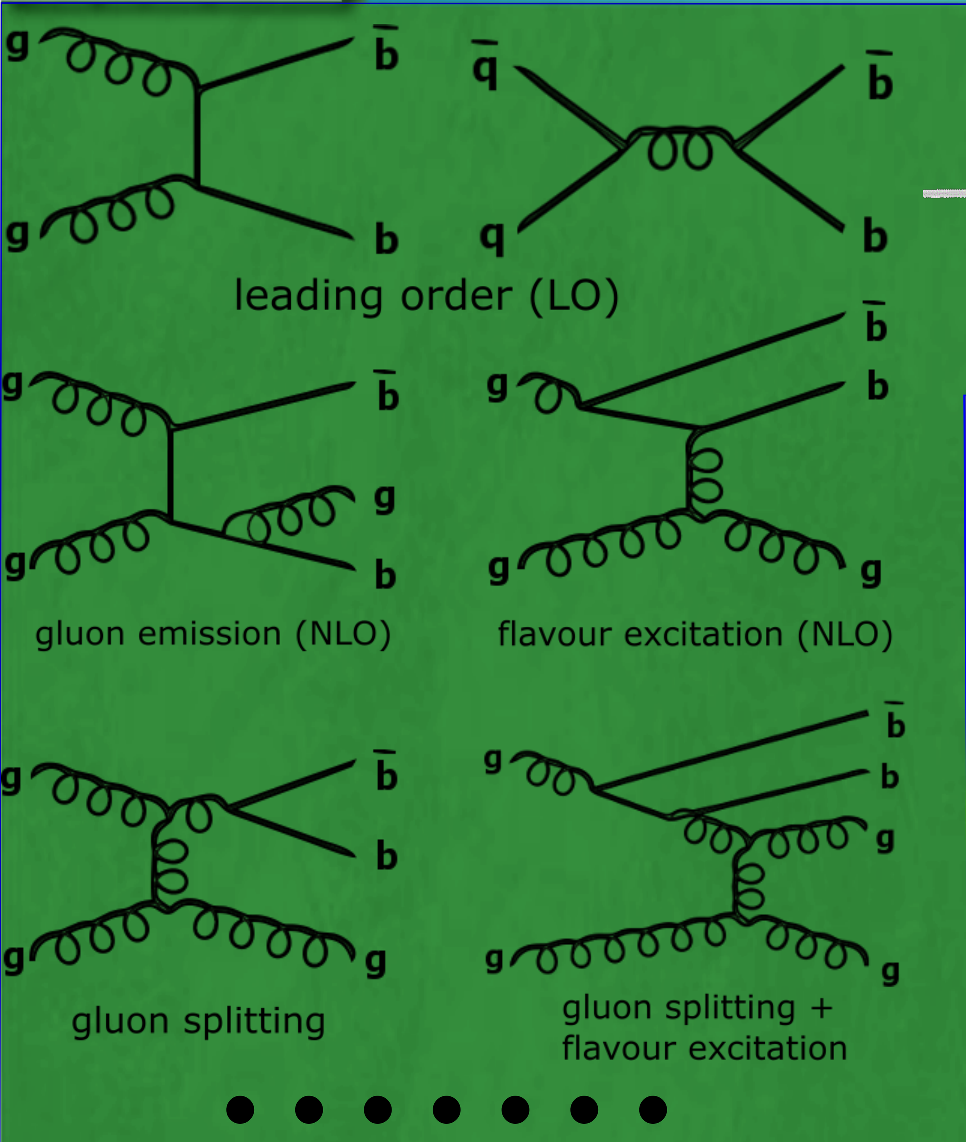
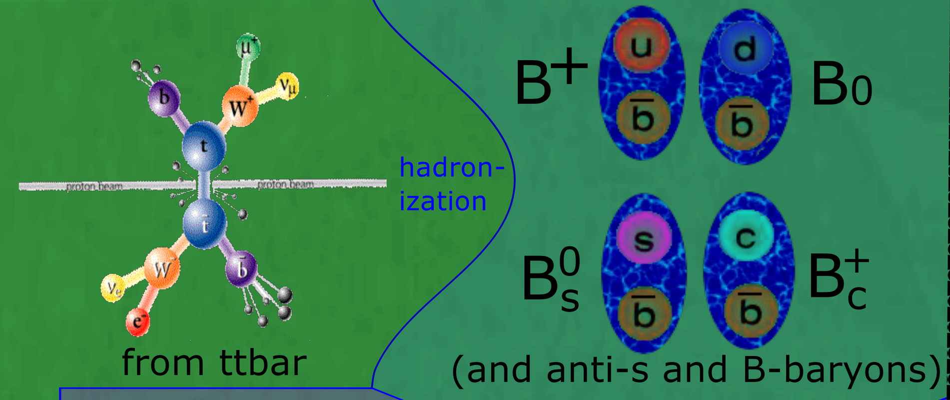
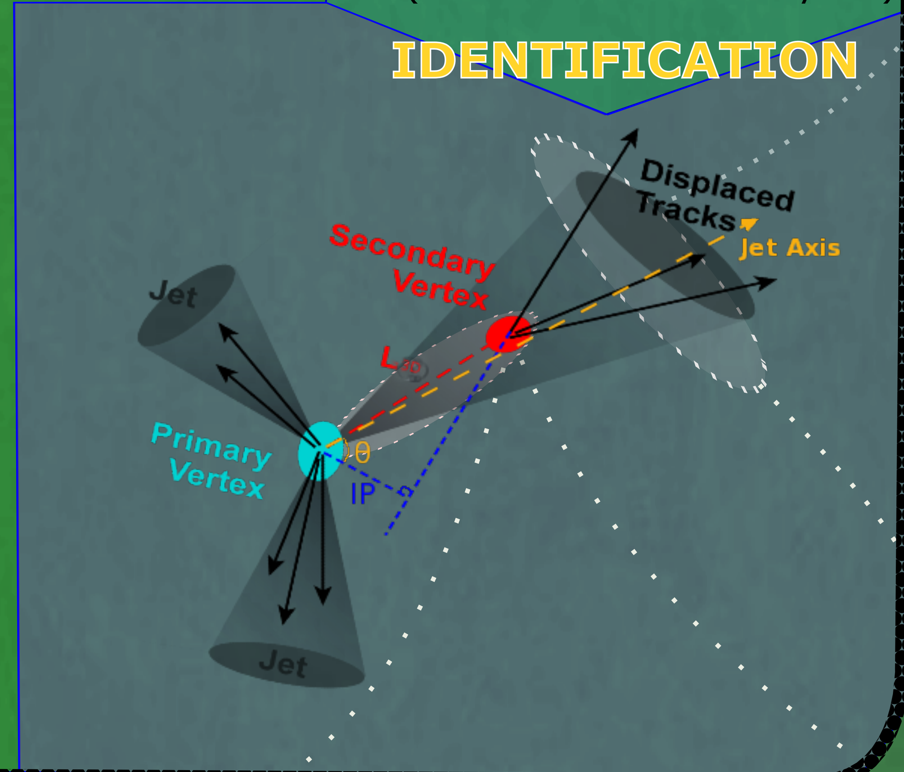
Identification of b-quark jets in the CMS experiment

Jinzhong Zhang - Northeastern University - on behalf of the b-tagging validation (BTV) group in CMS







Indentification Algorithms/Discriminators

Track Counting (TC), Jet (B) Probability (JP,JBP)

Standard track selectons:
- number of silicon hits≥8
- number of pixel hits≥2
- transverse impact parameter $d_{xy}<0.2$cm
- longitudinal impact parameter $d_z<17$cm
- $p_T>1.0$GeV · χ^2/ndof of the track fit < 5.0
- inside the jet cone: ΔR<0.5

Additional track selections:
to reduce pile-ups, at the closest point to the jet axis
- the distance to the jet axis < 700μm
- the distance to primary vertex < 5cm

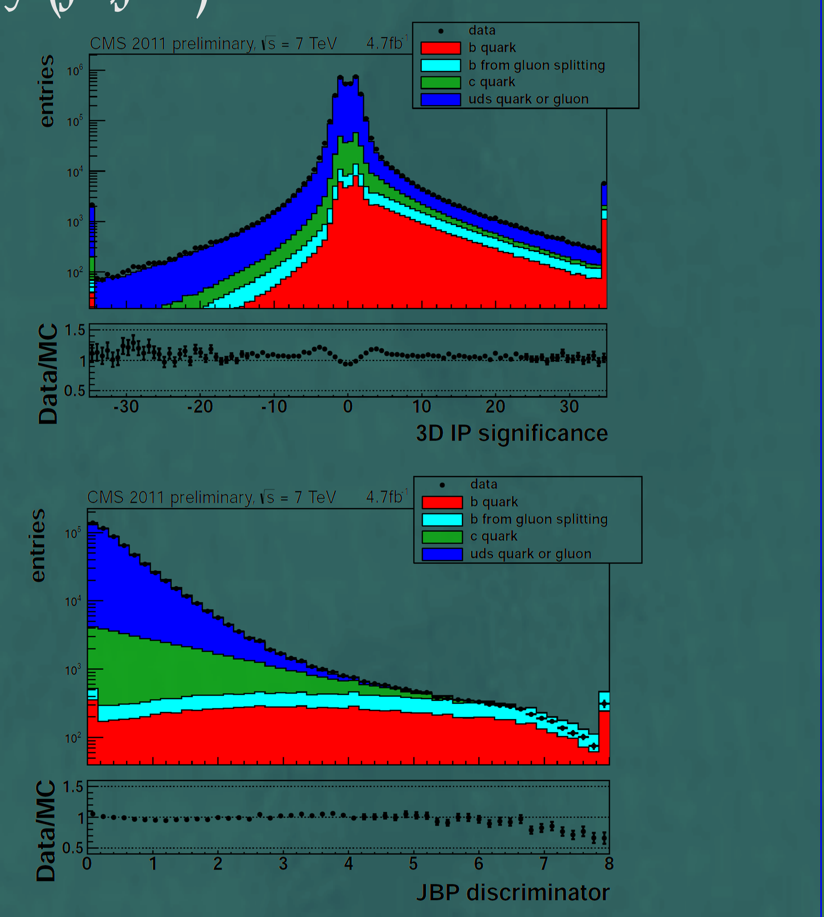
Track Counting discriminator:
the IP significance of the second or the third ranked track

Jet Probablility discriminator:

$$P_{jet} = \Pi \cdot \sum_{i=0}^{N-1} \frac{(-\ln \Pi)^i}{i!} \quad \text{with} \quad \Pi = \prod_{i=1}^{N} \max(P_i, 0.005)$$

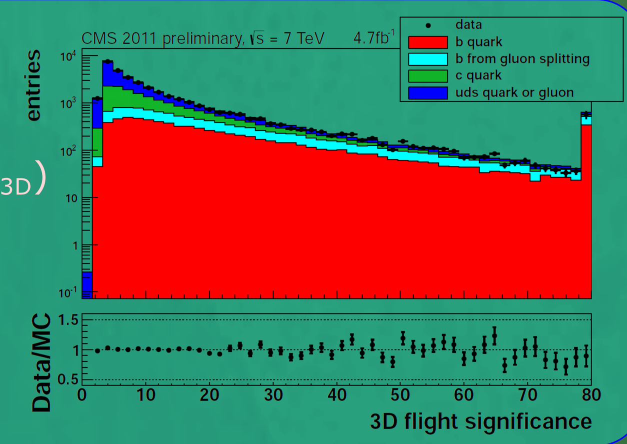
P_i is the compatibility of track i with the primary vertex.
$-1/4 \cdot \log_{10}(P_{jet})$ is plotted, using only positive IP tracks.

Jet B Probablility discriminator:
It gives more weight to the tracks with the highest IP significance up to a maximum of four such tracks, matching the average number of reconstructed charged particles from B-hadron decays.



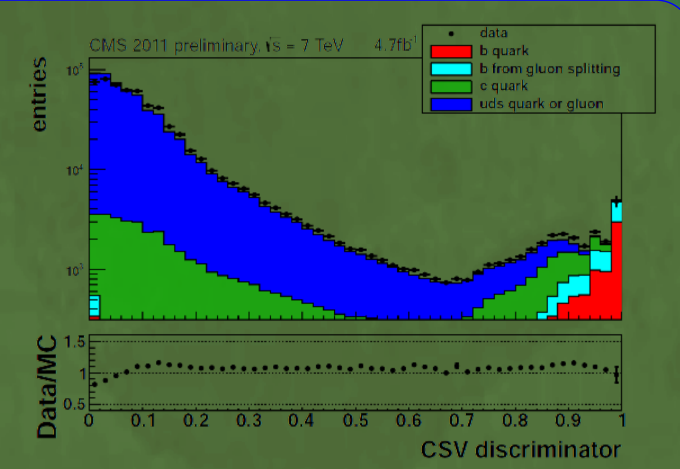
Simple Secondary Vertex (SSV)

- Reject vertices which could be compatible with primary vertex: less than 65% tracks are shared with primary vertex, $L_{3D}> 3\sigma(L_{3D})$
- Reject interaction vertices and decays of long-lived mesons: L_{3D}>2.5cm, M_{SV}<6.5GeV, $M_{SV} \neq M_{K,K^0}$
- ΔR of the flight direction (red) and the jet axis (orange) < 0.5
- Discriminator: $\log[1+|L_{3D}|/\sigma(L_{3D})]$



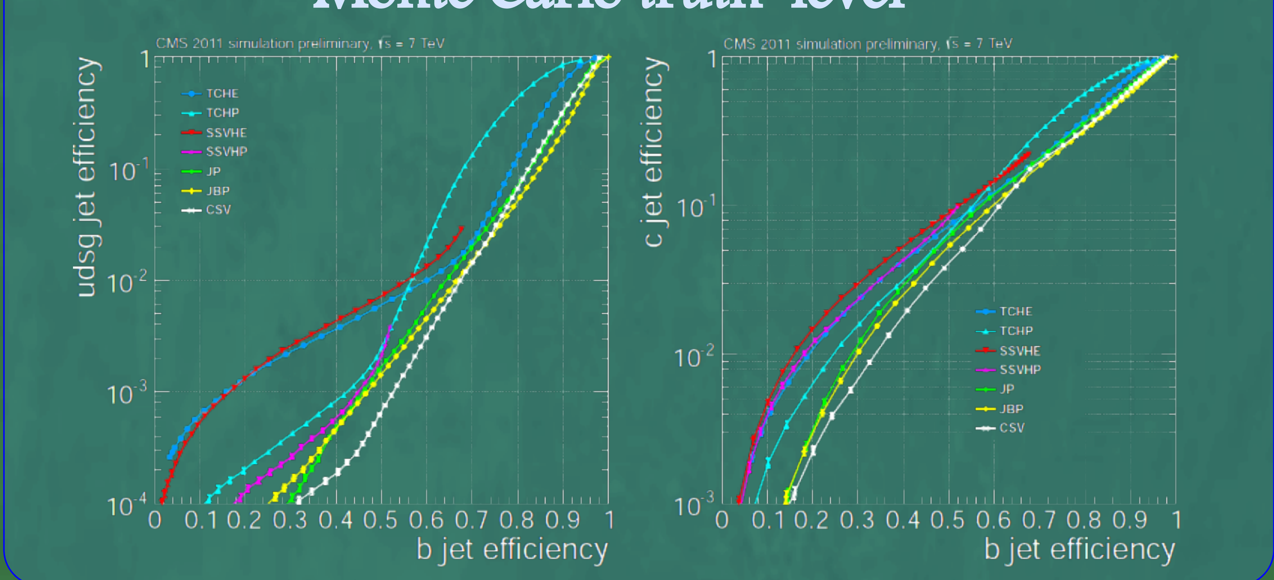
Combined Secondary Vertex (CSV)

It inputs the following variables into a neural network algorithm. The discriminator is the output.
- the real, "pseudo," or "no" vertex categories
- the 2D flight distance significance
- the vertex mass
- the number of tracks at the vertex
- the number of tracks in the jet;
- the 3D signed IP significances for each track in the jet.
- the ratio of the energy carried by tracks at the vertex with respect to all tracks in the jet;
- the pseudo-rapidity of the tracks at the vertex with respect to the jet axis
- the 2D IP significance of the first track that raises the invariant mass above the charm threshold of 1.5 GeV when subsequently summing up tracks ordered by decreasing IP significance;



Efficiency Measurements

Monte Carlo truth-level



Measure From Data Using Muon Jets

Muon selection:
Global Muon, p_T>5GeV, |η|<2.4, number of muon hits≥1,
number of muon segment≥1, χ2/ndof of the global muon<10,
number of silicon hits≥10, number of pixel hits≥1, χ2/ndof of the tracker track<10,
longitudinal impact parameter dz<1cm

Muon jet:
the muon is inside the jet cone: ΔR<0.4

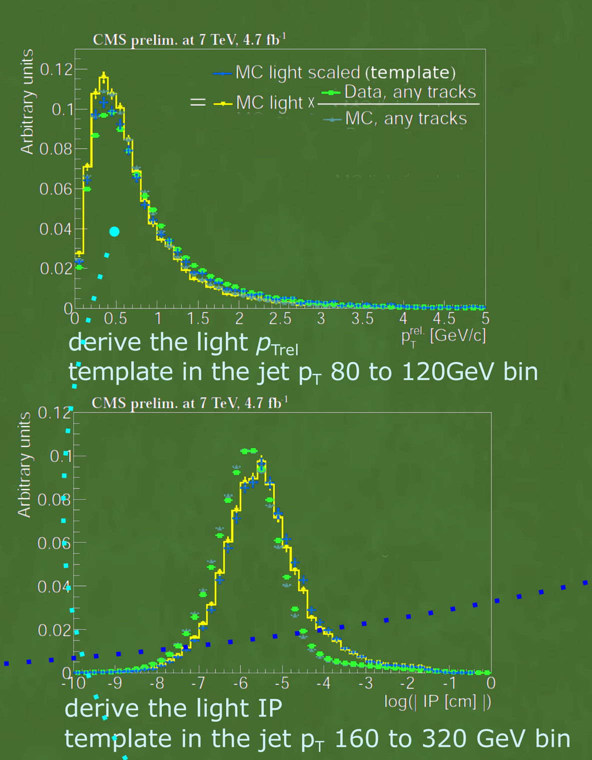
Common Systematic Errors Calculation:
pile-up: vary the average value of the pile-up in data by ∓ 10%
gluon splitting: investigate Monte Carlo sample where the number of events with gluon splitting was artificially changed by 50%.
● : vary p_T cut up to 9GeV

Measurement with kinematic properties of the muon jets

Steps:
(1) the event has another jet fulfilling the Track Counting High Purity (third ranked track) b-tagging criterion at "medium" b-tagging efficiency working point (TCHPM).
(2) Separate the muon jets into tagged and untagged subsamples by a discriminator working point whose efficiency is to be measured.
(3) For the two subsamples separately, fit the spectra of muon jets p_{Trel} or IP3D using by templates for b, c and udsg jets derived from simulation or inclusive jet data using root::TFractionFitter. The fraction of b jets is gotten in this step as (f_b^{tag}, f_b^{untag}).
p_{Trel} is the transverse momentum of muon associated with the jet direction.
(4) use the equtation to get the result:

$$\varepsilon_b^{tag} = \frac{f_b^{tag} \cdot N_{data}^{tag}}{f_b^{tag} \cdot N_{data}^{tag} + f_b^{untag} \cdot N_{data}^{untag}}$$

Systematic Errors:
● : use (TCHEL,TCHEM, TCHPM) and repeat.
● : vary the predicted ratio by ∓ 20% and repeat.



Measurement with JP as a reference tagger

It can be applied on any jet (not only muon jet) since most jets have JP information.
Steps:
(1) aquire C_b, the fraction of b jets with JP information (before tagging) in the Monte Carlo.
(2) fit the JP distribution by templates for b,c and udsg jets using root::TFractionFitter to get the fraction of b jets $f_b^{beforetag}$
(3) tag with a discriminator working point whose efficiency is to be measured
(4) repeat step 2 on step 3 tagged sample and get the fraction of b jets f_b^{tag}
(5) the efficiency is the ratio of $(f_b^{tag} \cdot N_{data}^{tag})$ and $(f_b^{beforetag} \cdot N_{data}^{beforetag})/C_b$

Systematic Errors:
● : the difference between muon jets and inclusive jets is considered as a systemetic error
● : estimate the systematic uncertainty of the half residual correction
● : instead of JP, use CSV as a reference tagger and then compare

Measurement with the System8 method

Steps:
(1) take all muon jets as sample 1 (n);
(2) take the muon jets in whose event another jet is tagged by the Track Counting High Purity b-tagging criterion at "low" working point (TCHPL) as sample 2 (p);
(3) tag the muon jet by p_{Trel}>0.8GeV or a discriminator working point whose efficiency is to be measured.
(4) solve the following eight equations and get the efficiency:

$$\begin{aligned}
n &= n_b + n_{cl} \\
p &= p_b + p_{cl} \\
n^{tag} &= \varepsilon_b^{tag} n_b + \varepsilon_{cl}^{tag} n_{cl} \\
p^{tag} &= \beta^{tag} \varepsilon_b^{tag} p_b + \alpha^{tag} \varepsilon_{cl}^{tag} p_{cl} \\
n^{p_{Trel}} &= \varepsilon_b^{p_{Trel}} n_b + \varepsilon_{cl}^{p_{Trel}} n_{cl} \\
p^{p_{Trel}} &= \beta^{p_{Trel}} \varepsilon_b^{p_{Trel}} p_b + \alpha^{p_{Trel}} \varepsilon_{cl}^{p_{Trel}} p_{cl} \\
n^{tag, p_{Trel}} &= \beta^{n} \varepsilon_b^{tag} \varepsilon_b^{p_{Trel}} n_b + \alpha^{n} \varepsilon_{cl}^{tag} \varepsilon_{cl}^{p_{Trel}} n_{cl} \\
p^{tag, p_{Trel}} &= \beta^{p} \varepsilon_b^{tag} \varepsilon_b^{p_{Trel}} p_b + \alpha^{p} \varepsilon_{cl}^{tag} \varepsilon_{cl}^{p_{Trel}} p_{cl}
\end{aligned}$$

● variables to be solved
● corratlation factors between the sample1 and sample2 which are obtained from MC simulation

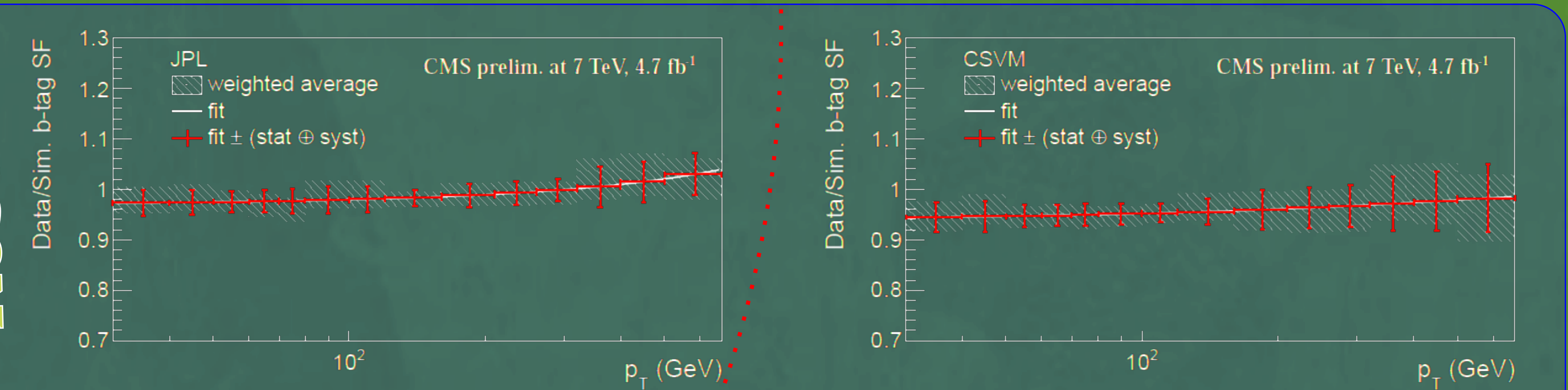
variables:
n: number of events of sample1;
p: number of events of sample2;
ε: efficiency;

subscripts:
b: b jet; cl: non-b jet;

superscripts:
tag: tagged by the discriminator working point to be measured;
p_{Trel}: tagged by p_{Trel}>0.8GeV.

Systematic Errors:
● : use (TCHEL,TCHEM, TCHPM) and repeat.
● : vary from 0.5 to 1.2GeV and take the largest discrepancy.

Result



The result is shown as the ratio of the efficiencies measured in data and the Monte Carlo truth efficiency, SFb. The errors are presented in statistical+systematic format.

b tagger	SF_b (PtRel)	SF_b (System8)	SF_b (JP)	SF_b(comb.)
JPM	$0.90 \pm 0.01 \pm 0.03$	$0.91 \pm 0.03 \pm 0.06$	$0.99 \pm 0.02 \pm 0.05$	0.92 ± 0.03
JBPM	$0.92 \pm 0.01 \pm 0.02$	$0.94 \pm 0.03 \pm 0.08$	$0.99 \pm 0.02 \pm 0.05$	0.92 ± 0.03
TCHEM	$0.94 \pm 0.01 \pm 0.03$	$0.97 \pm 0.03 \pm 0.07$	$0.98 \pm 0.01 \pm 0.03$	0.95 ± 0.02
TCHPM	$0.95 \pm 0.01 \pm 0.03$	$0.93 \pm 0.02 \pm 0.09$	$0.97 \pm 0.01 \pm 0.02$	0.96 ± 0.02
SSVHEM	$0.93 \pm 0.01 \pm 0.02$	$0.91 \pm 0.03 \pm 0.05$	$0.97 \pm 0.01 \pm 0.02$	0.95 ± 0.02
CSVM	$0.93 \pm 0.01 \pm 0.02$	$0.95 \pm 0.03 \pm 0.06$	$0.97 \pm 0.01 \pm 0.03$	0.95 ± 0.02

80GeV<jet p_T<120GeV

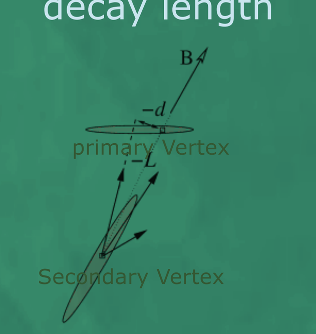
SF_b in multijet events	SF_b in $t\bar{t}$ events
0.92 ± 0.03	0.95 ± 0.03
0.92 ± 0.03	0.93 ± 0.04
0.95 ± 0.03	0.96 ± 0.04
0.94 ± 0.03	0.93 ± 0.04
0.95 ± 0.03	0.96 ± 0.04
0.95 ± 0.03	0.97 ± 0.04

overall and ttbar results

Mistag Rate Measurements

The definition of negative discriminator:
negative IP : when θ(see the top plot)>90°
TC has the same sign with IP.
JP negative is formed from only negative IP tracks.
SV is negative when it has negative decay length.
CSV: when SV is found, same sign with SV;
otherwise same sign with TC.

negative decay length

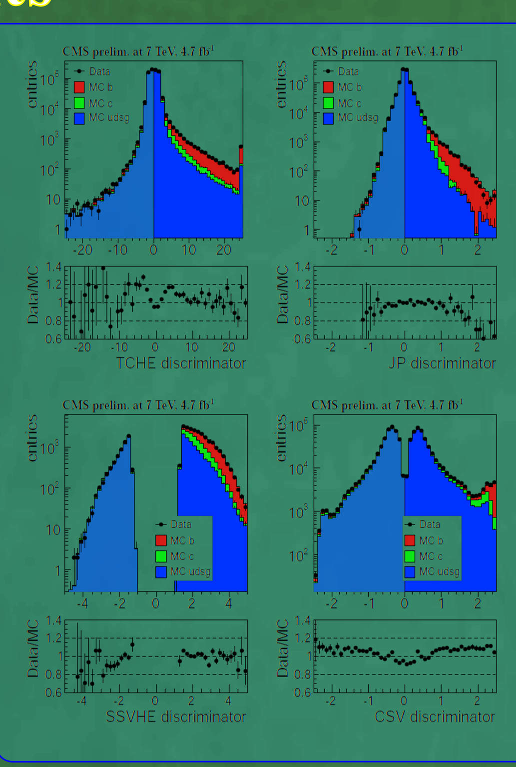


Get the efficiency directly by:

$$\varepsilon_{data}^{mistage} = \varepsilon_{data}^{nagative} \cdot (\varepsilon_{MC}^{mistag}/\varepsilon_{MC}^{nagative})$$

Result:
It is shown as the ratio of the efficiencies measured in data and the Monte Carlo truth efficiency $(\varepsilon_{data}^{negative}/\varepsilon_{MC}^{negative})$.

b tagger	mistag rate (±stat)	scale factor (±stat ± syst)
JPM	0.0109 ± 0.0002	$1.02 \pm 0.02 \pm 0.16$
JBPM	0.0112 ± 0.0001	$0.94 \pm 0.01 \pm 0.11$
TCHEM	0.0286 ± 0.0003	$1.20 \pm 0.01 \pm 0.14$
TCHPM	0.0306 ± 0.0003	$1.24 \pm 0.01 \pm 0.12$
SSVHEM	0.0209 ± 0.0002	$0.93 \pm 0.01 \pm 0.08$
CSVM	0.0152 ± 0.0002	$1.10 \pm 0.01 \pm 0.11$



Systematic Errors:
b and c fractions, gluon fraction, long lived K_s^0 and Λ decays, photon conversion and nuclear interactions, mismeasured tracks, the ratio of the number of negative over positive tagged jets, pile-up, and event sample.