CMS Physics Analysis Summary

Contact: cms-pog-conveners-jetmet@cern.ch 2017/03/24

Jet algorithms performance in 13 TeV data

The CMS Collaboration

Abstract

The performance of jet algorithms with data collected by the CMS detector at the LHC in 2015 with a center-of-mass energy of 13 TeV, corresponding to 2.3 fb^{-1} of integrated luminosity, is reported. The criteria used to reject jets originating from detector noise are discussed and the efficiency and noise jet rejection rate are measured. A likelihood discriminant designed to differentiate jets initiated by light-quark partons from jets initiated from gluons is studied. A multivariate discriminator is built to distinguish jets initiated by a single high p_T quark or gluon from jets originating from the overlap of multiple low p_T particles from non-primary vertices (pileup jets). Algorithms used to identify large radius jets reconstructed from the decay products of highly Lorentz boosted W bosons and top quarks are discussed, and the efficiency and background rejection rates of these algorithms are measured.

1 Introduction

The CMS experiment makes use of a large variety of algorithms to identify the origin of particle jets measured in the detector. In this note, the performance of such jet identification algorithms with data collected by the CMS detector [1] at the LHC with a center-of-mass energy of 13 TeV is reported.

The identification and rejection of jets originating from noise and reconstruction failures is critical to all CMS analyses. The particle flow (PF) jet ID, described in this note, accomplishes this while retaining 98-99% of real jets by taking into account the PF jet constituent fractions and other variables sensitive to hadronic calorimeter (HCAL) and electromagnetic calorimeter (ECAL) noise. The efficiency for physical jets is measured in data using a tag-and-probe procedure in dijet events. The background rejection rate is measured using a noise enriched Minimum Bias event selection.

As known both from first principles and a large collection of experimental measurements, hadronic jets initiated by gluons exhibit differences with respect to jets initiated by light-flavor quarks (u, d, s). These differences have been exploited to construct a probability tagger capable of discriminating jets initiated by light-quark partons from those initiated by gluons [2, 3]. Such a tagger can be used as a tool in analyses that rely on the identification of an exclusive final state which includes a fixed number of hadronic jets generally originating from light-quarks. In many cases these hadronic final states suffer from overwhelming backgrounds from multi-jet QCD production or from electroweak backgrounds with hard initial or final state gluon radiation. The tagging of quark jets is also useful in the mass reconstruction of hadronically decaying objects, where the resolution is generally degraded due to combinatorial backgrounds.

Pileup jets are the result of hard interactions from non-primary vertices or of the overlap of multiple low p_{T} particles from non-primary vertices, leading to broad jets with non-negligible p_{T}. A pileup jet discriminator is built, based on jet-shape and tracking observables, to reject these objects. The identification and rejection of pileup jets is applied in analysis context, for example, when applying a jet veto or when measuring the missing transverse momentum. The rejection of pileup jets has been reported previously by CMS [4] and ATLAS [5].

At sufficiently large Lorentz boost the final state hadrons from the $\mathrm{W} \to \bar{q}q'$ decay merge into a single jet, and the traditional analysis techniques relying on resolved jets are no longer applicable. In such cases, the analysis of jet substructure can be used to identify those jets arising from decays of W bosons, which is equally applicable to the tagging of Z and H bosons. We study the performance of softdrop [6] grooming, N-subjettiness [7] and a "DDT" transformation [8] of N-subjettiness in data and simulation. Measurements of jet substructure observables related to identification of W bosons as well as performance studies of tagging algorithms have been previously reported by CMS [9–12] and ATLAS [13–16].

Similarly, the decay products of highly Lorentz boosted hadronically decaying top quarks will be closely collimated and can merge into a single jet. Numerous jet substructure techniques have been developed to identify these "top jets". We explore softdrop grooming, N-subjettiness, and the HEPTopTagger Version 2 [17] (HTT V2) algorithm. These techniques amongst other top tagging variables have been previously explored at CMS [18–20] and ATLAS [21–23].

2 The CMS detector

The central feature of the CMS detector is a 3.8 T superconducting solenoid of 6 m internal diameter. A complex silicon tracker, a crystal electromagnetic calorimeter (ECAL), and a hadron calorimeter (HCAL) are located within the magnetic field volume. A muon system is installed outside the solenoid, and embedded in the steel return yoke. The CMS tracker consists of 1440 silicon pixel and 15 148 silicon strip detector modules. The ECAL consists of 75 848 lead tungstate crystals, which provide coverage in pseudorapidity of $|\eta| < 1.48$ in the central barrel region and $1.48 < |\eta| < 3.00$ in the two forward endcap regions. The muon system includes barrel drift tubes covering the pseudorapidity range $|\eta| < 1.2$, endcap cathode strip chambers ($0.9 < |\eta| < 2.5$), and resistive plate chambers ($|\eta| < 1.6$). A more detailed description of the CMS detector, together with a definition of the coordinate system used and the relevant kinematic variables, can be found in Ref. [1].

3 Event reconstruction

3.1 Jet reconstruction

Jets are reconstructed by clustering particle candidates obtained using the particle flow (PF) algorithm [24–26]. The PF procedure identifies each individual particle (a PF candidate) through an optimized combination of all subdetector information. The energy of photons is obtained directly from the ECAL measurement, corrected for suppression effects of energies from calorimetric channels with small signals (referred to as zero-suppression) [27]. The energy of an electron is determined from a combination of the track momentum at the main interaction vertex, the corresponding ECAL cluster energy, and the energy sum of all bremsstrahlung photons associated with the track. The energy of a muon is obtained from the corresponding track momentum. The energy of a charged hadron is defined either by the combined fit of the tracker and calorimeter information or from tracker information only depending on the compatibility of the calibrated calorimeter cluster energy with measured track energy. Finally, the energy of a neutral hadron is obtained from the calibrated energies in ECAL and HCAL.

The PF candidates are clustered into jets using the anti-k_T algorithm [28] and the Cambridge-Aachen algorithm [29, 30], as implemented in FASTJET version 3.0.1 [31]. In this note three different choices of clustering algorithm and distance parameter are used: anti-k_T $R = 0.4$ (AK4) jets are used for jet ID studies, quark gluon jet discrimination, pileup jet identification, and in the event selection for W-tagging and top-tagging studies, anti-k_T $R = 0.8$ (AK8) jets are used to reconstruct W-jets and high p_T top jets, and Cambridge-Aachen $R = 1.5$ (CA15) jets are used to reconstruct low p_T top jets. All jet substructure observables are computed using PF candidates calibrated prior to jet clustering. However, the resulting jets require another small correction to the jet momentum and energy that accounts for tracking inefficiencies and threshold effects [32].

3.2 Pileup mitigation for jets

We consider two approaches to mitigate the effect of multiple interactions in the same bunch crossing, the so-called pileup (PU). In the first approach, charged hadrons associated with vertices other than the primary interaction vertex, chosen to be the one with the highest sum p_T over its associated tracks, are removed from the list of PF candidates. The procedure is referred to as charged-hadron subtraction (CHS) [33] and strongly reduces the dependence of the jet energy and substructure reconstruction on pileup. An event-by-event jet-area-based correction [32, 34, 35] is applied to the jet 4-momenta to remove the remaining energy due to

remaining neutral and charged particles originating from PU vertices. This approach is used for jet ID studies, quark gluon jet discrimination, pileup jet identification, and in the event selection for W-tagging and top-tagging studies.

The second approach, called pileup per particle identification (PUPPI) [36], attempts to use local shape information, event pileup properties and tracking information together to mitigate the effect of pileup on jet observables. PUPPI thus operates at the PF candidate level, before any jet clustering is performed. A local variable α is computed which contrasts the collinear structure of QCD with the soft diffuse radiation coming from pileup interactions. The α variable is used to calculate a weight which encodes the probability that an individual particle originates from a pileup collision. These per-particle weights are used to rescale the particles four-momenta to correct for pileup, superseding the need for jet-based pileup corrections.

As discussed in [36], various definitions of the discriminating variable α are possible. We adopted a configuration with a different definition of α for particles in the central ($|\eta| < 2.5$) and forward region ($|\eta| > 2.5$) of the detector, where tracking information is not available. The choice is optimized in order to obtain the best discriminating power between particles originating from the hard scattering vertex and pileup vertices in the pileup scenario under study. In the central region, the α variable for a given particle i is defined as

$$\alpha_i = \log \sum_{\substack{j \in Ch,PV \\ j \neq i}} \left(\frac{p_{T,j}}{\Delta R_{ij}} \right)^2 \Theta(R_0 - \Delta R_{ij}) \tag{1}$$

where Θ is the step function, i refers to the particle in question and j to the neighboring charged particles from the primary vertex within a cone of radius R_0. We consider charged particles as coming from the primary vertex if their track is associated to the leading vertex of the event or is unassociated but with d_z <0.3 cm, where d_z is the distance along the z axis with respect to the leading vertex. In the forward region, the same variable is used, but based on all particles rather than only charged particles associated to the primary vertex:

$$\alpha_i = \log \sum_{j \neq i} \left(\frac{p_{T,j}}{\Delta R_{ij}} \right)^2 \Theta(R_0 - \Delta R_{ij}) \tag{2}$$

where i refers to the particle in question and j to all neighboring particles within R_0. A χ^2 approximation

$$\chi_i^2 = \frac{(\alpha_i - \bar{\alpha}_{PU})^2}{RMS_{PU}^2} \tag{3}$$

where $\bar{\alpha}_{PU}$ is the median value of the α_i distribution for pileup particles in the event and RMS_{PU} is the corresponding RMS, is used to determine the probability of a particle to be from pileup. In the tracker region, $\bar{\alpha}_{PU}$ and RMS_{PU} are calculated using all charged pileup particles (i.e. all charged particles not from PV), while in the forward region they are calculated using all the particles in the event. The pseudorapidity dependence of α_{PU} and RMS_{PU} is accounted for by computing their values separately in three pseudorapidity bins ($0 < |\eta| < 2.5$, $2.5 < |\eta| < 3$ and $|\eta| > 3$). Particles are then assigned a weight given by $w_i = F_{\chi^2,NDF=1}(\chi_i^2)$ where $F_{\chi^2,NDF=1}$ is the cumulative distribution function of the χ^2 distribution with one degree of freedom.

The choice of algorithm parameters are close to the initial suggestion in [36]. The radius of the cone R_0 is set to 0.4. Particles with weights w_i smaller than 0.01 are rejected. In addition a cut on the minimum scaled p_T of the neutral particles is applied: $w_i \cdot p_{T,i} > (A + B \cdot n_{PV})$ GeV, where n_{PV} is the number of reconstructed vertices in the event, and A and B are tuneable

parameters which are tuned separately in three pseudorapidity bins. In the pseudorapidity regions $0 < |\eta| < 2.5$ and $2.5 < |\eta| < 3$ the parameters are chosen such that jet mass and p_{T} resolution are optimized, and in the forward region $|\eta| > 3$ the parameters are chosen such that MET resolution is optimized. No additional event-by-event pileup correction (as described in Ref. [32]) is needed for jets clustered from these weighted inputs. This approach is used for W-tagging and top-tagging studies.

3.3 Lepton reconstruction

Muons are reconstructed using the information collected in the muon detectors and the inner tracking detectors, and are measured in the range $|\eta| < 2.4$. Tracks associated with muon candidates must be consistent with a muon originating from the leading primary vertex, and are required to satisfy a set of identification requirements [37]. The muon isolation variable is computed as the sum of the transverse energy of the particles inside a cone of radius $\Delta R = 0.3$ around the muon direction divided by the muon transverse momentum. The isolation variable is corrected for the expected contribution from pileup, and is required to be lesser than 0.1 for isolated muons.

Electrons are reconstructed in the range $|\eta| < 1.442$ and $1.56 < |\eta| < 2.5$ by combining tracking information with energy deposits in the ECAL. Candidates are identified [38] using information on the spatial distribution of the shower, the track quality, and the spatial match between the track and electromagnetic cluster, the fraction of total cluster energy in the HCAL, and the level of activity in the surrounding tracker and calorimeter regions.

4 Data and simulated samples

The pp collision data collected by CMS in 2015 with the detector in a fully operational state is used for this study, amounting to 2.3–2.6 fb^{-1} of integrated luminosity.

Samples of simulated Monte Carlo (MC) events are used to evaluate the performance of the jet algorithms discussed in this note. Studies of the quark/gluon discriminator and pileup jet ID utilize Z+jets events produced with the leading-order (LO) mode of aMC@NLO v5.2.2.2 [39] generator, called MADGRAPH in the following, that is interfaced with parton shower simulations from PYTHIA v8.205 [40, 41] using the Lund string fragmentation model [42, 43] for jets. Dijet events are produced with PYTHIA in standalone mode. The PYTHIA parameters for the underlying event were set according to the CUETP8M1 tune [44, 45]. Studies of top tagging and W-tagging algorithms are preformed using $t\bar{t}$ events simulated with POWHEG [46–48]. The validation of these algorithms also requires the simulation of single top quark production via the s-channel, t-channel, and tW processes, W and Z boson production in association with jets, and the $tt + W$ and $tt + Z$ processes. The simulation of these events is accomplished using the MADGRAPH generator. The NNPDF 3.0 [49] parton distribution functions (PDF) are used in all generated samples.

The studies performed on the simulation assign a flavor to each reconstructed jet. In perturbative QCD, flavor is not an infrared-safe observable, and is therefore not well defined beyond the tree level. Jet flavor is defined by accessing the generator information (partons): the reconstructed jet is matched to colored generator partons (quarks and gluons) by minimizing their distance (ΔR) in the $\eta - \phi$ plane. If a match with $\Delta R[\text{jet, closest parton}] < R_{min}$ is found, the jet flavor is defined to be the flavor of the closest parton. When assigning flavor to jets used in quark/gluon discrimination, an inclusive value of $R_{min} = 0.4$ is chosen. If no parton is found satisfying this criteria the jet flavor is considered "undefined". In assigning flavor to the jets

used in training the pileup jet discriminator, a value of $R_{min} = 0.2$ is chosen in order to increase the purity of this assignment. Pileup jets are defined to be reconstructed jets which are not matched to a simulated parton or simulated jet, by requiring the reconstructed jet to be separated by $\Delta R > 0.2$ from the closest parton and $\Delta R > 0.3$ from the closest simulated jet.

5 Noise jet identification

The PF algorithm reconstructs muon, electron (or "charged EM"), photon (or "neutral EM"), charged hadron and neutral hadron candidates. The fractions of the jet energy carried by certain types of PF candidates clustered into a jet (PF jet energy fractions), along with the number of PF candidates clustered into a jet are used in order to discriminate between noise jets and physical jets. The jet energy fraction and multiplicity variables are sensitive to different sources of noise from the hadronic (HCAL) and electromagnetic (ECAL) calorimeters. Table 1 presents these particle flow jet identification (PF jet ID) criteria. Note that the "charged" variables extend to $|\eta| < 2.4$ since there is no tracker coverage outside of this region, whereas the "neutral" variables extend to the whole η region up to $|\eta| < 5$. Three PF jet ID working points are defined: "loose", "tight" and "tight lepton veto". The "loose" and "tight" working points are designed to remove jets originating from calorimetric noise, while the "tight lepton veto" working point additionally rejects the potential background from mis-reconstructed electron and muon candidates, effectively resolving also the ambiguity between isolated lepton candidates and jets reconstructed from single lepton candidates.

Table 1: The PF jet ID criteria for the whole η region from -5 up to 5.

Jet Variables	$\|\eta\|$ range	Loose	Tight	Tight Lepton Veto
Charged Hadron Fraction	$\|\eta\| < 2.4$		>0.0	
Charged Multiplicity	$\|\eta\| < 2.4$		>0	
Charged EM Fraction	$\|\eta\| < 2.4$	<0.99	<0.99	<0.9
Muon Fraction	$\|\eta\| < 2.4$			<0.8
Neutral Hadron Fraction	$\|\eta\| < 2.7$	<0.99	<0.9	<0.9
Neutral EM Fraction	$\|\eta\| < 2.7$	<0.99	<0.9	<0.9
	$2.7 < \|\eta\| < 5$		<0.9	
Neutral Multiplicity	$2.7 < \|\eta\| < 3$		>2	
	$3 < \|\eta\| < 5$		>10	

Two event selections are defined in order to study the PF jet ID criteria performance: a noise enriched minimum bias selection and a physical jet enriched dijet selection.The dijet events are selected with both a prescaled HT>350 GeV trigger and an unprescaled HT>800 GeV trigger, with leading jet $p_T > 60$ GeV, second leading jet $p_T > 30$ GeV, and with azimuthal angle between the two leading jets $\Delta\phi > 2.7$. The pairwise dijet mass is required to be greater than 600 GeV for the prescaled trigger and greater 1200 GeV for the unprescaled trigger in order to ensure a trigger efficiency of $> 99\%$. The noise enriched minimum bias events are selected with a minimum bias trigger, with no further event selection. The majority of the events selected are composed of noise induced clusters, originating from instrumental noise and/or mis-reconstructions. By comparing the missing transverse energy (E_T^{miss}) over the sum of the p_T's of all reconstructed objects (Σp_T) in dijet and minimum bias events, one can see that the majority of events in the minimum bias selection originate from non-physical noise sources (Fig. 1). In Fig. 2, the PF jet variables used by the PF jet ID criteria are shown for leading jet of the back-to-back pair in dijet events and all the jets in minimum bias events in the central $|\eta|$ bin ($0< |\eta| <0.5$).

The PF jet ID efficiency is estimated using a tag-and-probe technique. One of the two leading jets from the dijet selection is chosen randomly and required to satisfy the tight PF jet ID criteria. The jet constitutes the "tag", and the opposite jet in the event is the "probe" jet. The efficiency is then defined as the number of probe jets satisfying the tight PF jet ID criteria divided by the total number of probe jets. For the central $|\eta|$ bin 0-0.5, the efficiency is measured to be 99.8%.

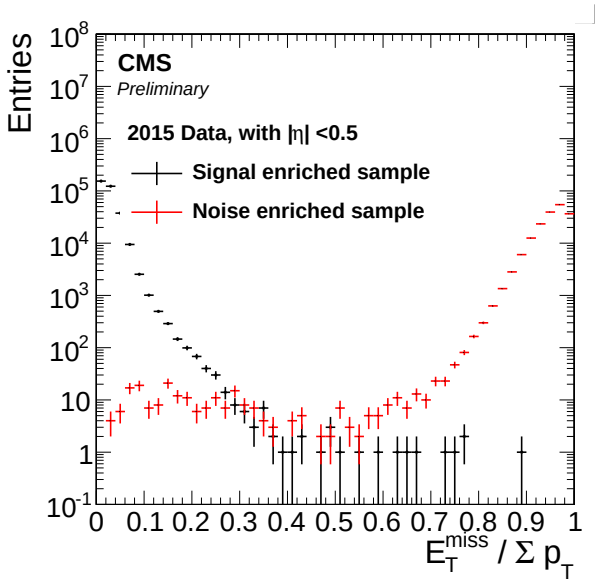


Figure 1: The distributions of $E_{\mathrm{T}}^{\mathrm{miss}}$ over Σp_T for signal enriched back-to-back dijet events (black) and for noise enriched events from a minimum bias selection (red) before applying the PF jet ID.

The PF jet ID efficiency for the tight working point is shown in Fig. 3 as a function of p_{T} and η. The errors have been calculated using Wilson intervals of binomial errors.

The noise jet background rejection rate is defined as the fraction of jets with $p_{\mathrm{T}} > 30$ GeV in the minimum bias selection which fail the PF jet ID criteria. The rate for the tight PF jet ID criteria is measured to be 99.999% in the central eta region. The PF jet ID noise jet rejection rate for the tight working point is shown in Fig. 4 as a function of p_{T} and η.

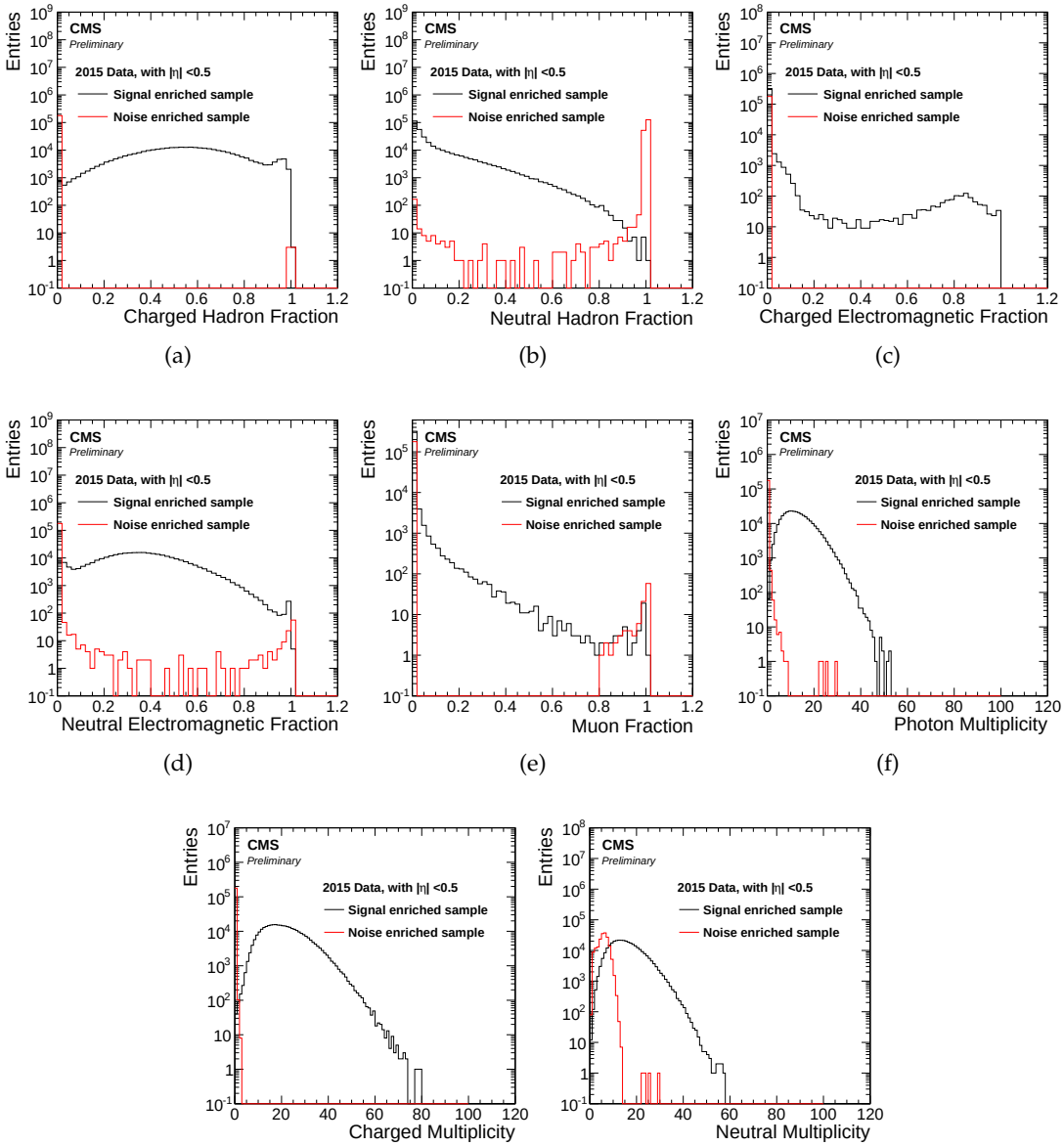
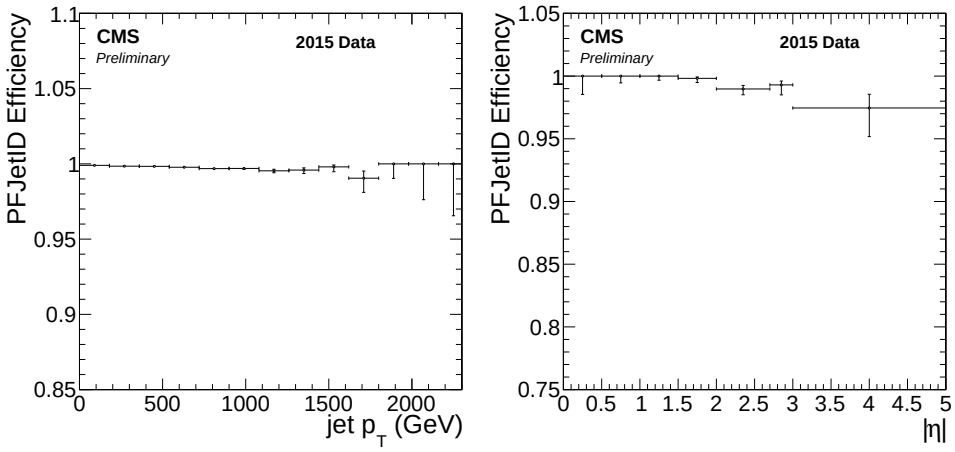
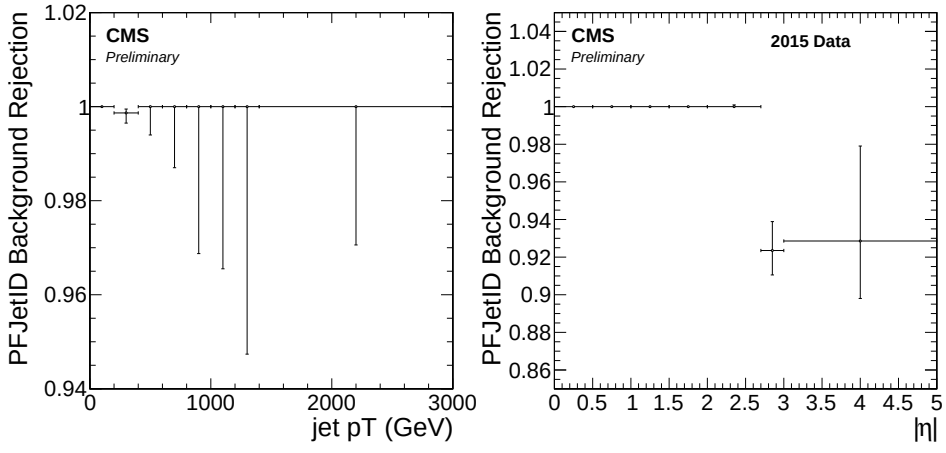


Figure 2: Distributions of PF jet variables for central jets ($|\eta| < 0.5$) as measured in signal enriched back-to-back dijet events (black) and for noise enriched events from a minimum bias selection (red) before applying the PF jet ID. The quanties plotted are: (a) charged hadron energy fraction, (b) netural hadron energy fraction, (c) charged electromagnetic energy fraction, (d) neutral electromagnetic energy fraction, (e) muon energy fraction, (f) photon multiplicity, (g) charged multiplicity, (h) neutral multiplicity.



(a) (b)

Figure 3: The tight PF jet ID efficiency (a) as a function of p_{T} for central jets ($|\eta| < 0.5$) and (b) as a function of $|\eta|$ for 30$<$ p_{T} $<$100 GeV.



(a) (b)

Figure 4: The noise jet background rejection rate for the tight PF jet ID criteria (a) as a function of p_{T} for central jets ($|\eta| < 0.5$) and (b) as a function of η for 30$<$ p_{T} $<$100 GeV.

6 Quark and gluon jet identification

6.1 Observables

The separation of quark and gluon jets based on internal jet properties has been extensively studied and validated with CMS 2012 Run 1 data [2]. Several CMS Run 1 analyses have implemented methods to discriminate between quark- and gluon-induced jets to improve Higgs boson searches [50, 51] and electroweak measurements [52, 53]. In what follows, the performance of the same Run 1 discriminating variables and likelihood discriminator approach has been studied in the 2015 Run 2 data. The present study is performed using AK4 jets with PF+CHS inputs which differs from the Run 1 study which used AK5 jets with PF inputs.

The discriminating variables used for quark/gluon discrimination are based on the PF candidate jet constituents that are either well associated to the primary interaction vertex (for charged constituents), or have a transverse momentum p_{T} >1 GeV (for neutral constituents), ensuring both robustness with respect to particle reconstruction and pileup contamination, and sensitivity to the jet fragmentation differences. The three discriminating variables chosen based on these inputs are described below.

- The jet constituents multiplicity. This variable is particularly discriminating for high-p_{T} jets as the average gluon-to-quark multiplicity ratio grows with energy and would ideally converge to $C_A/C_F = 9/4$.
- The jet minor angular opening (σ_2) of the p_{T}^2-weighted constituents distribution in the $\eta - \phi$ plane [2]. This variable is particularly useful for jets with lower p_{T} where gluon-jets are substantially wider than quark-jets.
- The jet fragmentation distribution $p_{\mathrm{T}}D$ defined as $p_{\mathrm{T}}D = \frac{\sqrt{\sum_i p_{\mathrm{T},i}^2}}{\sum_i p_{\mathrm{T},i}}$ where the sum runs over the jet constituents. This variable takes values between zero and one, higher for quark-jets, and provides very good discrimination power for the full p_{T} spectrum considered.

6.1.1 Discriminator

The final discriminator is a likelihood built from the product of the probability density functions (PDFs) of the three variables described above: the jet constituent multiplicity, $p_{\mathrm{T}}D$ and the σ_2. The PDFs are built from jets in simulated QCD dijet events, which have been successfully tagged as light-quark or gluon jets.

The PDFs are computed separately in eight exclusive pseudorapidity regions with upper boundaries $|\eta|$=1.3, 1.5, 1.8, 2.1, 2.4, 2.7, 3.0, 4.7. In order to take into account the dependence of the means and shapes of the variables both on the jet p_{T} and the amount of PU in the event, the PDFs are computed double-differentially in bins of jet p_{T} and ρ, where ρ is the average p_{T}-density per unit area (in η-ϕ space) in an event [34] and is closely correlated to the amount of PU in an event. The transverse momentum binning is logarithmically spaced with boundaries at jet $p_{\mathrm{T}} = 20 \cdot 100^{i/20}$ GeV, for $i = 0, \ldots, 20$. The binning in ρ is linearly spaced with boundaries at $\rho = 8, 11, 14, 17$ GeV.

The binning of the likelihood in this analysis differs from the Run 1 study in order to improve discrimination performance in some kinematic regions. Discriminator peformance has also been improved by adding a tunable weight to each variable used in the likelihood, that weights the variables according to their discrimination power as a function of p_{T}. At high p_{T}, where discrimination power comes mainly from the multiplicity variable that is weighted higher compared to the other two variables, order of 2% in efficiency can be gained at same

background rejection rate.

The input variables to the quark-gluon likelihood and the final likelihood discriminant (LD) are shown for light-quark and gluon flavor (see Sec. 4) simulated jets in Fig. 5. The identification efficiency to tag quark jets as a function of the gluon jet rejection is shown in Figs. 6 (a) and (b).

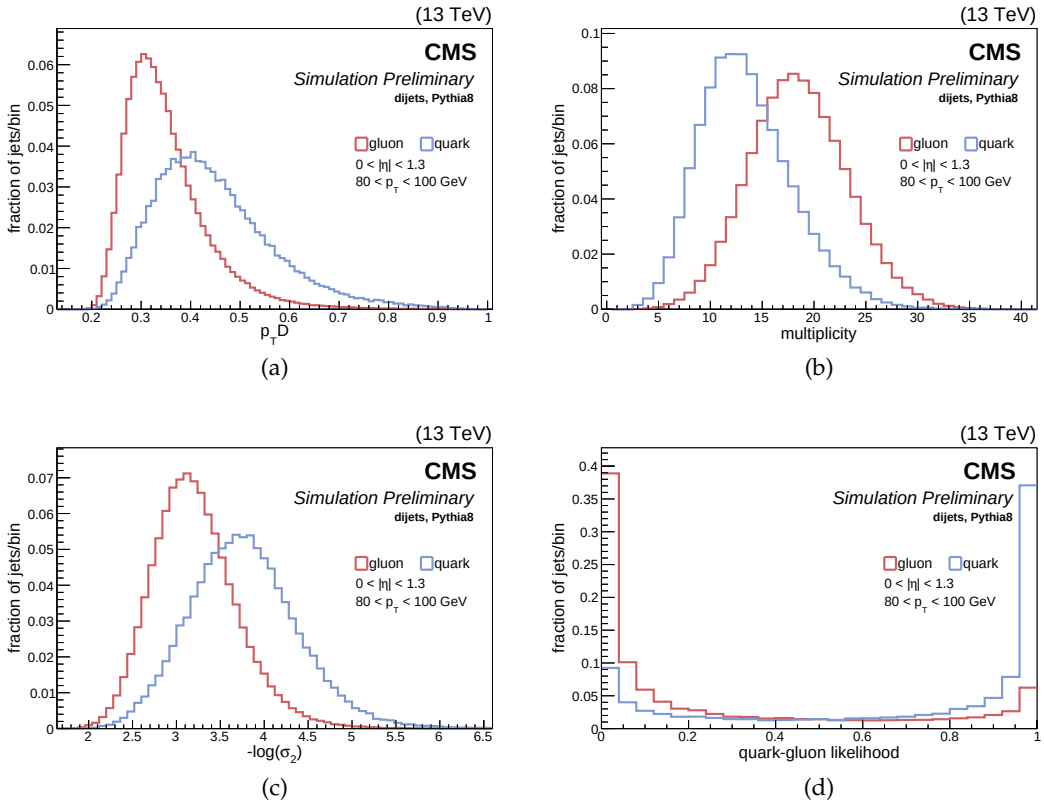


Figure 5: Quark-gluon discrimination variables from simulation: (a) $p_{\mathrm{T}}D$ (b) multiplicity (c) σ_2 (d) the quark-gluon likelihood.

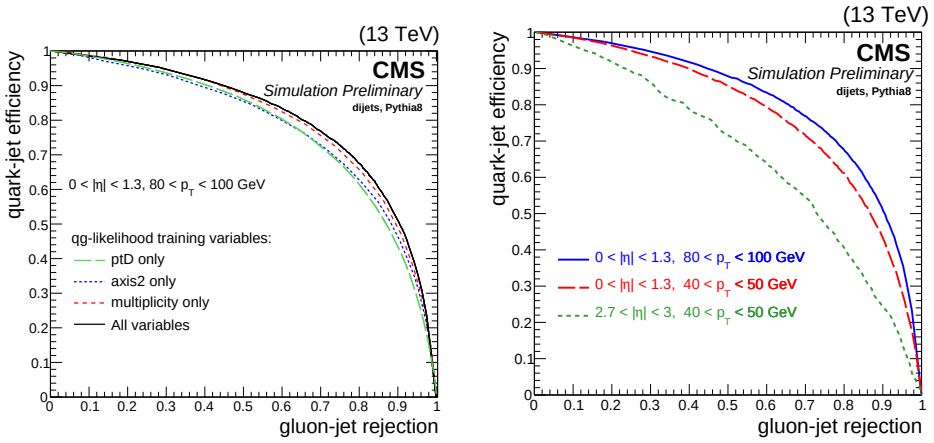
6.2 Validation on Z+jets and dijet events

The discriminator is validated in data using two event selections: Z+jets events, which are quark-enriched; and dijet events, which are gluon-enriched. By the simultaneous use of these two control samples, the performance of the discriminator can be verified on both parton flavors, and across the whole phase space. The following sections detail the event selection and the obtained results in these two control samples. In what follows, all MC distributions are normalized to the integral of the data, as the interest lies in a comparison of the variable shapes.

The Z+jets control sample offers a relatively pure sample of quark jets in which more than 70% of hard ($p_{\mathrm{T}} > 100$ GeV) and central ($|\eta| < 2$) jets are expected to originate from light-quark hadronizations.

Data events containing a pair or muons consistent with a Z boson decay are used for this study. The sample is collected by single muon trigger paths with a p_{T} threshold of 20 GeV, and corresponds to an integrated luminosity of 2.3 fb^{-1}. The event selection further requires:

- the presence of two muons of opposite charge with p_{T} >20 GeV;
- the dimuon invariant mass to fall in the 70-110 GeV range;



(a) (b)

Figure 6: Quark jet tagging efficiency as a function of the gluon jet rejection rate: (a) individual variable discrimination rate compared to the full likelihood (b) likelihood performance in different kinematic regions.

- the dimuon system and the jet with the highest transverse momentum to be back-to-back in the transverse plane by requiring their azimuthal difference to be greater than 2.1 rad;
- the subleading jet in the event to have a p_{T} smaller than 30% of that of the dimuon system.

The leading jet in the event is considered to validate the discriminating variables, and the resulting likelihood.

A data-MC comparison of the input variables to the quark/gluon likelihood, as measured in the Z+jets control sample, is shown in Fig. 7, and the discriminator output is shown in Fig. 8. Similar level of agreement between data and simulation is observed in the other kinematic regions.

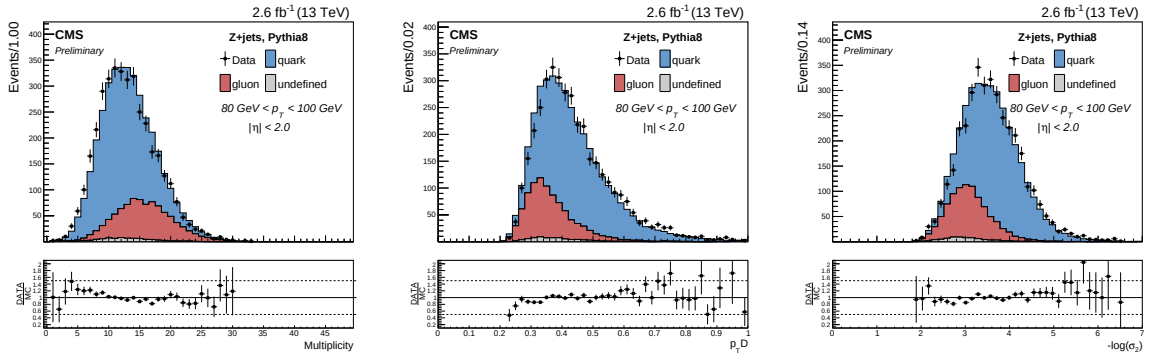


Figure 7: Data-MC comparisons, for jets with $80 < p_{\mathrm{T}} < 100$ GeV and $|\eta| < 2$ in Z+jets events, of the three input variables used in the discriminator: multiplicity (right), $p_{\mathrm{T}}D$ (center) and σ_2 (right). The data (black markers) are compared to the MADGRAPH/PYTHIA simulation, on which the different components are shown: quarks (blue), gluon (red) and unmatched/pileup (grey).

The dijet data control sample has been collected using prescaled zero bias trigger paths, collected uniformly during the data taking period, and corresponding to an integrated luminosity

of 23 nb^{-1}, that takes into account the large prescale factors deployed. These trigger paths allow us to reach the low-p_{T} regime without suffering from trigger biases as would be the case for jet-based triggers. The offline event selection further requires

- two jets with $p_{\text{T}} > 30$ GeV;
- the two p_{T}-leading jets to be back-to-back in the transverse plane by requiring their azimuthal difference to be greater than 2.5 rad;
- the third jet in the event to have a p_{T} less than 30% of the average p_{T} of the two leading jets.

In order to minimize jet migration effects due to jet energy resolution, which could have a large impact at low transverse momentum, a dijet tag-and-probe approach is pursued: one of the two jets (the tag jet) is used to identify the p_{T} interval, while the other jet (the probe jet) is used to identify the η region and to fill the histograms with its properties (either the input variables or the discriminant value). This is done twice per event, so that each jet is alternatively a tag jet and a probe jet.

A comparison of the discriminator output in data and MC dijet events is shown in Fig. 8 (right) for jets with $|\eta| < 2$ and $80 < p_{\text{T}} < 100$ GeV. A significant disagreement is visible among the observed data and Monte Carlo predicted distributions, and a correction procedure, that will serve also as an assessment of systematic uncertainties, will be discussed in the next section.

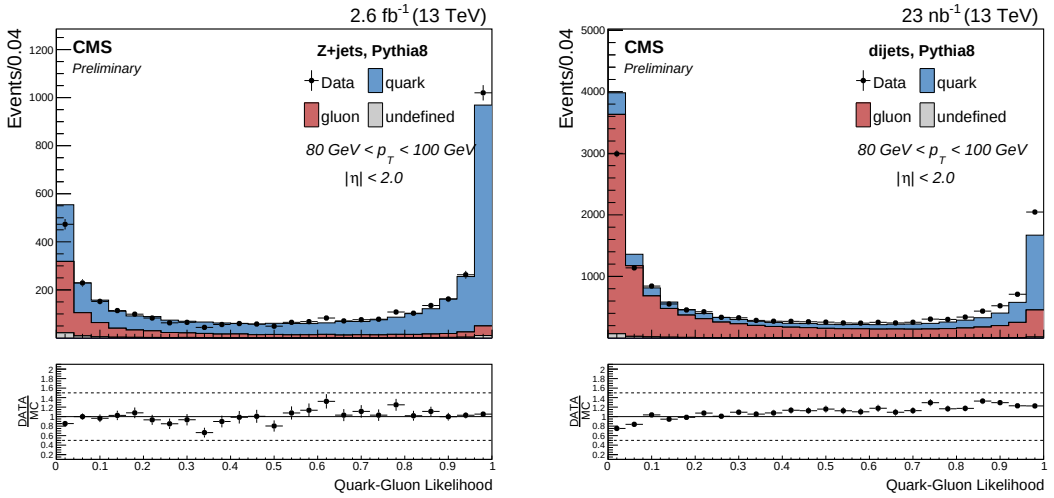


Figure 8: Data-MC comparison for the quark-gluon discriminant in Z+jets (left) and dijet (right) events for jets in the central region with $80 < p_{\text{T}} < 100$ GeV. The data (black markers) are compared to the MADGRAPH/PYTHIA simulation, on which the different components are shown: quarks (blue), gluon (red) and unmatched/pileup (grey).

6.3 Systematic uncertainties

In order to allow the estimation of systematic uncertainties in an analysis context, we estimate the shape uncertainty on the likelihood discriminant output using a generally applicable recipe that takes into account the discriminator shape variations observed in the validation of the simulated samples. For this purpose the data and MC samples are used to define shape differences of the discriminant output, separately for quark and gluon jets. The shape differences are used to quantify the combined effect of the systematic uncertainties of the MC.

In contrast to the more simplistic approach of Ref. [2], we attempt to reshape the MC distri-

butions of the quark and gluon components with appropriate weights as a function of the jet likelihood output by constraining them to the yields observed in data. In order to obtain the weights, two samples with different quark and gluon fractions are used simultaneously. Using the dijet (jj) and Z+jet (Zj) samples, we obtain the weights w_i^q and w_i^g to be applied to the quark and gluon jet MC components in each likelihood output bin i by solving the following linear system in each kinematic region defined by the p_T and η bin:

$$\begin{aligned} N_i^{\text{DATA}}(jj) &= w_i^q N_i^q(jj) + w_i^g N_i^g(jj) + N_i^{un}(jj) \\ N_i^{\text{DATA}}(Zj) &= w_i^q N_i^q(Zj) + w_i^g N_i^g(Zj) + N_i^{un}(Zj) \end{aligned}$$

where N_i^q, N_i^g and N_i^{un}, represent the MC-predicted yields in bin i respectively for the quark, gluon and unidentified jets. In practice the output distributions (for example Fig. 8) are used to solve the above linear system for each likelihood output bin. The resulting weights w_i^g and w_i^q extracted simultaneously from dijet and Z+jet data, are then fitted with polynomial functions of seventh and third degree respectively for gluon- and quark-jets, in order to obtain a smooth interpolation of the weighting coefficients.Applying these to the simulated distributions, the fitted polynomial functions are then applied back as weights to the MC distributions, and the resulting reshaped distributions are shown in Fig. 9.

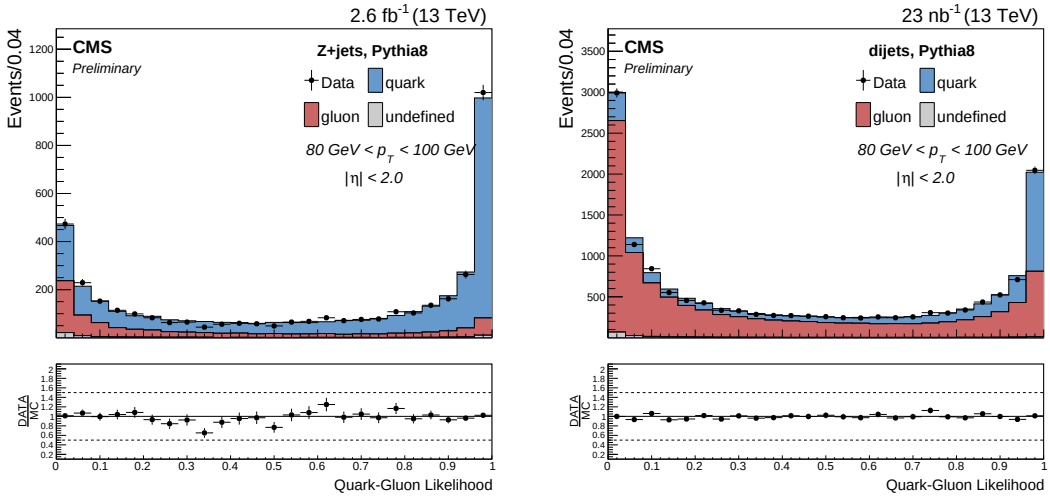


Figure 9: Data-MC comparison for the quark-gluon discriminant in Z+jets (left) and dijet (right) events for jets in the central region with $80 < p_T < 100$ GeV, after the data-driven systematics reshaping procedure. The data (black markers) are compared to the reshaped MADGRAPH/PYTHIA simulation, on which the different components are shown: quarks (blue), gluon (red) and unmatched/pileup (grey).

In order to quantify the systematic change in performances after the data-driven reshaping, the efficiencies to select gluon- and quark-jets with using a fixed cut on the likelihood output are evaluated before and after the MC reshaping. The comparisons of efficiencies are shown in Fig. 10 as a function of the jet transverse momentum, as evaluated on the dijet event sample. The absolute changes in performance are below 1% for quark-jets but can reach the 10% level for gluon jets.

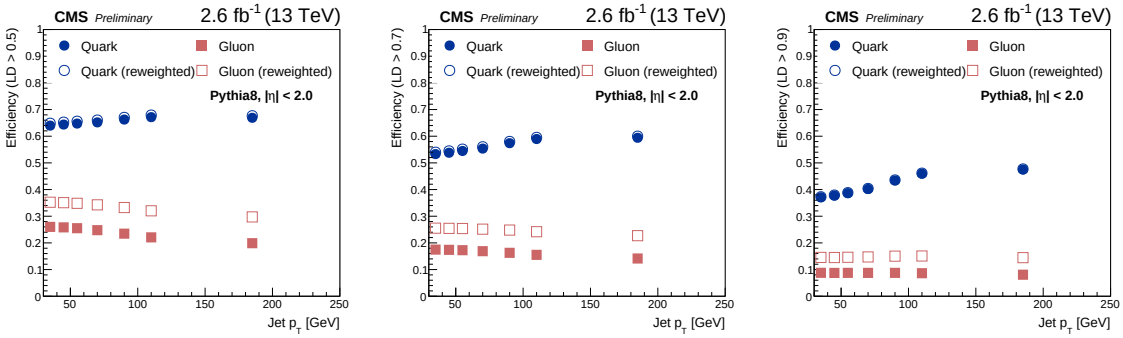


Figure 10: Gluon- and quark-jet selection efficiencies by applying a fixed cut on the likelihood output LD>0.5 (left), 0.7 (center) and 0.9 (right). Efficiencies are evaluated in dijet events, as a function of the jet p_{T}, before and after the reshaping of the outputs.

6.4 Comparison of performances with the HERWIG showering model

The discriminative power of the quark/gluon likelihood is subject to systematic uncertainties from the details of the modeling of the parton shower To further expand the systematic uncertainties studies, a full set of comparisons have been produced for both dijet and Z+jet event samples, using, instead of PYTHIA8, the HERWIG++ (version 2.7.0) [54, 55] showering model, with the CUETHS1 tune. Note that for this comparison both the PYTHIA and HERWIG++ versions and parameter tunes are updated with respect to those used for the Run 1 data studies in Ref. [2].

The procedure outlined in Sec. 6.3 to reshape the Monte Carlo predictions based on the data distributions of the likelihood output has been applied to the HERWIG++ samples. The performance differences between the two showering options, with and without the data-driven reshaping, are again evaluated as the efficiencies to select gluon- and quark-jets with using a fixed cut on the likelihood output. The full comparisons of efficiencies are shown in Fig. 11 as a function of the jet transverse momentum, as evaluated on the dijet or Z+jets event samples.

Before reshaping, the data-driven corrections differences between the HERWIG++ and PYTHIA8 simulations are sizeable with larger efficiency differences in the case of gluon jets. After the data-driven reshaping the efficiencies predicted by either HERWIG++ or PYTHIA8 are within the percent level of disagreement.

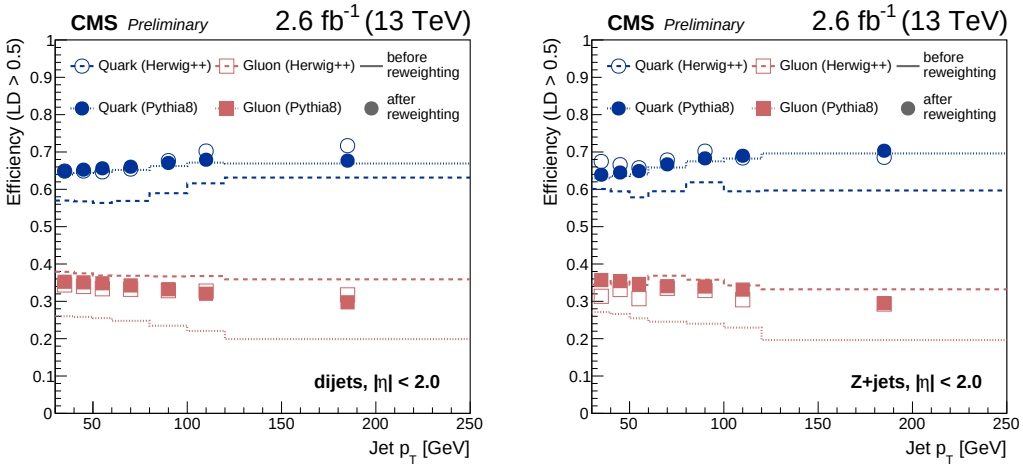


Figure 11: HERWIG++ (v2.7.0 with CUETHS1 tune)) and PYTHIA8 (v8.205 with CUETP8M1 tune) gluon- and quark-jet selection efficiencies by applying a fixed cut on the likelihood output LD>0.5. Efficiencies are evaluated in dijet events (left) or Z+jet events (right), as a function of the jet p_{T} with or without the data-driven reshaping of the outputs.

7 Pileup jet identification

7.1 Observables

Pileup collisions at the LHC deposit energy randomly throughout the CMS detector. In some cases, many low p_{T} energy deposits can overlap, resulting in one high p_{T} jet (known as a "pileup jet"). Pileup jets can be identified and rejected by utilizing jet-shape and tracking observables. In removing charged hadrons from pileup vertices before clustering jets, the CHS procedure effectively reduces the rate of PU jets. Without this procedure, the PU jet rate in the region $|\eta| < 1.3$ would increase by a factor of 5.

The identification of pileup jets is based on two observations:

1. The majority of tracks associated with pileup jets come from non-primary vertices.
2. Pileup jets originate from overlapping particles from pileup collisions and therefore tend to be more broad and diffuse than jets originating from one single quark or gluon from the hard scatter.

Based on these observations, numerous variables have been developed which can identify pileup jets [4]. Track-based variables are defined based on observation (1), while jet shape observables are defined based on observation (2). The pileup jet identification (ID) algorithm at CMS relies on 14 variables in the central tracking region and 12 variables in the forward region. These variables are combined with a boosted decision tree (BDT).

The track-based variables include β and $N_{vertices}$, where β is the sum p_{T} of all PF charged candidates in the jet originating from the primary vertex divided by the sum p_{T} of all charged candidates in the jet:

$$\beta = \frac{\sum_{i \in PV} p_{Ti}}{\sum_i p_{Ti}} \tag{4}$$

The β variable provides the strongest discrimination of any variable included in the likelihood, but is available only within the tracking volume. The inclusion of the $N_{vertices}$ variable allows the BDT to determine the optimal discriminating variables as the pileup is increased.

The jet-shape variables included in the BDT are as follows: $\langle \Delta R^2 \rangle$, f_{ring0}, f_{ring1}, f_{ring2}, f_{ring3}, $p_{\mathrm{T}}^{\mathrm{lead}} / p_{\mathrm{T}}^{\mathrm{jet}}$, $|\vec{m}|$, N_{total}, $N_{charged}$, major axis (σ_1), minor axis (σ_2), and p_{T}^{D}.

The first jet-shape variable is defined as

$$\langle \Delta R^2 \rangle = \frac{\sum_i \Delta R_i^2 p_{Ti}^2}{\sum_i p_{Ti}^2} \tag{5}$$

where the sum runs over all PF candidates inside the jet and $\Delta R = \sqrt{\Delta\eta^2 + \Delta\phi^2}$ is the distance of the PF candidate with respect to the jet axis. Jets originating from pileup tend to have larger $\langle \Delta R^2 \rangle$.

The annulus variables, f_{ringX} are defined as the the fractional $\sum p_{\mathrm{T}}$ of particle flow candidates in an annulus around the jet direction. f_{ring0} corresponds to $0 < \Delta R < 0.1$, f_{ring1} corresponds to $0.1 < \Delta R < 0.2$, etc. Pileup jets tend to have a higher energy fraction in the large R annulus, while quark and gluon jets tend to deposit more energy close to the jet axis.

The leading constituent p_T ratio p_T^{lead}/p_T^{jet} is defined as the p_T of the highest p_T jet constituent divided by the jet p_T. The pull magnitude variable is defined in [56, 57] as

$$|\vec{m}| = \left| \sum_i \frac{p_T^i |r_i|}{p_T^{jet}} \vec{r}_i \right| \quad (6)$$

where the sum runs over all PF candidates inside the jet and $\vec{r}_i = (\Delta y_i, \Delta\phi_i) = \vec{c}_i - \vec{J}$, where $\vec{J} = (y_J, \phi_J)$ is the location of the jet and $\vec{c}_i$ is the position of a particle with transverse momentum p_T^i. The multiplicity variables $N_{charged}$ and N_{total} are defined as the number of charged particles in the jet and the total number of particle in the jet, respectively.

The minor axis (σ_2), and p_T^D variables are defined in Section 6.1. The pileup discriminant also utilizes the major axis σ_1, defined as the major angular opening angle of the p_T^2-weighted constituents distribution in the η-ϕ plane. In the case of pileup jet identification, all the neutral candidates of a jet are used when calculating these variables, whereas for the CMS quark-gluon discriminator neutral candidates having a $p_T > 1\,\text{GeV}$ are used in order to make the quark-gluon discriminator more resistant to the effects of pileup contamination. As pileup jets tend to have lower p_T^D than gluon jets, the addition of this variable enhances the gluon-pileup separation, particularly at high η.

7.2 Pileup jet identification training

The pileup jet ID BDT is trained using the variables defined in Section 7.1. The BDT training and optimization of the working points are done separately in four regions corresponding to the four different regions of the calorimeters: the tracker volume ($|\eta| < 2.5$), the tracker-endcap transition region ($2.5 < |\eta| < 2.75$), the endcap region ($2.75 < |\eta| < 3.0$) and the HF region ($3.0 < |\eta| < 5.0$). The tracker volume corresponds to the region where the jet axis is contained in the tracker acceptance, thus at least half the tracks inside the jet are reconstructed. The transition region corresponds to the region where part of the jet is typically within the tracker volume and thus tracking variables can still be used, however their behavior is different to those within the tracker volume. The endcap region corresponds to the region where the HCAL and ECAL endcap are still present. The HF region corresponds to the region where the central jet axis lies in HF.

The training is done using a Z+jets MC simulation, since the quark-gluon composition of the jets in this sample is representative of for the most important use case of pileup jet identification, namely rejection of background in vector boson fusion topologies of standard model Higgs boson production. Pileup jets, quark jets, and gluon jets are defined based on matching the jet to generated partons from the hard scattering and simulated jets. Quark and gluon jets are defined as jets which are with ΔR <0.2 of a hard process quark or gluon. Pileup jets are defined as jets which are not matched to a simulated jet (ΔR >0.3), nor are they matched with a parton within ΔR <0.2. Any jet not falling into these categories is defined as "rest" and not used for the training of the BDT.

For central jets with $|\eta| < 2.5$ and 30< p_T <50 GeV, the pileup jet identification BDT rejects 89% of pileup jets while maintaining 96% of gluon jets. The rejection rate as a function of the quark and gluon jet efficiency is shown in Fig. 12.

7.3 Validation in data

The performance of the pileup jet ID BDT is evaluated using a sample of Z+jets events, with the Z boson decaying to muons. This allows for a clean definition of the recoiling p_T, for which

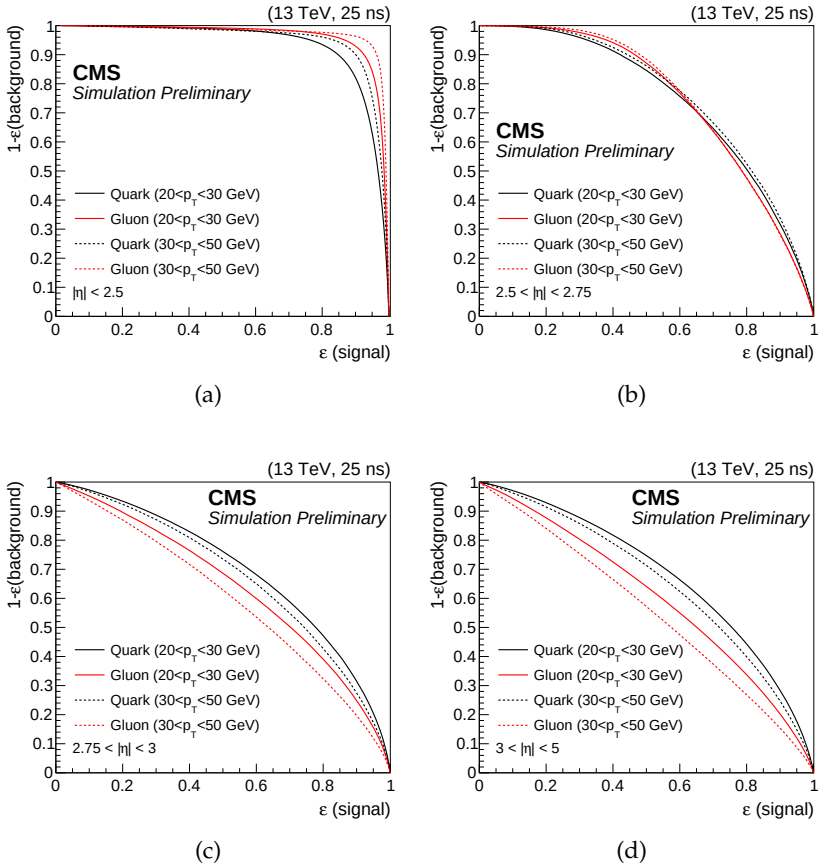


Figure 12: Fraction of rejected pileup jets as a function of the fraction of true quark and gluon jets which are correctly tagged for jets. Curves are shown for both quark initiated and gluon initiated jets in bins of $20 < p_T < 30$ GeV and $30 < p_T < 50$ GeV in four different $|\eta|$ regions: (a) $|\eta| < 2.5$, (b) $2.5 < |\eta| < 2.75$, (c) $2.75 < |\eta| < 3$, (d) $3 < |\eta| < 5$.

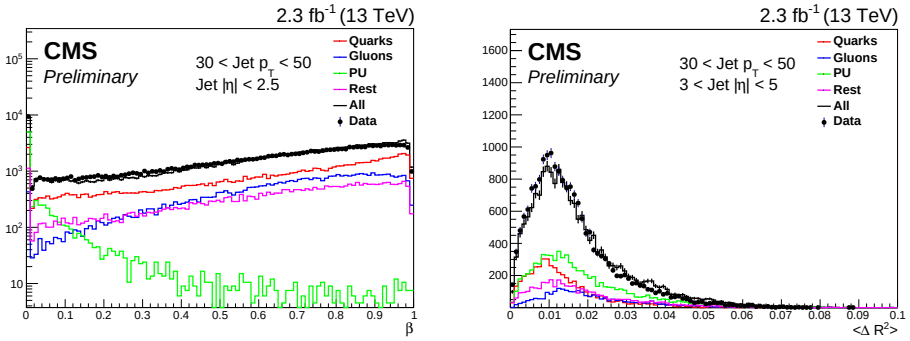
jets can be balanced against. Data are compared to a Drell-Yan MC sample simulated with MADGRAPH, and PYTHIA8 used for showering.

Events are required to pass the di-muon trigger, with thresholds on the muon transverse momenta of 17 GeV and 8 GeV respectively. $Z\rightarrow \mu\mu$ events are selected by requiring two isolated muons with $p_T > 20$ GeV and $|\eta| < 2.4$, with an invariant mass in a window of 30 GeV around the nominal Z mass. The muons must have opposite charge. The analysis is completed on all AK4 PF+CHS jets with $p_T > 20$ GeV and $|\eta| < 5.0$ which are separated from the muons by $\Delta R >$0.4.

Two examples of variable inputs to the BDT are included in Fig. 13, and the final pileup jet ID discriminant is shown in Fig. 14. The simulation is found to describe the data sufficiently well to allow for a correction strategy based on data/MC efficiency scale factors.

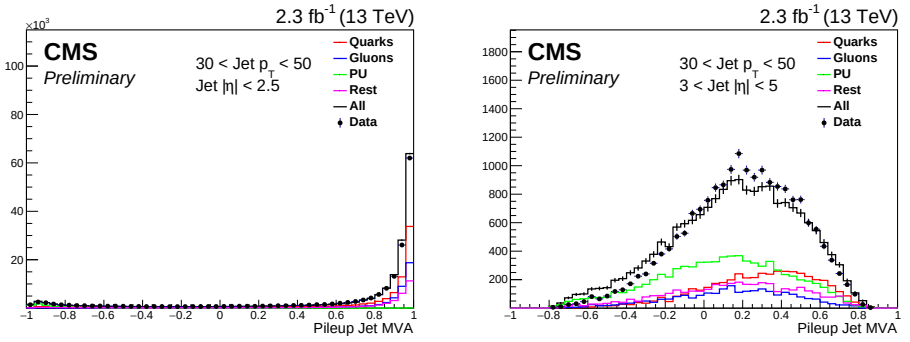
7.4 Data/MC scale factors for efficiencies

The efficiency of the pileup jet identification criteria on real jets is checked using a tag-and-probe method on a control sample of $Z(\rightarrow \mu\mu)$+jets events, where the jet recoiling against the Z is used as a probe. In order to reduce the pileup contamination on the probe side, requirements on the balancing between the Z and the hardest jet momenta are applied: the absolute



(a) (b)

Figure 13: Pileup jet MVA discriminant (left) β measured in central region (right) $\langle \Delta R^2 \rangle$ measured in forward region.



(a) (b)

Figure 14: Pileup jet MVA discriminant for (left) central and (right) forward jets.

azimuthal separation $|\Delta\phi(Z, j)|$ between the Z and the jet must be larger than 2.5 and the ratio between the jet p_T and the Z p_T must be between 0.5 and 1.5. Under the assumption that the $\Delta\phi(Z, j)$ distribution is flat for pileup jets, the residual background due to pileup jets in the control sample (both before and after applying the pileup jet id) is estimated from the pileup enriched region with $|\Delta\phi(Z, j)| < 1.5$. The efficiency on real jets is therefore computed as:

$$\epsilon = \frac{N_{passId,sig} - k \cdot N_{passId,bkg}}{N_{all,sig} - k \cdot N_{all,bkg}} \tag{7}$$

where $N_{all,sig}$ is the total number of jets in the control region ($|\Delta\phi(Z, j)| > 2.5$) , $N_{all,bkg}$ is the total number of jets in the pileup enriched region ($|\Delta\phi(Z, j)| < 1.5$), $N_{passId,sig}$ is the number of jets in the control region passing the jet identification, $N_{passId,bkg}$ is the number of jets passing the jet identification in the pileup enriched region and, finally, $k = (\pi - 2.5)/1.5$ is a phase space factor defined by the allowed angular regions used in the tag and probe method to extrapolate the number of pileup jets from the pileup enriched region to the control sample.

The results of the efficiency measured in data and MC simulation and of their ratio are reported in Fig. 15. As shown, the agreement between data and simulation is within 2-10% depending on the jet pseudorapidity and transverse momentum range. The ratios of data to simulation in bins of jet p_T and η are used as scale factors to correct for mismodeling in the simulation. The largest data/MC scale factors are observed for the forward region as a consequence of the

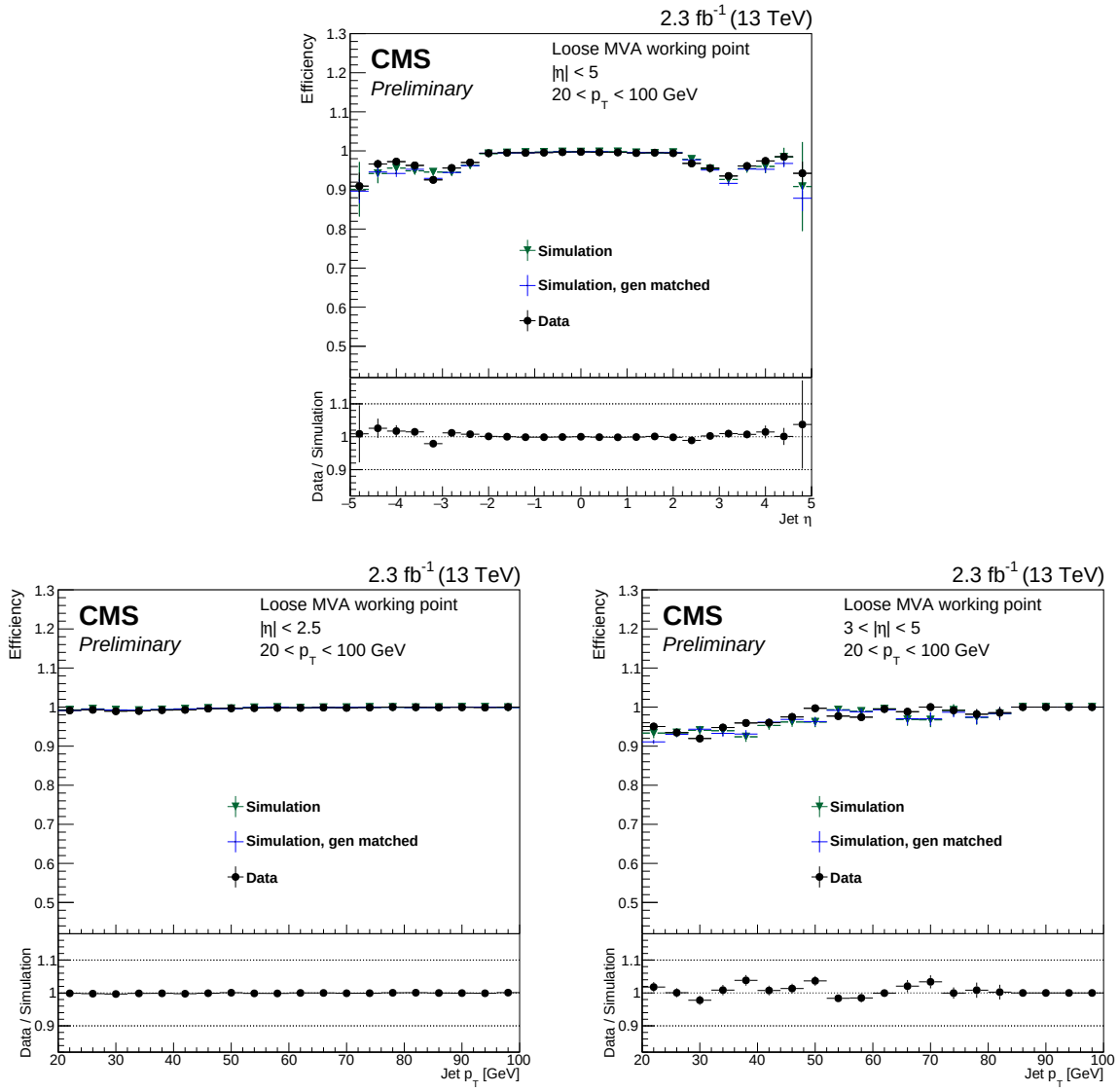


Figure 15: Data-MC comparison of the MVA (loose working point) pileup jet identification efficiency on the Z($\rightarrow \mu\mu$)+jets sample for PF jets with $p_{\mathrm{T}} > 20$ GeV: the efficiency is shown as a function of the jet pseudorapidity (top) and as a function of p_{T} for jets with $|\eta| < 2.5$ (bottom-left) and $3 < |\eta| < 5$ (bottom-right).

data/MC differences on the pileup discriminator shown in Fig. 14 (right).

8 W jet identification

8.1 Observables

When a hadronically decaying W boson has sufficient transverse momentum, the two associated jets can be merged into a single jet in reconstruction. These "W jets" are formed when W bosons are produced at large Lorentz boost ($p_T > 200$ GeV) and the final state hadrons from the W$\to \bar{q}q'$ decay become highly collimated W jets can be identified with jet substructure observables. The W jet mass is similar to that of the W boson mass and the internal structure of the jet contains two distinct subjets.

The anti-k_T jet clustering algorithm with $R = 0.8$ (AK8) is used to reconstruct W jets. We study AK8 jets with both PF+CHS and PF+PUPPI inputs. Subsequent to the AK8 clustering, the constituents of these jets are reclustered with the Cambridge-Aachen algorithm [29, 30]. Two different jet grooming algorithms are investigated in this note. The first is the "jet pruning" algorithm [58, 59], which attempts to remove soft and wide-angle contributions to jets during reclustering, controled by a soft threshold parameter z_{cut} set to 0.1 here and an angular separation threshold of $\Delta R > m_{jet}/p_{T,jet}$. This algorithm was used extensively during Run 1 of the LHC at CMS. Following theoretical work [60, 61] that aimed to understand and calculate jet mass observables in QCD multijet events, a new algorithm was developed to accomplish jet grooming in a theoretically calculable way, the "soft drop" algorithm [6]. In addition to the soft threshold parameter z_{cut} set to 0.1, the soft drop algorithm depends on an angular exponent β that is set to 0 here. In the case of $\beta = 0$, the soft drop algorithm is equivalent to the modified mass-drop tagger (mMDT) as detailed in [60, 62]. The softdrop algorithm is primarily aimed at separating W-jets from q/g-jets and does not fully reject contributions from underlying event and pileup. Therefore, a pileup suppression algorithm that corrects also the shape of a jet is promising. Here, we study softdrop in combination with the PUPPI algorithm [36].

In addition to these jet grooming algorithms, the "N-subjettiness" of jets [7] is also extensively used to identify boosted vector bosons that decay hadronically. This observable measures the distribution of jet constituents relative to candidate subjet axes in order to quantify how well the jet can be divided into N subjets. Subjet axes are determined by a one-pass optimization procedure which minimizes N-subjettiness [63]. The direction of all the jet constituents w.r.t. the closest of the the N subjets are then used to compute the "N-subjettiness" as $\tau_N = \frac{1}{d_0}\sum_k p_{T,k} \times \min(\Delta R_{1,k}, \Delta R_{2,k}, ..., \Delta R_{N,k})$ with the normalization factor $d_0 = \sum_k p_{T,k} \times R_0$, where R_0 is the clustering parameter of the original jet, $p_{T,k}$ is the p_T of the k-th jet constituent and $\Delta R_{n,k} = \sqrt{(\Delta\eta_{n,k})^2 + (\Delta\phi_{n,k})^2}$ is its distance to the n-th subjet. In particular, the ratio of the "2-subjettiness" to the "1-subjettiness" ($\tau_2/\tau_1 = \tau_{21}$) has excellent capability of separating jets originating from boosted vector bosons from jets originating from quarks and gluons.

Finally, we study the "DDT" transformation of τ_{21} as presented in [8]. This approach takes the linear correlation of τ_{21} against a QCD-invariant type variable to transform it to be flat. The transformation of the variable is given as the trivial linear transformation:

$$\tau_{21}^{DDT} = \tau_{21} - M \times \rho^{DDT} \tag{8}$$

The slope $M = -0.063$ is obtained from the profile distribution of τ_{21} against the variable $\rho^{DDT} = \log(m^2/p_T/\mu)$, where $\mu = 1\,\text{GeV}$. The groomed jet mass distribution for QCD jets is a smoothly falling distribution, but after applying a τ_{21} selection the groomed jet mass is shaped such that it peaks at some value. The transformed variable τ_{21}^{DDT} maintains the smoothly falling groomed jet mass distribution after selection.

In the following sections we examine the simulated and observed p_T and PU dependence of the

W tagging efficiency and then extract data-to-MC scale factors based on merged W jets from $\bar{t}t$ production. In addition, a measurement of the mistag rate is provided.

8.2 Performance in simulation

In this section we examine the p_T and PU dependence of the W tagging efficiency in the MC simulation. The figure of merit for comparing different substructure observables is the background rejection efficiency as a function of signal efficiency ("receiver operating characteristic", or the ROC curve). Figure 16 shows the performance of the observables under study. The efficiency is measured for the joint condition on m_{jet} and τ_2/τ_1, demonstrating the impact of these discriminants.

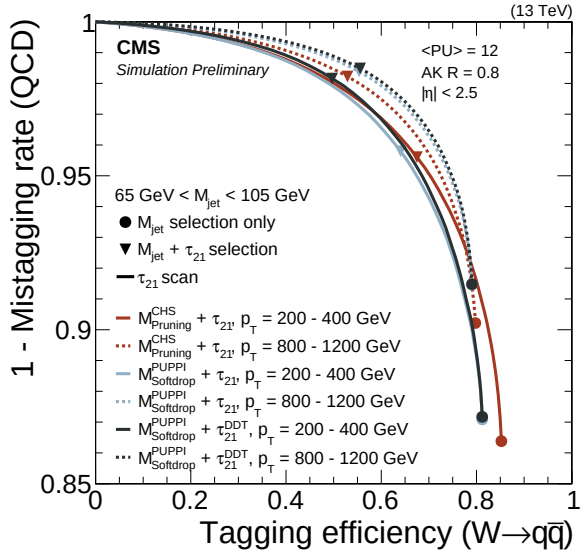


Figure 16: Performance of several discriminants in the background-signal efficiency plane. The baseline selection for W tagging requiring a PF+CHS pruned or PF+PUPPI softdrop jet mass of $65 < m_{\text{jet}} < 105$ GeV, and N-subjettiness ratio (PF+CHS inputs) of $\tau_2/\tau_1 < 0.45$ or N-subjettiness ratio (PF+PUPPI inputs) of $\tau_2/\tau_1 < 0.4$ or $\tau_{21}^{\text{DDT}} < 0.52$ are indicated with symbols.

Figure 17 shows the efficiency of the baseline selection ($65 < m_{\text{jet}} < 105$ GeV and PF+CHS $\tau_2/\tau_1 < 0.45$ or PF+PUPPI $\tau_2/\tau_1 < 0.4$ or $\tau_{21}^{\text{DDT}} < 0.52$.) determined from a WW simulation, requiring at generator level that the quarks from the W decay are within $\Delta R < 0.8$ of the reconstructed jet. The efficiency is given as a function of (a) jet p_T and (b) the number of reconstructed vertices, reflecting the contribution from pileup. The corresponding mistag rate is shown in (c) and (d). High efficiency up to at least $p_T = 2.5$ TeV is achieved with recent improvements to the CMS particle flow algorithm exploiting the full potential of the CMS ECAL granularity to resolve jet substructure [11]. While the PF+CHS pruned jet mass and PF+PUPPI softdrop jet mass selection yield very similar efficiency at high p_T, the PF+PUPPI softdrop jet mass has lower efficiency in the range $200 < p_T < 300$ GeV. This difference is an effect of the PUPPI pileup suppression and not due to the differences between the pruning and softdrop algorithms. It is also observed that PF+CHS τ_2/τ_1 and PF+PUPPI τ_2/τ_1 exhibit a decrease in mistag rate and efficiency as a function of p_T, τ_{21}^{DDT} conserves an approximately constant mistag rate and an increasing efficiency as a function of p_T.

The efficiency of the m_{jet} selection as a function of the number of reconstructed vertices, shown in Fig. 17 (b), is constant as a function of number of primary vertices (PV), whereas the additional τ_2/τ_1 selection efficiency drops from 60% at 0 PV to 40% at 30 PV. However, the mistag-

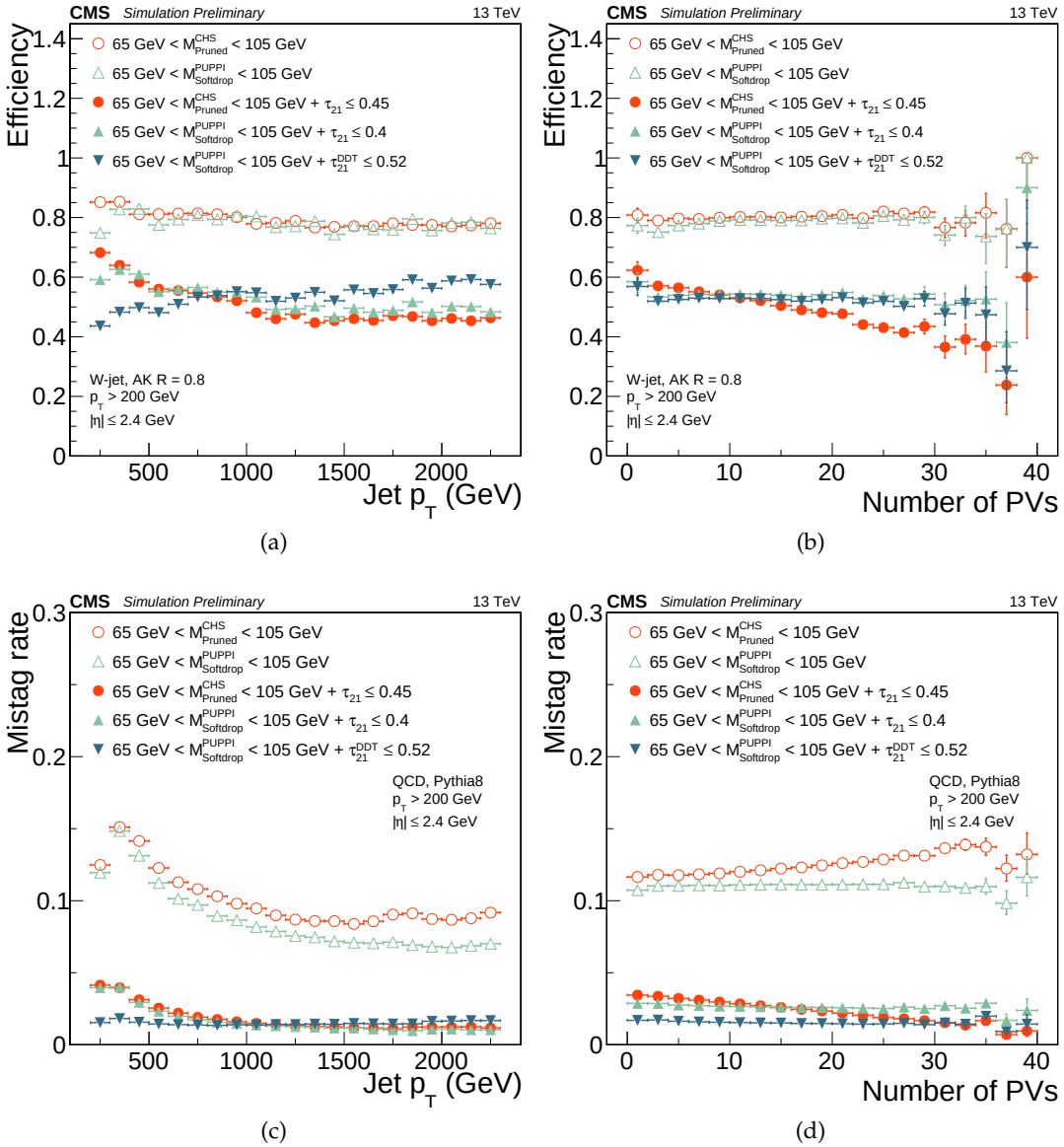


Figure 17: Efficiency of the PF+CHS pruned jet mass and PF+PUPPI softdrop jet mass selection and the combined (PUPPI) τ_2/τ_1 (DDT) and m_{jet} selection on WW signal samples as a function of (a) p_{T} and (b) the number of reconstructed vertices. Reconstructed jets enter (the demoninator and numberator of) the efficiency only if at generator level both quarks from the W decay are within $\Delta R < 0.8$ of the jet axis. (c) Mistag rate of the PF+CHS pruned jet mass and PF+PUPPI softdrop jet mass selection and the combined (PUPPI) τ_2/τ_1 (DDT) and m_{jet} selection on WW signal samples as a function of (c) p_{T} and (d) the number of reconstructed vertices. The error bars represent the statistical uncertainty in the MC simulation and the horizontal ones the binning.

ging of the background also decreases with pileup for the same selection, yielding similar discrimination. Efficiency and mistagging rate are affected by pileup in the same way, since additional pileup shifts the PF+CHS τ_2/τ_1 distribution towards higher values (towards background like) for both signal and background. Therefore, the same signal efficiency can be reached at the same background rejection rate for up to 30 reconstructed vertices by merely adjusting the τ_2/τ_1 selection. For PF+PUPPI $\tau_2/\tau_1 < 0.4$ there is no such dependence on the number of primary vertices.

8.3 Efficiency scale factors and mass scale/resolution measurement

A lepton+jets semileptonic $\mathrm{t\bar{t}}$ selection is used to study a pure sample of hadronic W-jets in data. This sample is then used to measure the W-jet tagging efficiency data-simulation scale factor. The selection is defined by requiring exactly one isolated energetic lepton (electron or muon), large missing transverse momentum, at least one b-tagged AK4 jet, and one high p_T AK8 jet in the final state. The reconstructed muon candidates are required to have transverse momenta $p_\mathrm{T} > 53\,\mathrm{GeV}$ and $|\eta| < 2.1$. The reconstructed electron candidates are required to have transverse momenta $p_\mathrm{T} > 120\,\mathrm{GeV}$ and $0 < |\eta| < 1.442$ or $1.56 < |\eta| < 2.5$ (therefore excluding the transition region between the ECAL barrel and endcaps). Events are required to have $E_\mathrm{T}^\mathrm{miss} > 40\,\mathrm{GeV}$ for the muon channel and $E_\mathrm{T}^\mathrm{miss} > 80\,\mathrm{GeV}$ for the electron channel in order to suppress contribution from QCD multijet background. The event must contain at least one AK4 jet with $p_\mathrm{T} > 30\,\mathrm{GeV}$ and $|\eta| < 2.4$. If an AK4 jet is within $\Delta R < 0.3$ of any tight electron or tight muon, or $\Delta R < 0.8$ of any selected AK8 jets, the jet is not used in the analysis. A jet is considered to be b-tagged if it passes the "medium" working point, corresponding to a misidentification probability of 1%, of the particle flow inclusive CSV algorithm [64]. The p_T of the reconstructed leptonic W must be greater than 200 GeV, while the AK8 jet must satisfy $p_\mathrm{T} > 200$ GeV. Additional requirements to reduce the combinatorial background from $\mathrm{t\bar{t}}$ improve the precision of the determined scale factor. Therefore, the angular distance ΔR between the W jet candidate and the closest b-tagged AK4 jet is required to be less than 2.0, i.e. less than the average separation for boosted top quarks with $p_\mathrm{T} > 300$ GeV [12].

The lepton+jets $\mathrm{t\bar{t}}$ sample is used to extract data-to-simulation scale factors for the W jet efficiency. These factors are meant to correct the description of the W-tagging efficiency in the simulation. They depend on the definition of the W-tagger as well as the MC generator used for simulation. We extract data-to-simulation scale factors for several selections on PF+CHS τ_2/τ_1 and PF+PUPPI τ_2/τ_1, and for the jet mass scale and resolution based on a simulation using POWHEG interfaced with PYTHIA8. We are concerned only with the efficiency for the pure W jet signal, and must therefore subtract background contributions when measuring the scale factors. The PF+CHS pruned jet mass or PF+PUPPI softdrop jet mass distribution is used to discriminate the pure W jet signal from background contributions. The generated W boson in the $\mathrm{t\bar{t}}$ simulation provides a model of the contribution from the W jet peak in the jet mass. The contribution from combinatorial background is derived from $\mathrm{t\bar{t}}$ simulation as well, by splitting the $\mathrm{t\bar{t}}$ simulation into events with pure W jets ("merged"), which are matched in ΔR to generator level $W \to q\bar{q}$ decays and other jets ("unmerged") from $\mathrm{t\bar{t}}$.

The scale factors (SF) for the efficiency of the selection on PF+CHS τ_2/τ_1 and PF+PUPPI τ_2/τ_1 are extracted by estimating the ratio of the selection efficiencies on data and simulation. The PF+CHS pruned jet mass or PF+PUPPI softdrop distribution of events that pass and fail the PF+CHS τ_2/τ_1 or PF+PUPPI τ_2/τ_1 selection are fitted simultaneously to extract the selection efficiency on the pure W jet component as shown in Fig. 18. In simulation a slight shift in mass is found to be primarily due to extra radiation in the W jet from the nearby b quark. We find the "pass" sample agrees well between the data and simulation while the "fail" sample is not

as well modeled, particularly when the failing jet is not a fully merged W boson but a quark or gluon jet. This is compensated in the computation of the data-to-MC scale factor, which is summarized in Table 2. When using a $t\bar{t}$ simulation where MADGRAPH is interfaced with HERWIG++ rather than PYTHIA8, the estimated W-tagging scalefactor is typically closer to unity. For pruning and a $\tau_2/\tau_1 < 0.45$ selection the resulting scalefactor is 1.02±0.06 (HERWIG++) rather than and 0.94±0.06 (PYTHIA8). Systematic effects to this scale factor are described later in this section.

Table 2: Data-to-simulation scale factors for the W-tagging procedure, as extracted from a top enriched data sample and from simulation, for both categories (high purity and low purity) for two different working points. The systematic uncertainties on the scale factor due to the simulation of the $t\bar{t}$ topology and the choice of the signal and background fit model are listed as well.

Category	Definition	W scale factor
High-purity Pruning	$(\tau_2/\tau_1 < 0.45)$	$0.94 \pm 0.05\ (\text{stat}) \pm 0.03\ (\text{sys}) \pm 0.003\ (\text{sys})$
Low-purity Pruning	$(0.45 < \tau_2/\tau_1 < 0.75)$	$1.27 \pm 0.25\ (\text{stat}) \pm 0.13\ (\text{sys}) \pm 0.008\ (\text{sys})$
High-purity Pruning	$(\tau_2/\tau_1 < 0.6)$	$0.98 \pm 0.03\ (\text{stat}) \pm 0.003\ (\text{sys}) \pm 0.02\ (\text{sys})$
High-purity PUPPI softdrop	$(\tau_2/\tau_1 < 0.4)$	$0.97 \pm 0.06\ (\text{stat}) \pm 0.04\ (\text{sys}) \pm 0.06\ (\text{sys})$
Low-purity PUPPI softdrop	$(0.4 < \tau_2/\tau_1 < 0.75)$	$1.12 \pm 0.24\ (\text{stat}) \pm 0.17\ (\text{sys}) \pm 0.12\ (\text{sys})$

To extract corrections to the jet mass scale and resolution, we use the mean $\langle m \rangle$ and resolution σ value of the Gaussian component of the fitted function of the W bosons in the passed sample. The fits are shown for the PF+CHS $\tau_2/\tau_1 < 0.45$ selection in Fig. 18 (a) and for the PF+PUPPI $\tau_2/\tau_1 < 0.40$ selection in Fig. 18 (c), and the resulting parameters are summarized in Table 3. We find that the W jet mass scale in data is 1-2% smaller than in simulation and the W jet mass resolution differs by order of 10% though not statistically significant.

Table 3: Summary of the fitted W-mass peak fit parameters.

Parameter	Data	Simulation	Data/Simulation
pruning $\langle m \rangle$	80.9 ± 0.6 GeV	82.5 ± 0.1 GeV	0.980 ± 0.007
pruning σ	6.7 ± 0.7 GeV	7.5 ± 0.3 GeV	0.89 ± 0.10
PUPPI softdrop $\langle m \rangle$	80.3 ± 0.8 GeV	81.9 ± 0.01 GeV	0.98 ± 0.01
PUPPI softdrop σ	9.0 ± 0.9 GeV	8.5 ± 0.4 GeV	1.07 ± 0.12

A detailed description of the various sources of systematic on the τ_2/τ_1 scale factor can be found in Ref. [9]. Leading systematic effects are due to the simulation of the $t\bar{t}$ topology used to derive the data-to-simulation scale factors. We evaluated an uncertainty due to the simulation of the $t\bar{t}$ topology (nearby jets, p_T spectrum) by comparing the estimated $\tau_2/\tau_1 < 0.45$ selection efficiency in $t\bar{t}$ samples from POWHEG at NLO in QCD interfaced with PYTHIA8 and MADGRAPH at LO in QCD interfaced with PYTHIA8. This uncertainty amounts to 3-17% and is listed in Table 2. In addition, we evaluated an uncertainty due to parton showering by comparing the estimated $\tau_2/\tau_1 < 0.45$ selection efficiency in $t\bar{t}$ samples from POWHEG interfaced with PYTHIA8 and POWHEG interfaced with HERWIG++. This uncertainty amounts to 8.6% and quantifies the discrepancy between the jet substructure modeling of PYTHIA8 and HERWIG++. It is only relevant for analyses applying the data-to-simulation scale factors derived with PYTHIA8 to simulation based on HERWIG++ showering.

The uncertainty on the p_T dependence of the scale factor, when using it for higher momenta jets than the range of the $t\bar{t}$ control sample of $p_T \sim 200$ GeV was studied with WW signal simulation showered by PYTHIA8 and HERWIG++. The difference between PYTHIA8 and HERWIG++

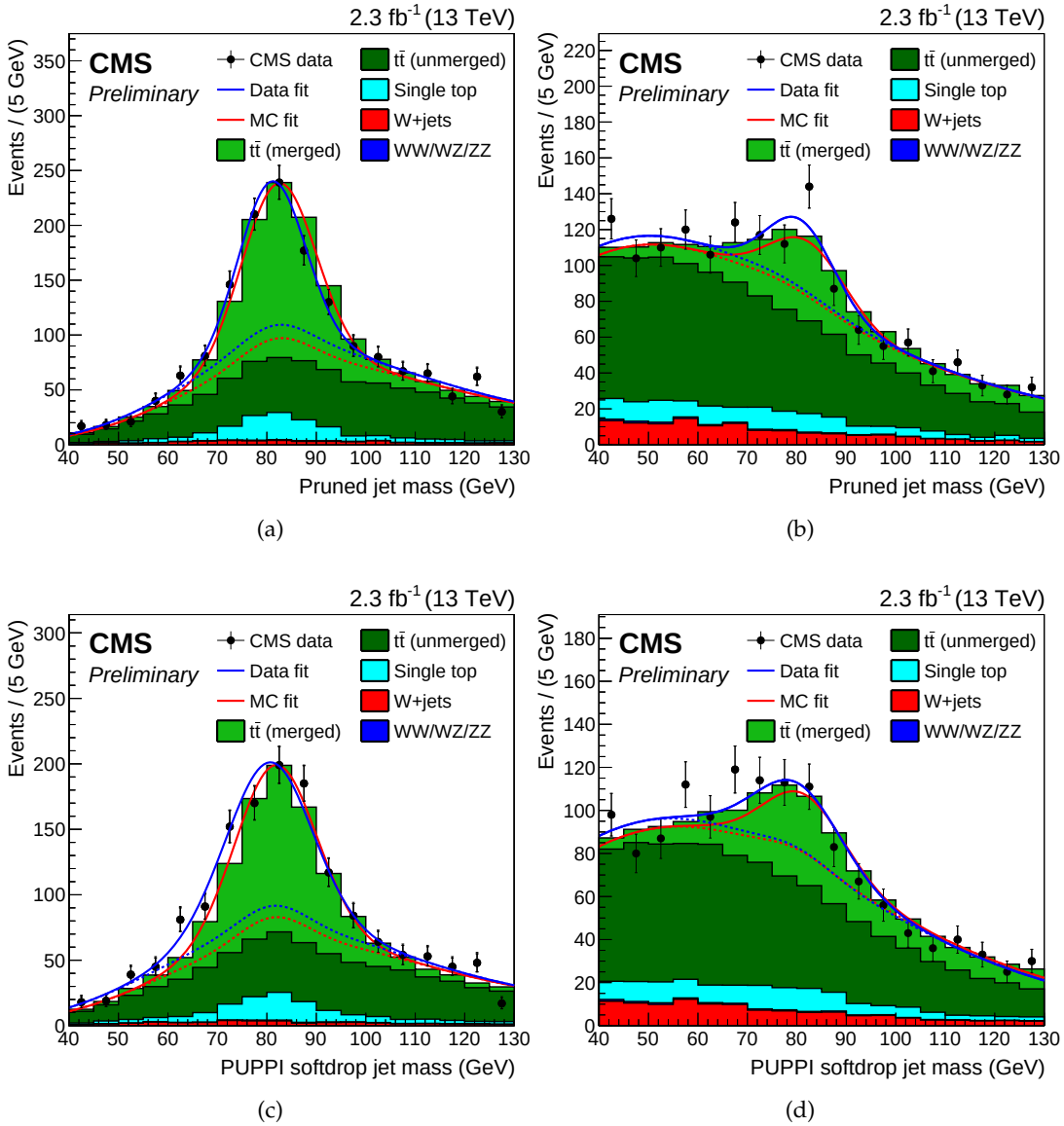


Figure 18: PF+CHS pruned jet mass distribution that (a) pass and (b) fail the PF+CHS $\tau_2/\tau_1 < 0.45$ selection in the $t\bar{t}$ control sample. PF+PUPPI softdrop jet mass distribution that (c) pass and (d) fail the PF+PUPPI $\tau_2/\tau_1 < 0.40$ selection. The result of the fit to data and simulation are shown, respectively, by the solid and long-dashed line and the background components of the fit are shown as dashed-dotted and short-dashed line.

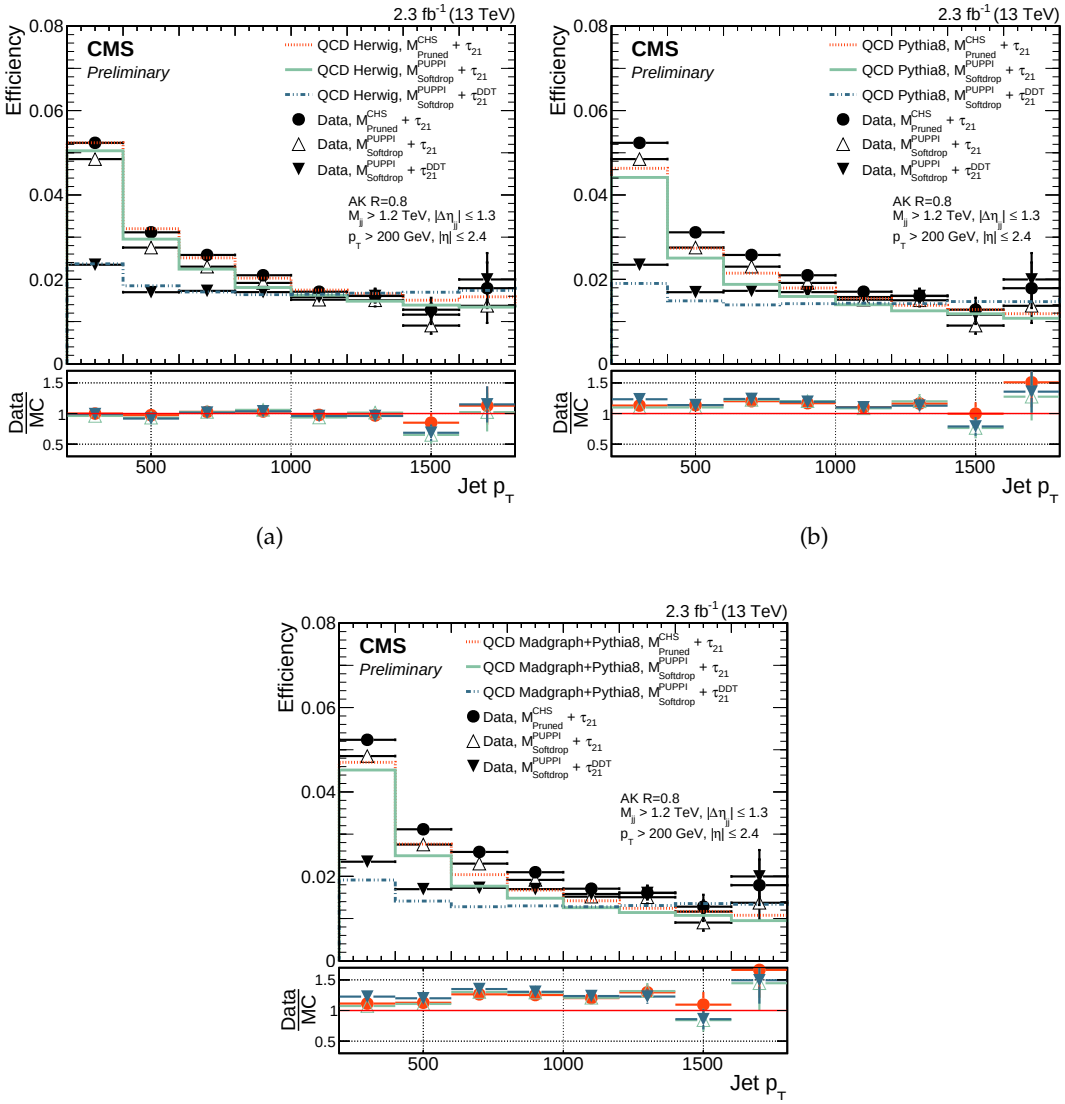
relative to the difference in the case of $p_T \sim 200$ GeV, is found to be $4.53\% \times \ln(p_T/200\,\text{GeV})$ ($5.90\% \times \ln(p_T/200\,\text{GeV})$) for $\tau_{21} < 0.6$ ($\tau_{21} < 0.45$), and is considered as the p_T extrapolation uncertainty of τ_2/τ_1 scale factor.

Potential systematic effects due to the choice of the signal and background fit model have been evaluated, by comparing the estimated efficiency on simulated $t\bar{t}$ samples with two different fit models. In the default model, the signal is purely modeled by a Gaussian peak, while the tails of the signal peak distribution are absorbed by the background fit model. In the alternative model, the signal is modeled by a Gaussian peak with tails including the non-peaking part of the W jets obtained from generator matched $t\bar{t}$ simulation. The estimated efficiencies obtained with those two methods, corrected for the fraction of W jets in the tails, agree within 0.3-12% and are listed as systematic uncertainty in Table 2.

Contributions from lepton identification, b tagging, E_T^{miss} scale and underlying event are at $<$ 0.5% level and negligible. Uncertainties due to the jet energy scale/resolution, pileup effects of order $1-2\%$ depend on the event topology and jet mass selection, and are thus evaluated in the context of individual analysis, thus not reported here. Uncertainties due to the W-polarization (2.0%) depend on the signal topology, thus applied only to cases where a significantly different polarization from the W bosons in the $t\bar{t}$ sample is expected.

8.4 Mistagging rate measurement

A dijet sample is used to measure the rate of mistagged W jets, or mistags. The mistagging rate is measured in data and compared to simulation. As discussed previously, the W tagger selection requires $65 < m_{jet} < 105$ GeV and $\tau_2/\tau_1 < 0.45$. Figure 19 shows the fraction of jets passing the m_{jet} and τ_2/τ_1 requirements, as a function of p_T and of the number of reconstructed vertices. The mistagging rate of the m_{jet} and τ_2/τ_1 requirements in data is reproduced better by HERWIG++ than by PYTHIA8. The p_T dependence in data is well reproduced by all generators.



(a) (b)

(c)

Figure 19: Fraction of jets passing the m_{jet} and τ_2/τ_1 selections in a dijet data sample and in simulation as a function of p_{T}, comparing (a) HERWIG++, (b) PYTHIA8 and (c) PYTHIA8 with MADGRAPH as matrix-element generator. The data over simulation ratio is shown for the combination of the m_{jet} and τ_2/τ_1 selections.

9 Top jet identification

9.1 Observables

A top jet is formed when the decay products of a highly Lorentz boosted hadronically decaying top quark merge into a single jet. Top jets can be identified with jet substructure observables based on the kinematics of the t→bW→ $\bar{q}q'b$ decay: the top jet mass is similar to that of the top quark mass, the internal structure of the jet contains three distinct subjets, and a pair of subjets has a combined mass approximately equal to the W boson mass.

We study two methods of top jet reconstruction, depending on the jet p_{T}. For less collimated low-p_{T} jets, up to $\approx$ 500 GeV, the initial reconstruction is done using the Cambridge-Aachen algorithm with R = 1.5 ("CA15") with PF+CHS jet reconstruction. With increasing p_{T}, the top quark's decay products become more collimated and jets can be reconstructed with the same AK8 algorithm as is used for W jets. We study AK8 jets with both PF+CHS and PF+PUPPI inputes. For the reconstruction of high p_{T} top candidates the softdrop ($\beta = 0, z = 0.1$) mass together with the "N-subjettiness" ratio τ_3/τ_2 (τ_{32}) calculated from all constituents of the AK8 jet is used.

Low p_{T} top candidates are further processed by the HEPTopTagger version 2 (HTT V2) algorithm [17, 65, 66]. This algorithm uses a combination of massdrop unclustering and filtering steps to identify three subjets corresponding to the three quarks from the top decay. The discriminating variables are the reconstructed top quark mass m_{123}, defined as the invariant mass calculated from the four-vector sum of the three subjets as well as the reconstructed W boson to top quark mass ratio

$$f_{Rec} = \min_{ij} \left| \frac{\frac{m_{ij}}{m_{123}}}{\frac{m_{\mathrm{W}}}{m_{\mathrm{t}}}} - 1 \right|. \tag{9}$$

Here the invariant mass of the pair of subjets i and j is denoted by m_{ij} while m_{W} and m_{t} are the mass of the W and top, respectively.

The procedure is repeated multiple times for successively smaller R parameters (in decrements of 0.1) until the "optimal" value of R is found. The optimal R value is defined as the smallest value of R so that $(m_{123}^{R=1.5} - m_{123}^{Opt.})/m_{123}^{R=1.5} < 0.2$. The mass and f_{Rec} at optimal R are then used to identify top quark candidates. Additionally the "N-subjettiness" ($\tau_3/\tau_2 = \tau_{32}$) calculated using the constituents of the CA15 jet after applying the softdrop algorithm ($\beta = 1, z = 0.2$) is used to further discriminate top quarks from background jets. These parameters differ from the softdrop settings used for the reconstruction of AK8 jets and were found to yield the best performance for large-R jets [20].

Because top quark decays almost always contain at least one b quark jet, the use of b jet tagging can further improve discrimination power. Both at low p_{T} and high p_{T} we use the multivariate Combined Secondary Vertex (CSV) [64] algorithm. This method combines information on the impact parameter significance of charged-particle tracks as well as the presence and properties of reconstructed decay vertices using an artificial neural network. For CA15 jets the discriminator is calculated for the three subjets produced by the HTT V2 algorithm, while for AK8 jets the subjets identified by the softdrop technique are used. In both cases the highest subjet CSV value is used for discrimination.

9.2 Top jet measurements with a semi-leptonic $\mathrm{t\bar{t}}$ selection

Semi-leptonic $\mathrm{t\bar{t}}$ events are selected and used to study a pure sample of boosted top jets in data. The event selection requires exactly one muon and at least one AK4 jet in the same hemisphere

of the event. These requirements account for the leptonic decay of the top. An AK8 or CA15 jet with high transverse momentum is required in the hemisphere opposite the muon. The detailed selection is described below.

Only events passing the unprescaled, single muon trigger are considered. The event must contain at least one good primary vertex. Reconstructed muons are required to have $p_T > 50$ GeV and $|\eta| < 2.1$ and to fulfill tight muon identification criteria. Additional muons in the event are vetoed.

The muon divides the event into two hemispheres:

- the hadronic hemishpere: $|\phi - \phi_\mu| > \frac{2}{3}\pi$;
- the leptonic hemisphere: $|\phi - \phi_\mu| < \frac{2}{3}\pi$.

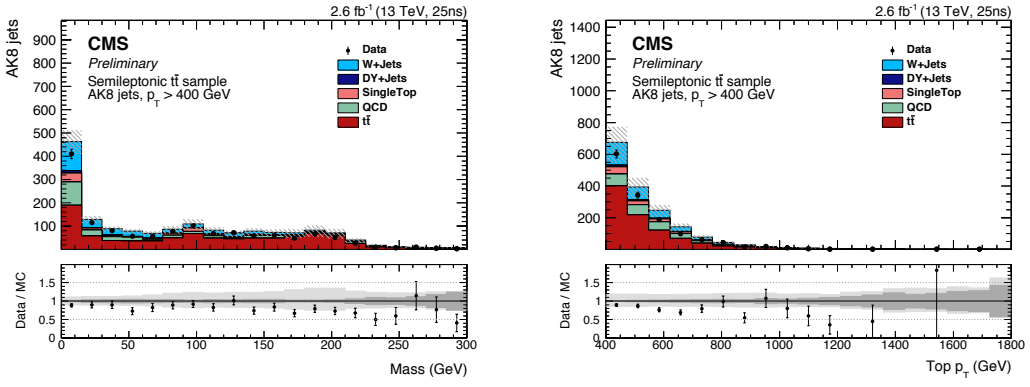
At least two AK4 jets with $p_T > 30$ GeV and $|\eta| < 2.4$ are required. The hadronic hemisphere must contain one AK8 jet with $p_T > 400$ GeV and $|\eta| < 2.4$. This jet represents the probe jet, used to evaluate the top tag scale factors. For low p_T studies, a CA15 jet with $p_T > 150$ GeV and $|\eta| < 2.4$ is required instead.

In order to improve the agreement between data and MC for the $t\bar{t}$ process a jet p_T reweighting is applied [67, 68]. The event weight is calculated for all events in the ttbar sample as given by

$$w = \sqrt{\exp(a + b \cdot p_{T,t}) \cdot \exp(a + b \cdot p_{T,\bar{t}})}\,,$$

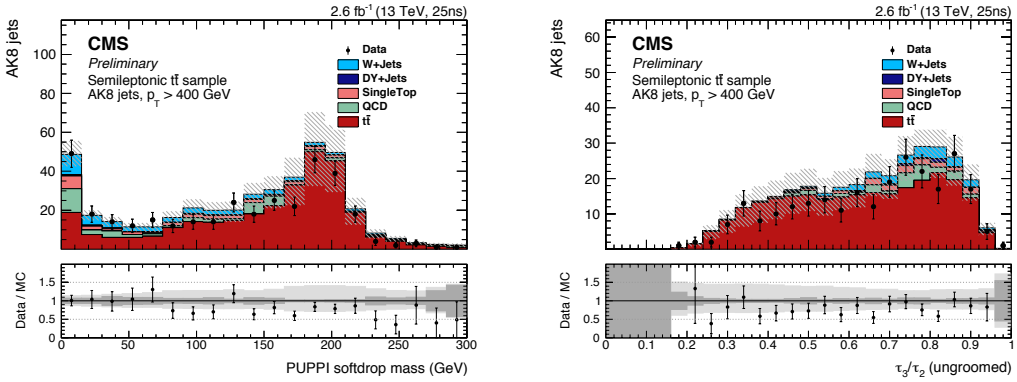
where the parameters $a = 0.156$ and $b = -0.00137$ are used. The impact of this technique on the measured scale factors is negligible, on the order of 1-2%.

The selected sample is largely enriched with $t\bar{t}$ events. Residual background contributions are given by W and Z boson production in association with jets and by single top quark production. All the simulated processes are normalized to their cross section. Control distributions for the selected data sample are shown in Fig. 20.



(a) (b)

Figure 20: Kinematic distributions for the AK8 jet associated to the boosted top hadronic decay in selected semi-leptonic $t\bar{t}$ events. The $t\bar{t}$ MC and the selected backgrounds are stacked. The distributions are: (a) Corrected softdrop mass, (b) Transverse momentum. A hashed band indicates the sum in quadrature of the statistical and systematic uncertainties of the simulation. The ratio of data to simulation is displayed below the distribution. A dark shaded and a light shaded band indicate the statistical uncertainty of the simulation and the systematic uncertainty of the modeling of top p_T spectrum, respectively.



(a) (b)

Figure 21: Top tagging variable distributions for the AK8 jet associated to the boosted top hadronic decay in selected semi-leptonic $t\bar{t}$ events for the loose (3.0% nominal background mistag rate) high p_T softdrop/PUPPI working point. The $t\bar{t}$ MC and the selected backgrounds are stacked. (a) Corrected softdrop mass (b) ungroomed N-subjettiness. A hashed band indicates the sum in quadrature of the statistical and systematic uncertainties of the simulation. The ratio of data to simulation is displayed below the distribution. A dark shaded and a light shaded band indicate the statistical uncertainty of the simulation and the systematic uncertainty of the modeling of top p_T spectrum, respectively.

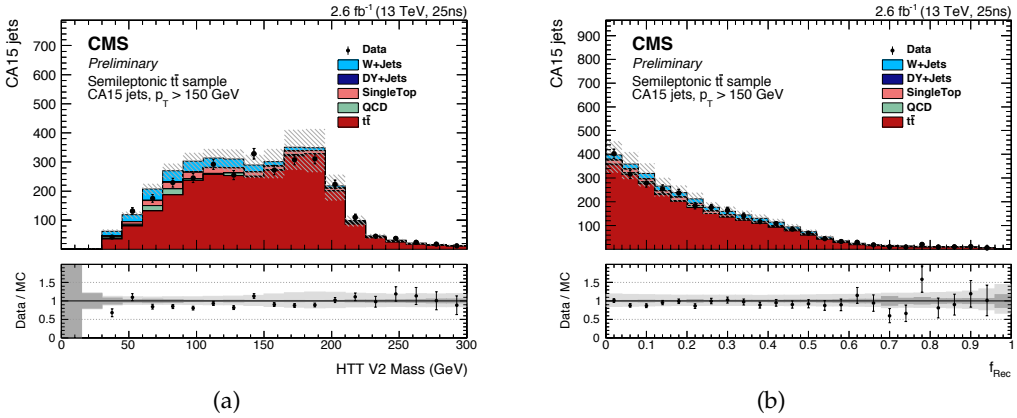
The softdrop mass and the N-subjettiness distribution are presented in Fig. 21. The distributions are made for AK8 jets after applying the PUPPI procedure. In addition to the event selection, jets are required to pass the selection criteria for the loose (3.0% nominal background mistag rate) working points as listed in Table. 4. For the mass (N-subjettiness) distribution the cut on the mass (N-subjettiness) is omitted.

The corresponding distributions for the HTT V2 algorithm are presented in Fig. 22. The reconstructed HTT V2 mass is shown in Fig. 22 (a), the mass-ratio variable f_{Rec} in Fig. 22 (b) and N-subjettiness in Fig. 22 (c).

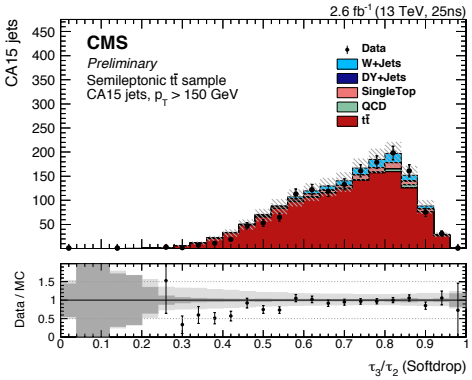
The top tagging efficiency for different selection criteria on substructure variables for AK8 jets is shown in Fig. 23. All selections, excluding selecting top candidates only according to the N-subjettiness, are stable within 5% points for a jet transverse momentum between 400 GeV and 1.6 TeV.

9.2.1 Scale factor measurement

The scale factors are evaluated comparing the top-tagging efficiencies in data and $t\bar{t}$ MC for each individual top-tagging working points. The top-tagging efficiency is measured using a semi-leptonic $t\bar{t}$ event selection. A small fraction of the jets in this event selection originate from non-$t\bar{t}$ sources. This non-$t\bar{t}$ contamination is subtracted based on simulated events. The efficiency is then defined as the number of jets passing the top tagging requirements divided by the total number of jets in the event selection. The scale factor is defined as the ratio of the efficiency in data and the efficiency in $t\bar{t}$ MC. The scale factors are evaluated both inclusively as well as for specific p_T ranges and are summarized in Table 4 and Table 5.



(a) (b)



(c)

Figure 22: Top tagging variable distributions for the CA15 jet associated to the boosted top hadronic decay in selected semi-leptonic $t\bar{t}$ events for the loose (1.0% nominal background mistag rate) low p_T HTT V2/CHS working point. The $t\bar{t}$ MC and the selected backgrounds are stacked. The distributions are: (a) HTT V2 mass. (b) f_{Rec}. (c) softdrop groomed N-subjettiness A hashed band indicates the sum in quadrature of the statistical and systematic uncertainties of the simulation. The ratio of data to simulation is displayed below the distribution. A dark shaded and a light shaded band indicate the statistical uncertainty of the simulation and the systematic uncertainty of the modeling of top p_T spectrum, respectively.

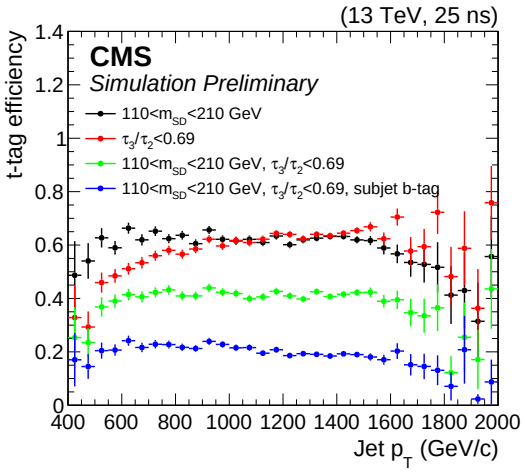


Figure 23: Top tagging efficiency as measured in $t\bar{t}$ simulation.

Table 4: Result scale factors for the for softdrop high p_{T} working points and different p_{T} ranges. The reported uncertainties are statistical only. The scale factors are labelled by expected background mistag rate $\epsilon(B)$ and expected signal efficiency $\epsilon(S)$ on MC simulated events for the selection including the b-tag requirement. For comparison the inclusive scale factor is also presented without this cut.

				subjet b-tag > 0.46			no subjet b-tag cut
Input	$\epsilon(B)$	$\epsilon(S)$	p_{T} bin [GeV]	400 - 550	550 - ∞	incl. (400 - ∞)	incl. (400 - ∞)
PUPPI	0.1 %	$\approx$ 15 %	$M_{SD}[105,210]$, $\tau_{32} < 0.46$	0.96 ± 0.18	1.32 ± 0.40	1.04 ± 0.16	0.98 ± 0.15
	0.3 %	$\approx$ 30 %	$M_{SD}[105,210]$, $\tau_{32} < 0.54$	1.02 ± 0.15	1.14 ± 0.32	1.05 ± 0.14	1.01 ± 0.13
	1.0 %	$\approx$ 45 %	$M_{SD}[105,210]$, $\tau_{32} < 0.65$	1.06 ± 0.13	1.07 ± 0.28	1.06 ± 0.12	1.06 ± 0.11
	3.0 %	$\approx$ 55 %	$M_{SD}[105,210]$, $\tau_{32} < 0.80$	1.04 ± 0.11	1.05 ± 0.26	1.05 ± 0.10	1.02 ± 0.09
CHS	0.1 %	$\approx$ 15 %	$M_{SD}[105,220]$, $\tau_{32} < 0.50$	0.76 ± 0.14	1.10 ± 0.30	0.85 ± 0.13	0.86 ± 0.13
	0.3 %	$\approx$ 25 %	$M_{SD}[105,220]$, $\tau_{32} < 0.57$	0.82 ± 0.13	1.00 ± 0.25	0.97 ± 0.11	0.88 ± 0.11
	1.0 %	$\approx$ 45 %	$M_{SD}[105,220]$, $\tau_{32} < 0.67$	0.90 ± 0.11	1.03 ± 0.21	0.94 ± 0.10	0.93 ± 0.09
	3.0 %	$\approx$ 60 %	$M_{SD}[105,220]$, $\tau_{32} < 0.81$	0.88 ± 0.09	1.09 ± 0.19	0.94 ± 0.08	0.96 ± 0.08

Table 5: Result scale factors for the for HEPTopTaggerV2/CHS low p_{T} working points and different p_{T} ranges. The reported uncertainties are statistical only. The scale factors are labelled by expected background mistag rate $\epsilon(B)$ and expected signal efficiency $\epsilon(S)$ on MC simulated events for the selection including the b-tag requirement. For comparison the inclusive scale factor is also presented without this cut.

			subjet b-tag > 0.46				no subjet b-tag cut
$\epsilon(B)$	$\epsilon(S)$	p_{T} bin [GeV]	150 - 400	400 - 550	550 - ∞	incl. (150 - ∞)	incl. (150 - ∞)
0.1 %	$\approx$ 15 %	M[130,185], $\tau_{32,SD}$ <0.55,f_{Rec} <0.17	0.62 ± 0.06	0.85 ± 0.20	1.47 ± 0.47	0.68 ± 0.06	0.73 ± 0.06
0.3 %	$\approx$ 25 %	M[115,180], $\tau_{32,SD}$ <0.62,f_{Rec} <0.27	0.87 ± 0.05	0.83 ± 0.17	1.26 ± 0.39	0.87 ± 0.05	0.91 ± 0.04
1.0 %	$\approx$ 35 %	M[110,185], $\tau_{32,SD}$ <0.93,f_{Rec} <0.20	0.87 ± 0.03	0.91 ± 0.11	1.50 ± 0.30	0.91 ± 0.03	0.95 ± 0.02
3.0 %	$\approx$ 45 %	M[85,280], $\tau_{32,SD}$ <0.97,f_{Rec} <0.47	0.90 ± 0.02	1.02 ± 0.06	1.20 ± 0.16	0.92 ± 0.01	0.98 ± 0.01

10 Summary

The performance of jet and jet substructure algorithms has been studied in data collected by the CMS experiment at the LHC with a center-of-mass energy of 13 TeV. The rejection rate of jet identification criteria against noise has been measured using a noise enriched minimum bias event selection, while the efficiency for identifying real physical jets has been meaured in data using a tag-and-probe procedure. The background rejection rejection rate has been measured to be greater than 99.999% in the barrel region and greater than 92% in the forward detector region. A multivariate BDT which uses vertex and jet shape information to discriminate pileup jets has been discussed, and its performance has been measured in data and in simulation. For central jets with $|\eta| < 2.5$ and $30 < p_{\mathrm{T}} < 50$ GeV, the pileup jet identification BDT rejects 89% of pileup jets while maintaining 96% of gluon jets. A likelihood-based tagger which relies on the internal structure of jets to discriminate jets initiated by light-quark partons from those initiated by gluons has been studied. A recipe to evaluate the systematic uncertainties associated to the use of the quark/gluon discriminator has been given, based on the observed data versus MC differences in the validation samples. The efficiency and mistag rate of W tagging and top tagging algorithms has been discussed, and scale factors have been measured. A new W tagger based on DDT corrected N-subjettiness has been studied and found to yield a mistag rate that is independent of p_{T}. W tagging and top tagging techniques relying on PUPPI pileup suppression have been validated in data for the first time and were found to maintain W and top tagging performance up to at least 40 simultaneous interactions.

References

[1] CMS Collaboration, "The CMS experiment at the CERN LHC", *JINST* **3** (2008) S08004, doi:10.1088/1748-0221/3/08/S08004.

[2] CMS Collaboration, "Performance of quark/gluon discrimination in 8 TeV pp data", CMS Physics Analysis Summary CMS-PAS-JME-13-002, 2013.

[3] ATLAS Collaboration, "Light-quark and gluon jet discrimination in pp collisions at $\sqrt{s} = 7$ TeV with the ATLAS detector", *Eur. Phys. J.* **C74** (2014), no. 8, 3023, doi:10.1140/epjc/s10052-014-3023-z, arXiv:1405.6583.

[4] CMS Collaboration, "Pileup Jet Identification", CMS Physics Analysis Summary CMS-PAS-JME-13-005, 2013.

[5] ATLAS Collaboration, "Tagging and suppression of pileup jets with the ATLAS detector", ATLAS Note ATLAS-CONF-2014-018, 2014.

[6] A. J. Larkoski, S. Marzani, G. Soyez, and J. Thaler, "Soft Drop", *JHEP* **1405** (2014) 146, doi:10.1007/JHEP05(2014)146, arXiv:1402.2657.

[7] J. Thaler and K. Van Tilburg, "Maximizing Boosted Top Identification by Minimizing N-subjettiness", *JHEP* **1202** (2012) 093, doi:10.1007/JHEP02(2012)093, arXiv:1108.2701.

[8] J. Dolen et al., "Thinking outside the ROCs: Designing Decorrelated Taggers (DDT) for jet substructure", arXiv:1603.00027.

[9] CMS Collaboration, "Identification techniques for highly boosted W bosons that decay into hadrons", *JHEP* **12** (2014) 017, doi:10.1007/JHEP12(2014)017, arXiv:1410.4227.

[10] CMS Collaboration, "Studies of jet mass in dijet and W/Z+jet events", *JHEP* **05** (2013) 090, doi:10.1007/JHEP05(2013)090, arXiv:1303.4811.

[11] CMS Collaboration, "V Tagging Observables and Correlations", CMS Physics Analysis Summary CMS-PAS-JME-14-002, 2014.

[12] CMS Collaboration, "Study of jet substructure in pp Collisions at 7 TeV in CMS", CMS Physics Analysis Summary CMS-PAS-JME-10-013, 2010.

[13] ATLAS Collaboration, "Identification of boosted, hadronically decaying W bosons and comparisons with ATLAS data taken at $\sqrt{s} = 8$ TeV", *Eur. Phys. J.* **C76** (2016), no. 3, 154, doi:10.1140/epjc/s10052-016-3978-z, arXiv:1510.05821.

[14] ATLAS Collaboration, "A new method to distinguish hadronically decaying boosted Z bosons from W bosons using the ATLAS detector", *Eur. Phys. J.* **C76** (2016), no. 5, 238, doi:10.1140/epjc/s10052-016-4065-1, arXiv:1509.04939.

[15] ATLAS Collaboration, "Jet mass and substructure of inclusive jets in sqrt(s) = 7 TeV pp collisions with the ATLAS experiment", *JHEP* **05** (2012) 128, doi:10.1007/JHEP05(2012)128, arXiv:1203.4606.

[16] ATLAS Collaboration, "Performance of jet substructure techniques for large-R jets in proton-proton collisions at sqrt(s) = 7 TeV using the ATLAS detector", *JHEP* **09** (2013) 076, doi:10.1007/JHEP09(2013)076, arXiv:1306.4945.

[17] G. Kasieczka et al., "Resonance Searches with an Updated Top Tagger", arXiv:1503.05921.

[18] CMS Collaboration, "Jet Substructure Algorithms", CMS Physics Analysis Summary CMS-PAS-JME-10-013, 2011.

[19] CMS Collaboration, "Boosted Top Jet Tagging at CMS", CMS Physics Analysis Summary CMS-PAS-JME-13-007, 2014.

[20] CMS Collaboration, "Top Tagging with New Approaches", CMS Physics Analysis Summary CMS-PAS-JME-15-002, 2016.

[21] ATLAS Collaboration, "Identification of high transverse momentum top quarks in pp collisions at $\sqrt{s}$ = 8 TeV with the ATLAS detector", arXiv:1603.03127.

[22] ATLAS Collaboration, "Boosted hadronic top identification at ATLAS for early 13 TeV data", ATLAS Note ATL-PHYS-PUB-2015-053, 2015.

[23] ATLAS Collaboration, "Performance of boosted top quark identification in 2012 ATLAS data", ATLAS Note ATLAS-CONF-2013-084, 2013.

[24] CMS Collaboration, "Jet performance in pp collisions at $\sqrt{s}$ = 7 TeV", CMS Physics Analysis Summary CMS-PAS-JME-10-003, 2010.

[25] CMS Collaboration, "Particle–flow event reconstruction in CMS and performance for jets, taus, and $E_{\mathrm{T}}^{\mathrm{miss}}$", CMS Physics Analysis Summary CMS-PAS-PFT-09-001, 2009.

[26] CMS Collaboration, "Commissioning of the particle-flow event reconstruction with the first LHC collisions recorded in the CMS detector", CMS Physics Analysis Summary CMS-PAS-PFT-10-001, 2010.

[27] CMS Collaboration, "Energy calibration and resolution of the CMS electromagnetic calorimeter in pp collisions at $\sqrt{s}$ = 7 TeV", *JINST* **8** (2013) P09009, doi:10.1088/1748-0221/8/09/P09009, arXiv:1306.2016.

[28] M. Cacciari, G. P. Salam, and G. Soyez, "The anti-k_T jet clustering algorithm", *JHEP* **04** (2008) 063, doi:10.1088/1126-6708/2008/04/063, arXiv:0802.1189.

[29] Y. L. Dokshitzer, G. D. Leder, S. Moretti, and B. R. Webber, "Better jet clustering algorithms", *JHEP* **08** (1997) 001, doi:10.1088/1126-6708/1997/08/001, arXiv:hep-ph/9707323.

[30] M. Wobisch and T. Wengler, "Hadronization corrections to jet cross sections in deep-inelastic scattering", (1998). arXiv:hep-ph/9907280.

[31] M. Cacciari, G. P. Salam, and G. Soyez, "FastJet user manual", *Eur. Phys. J. C* **72** (2012) 1896, doi:10.1140/epjc/s10052-012-1896-2, arXiv:1111.6097.

[32] CMS Collaboration, "Jet energy scale and resolution in the CMS experiment in pp collisions at 8 TeV", *JINST* **12** (2017), no. 02, P02014, doi:10.1088/1748-0221/12/02/P02014, arXiv:1607.03663.

[33] CMS Collaboration, "Pileup Removal Algorithms", CMS Physics Analysis Summary CMS-PAS-JME-14-001, 2014.

[34] M. Cacciari, G. P. Salam, and G. Soyez, "The catchment area of jets", *JHEP* **04** (2008) 005, doi:10.1088/1126-6708/2008/04/005, arXiv:0802.1188.

[35] M. Cacciari and G. P. Salam, "Pileup subtraction using jet areas", *Phys. Lett. B* **659** (2008) 119, doi:10.1016/j.physletb.2007.09.077, arXiv:0707.1378.

[36] D. Bertolini, P. Harris, M. Low, and N. Tran, "Pileup Per Particle Identification", *JHEP* **10** (2014) 059, doi:10.1007/JHEP10(2014)059, arXiv:1407.6013.

[37] CMS Collaboration, "Performance of CMS muon reconstruction in pp collision events at $\sqrt{s} = 7$ TeV", *JINST* **7** (2012) P10002, doi:10.1088/1748-0221/7/10/P10002, arXiv:1206.4071.

[38] CMS Collaboration, "Performance of Electron Reconstruction and Selection with the CMS Detector in Proton-Proton Collisions at s = 8 TeV", *JINST* **10** (2015), no. 06, P06005, doi:10.1088/1748-0221/10/06/P06005, arXiv:1502.02701.

[39] J. Alwall et al., "The automated computation of tree-level and next-to-leading order differential cross sections, and their matching to parton shower simulations", *JHEP* **07** (2014) 079, doi:10.1007/JHEP07(2014)079, arXiv:1405.0301.

[40] T. Sjostrand, S. Mrenna, and P. Z. Skands, "PYTHIA 6.4 Physics and Manual", *JHEP* **05** (2006) 026, doi:10.1088/1126-6708/2006/05/026, arXiv:hep-ph/0603175.

[41] T. Sjostrand, S. Mrenna, and P. Z. Skands, "A Brief Introduction to PYTHIA 8.1", *Comput. Phys. Commun.* **178** (2008) 852–867, doi:10.1016/j.cpc.2008.01.036, arXiv:0710.3820.

[42] B. Andersson, G. Gustafson, G. Ingelman, and T. Sjostrand, "Parton Fragmentation and String Dynamics", *Phys. Rept.* **97** (1983) 31–145, doi:10.1016/0370-1573(83)90080-7.

[43] T. Sjostrand, "The Merging of Jets", *Phys. Lett. B* **142** (1984) 420–424, doi:10.1016/0370-2693(84)91354-6.

[44] P. Skands, S. Carrazza, and J. Rojo, "Tuning PYTHIA 8.1: the Monash 2013 Tune", *Eur. Phys. J.* **C74** (2014), no. 8, 3024, doi:10.1140/epjc/s10052-014-3024-y, arXiv:1404.5630.

[45] CMS Collaboration, "Underlying Event Tunes and Double Parton Scattering", CMS Physics Analysis Summary CMS-PAS-GEN-14-001, 2014.

[46] P. Nason, "A New method for combining NLO QCD with shower Monte Carlo algorithms", *JHEP* **11** (2004) 040, doi:10.1088/1126-6708/2004/11/040, arXiv:hep-ph/0409146.

[47] S. Frixione, P. Nason, and C. Oleari, "Matching NLO QCD computations with Parton Shower simulations: the POWHEG method", *JHEP* **11** (2007) 070, doi:10.1088/1126-6708/2007/11/070, arXiv:0709.2092.

[48] S. Alioli, P. Nason, C. Oleari, and E. Re, "A general framework for implementing NLO calculations in shower Monte Carlo programs: the POWHEG BOX", *JHEP* **06** (2010) 043, doi:10.1007/JHEP06(2010)043, arXiv:1002.2581.

[49] NNPDF Collaboration, "Parton distributions for the LHC Run II", *JHEP* **04** (2015) 040, doi:10.1007/JHEP04(2015)040, arXiv:1410.8849.

[50] CMS Collaboration, "Search for a Higgs boson in the decay channel H to ZZ(*) to q qbar ℓ^- l+ in pp collisions at $\sqrt{s} = 7$ TeV", *JHEP* **04** (2012) 036, doi:10.1007/JHEP04(2012)036, arXiv:1202.1416.

[51] CMS Collaboration, "Search for the standard model Higgs boson produced through vector boson fusion and decaying to $b\bar{b}$", *Phys. Rev.* **D92** (2015), no. 3, 032008, doi:10.1103/PhysRevD.92.032008, arXiv:1506.01010.

[52] CMS Collaboration, "Measurement of the hadronic activity in events with a Z and two jets and extraction of the cross section for the electroweak production of a Z with two jets in pp collisions at $\sqrt{s}$ = 7 TeV", *JHEP* **10** (2013) 062, doi:10.1007/JHEP10(2013)062, arXiv:1305.7389.

[53] CMS Collaboration, "Measurement of electroweak production of two jets in association with a Z boson in proton-proton collisions at $\sqrt{s} = 8$ TeV", *Eur. Phys. J.* **C75** (2015), no. 2, 66, doi:10.1140/epjc/s10052-014-3232-5, arXiv:1410.3153.

[54] P. Richardson and A. Wilcock, "Monte Carlo Simulation of Hard Radiation in Decays in Beyond the Standard Model Physics in Herwig++", *Eur. Phys. J.* **C74** (2014) 2713, doi:10.1140/epjc/s10052-014-2713-x, arXiv:1303.4563.

[55] J. Bellm et al., "Herwig++ 2.7 Release Note", arXiv:1310.6877.

[56] J. Gallicchio and M. D. Schwartz, "Seeing in Color: Jet Superstructure", *Phys. Rev. Lett.* **105** (2010) 022001, doi:10.1103/PhysRevLett.105.022001, arXiv:1001.5027.

[57] J. Gallicchio and M. D. Schwartz, "Quark and Gluon Jet Substructure", *JHEP* **04** (2013) 090, doi:10.1007/JHEP04(2013)090, arXiv:1211.7038.

[58] S. D. Ellis, C. K. Vermilion, and J. R. Walsh, "Techniques for improved heavy particle searches with jet substructure", *Phys.Rev.* **D80** (2009) 051501, doi:10.1103/PhysRevD.80.051501, arXiv:0903.5081.

[59] S. D. Ellis, C. K. Vermilion, and J. R. Walsh, "Recombination Algorithms and Jet Substructure: Pruning as a Tool for Heavy Particle Searches", *Phys.Rev.* **D81** (2010) 094023, doi:10.1103/PhysRevD.81.094023, arXiv:0912.0033.

[60] M. Dasgupta, A. Fregoso, S. Marzani, and G. P. Salam, "Towards an understanding of jet substructure", *JHEP* **1309** (2013) 029, doi:10.1007/JHEP09(2013)029, arXiv:1307.0007.

[61] M. Dasgupta, A. Fregoso, S. Marzani, and A. Powling, "Jet substructure with analytical methods", *Eur.Phys.J.* **C73** (2013), no. 11, 2623, doi:10.1140/epjc/s10052-013-2623-3, arXiv:1307.0013.

[62] J. M. Butterworth, A. R. Davison, M. Rubin, and G. P. Salam, "Jet substructure as a new Higgs search channel at the LHC", *Phys. Rev. Lett.* **100** (2008) 242001, doi:10.1103/PhysRevLett.100.242001, arXiv:0802.2470.

[63] J. Thaler and K. Van Tilburg, "Maximizing Boosted Top Identification by Minimizing N-subjettiness", *JHEP* **02** (2012) 093, doi:10.1007/JHEP02(2012)093, arXiv:1108.2701.

[64] CMS Collaboration, "Identification of b quark jets at the CMS Experiment in the LHC Run 2", CMS Physics Analysis Summary CMS-PAS-BTV-15-001, 2016.

[65] T. Plehn, G. P. Salam, and M. Spannowsky, "Fat Jets for a Light Higgs", *Phys. Rev. Lett.* **104** (2010) 111801, doi:10.1103/PhysRevLett.104.111801, arXiv:0910.5472.

[66] T. Plehn, M. Spannowsky, M. Takeuchi, and D. Zerwas, "Stop Reconstruction with Tagged Tops", *JHEP* **1010** (2010) 078, doi:10.1007/JHEP10(2010)078, arXiv:1006.2833.

[67] CMS Collaboration, "Measurement of differential cross sections for top quark pair production using the lepton+jets final state in proton-proton collisions at 13 TeV", *Submitted to: Phys. Rev. D* (2016) arXiv:1610.04191.

[68] CMS Collaboration Collaboration, "Measurement of the differential cross section for $\mathrm{t\bar{t}}$ production in the dilepton final state at $\sqrt{s} = 13$ TeV", Technical Report CMS-PAS-TOP-16-011, CERN, Geneva, 2016.