Identification of b-quark jets with the CMS experiment

Ivan Marchesini, University of Hamburg, on behalf of the CMS Collaboration

JINST 8 (2013) P04013

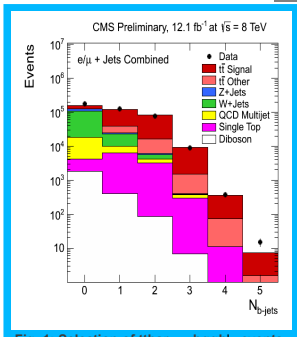

Fig. 1: Selection of ttbar → bqqblν events. The multiplicity of b-tagged jets can be exploited to powerfully reduce backgrounds not involving top quarks [CMS-PAS-TOP-12-027].

B-Tagging

- Identification ("tagging") of jets originating from the hadronization of b-quarks (b-jets).
- Selection tool for a wide range of processes, such as **top quark and Higgs decays**.
- Fundamental to **suppress main backgrounds** for several physics analyses, involving jets from gluons and light-flavor quarks (Fig. 1).

Observables (Fig. 2)

- **Lifetime**: due to the sizable life-time, the decay of the b-hadron is characterized by displaced tracks with a large impact parameter (IP) and a displaced secondary vertex (SV, multiplicity shown in Fig. 3), with a large flight distance. In order to take into account resolution effects the significance of these observables is used for b-tagging, given by the ratio observable/uncertainty.
- **High mass**, of about 5.2 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.
- **Muon/Electron** inside jet, from semi-leptonic b-hadron decay.

Detector

- CMS ideal for b-tagging: **excellent tracking** and well performing lepton identification.

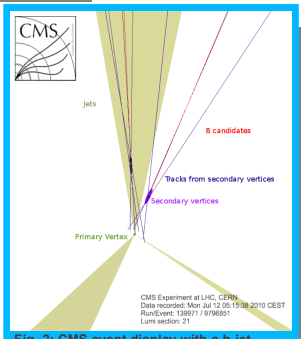

Fig. 2: CMS event display with a b-jet. Tracks displaced from the primary interaction and a secondary vertex are clearly visible.

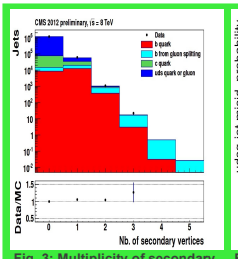

Fig. 3: Multiplicity of secondary vertices in QCD multijet events. The stacked, colored histograms indicate the contributions from jets of different flavors.

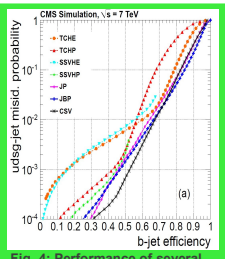

Fig. 4: Performance of several b-tagging algorithms in simulation. The light-flavor jets misidentification probability is shown as a function of the b-jet tagging efficiency.

B-Tagging Algorithms

The output of each b-tagging algorithm is a discriminator value, on which to cut to select different regions in the efficiency versus purity phase space. The performance in simulation is shown in Fig. 4. The curves are obtained varying the requirement on the discriminator. Three standard operating points of the algorithms are defined, "loose", "medium" (M), and "tight", being the discriminator values at which the misidentification probability of light-flavor jets according to simulations is 10%, 1%, or 0.1%, respectively. The choice of the working point is analysis-dependent. Algorithms implemented at CMS:

- **Track counting**: requires at least N tracks with an IP significance exceeding a given threshold. High efficiency version TCHE: N=2. High purity version TCHP: N=3.
- **Jet Probability** JP: estimates the likelihood that all tracks associated to the jet come from the primary vertex. The main advantage of this tagger is that it is calibrated directly on data, using displaced tracks. The *Jet B Probability* (JBP) version of the tagger gives more weight to the tracks with the highest IP significance, up to a maximum of four such tracks.
- **Simple Secondary Vertex**: requires a reconstructed SV with at least N associated tracks, with a flight distance significance greater than a certain threshold. High efficiency version SSVHE: N=2. High purity version SSVHP: N=3. The efficiency of this tagger is intrinsically limited by the efficiency of reconstructing a SV, of about 60-70%.
- **Combined Secondary Vertex** CSV: provides discrimination even when no SV found. Multivariate algorithm, using both SV and single track displacement information.
- **Soft Lepton Taggers** SL: rely on the properties of muons/electrons within the jet, produced during the semi-leptonic decay of the b-hadron.

Performance measurement

The CMS simulation reproduces the performance of the detector with a high precision. However, it is difficult to model perfectly all the quantities relevant for b-tagging and it is mandatory to measure the performance of the b-tagging algorithms directly from data. These measurements require samples of jets enriched in b-jets, which are either obtained selecting jets from top-quark decays or applying dedicated selections to a QCD multijet sample.

Sample of ttbar events, with at least one leptonic top-quark decay. Excellent channel: b-enriched, due to the almost exclusive top-quark decay into bW. Isolated lepton from the W decay: efficient background rejection. Some methods exploiting this channel:

- **bSample Method**: based on the selection of a b-enriched and a b-depleted region in the muon+jets ttbar channel. This is done cutting on different regions of the invariant mass of the muon and the b-jet associated to the leptonic W decay (Fig. 5). The discriminator for true b-jets is derived from the difference between the discriminator distributions in the two regions.
- **Flavor Tag Consistency**: requires consistency between the number of b-tagged jets in data and Monte Carlo in semi-leptonic ttbar events (Fig. 6), performing a likelihood fit, where the flavor composition of the jets is taken from simulations.
- **Flavor Tag Matching**: this method requires consistency between the observed and expected number of tags, based on di-leptonic ttbar decays.

Events including jets containing soft-muon. Large semi-leptonic branching fraction of b-hadrons: b-enriched sample. An additional b-tagged-jet ("away" jet) in the event can also be required, further enriching the sample in bb̄. Some methods based on this channel:

- **Lifetime Tag Method**: determines the fraction of b-jets before and after the tagging from a template fit to the JP discriminator distribution (Fig. 7). To evaluate the performance of the JP tagger itself, the CSV discriminator is used.
- **p_{Trel} Method**: the p_{Trel} distribution for the muon jet (Fig. 8) in data is fitted for fractions of b, c and light-flavor jets using a template from simulations, before and after tagging. To reduce backgrounds, an additional tagged "away" jet is required in the event.
- **System 8**: requires a jet with a muon with $p_{Trel} > 0.8$ GeV/c ("tag" tagger), where the "probe" tagger under study is applied. The correlation between the two taggers is taken from simulations. A system of 8 equations with 8 unknowns can be written (Fig. 9). There are two categories of jets, b-jets (*b*) and non b-jets (*cl*), and two categories of events, with and without a tagged away jet.

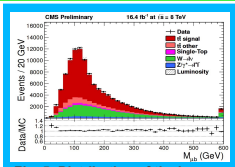

Fig. 5: Distribution of the b-jet/muon invariant mass in semi-leptonic ttbar events.

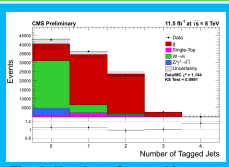

Fig. 6: Number of tagged jets per event in semi-leptonic ttbar events, for the medium operating point of CSV.

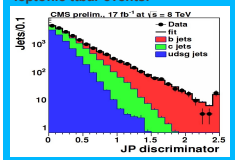

Fig. 7: Fits of the summed b-, c- and light-flavor templates to the JP discriminator distribution.

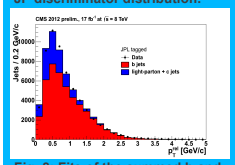

Fig. 8: Fits of the summed b and non-b templates from simulations to the muon p_{Trel}.

$n = n_b + n_{cl}$ muon-in-jet+away-jet

$p = p_b + p_{cl}$ muon-in-jet+tagged-away-jet

$n^{tag} = \epsilon_b^{tag} n_b + \epsilon_{cl}^{tag} n_{cl}$ apply "probe" tagger

$p^{tag} = \beta \epsilon_b^{tag} p_b + \alpha \epsilon_{cl}^{tag} p_{cl}$

$n^{pTrel} = \epsilon_b^{pTrel} n_b + \epsilon_{cl}^{pTrel} n_{cl}$ apply "tag" tagger

$p^{pTrel} = \delta \epsilon_b^{pTrel} p_b + \gamma \epsilon_{cl}^{pTrel} p_{cl}$

$n^{tag,pTrel} = K_b \epsilon_b^{tag} \epsilon_b^{pTrel} n_b + K_{cl} \epsilon_{cl}^{tag} \epsilon_{cl}^{pTrel} n_{cl}$ apply "tag" and "probe" taggers

$p^{tag,pTrel} = K_b \beta \delta \epsilon_b^{tag} \epsilon_b^{pTrel} p_b + K_{cl} \alpha \gamma \epsilon_{cl}^{tag} \epsilon_{cl}^{pTrel} p_{cl}$

Fig. 9: System8 method equations. The observables are on the left-hand side. The correlation factors from simulation are in Greek letters.

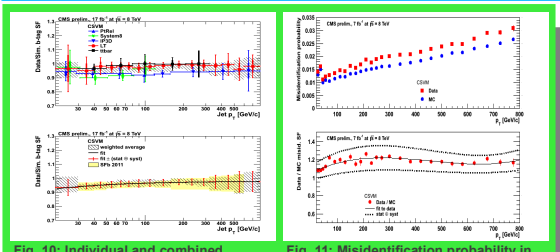

Fig. 10: Individual and combined measurements of the ratio of the b-jet tagging efficiencies in data to that in simulation for the CSVM tagger.

Fig. 11: Misidentification probability in data (red squares) and simulation (blue dots) for the CSVM tagger (top); misidentification scale factors (bottom).

Scale factors and misidentification probability

Residual differences between data and simulations in the b-tagging performance are quantified in terms of **scale factors**, defined as the ratio of the efficiency measured in data to the efficiency obtained in simulated samples. The scale factors are quantified as a function of the jet p_T (Fig. 10).

Scale factors need to be measured also for the **misidentification probability** of light-flavor jets. This measurement relies on inverted tagging algorithms, based on tracks with a negative IP or on reconstructed SV with a negative decay length. These "**negative taggers**" can be used in the same way as the tagging algorithms both in data and in simulations. The derived scale factors are generally expressed as a function of p_T and pseudorapidity of the jet (Fig. 11).