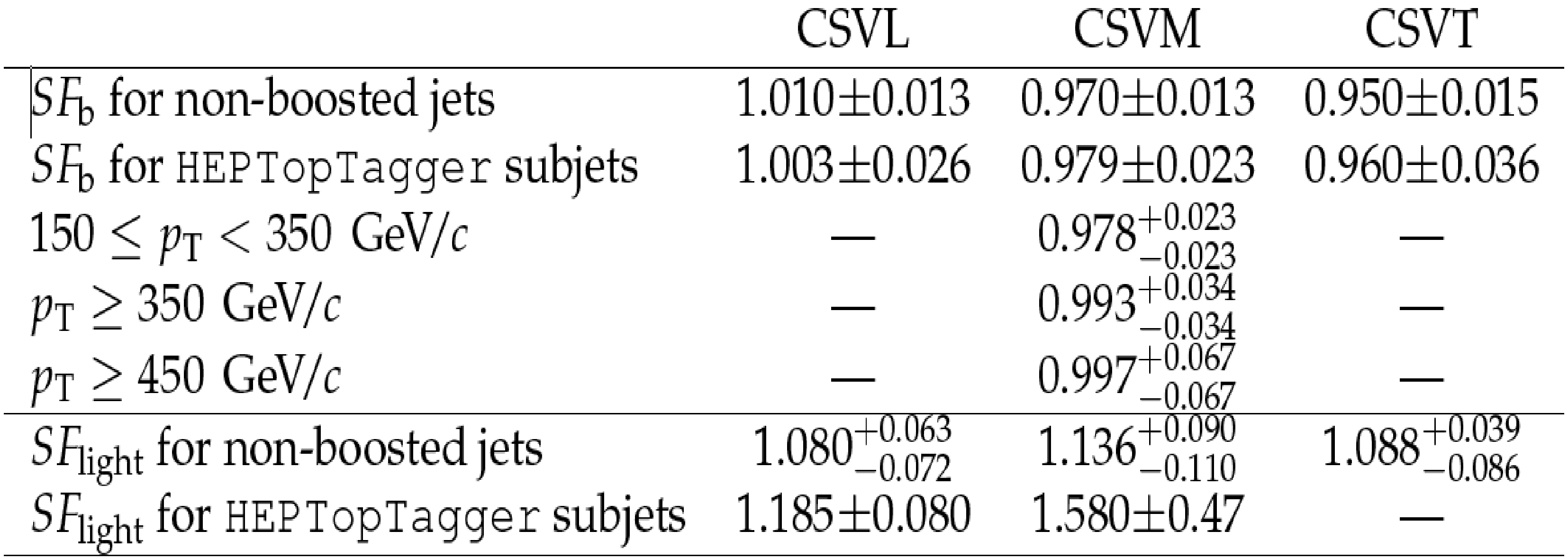
- Excellent data/MC agreement after fit of subjet b-tag multiplicity.



Post-fit distribution

# Scale Factors

- Measured $SF_b$ for boosted top subjets are in **agreement with standard $SF_b$** for AK5 jets.
- **No significant deviation at high top-$p_T$** of the measured $SF_b$.
- Mistag $SF_{light}$ are in **agreement with standard $SF_{light}$** for AK5 jets.



| | | CSVL | CSVM | CSVT |
|---|---|---|---|---|
| **$SF_b$** | $SF_b$ for non-boosted jets | 1.010±0.013 | 0.970±0.013 | 0.950±0.015 |
| | $SF_b$ for HEPTopTagger subjets | 1.003±0.026 | 0.979±0.023 | 0.960±0.036 |
| **$SF_b$ pT dependence** | $150 \leq p_T < 350$ GeV/$c$ | — | $0.978^{+0.023}_{-0.023}$ | — |
| | $p_T \geq 350$ GeV/$c$ | — | $0.993^{+0.034}_{-0.034}$ | — |
| | $p_T \geq 450$ GeV/$c$ | — | $0.997^{+0.067}_{-0.067}$ | — |
| **$SF_{light}$** | $SF_{light}$ for non-boosted jets | $1.080^{+0.063}_{-0.072}$ | $1.136^{+0.090}_{-0.110}$ | $1.088^{+0.039}_{-0.086}$ |
| | $SF_{light}$ for HEPTopTagger subjets | 1.185±0.080 | 1.580±0.47 | — |

# Conclusions

▶ First step into **integration of b-tagging and subtructure techniques** used in boosted topologies.

▶ Monte Carlo studies have identified **subjet b-tagging** as the optimal b-tagging technique in the boosted regime.

▶ Dedicated samples defined to study subjet b-tagging in boosted top and boosted Higgs-like topologies.

▶ A detailed study of track and secondary vertex variables for subjet b-tagging confirms a **similar level of data/MC agreement as for the standard b-tagging**.

▶ Standard scale factors for AK5 jets and **measured scale factors for the considered boosted topologies** show an excellent agreement.

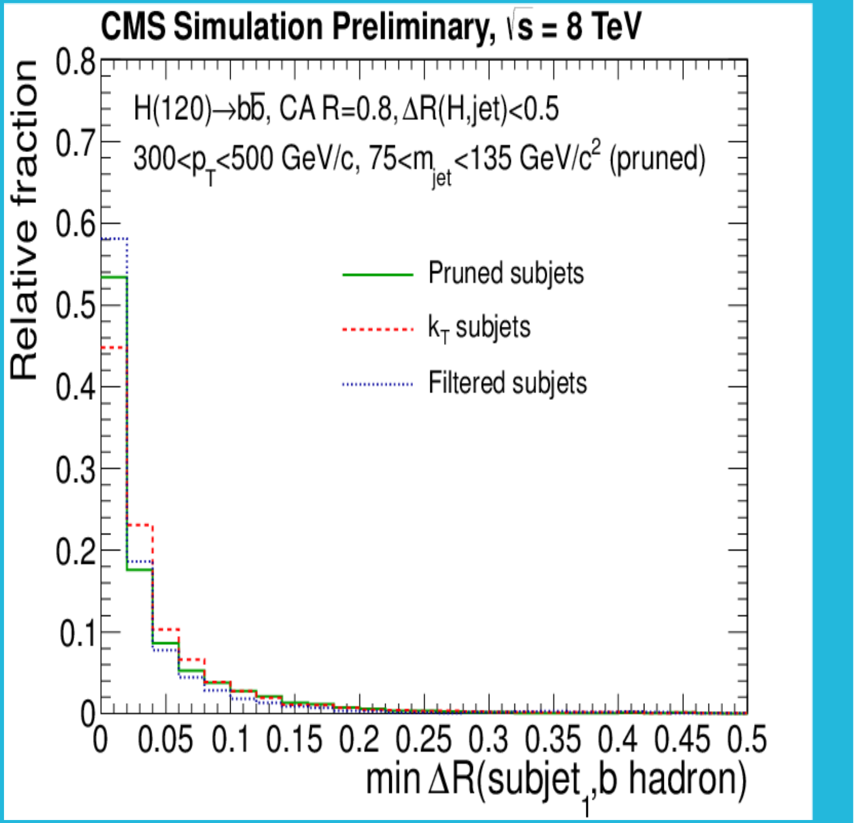
# Thank You!

**Public twiki:**
**https://twiki.cern.ch/twiki/bin/view/CMSPublic/PhysicsResultsBTV13001**

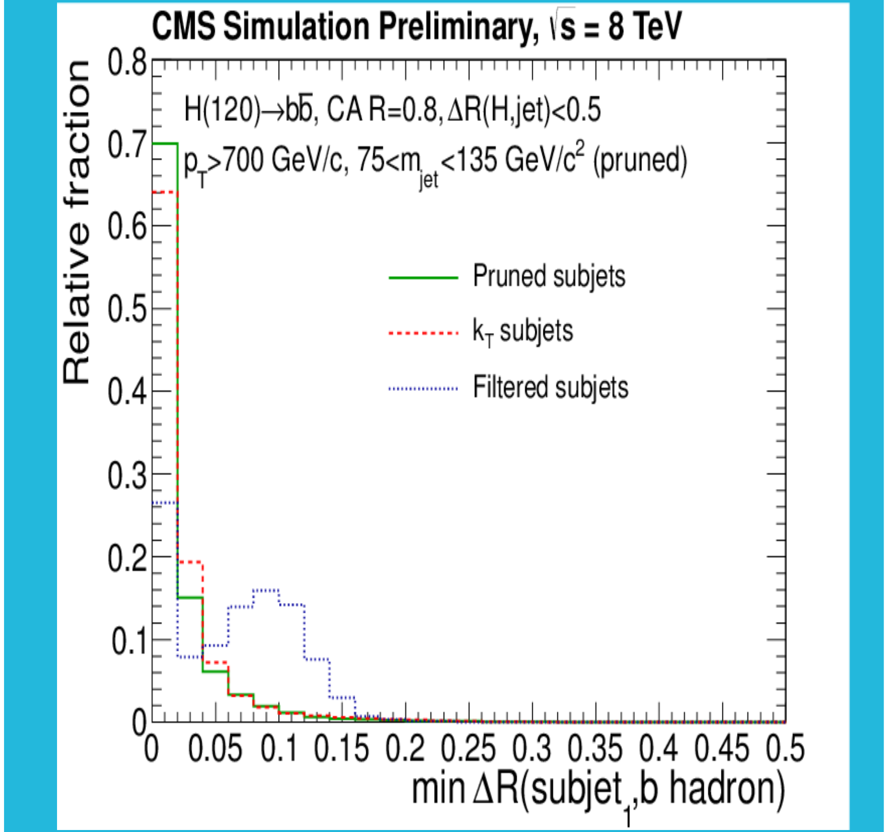
# Additional Slides

# Subjets Alignment: H→bb

- MC study on the **alignment of subjets** with b-hadron direction, for different subjet clustering techniques.
- Pruned subjets reproduce overall at best the b-hadron direction.
- Filtering with fixed cone size: dynamic cone size could improve.



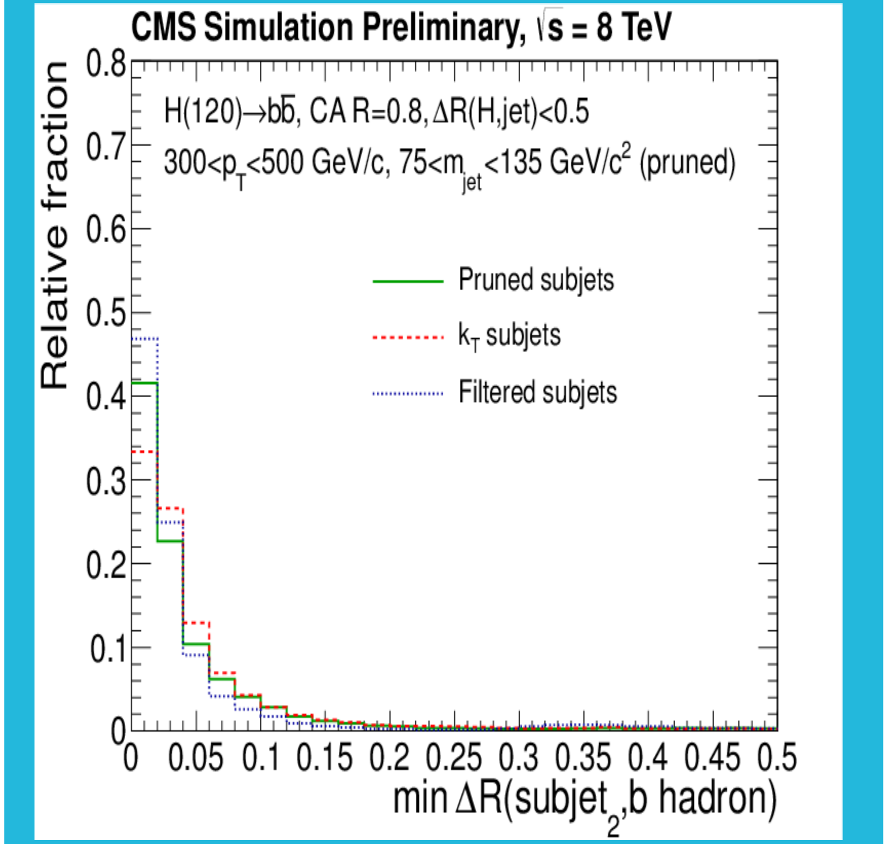
**300 < $p_T$ < 500 GeV**



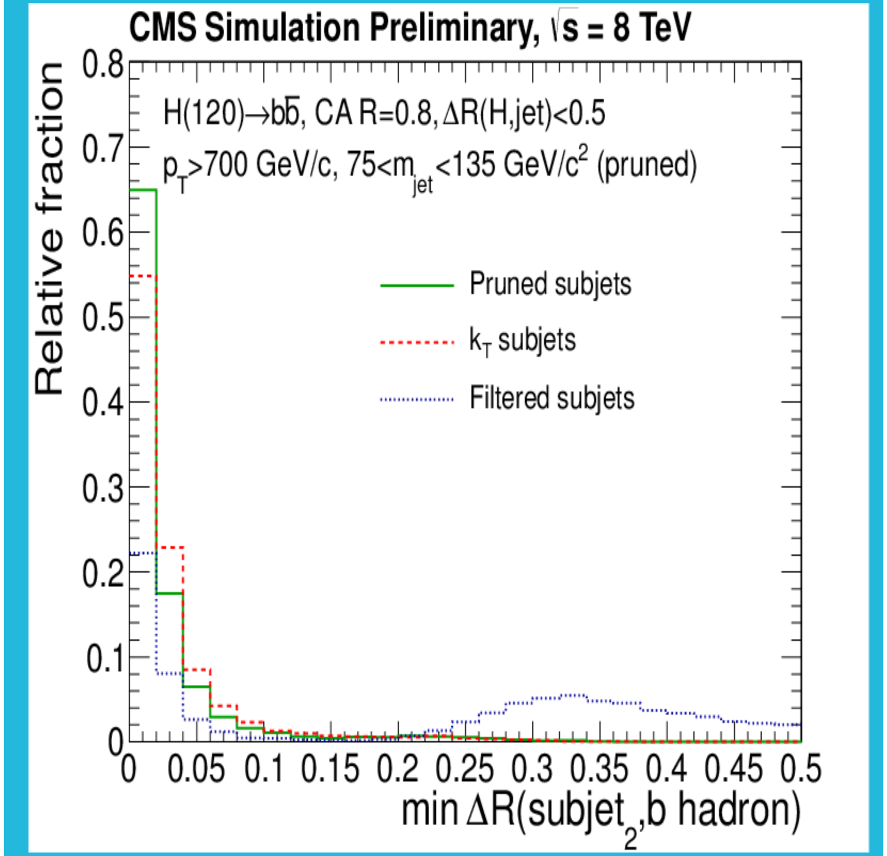
**$p_T$ > 700 GeV**

# Subjets Alignment: H→bb

- Second subjet.



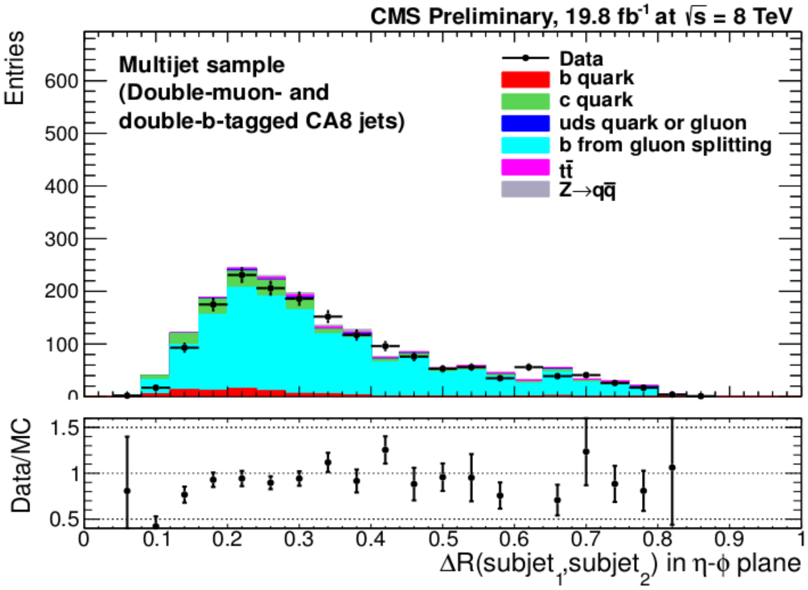
300 < $p_T$ < 500 GeV



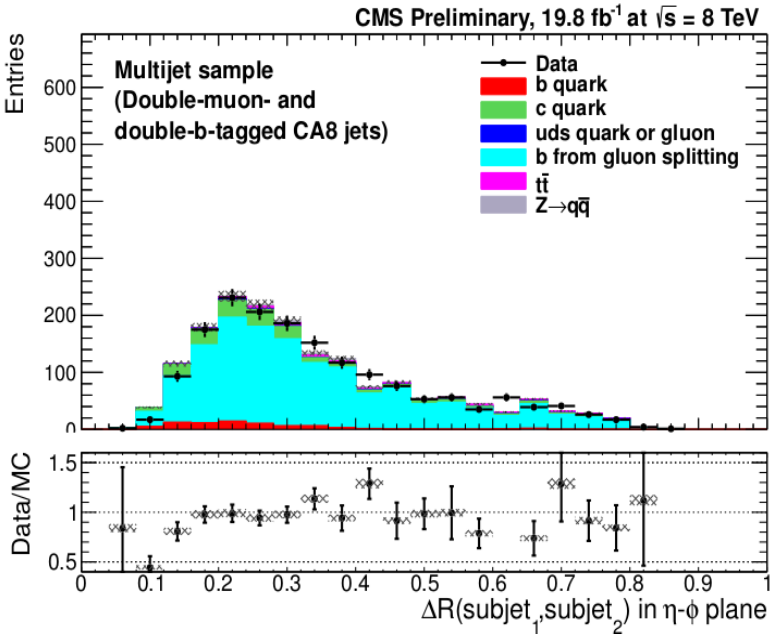
$p_T$ > 700 GeV

# Additional Cross-Check

- Select a sample of double-muon-tagged fat jets enriched in gluon splitting b-jets, likely to contain two b-quarks in a single fat jet, as a control sample.
- Require double CSVL b-tag.
- **Data/MC ratio consistent with unity** after applying standard SF for AK5 jets.



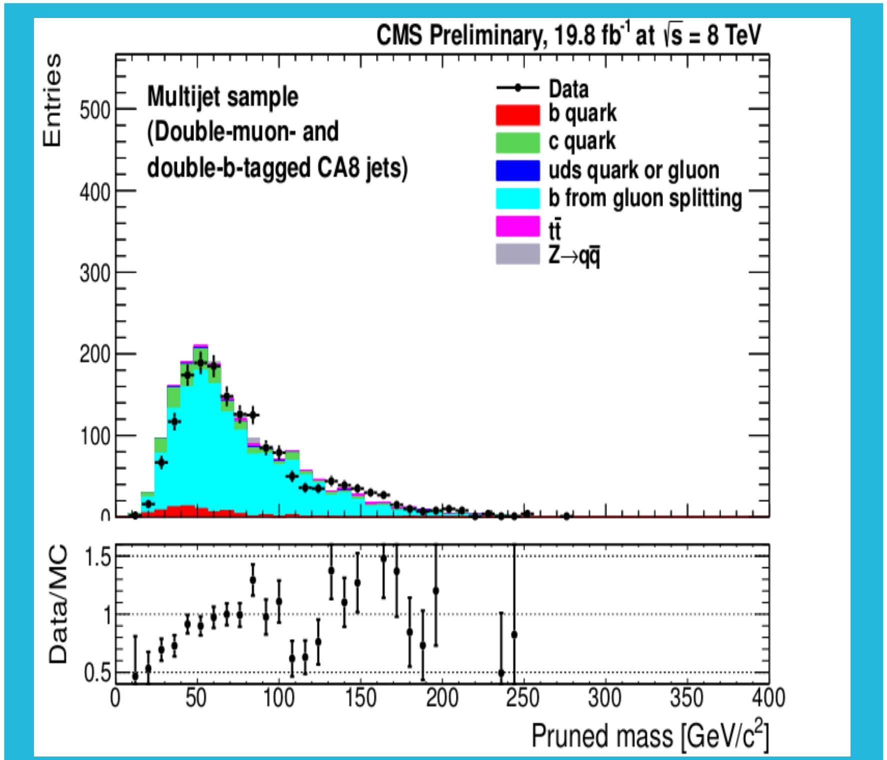
**before SF:**
**data/MC = 0.94 ±0.03 (stat.)**



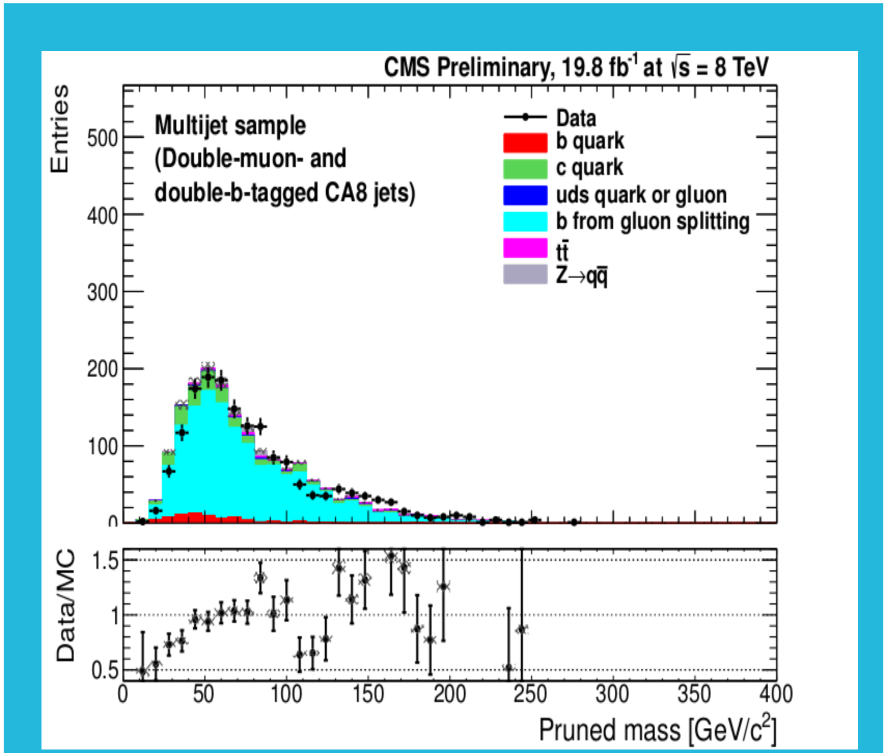
**after standard SF: (SF syst propagated)**
**data/MC = 0.98 ± 0.03 (stat.)$^{+0.04}_{-0.05}$ (syst.)**

# Pruned Mass: H→bb

- CA8 fat-jet sample, with **both subjets muon-tagged and CSVL b-tagged**.
- **Left**: no scale factors applied. **Right**: $SF_b$ applied. Only $SF_b$ applied, as the sample is largely heavy flavor dominated, due to the double muon-tag.



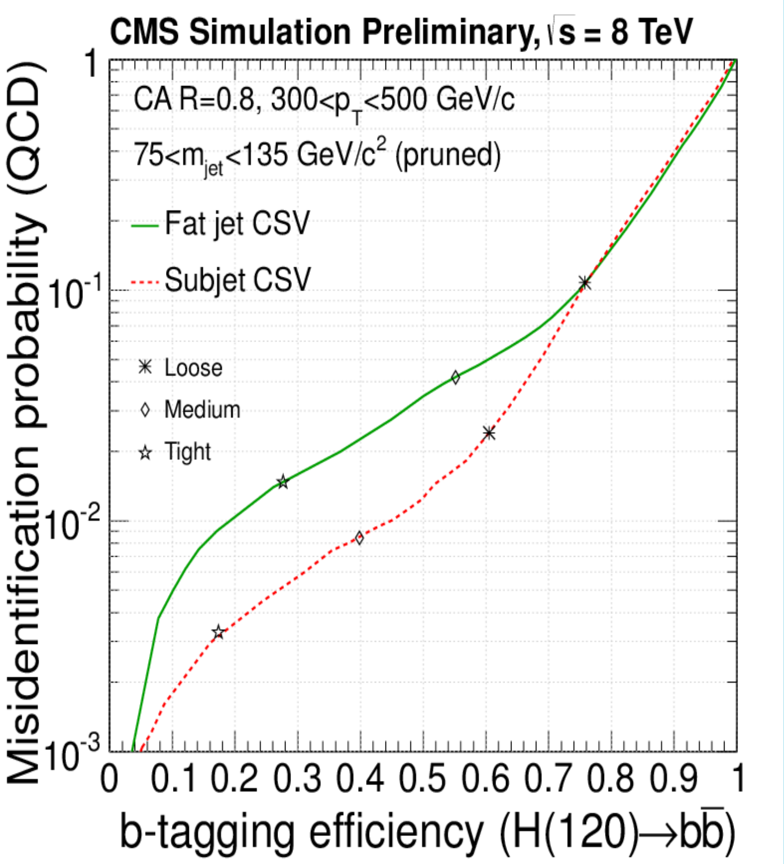
no SF



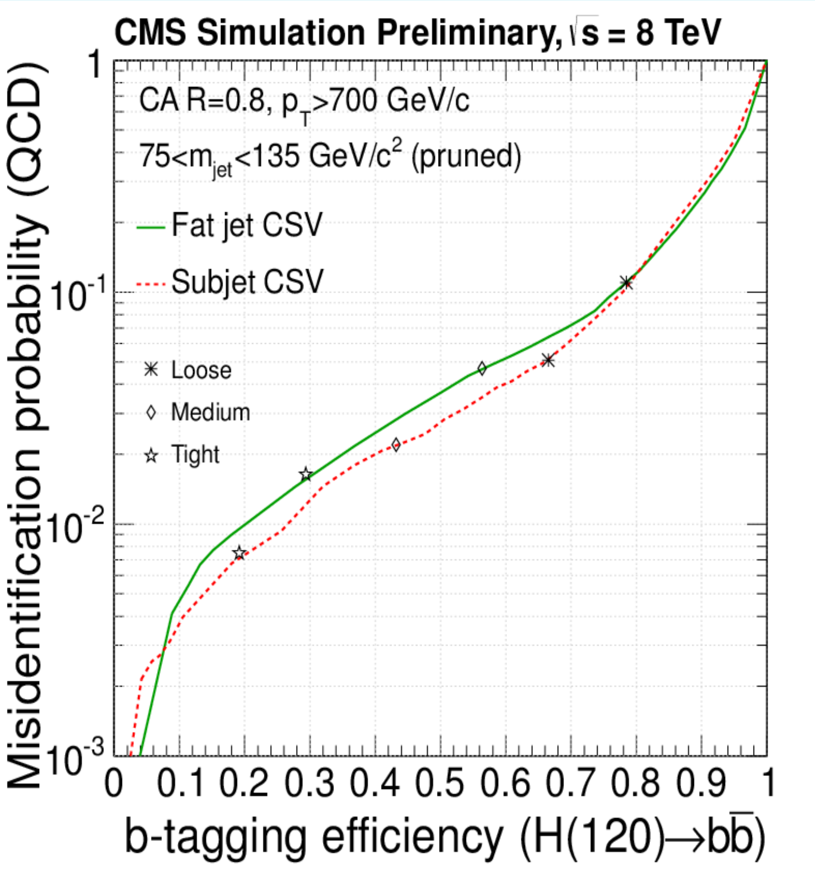
corrected for $SF_b$

# Standard Track Selection

- Performance of subjet b-tagging and fat-jet b-tagging in the Higgs channel, using standard track selection for the fat-jet b-tagging ($\Delta R<0.3$).



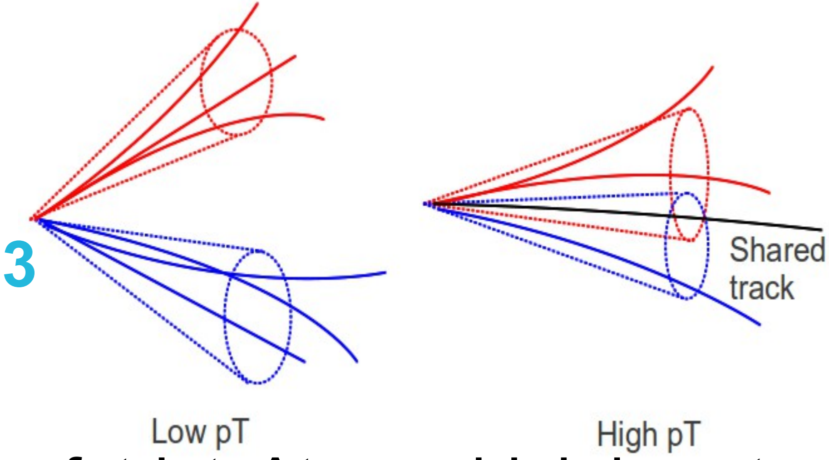
**medium boost region**



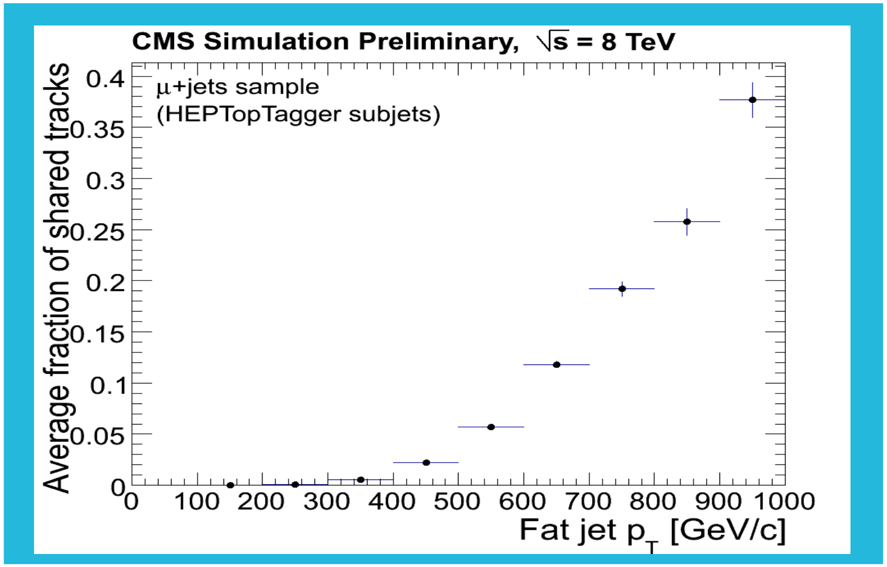
**large boost region**

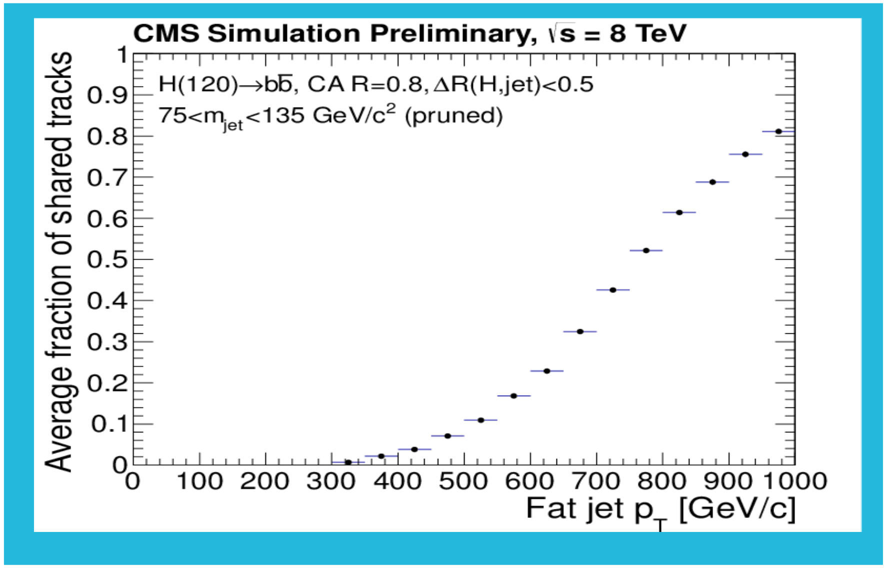
# Track Sharing

- Cross-check of **sharing of tracks selected** for b-tagging between subjets.
- Considere tracks in a cone of **$\Delta R<0.3$** around subjet axis (as used by CSV).



- Track-sharing increases with $p_T$ of the fat-jet. At very high boost, the level of track sharing becomes significantly large. One solution is to switch to fat-jet b tagging.





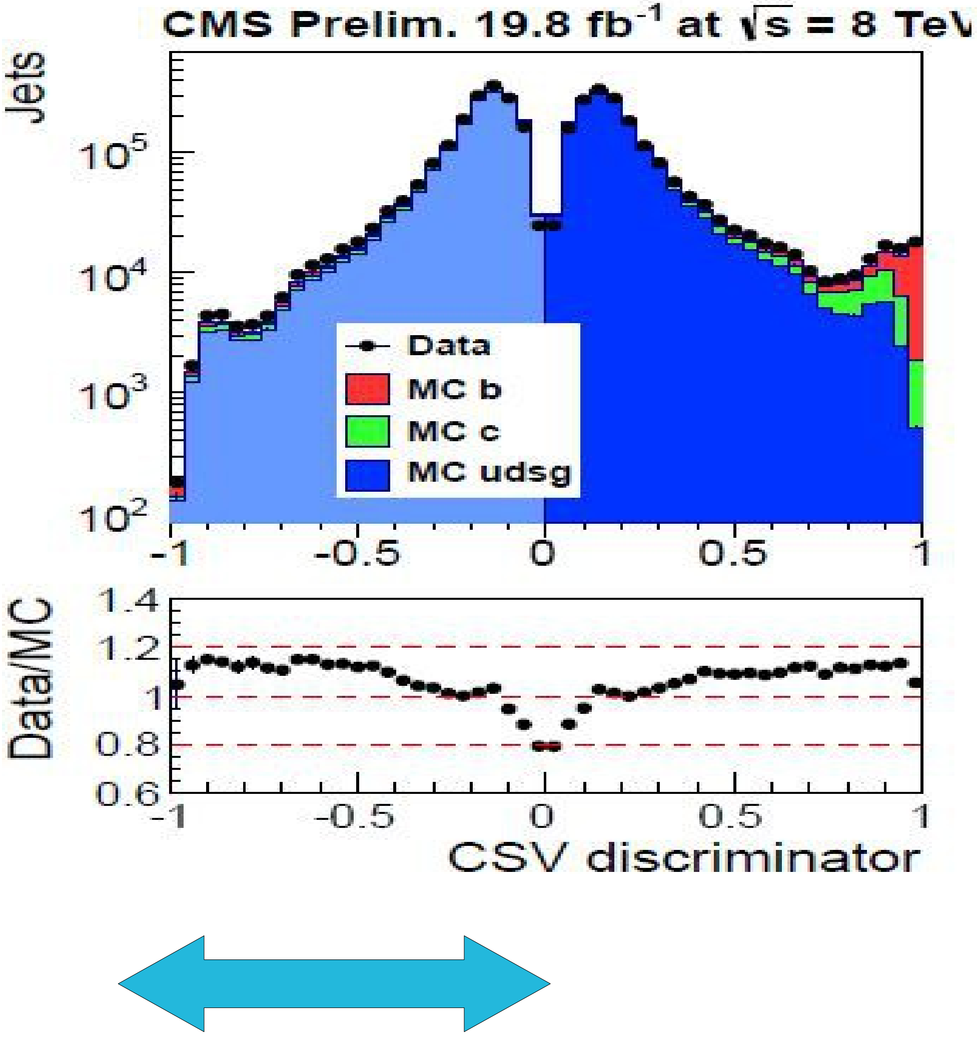
# Mistag SF

- Use tracks with negative IP or SV with negative decay length to define a **negative tagger for each tagger**.
- Scale factor for mistag obtained according to:

$$\begin{tikzcd}
\varepsilon_{MC}^{mistag} \arrow[r, "\mathbf{SF}_{\mathbf{mistag}}"] & \varepsilon_{data}^{mistag} \\
\varepsilon_{MC}^{neg\ tag} \arrow[u, "R_{MC}^{light}"] \arrow[r, "SF_{neg}"'] & \varepsilon_{data}^{neg\ tag} \arrow[u, "R_{data}^{light}"']
\end{tikzcd}$$

given by:

$$\mathbf{SF}_{\mathbf{mistag}} = SF_{neg\ tag} \cdot \frac{R_{light}^{data}}{R_{light}^{MC}}$$



**negative tagger**

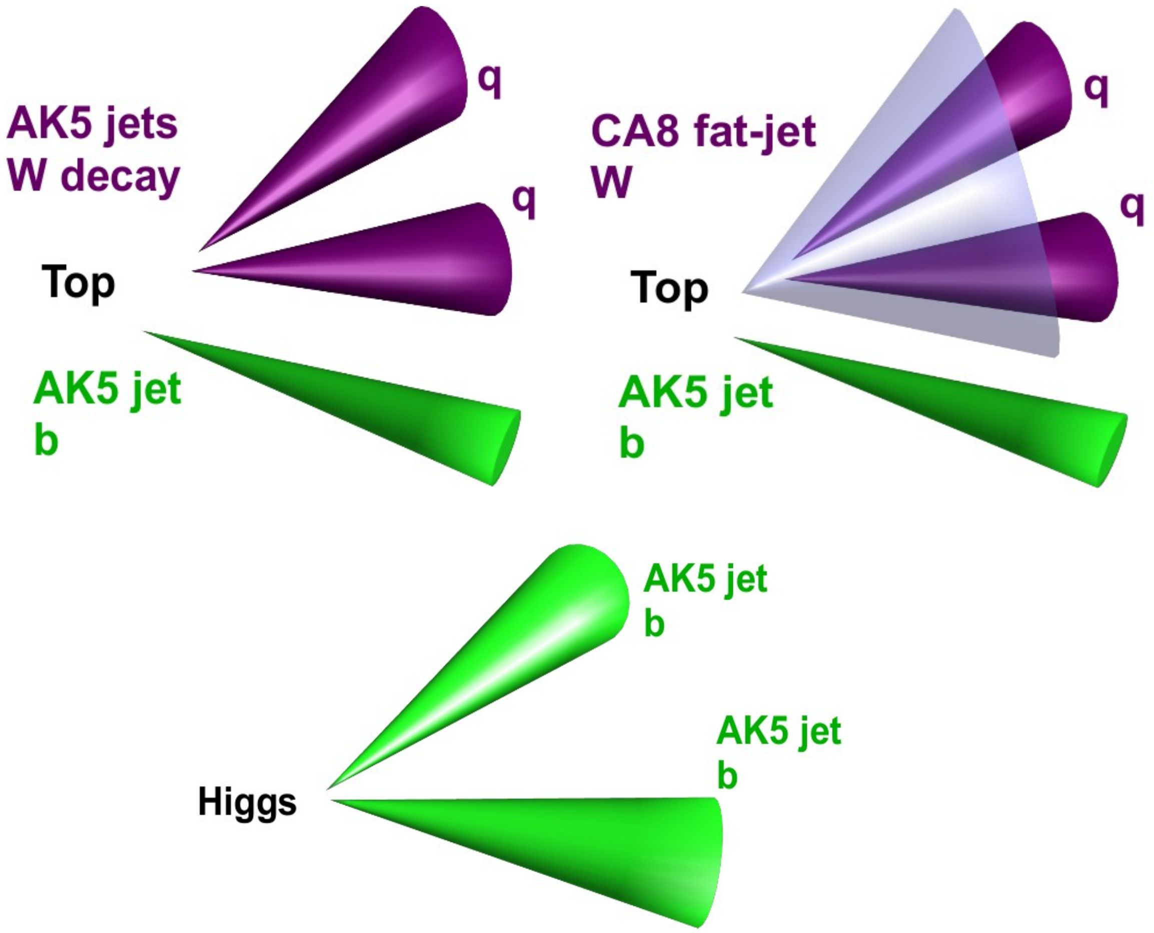
# Traditional B-Tagging

B-tagging at CMS traditionally developed on **isolated AK5 jets**, mostly suitable for the **non-boosted regime**.

**Hadronic top decay:**

**→we can apply standard b-tagging if b clustered in isolated AK5 jet**

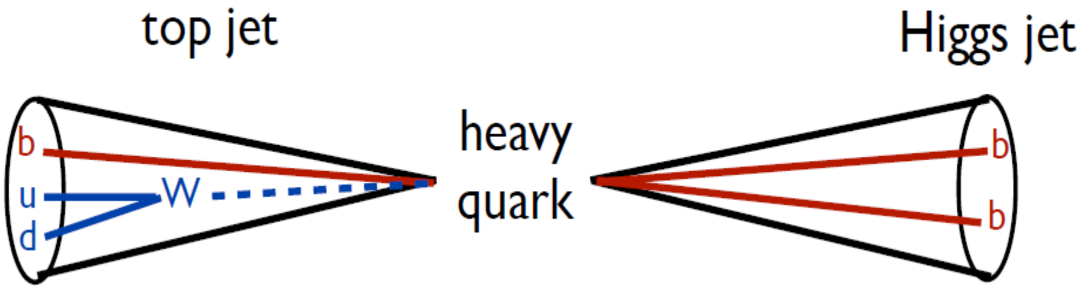
**→separate CA8 W fat-jet, or two AK5 jets from W decay**



**Higgs→b$\bar{b}$:**

**→traditional b-tagging possible if 2 separate AK5 b-jets**

# Motivation



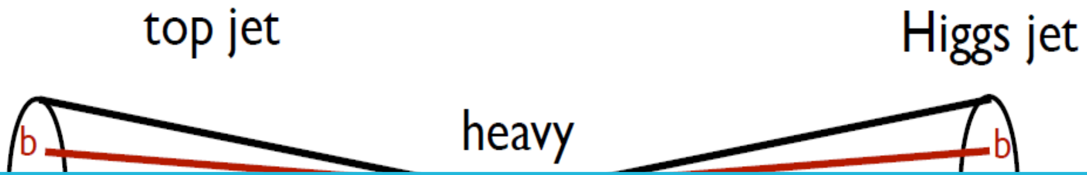
(similar for Z→hadrons)

- Several search channels feature boosted tops and Higgs:
  - searches for heavy T/B quarks from **vector-like 4th generation** of quarks, decay mode **T→tH, B→bH**;
  - boosted BSM resonances **Z'→tt**, **W'→tb**;
  - **RS graviton** and **BSM heavy Higgs** decaying into SM Higgs.

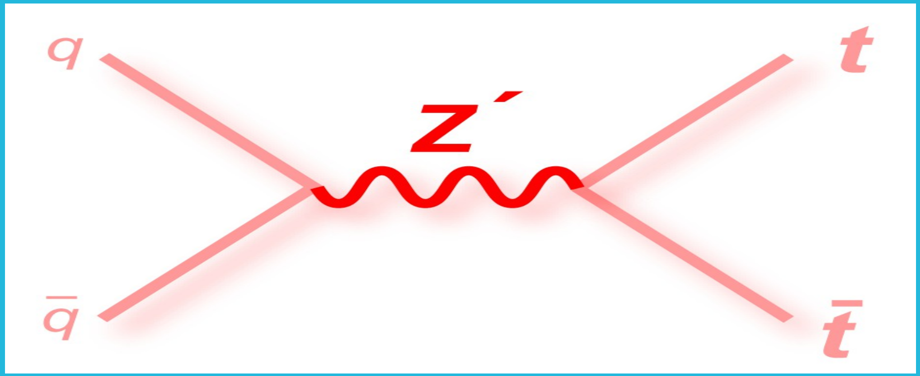
- 125 GeV Higgs gives large BR **H→bb**: double b-tagging useful selection tool. Higgs boosted, b's can overlay, non-standard b-tagging regime.

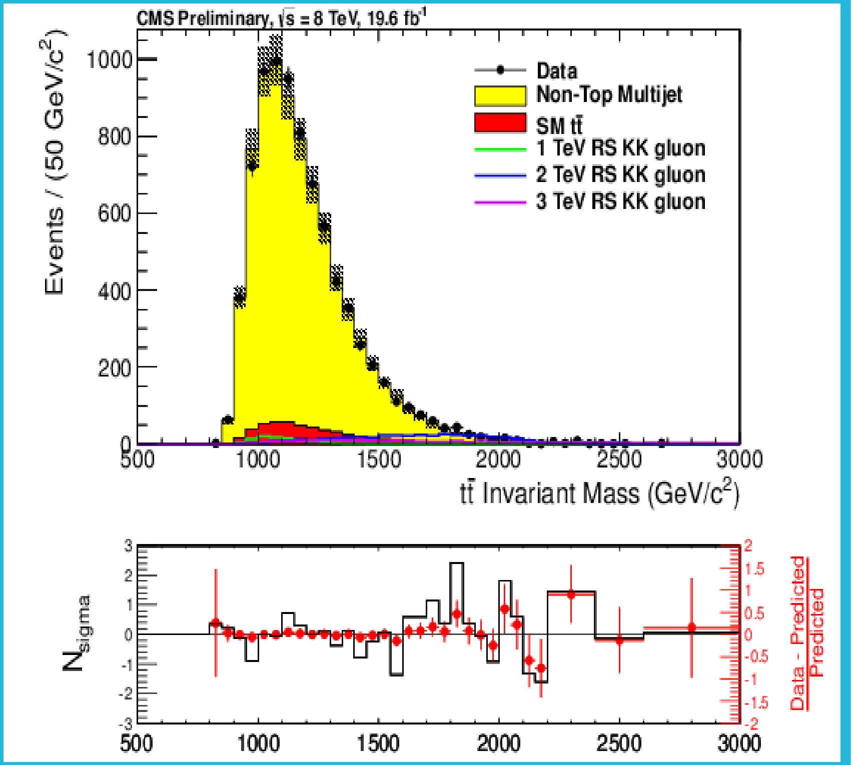
- Boosted tops: we have top-taggers, but additional b-tagging can **dramatically reduce QCD background**.

# Motivation



**Example: search for a BSM Z' → ttbar fully hadronic decay**





→CMS-TopTagger to identify tops

→High efficiency working point: dominating reducible QCD background

→Potential improvement by moving to higher purities: top+b-tagging

**CMS PAS B2G-12-005**

tagging regime:

▶Boosted tops: we have top-taggers, but additional b-tagging can **dramatically reduce QCD background**.

# Fit Modalities

▶ **2 parameters** fit:

→ $\sigma_{tt}$, $SF_b$ are free parameters. Fixed $SF_c = SF_b$ and fixed $SF_{light}$ to $SF_{light}$ for standard b-tagging on AK5 jets.

▶ 2 parameters **fit for different ranges of top fat-jet $p_T$**:

→ test deviations from standard SF in a very boosted sub-sample of tops.

▶ **3 parameters** fit:

→ $\sigma_{tt}$, $SF_b$ and $SF_{light}$ are free parameters. Fixed $SF_c = SF_b$.

**Systematic uncertainties** considered:

→ 2% subjets with no assigned flavor;

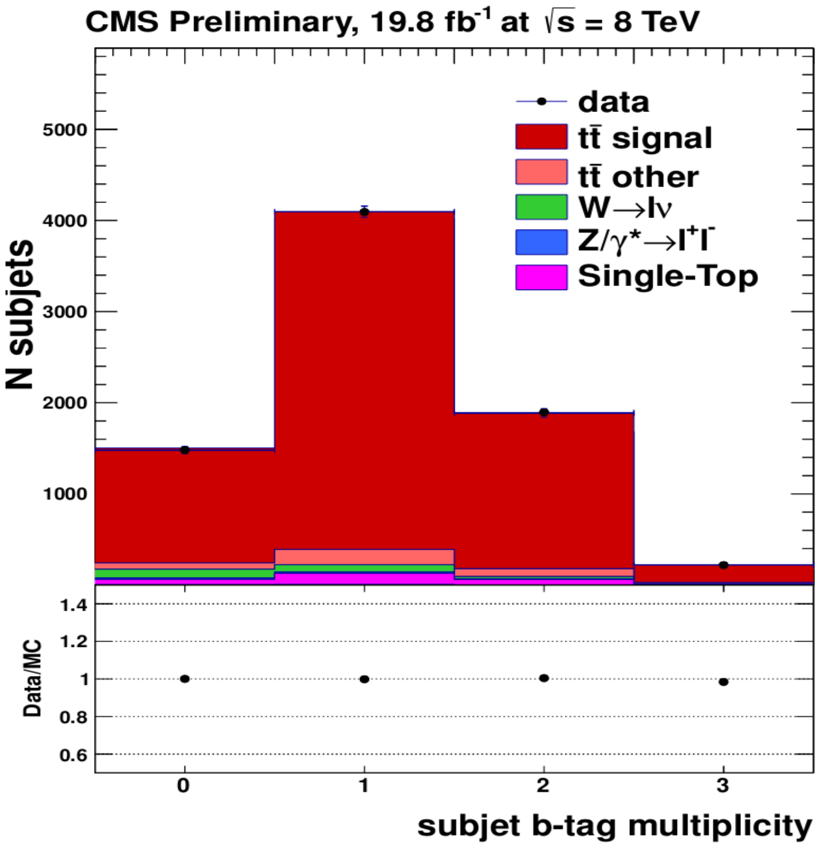
→ 50% uncertainty on background normalization, 15% on ttbar normalization;

→ uncertainties on $SF_c$ and on $SF_{light}$ (when fixed, 2 parameters fit).

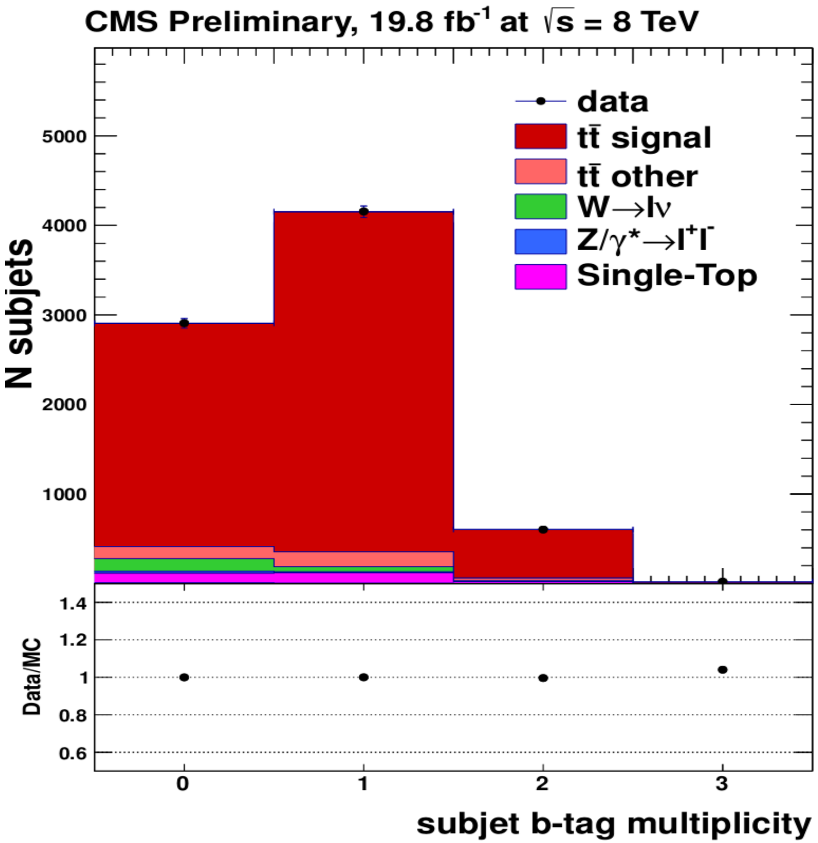
# Subjet b-tag Multiplicity

- After the fit very good agreement between data and Monte Carlo for the subjet b-tag multiplicity distributions (here for 3 parameters fit).

**Loose operating point**



**Medium operating point**



# Validation Sample: Higgs Channel

- Challenging definition of the control sample. Similar topology: **gluon splitting jets**, two closeby b's clustered in the same fat-jet.

- Event selection:
  - 1 CA8 jet, $p_T > 400$ GeV, $|\eta| < 2.4$;
  - $\Delta R(\text{subjets}) > m_{jet}/p_T$: remove infrared unsafe configurations;
  - MC samples: inclusive and muon-enriched QCD, tt, Z→qq.

- **Muon-tag** to b-enrich subjets sample: require muon with $p_T > 5$GeV within subjet cone.

- **3 samples with different flavor composition** considered:
  - inclusive sample of CA8 fat-jets;
  - sample of muon-tagged subjets of CA8 fat-jets;
  - sample of CA8 fat-jets enriched in gluon splitting, requiring **both subjets to contain soft-muon**: **Higgs-like sample**.

# B-tagging Observables

- Checking data/Monte Carlo agreement for b-tagging quantities. Presentation ordering:

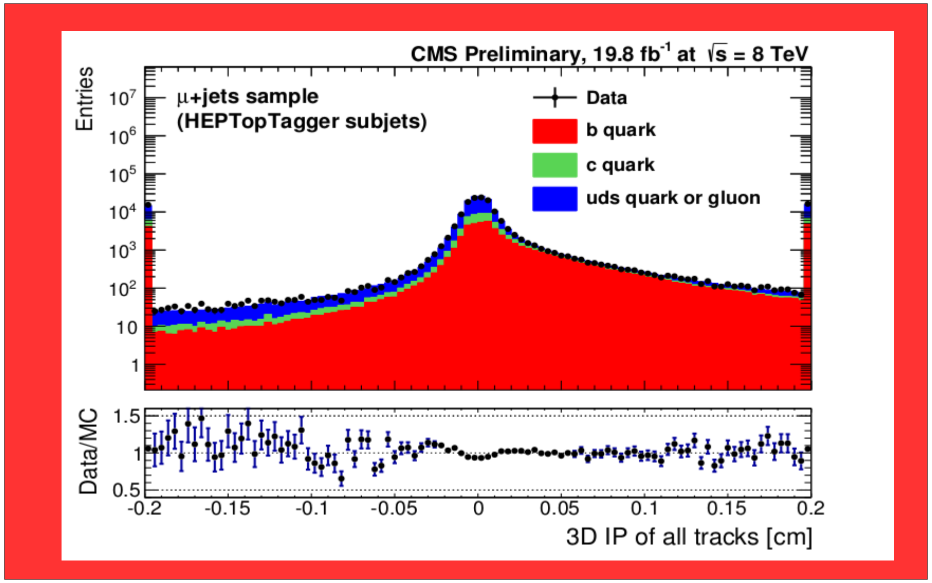
**Top channel validation:**
**HEPTopTagger**
**Subjets**

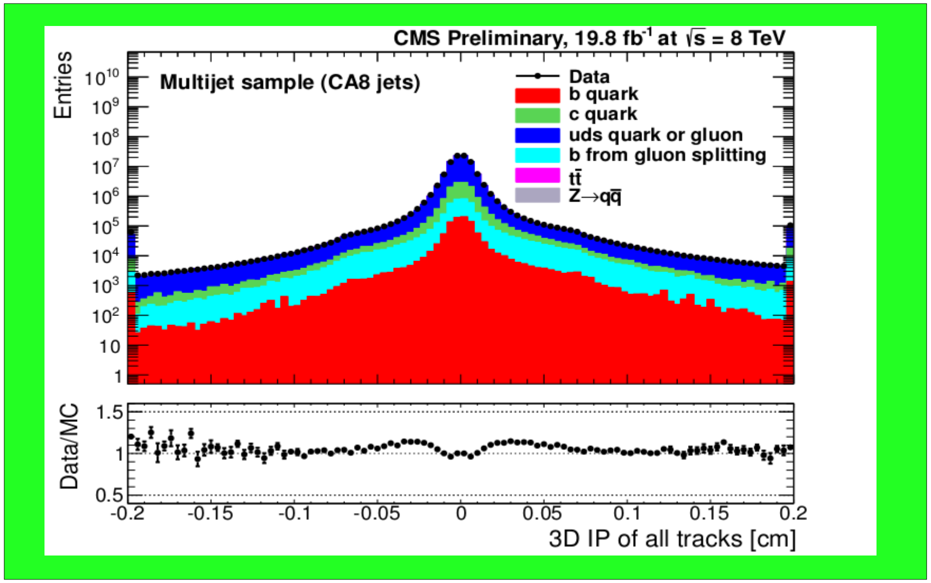
**Higgs channel validation:**
**Multijet sample**
**(CA8 jets)**

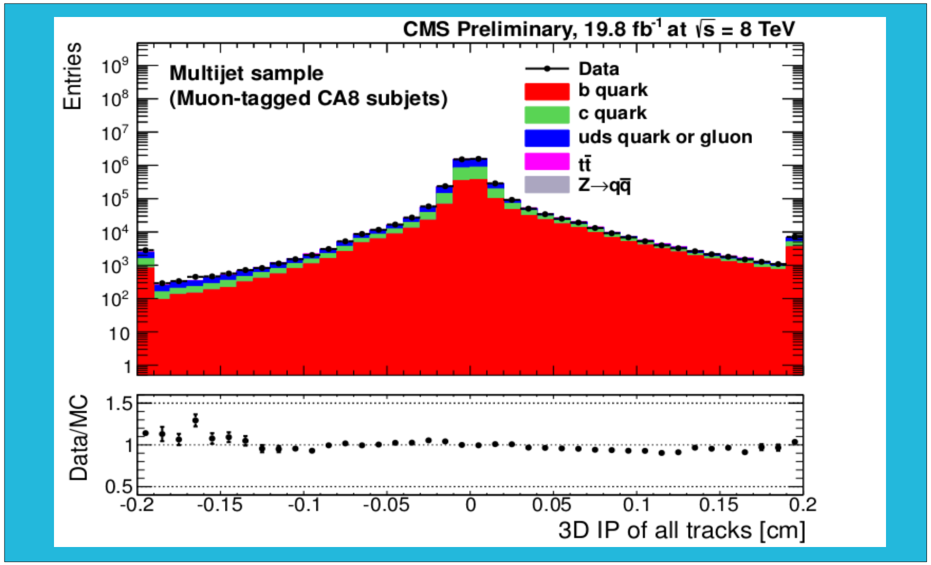
**Higgs channel validation:**
**Multijet sample**
**(CA8 muon-tagged subjets)**

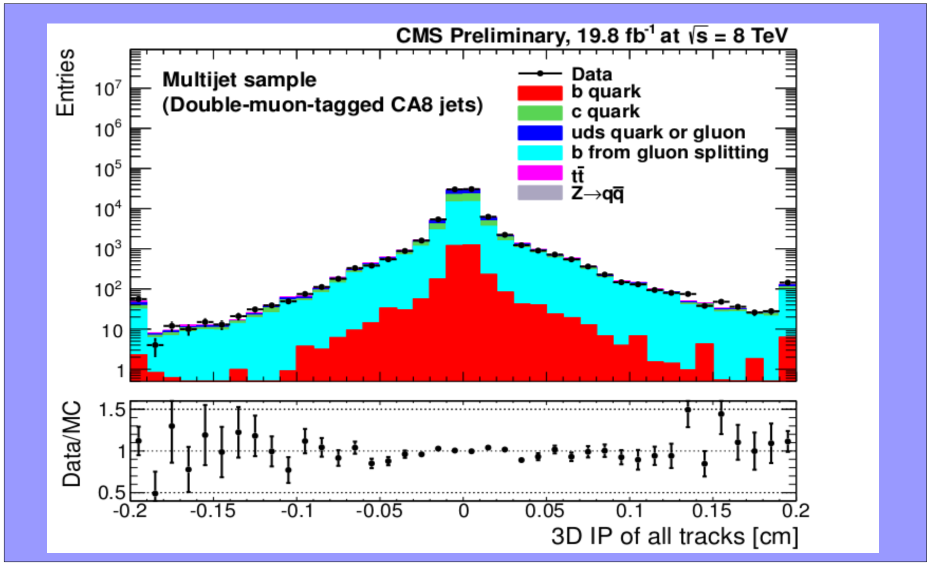
**Higgs channel validation:**
**Multijet sample**
**(double muon-tagged CA8 jets)**

# 3D Impact Parameter

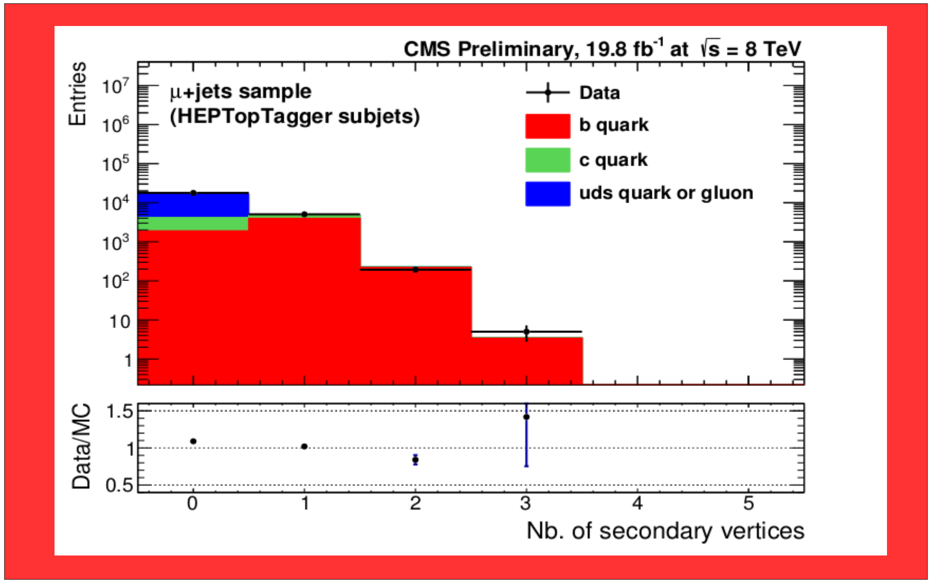


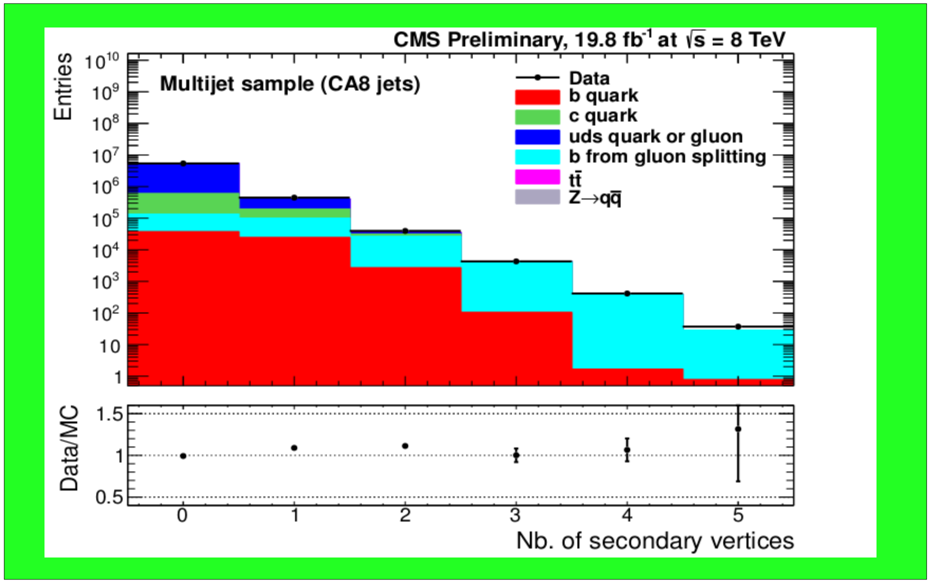


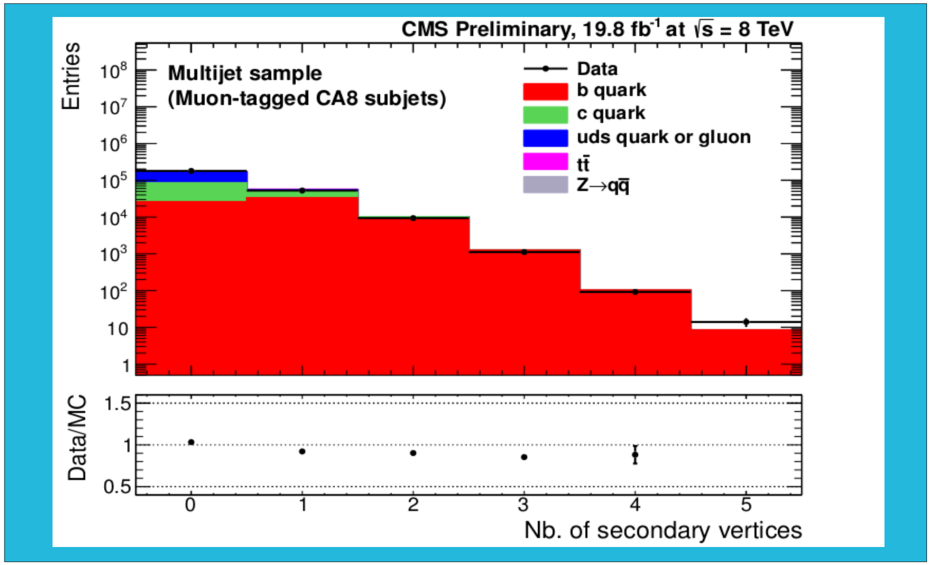


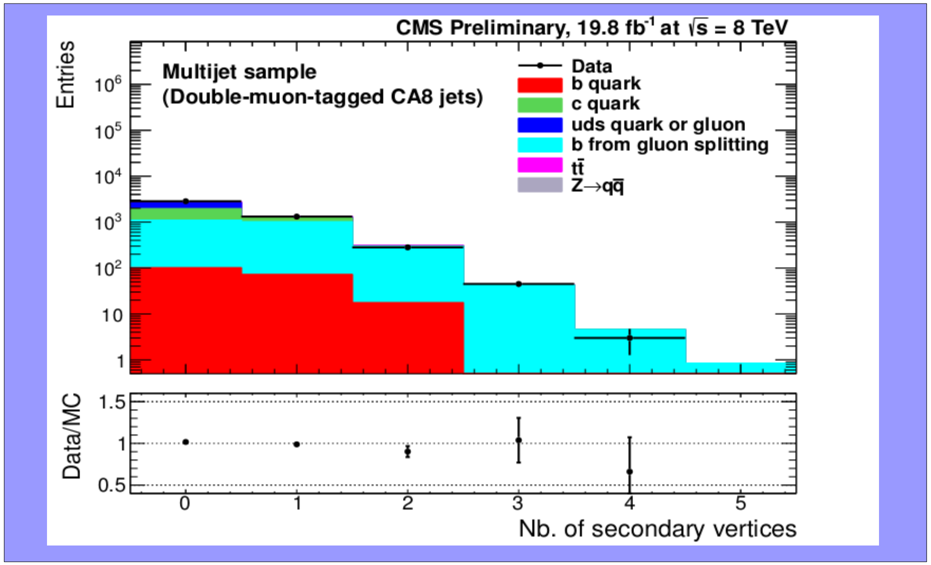


# Secondary Vertex Multiplicity

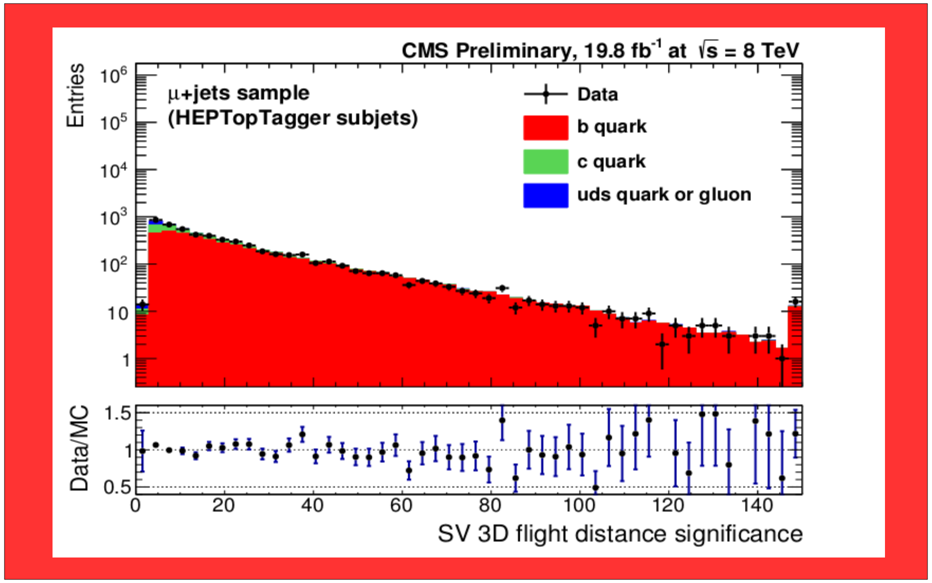


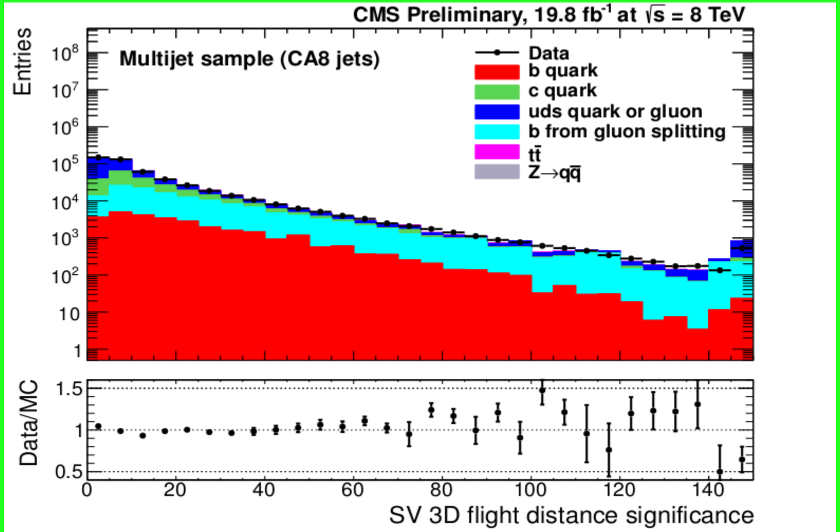


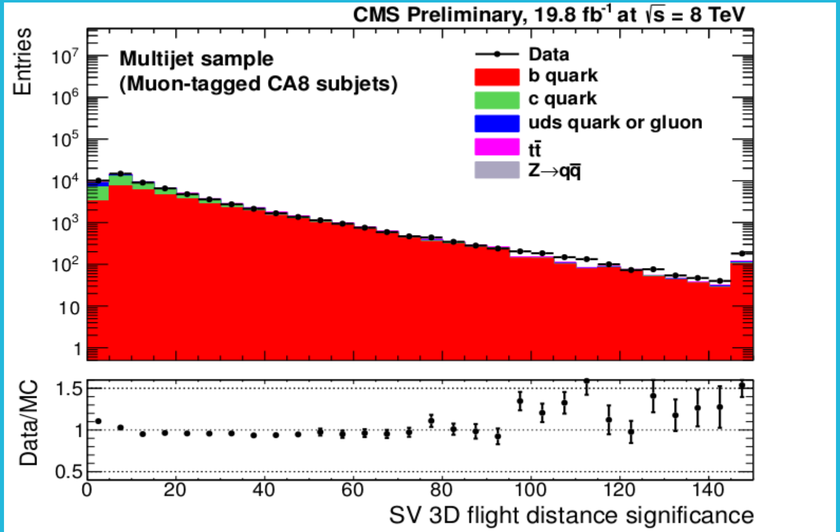


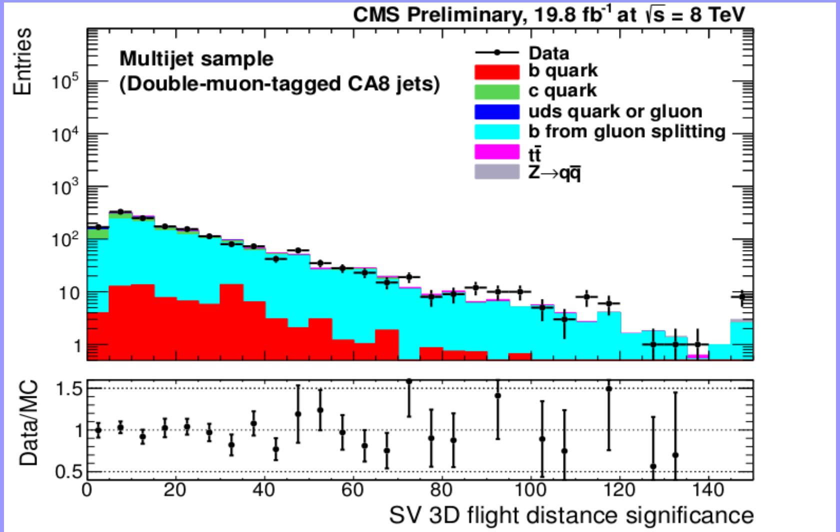


# SV Flight Distance Significance

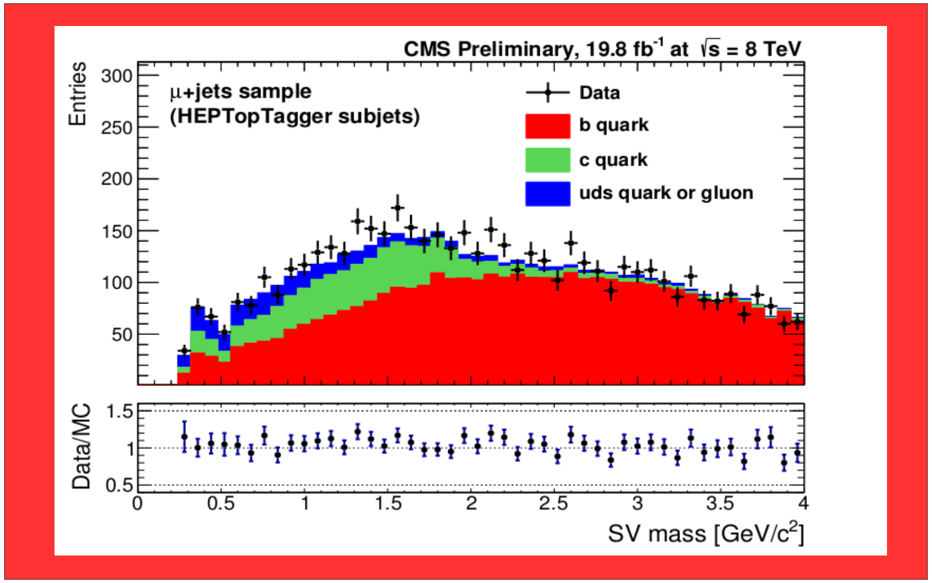


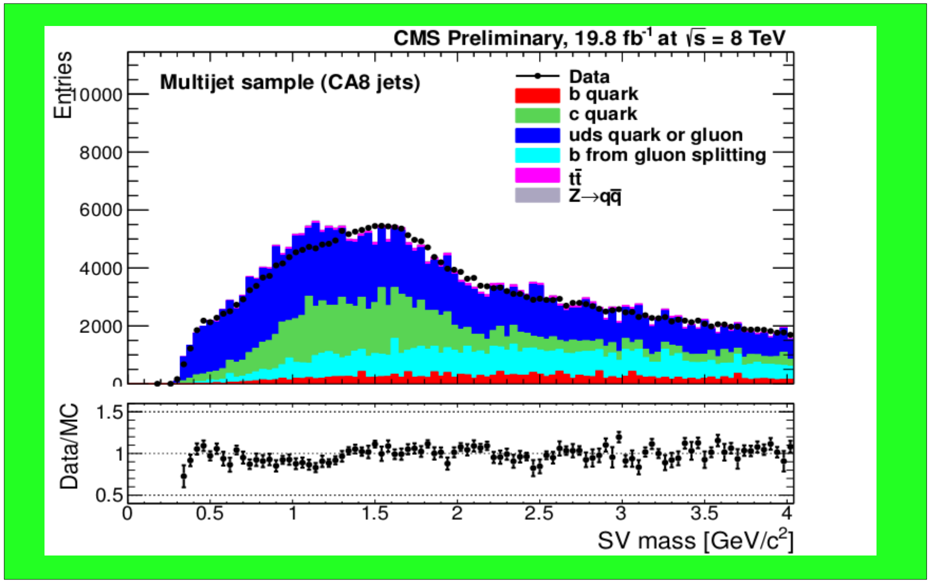


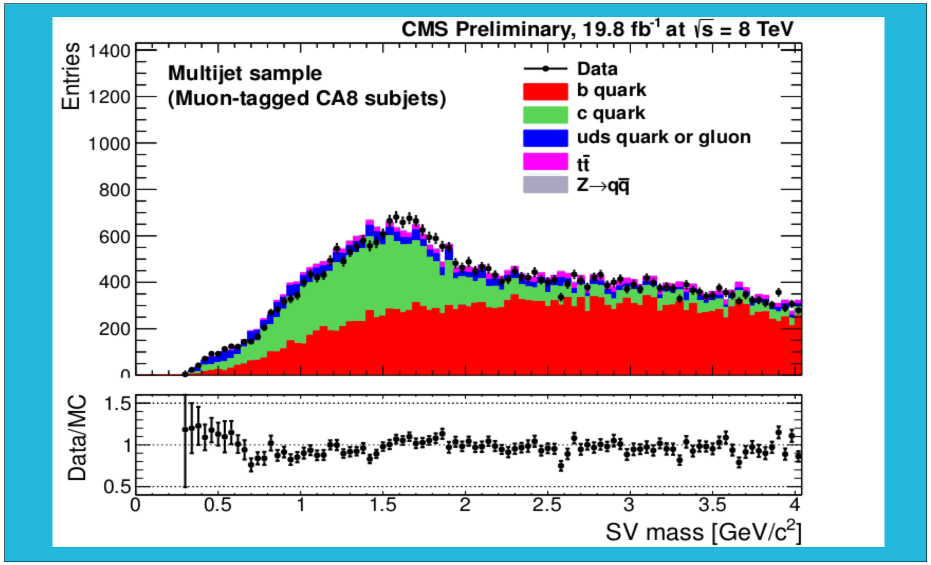


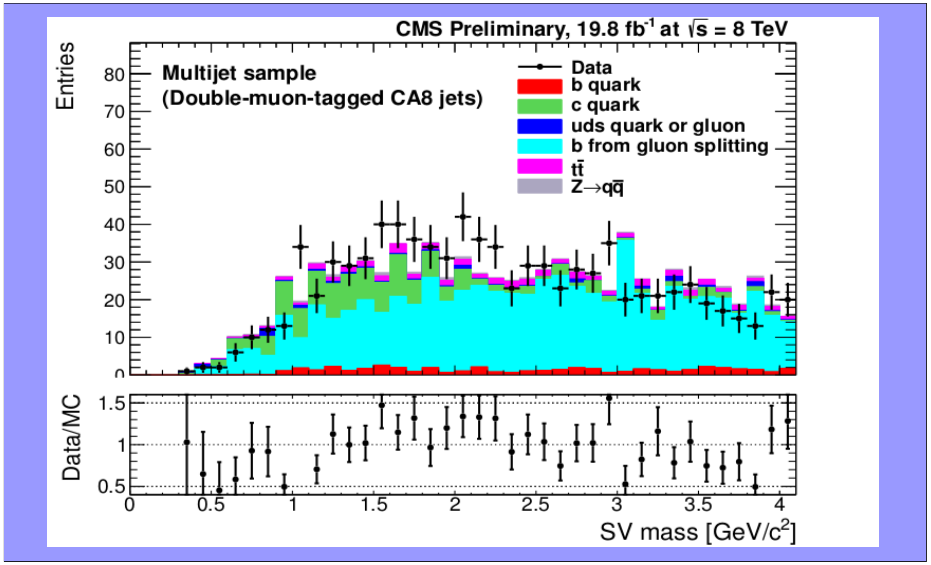


# Secondary Vertex Mass

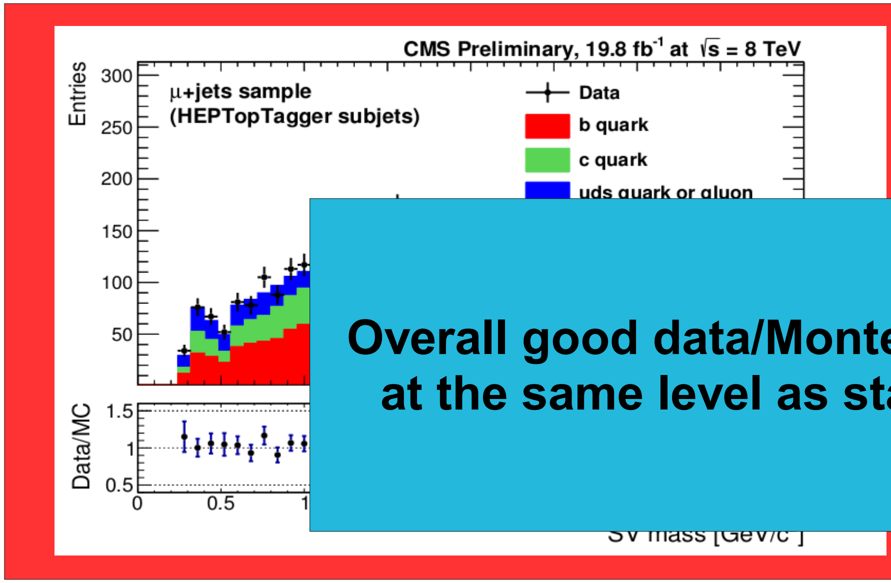


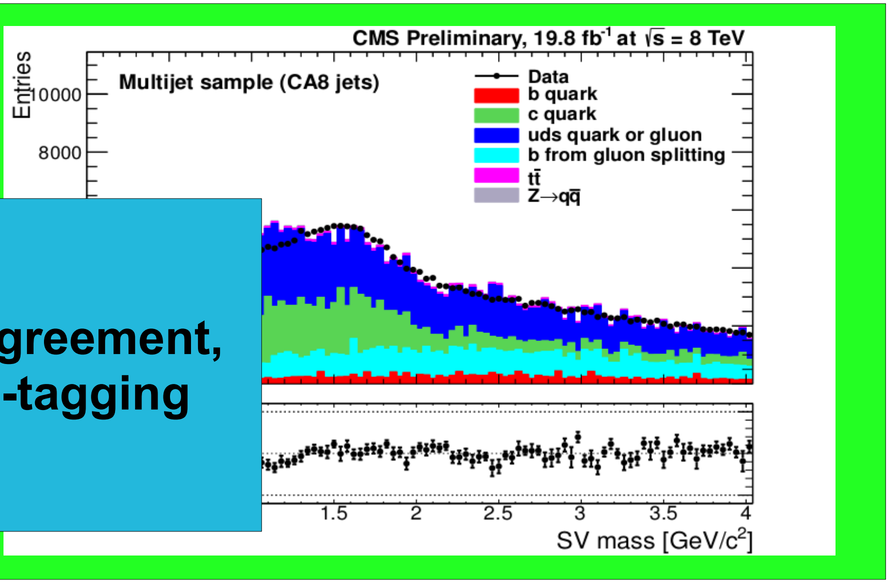




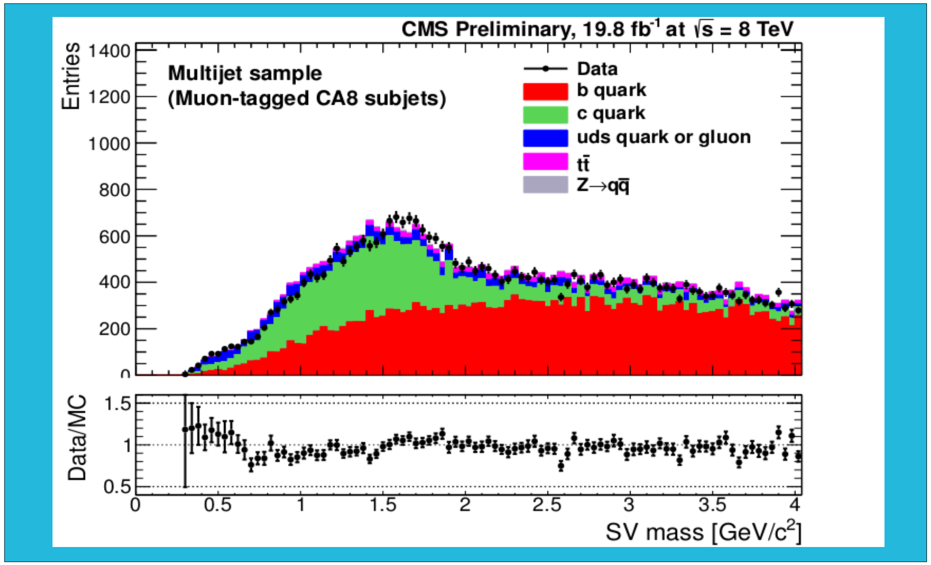


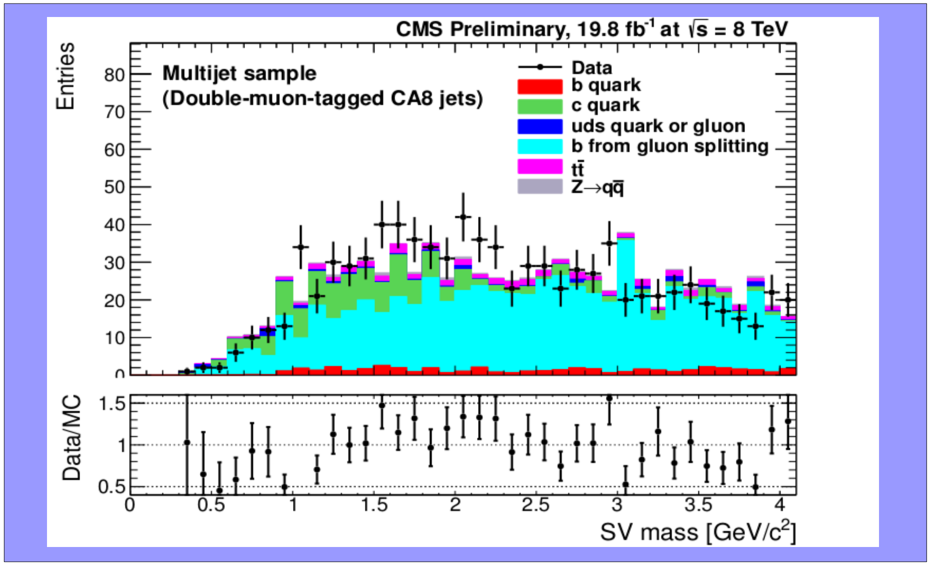
# Secondary Vertex Mass





**Overall good data/Monte Carlo agreement, at the same level as standard b-tagging**


CMS Preliminary, 19.8 fb$^{-1}$ at $\sqrt{s}$ = 8 TeV
Entries
Multijet sample
(Muon-tagged CA8 subjets)
Data
b quark
c quark
uds quark or gluon
$t\bar{t}$
$Z\rightarrow q\bar{q}$
Data/MC
SV mass [GeV/c$^2$]


CMS Preliminary, 19.8 fb$^{-1}$ at $\sqrt{s}$ = 8 TeV
Entries
Multijet sample
(Double-muon-tagged CA8 jets)
Data
b quark
c quark
uds quark or gluon
b from gluon splitting
$t\bar{t}$
$Z\rightarrow q\bar{q}$
Data/MC
SV mass [GeV/c$^2$]