Identification of b-quark jets in the CMS experiment

Sudhir Malik (on behalf of CMS collaboration)
University of Nebraska-Lincoln and LPC (LHC Physics Center), Fermilab, U.S.A.

What is CMS?

- **CMS** – **C**ompact **M**uon **S**olenoid detector at
- **LHC** – **L**arge **H**adron **C**ollider accelerator at
- **CERN -** Organisation Européenne pour la Recherche Nucléaire

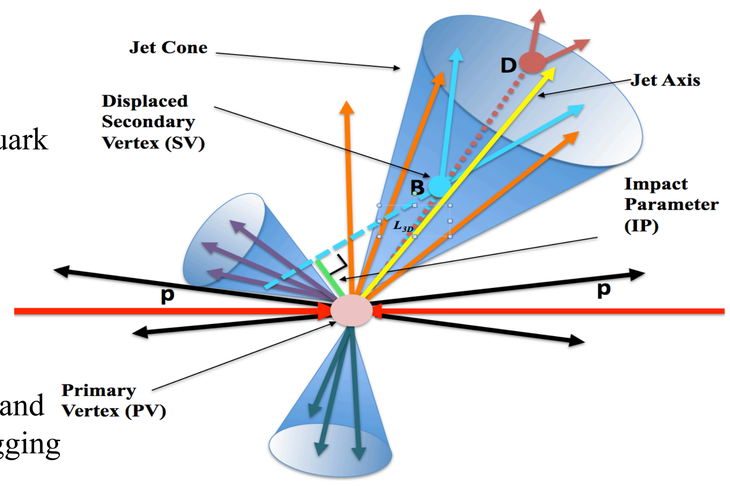
Highlights of CMS Experiment
- 3.8 Tesla Magnetic Field
- 12500 tonnes
- proton-proton collisions
- Search for Higgs Boson
- Extra dimensions
- Dark Matter
- Discover the Unexpected

- 40 countries
- 200 institutes
- 3500 collaborators

- Collision data taking since
- over 120 publications

What is b-tagging and why do we do it?

- Identification (or "tagging") of jets originating from b-quarks is called b-tagging
- Jets from b-quark hadronisation and decay (b jets) present in wide range of processes of interest – top quark decay, Higgs boson, supersymmteric particles
- Identifying b jets is vital to reduce the overwhelming background involving jets from gluons, light-flavour quarks and c-quark fragmentation
- b-quark properties can identify the hadronic jets into which they fragment
- CMS detector has a precise charged particle tracking and robust lepton identification => well matched for b-tagging



Algorithms and discriminators for b-jet identification

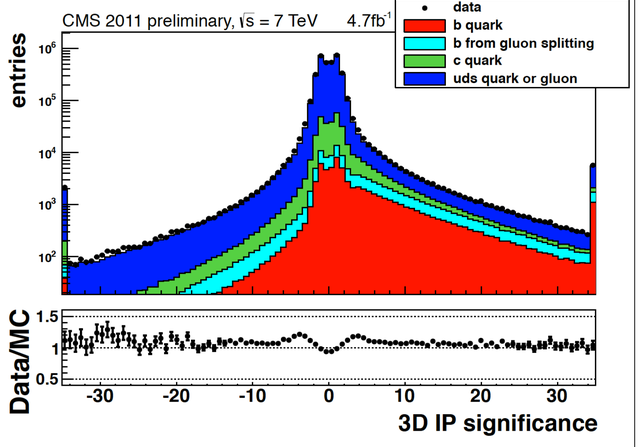
A variety of objects – **tracks, vertices, identified leptons** – are used to build observables which discriminate between b and light-flavor jets

Standard track selections:
- number of silicon hits≥8
- number of pixel hits≥2
- transverse impact parameter dxy<0.2cm
- longitudinal impact parameter dz<17cm
- pT>1.0GeV
- χ2/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

Impact Parameter (IP) of a track w.r.t. PV can be used to distinguish **b hadrons** from prompt tracks

Impact Parameter Significance $\mathbf{S_{IP}}$ (defined as ratio between IP and its estimated uncertainty) is used as observable



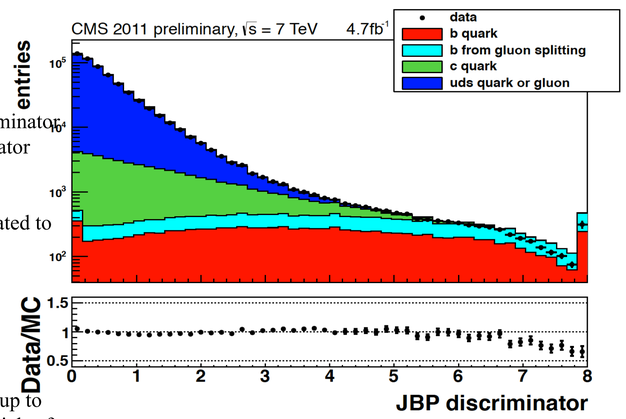
Algorithms that use **IP** and $\mathbf{S_{IP}}$ concept are:

Track Counting (TC) – sorts tracks in a jet by decreasing values of $\mathbf{S_{IP}}$
Its two versions are:
- **Track Counting High Efficiency (TCHE)** – uses $\mathbf{S_{IP}}$ of second track discriminator
- **Track Counting High Purity(TCHP)** – uses $\mathbf{S_{IP}}$ of third track as discriminator

Those that use info of several tracks in a jet
- **Jet Probability (JP)** – uses estimate of the likelihood that all tracks associated to the jet come from PV

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

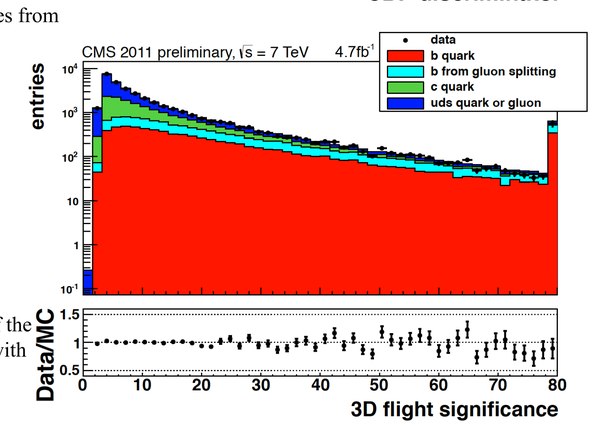
  - P_i is the compatibility of track I with PV
  - $-ln\ P_{jet}$ is plotted using only positive IP tracks



- **Jet B probability (JBP)** – gives more weight to tracks with the highest $\mathbf{S_{IP}}$ up to fours such tracks, matching the average number of reconstructed charged particles from B-hadron decays

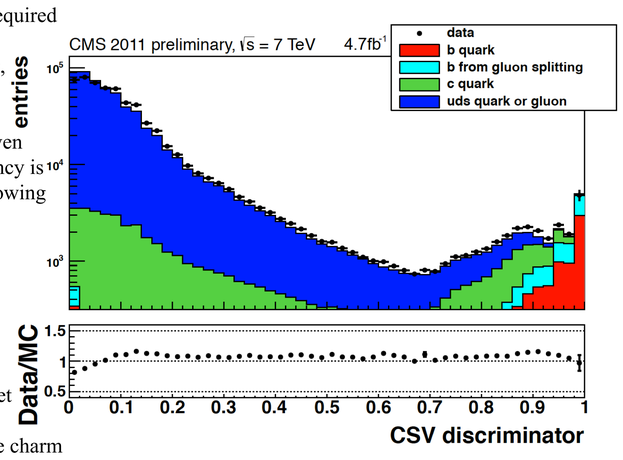
Secondary vertex candidate selection
- Reject vertices which could be compatible with primary vertex: less than 65% tracks are shared with primary vertex, $L_{3D} > 3\sigma(L_{3D})$
- Reject interaction vertices and decays of long-lived mesons: L_{3D}>2.5cm, M_{SV}<6.5GeV, $M_{SV} \neq M_{K,K^{\mp}}$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

Algorithms using secondary vertices are:



Simple Secondary Vertex (SSV) – uses significance of the flight distance (ratio of the flight distance to its estimated uncertainty). If several vertices are present the one with highest significance is used. Its two versions are:
- **Simple Secondary Vertex High Efficiency (SSVHE)** – uses vertices with at least two associated tracks
- **Simple Secondary Vertex High Purity (SSVHP)** – at least three tracks are required

Another algorithm uses the approach that involves the use of secondary vertices, together with the track based information and is called:

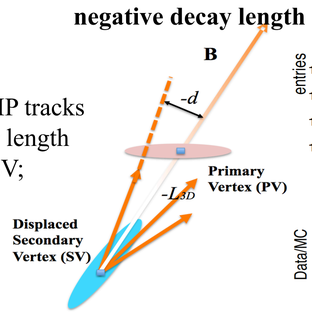


- **Combined Secondary Vertex (CSV) algorithm -** provides discrimination even when no secondary vertices are found, so maximum possible b-tagging efficiency is not limited by the secondary vertex reconstruction efficiency. It inputs the following variables into a neural network algorithm. The discriminator 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 w.r.t. 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;

Mistag Rate Measurement

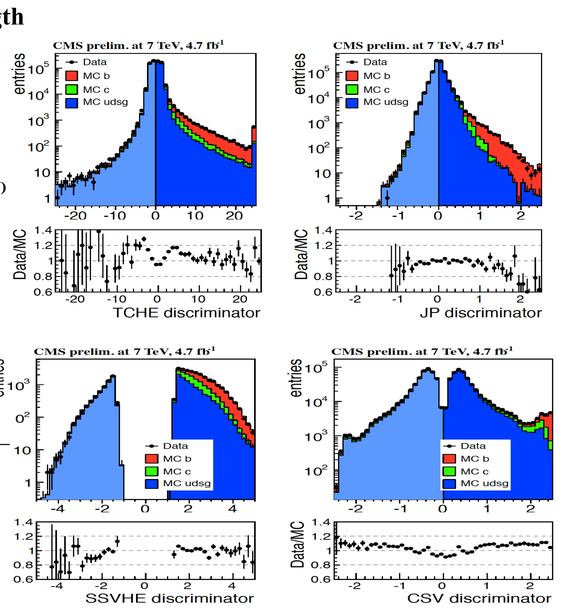
The definition of negative discriminator:
- negative IP : when θ(see the top plot)>90°
- TC has the same sign with IP
- JP negative is formed from only negative IP tracks
- SV is negative when it has negative decay length
- CSV: when SV is found, same sign with SV; otherwise same sign with TC.

negative decay length



Get the efficiency directly by:

$$\varepsilon^{mistag}_{data} = \varepsilon^{-}_{data} \cdot \varepsilon^{mistag}_{MC} / \varepsilon^{-}_{MC}$$



Result:It is shown as the ratio of the efficiencies measured in data and the Monte Carlo truth efficiency

b tagger	mistag rate ($\pm stat$)	scale factor ($\pm stat \pm syst$)
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 uncertainties: b and c fractions, gluon fraction, long lived K^0_s 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.

Efficiency measurements

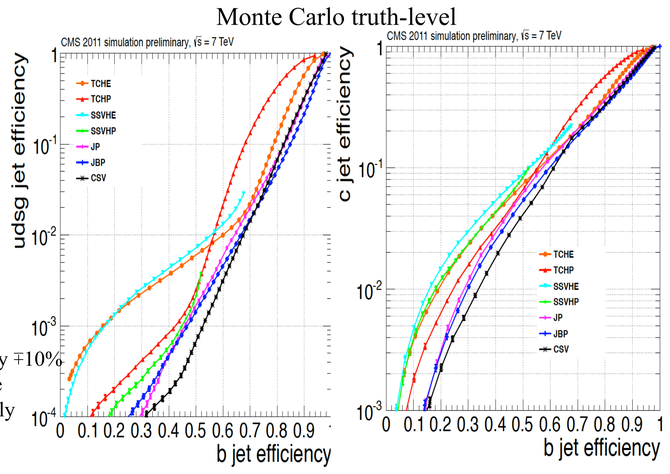
Measure from data using muons jets

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 track track<10, longitudinal impact parameter dz<1cm

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

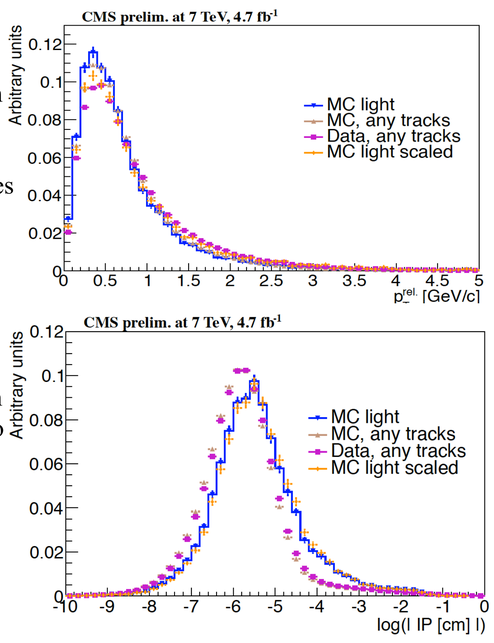
Common Systematic Errors Calculation
pile-up: vary the average value of the pile-up in data by ∓10%
gluon splitting: investigate Monte Carlo sample where the number of events with gluon splitting was artificially changed by 50%
Muon p_T: :vary p_T cut up to 9 GeV

Monte Carlo truth-level



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 templates of b, c and udsg jets derived from simulation or inclusive jet data.
The fraction of b jets is obtained in this step as (f_b^{tag}, f_b^{untag}). The p_{Trel} is the transverse momentum of muon associated with the jet direction use the equation to get the result:

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



(4) Systematic uncertainties:
: use (TCHEL,TCHEM, TCHPM) and repeat
: vary the predicted ratio by ∓20% and repeat

Measurement with the System8 method
Steps:
(1) take all muon jets as sample 1 (n)
(2) take the muon jets in those events where 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

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

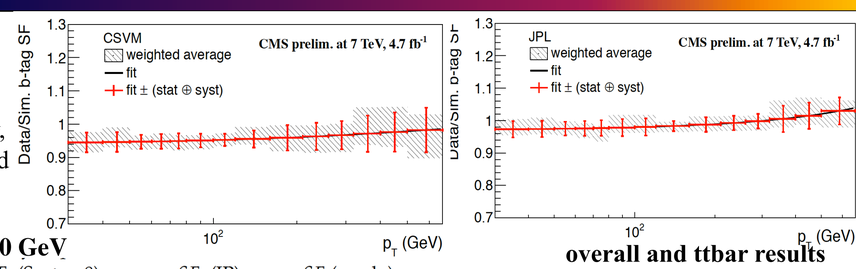
α's and β's are:
correlation factors between the sample1 and sample2 which are obtained from MC simulation
variables to solve:
n: number of events of sample1
p: number of events of sample2
ε: efficiency
subscripts:
b: b jet; cl: non-b jet
superscripts:
tag: tagged by the discriminator working point to be measured
p_{Trel}: tagged by p_{Trel}>0.8GeV

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

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) acquire C_b, the fraction of b jets with JP information (before tagging) in the Monte Carlo.
(2) fit t $f_b^{beforetag}$ ribution by templates for b,c and udsg jets to get the fraction of b jets
(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 $f_b^{tag}.N_{data}^{tag}$ and $(f_b^{beforetag}.N_{data}^{beforetag})/C_b$
Systematic uncertainties
: the difference between muon jets and inclusive jets is considered as a systematic error
: estimate the systematic uncertainty of the half residual correction
: instead of JP, use CSV as a reference tagger and then compare`

Results

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



80 GeV<jet p_T<120 GeV

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

overall and ttbar results

b tagger	SF_b in multijet events	SF_b in $t\bar{t}$ events
JPM	0.92 ± 0.03	0.95 ± 0.03
JBPM	0.92 ± 0.03	0.93 ± 0.04
TCHEM	0.95 ± 0.03	0.96 ± 0.04
TCHPM	0.94 ± 0.03	0.93 ± 0.04
SSVHEM	0.95 ± 0.03	0.96 ± 0.04
CSVM	0.95 ± 0.03	0.97 ± 0.04