Available on the CERN CDS information server **CMS PAS JME-14-002**

CMS Physics Analysis Summary

Contact: cms-pog-conveners-jetmet@cern.ch 2014/09/10

V Tagging Observables and Correlations

The CMS Collaboration

Abstract

We present studies of several techniques for improving identification of hadronically decaying W and Z bosons. The performance of various algorithms and the comparison between data and simulation are analyzed in events enriched in hadronically decaying W bosons. These studies use the full 2012 CMS dataset, corresponding to an integrated luminosity of $19.7 \pm 0.5\,\text{fb}^{-1}$ from proton-proton collisions produced by the LHC at $\sqrt{s} = 8\,\text{TeV}$. In anticipation of LHC Run II, performance studies of identification techniques in high pileup conditions and of substructure reconstruction for jets with high transverse momenta are performed using simulated events with highly boosted W bosons. We demonstrate that with just a few modifications to legacy techniques, performance for tagging weak vector bosons remain robust under extreme conditions that can be encountered in the upcoming run.

Contents

1 Introduction

Since the first observation of the W and Z bosons, electroweak physics has played an important role in testing and developing the Standard Model. Whether it be in the breakthrough discoveries of the top quark and the Higgs boson, or in the high precision measurements of the W boson mass and the Z line shape, weak vector boson (V) identification techniques have proved vital in furthering our understanding of particle physics at ever-higher energy scales. Traditionally, V tagging methods have depended largely on leptonic decay channels. Hadronic signatures are faced with the relatively poorer reconstruction of jets and large multijet backgrounds from QCD processes at hadron colliders.

Several recent developments have improved the tagging of hadronically decaying weak vector bosons. Many of these advances have resulted from the analysis of the internal components of a jet, i.e. its substructure. A more effective identification of hadronic V decays allows many analyses to profit from the substantially larger branching fraction of hadronic channels. This, in turn, may provide significant gains in searches for new physics.

The treatment of jets for V tagging differs depending on the boost of the boson. For W/Z bosons

with low transverse momentum (p_T), the daughter quarks produce two jets reconstructed in the detector. For this scenario, the properties of the individual jets as well as observables related to the dijet system are considered. At sufficiently high p_T, the quarks emerge very close to each other and only a single jet is reconstructed by typical jet clustering algorithms. In this case, jet substructure provides several observables that can be used to distinguish a merged V jet from single-parton backgrounds.

The unprecedented beam intensities and energies in Run II at the Large Hadron Collider (LHC) can generate conditions that adversely affect the performance of substructure quantities for V tagging. It will be common, for example, for 30 to 40 simultaneous proton-proton collisions (i.e. pileup interactions) to accompany a given hard-scatter interaction. Moreover, the increase in center of mass collision energy from 8 TeV to 13 TeV will raise the production rate and energies of highly boosted objects at the LHC. Such conditions challenge current jet reconstruction algorithms and can degrade the discriminating power of V tagging variables. It will be demonstrated, however, that simple adjustments are sufficient to achieve robust substructure reconstruction for jets above 1.5 TeV, and that many observables can be minimally modified to reduce the impact of pileup.

In this document, we build on previous studies [1] that investigate the performance of jet observables for the identification of vector bosons. We first consider the case of low-p_T W boson decays that result in resolved jets. We then proceed to the regime where quarks from W boson decays are unresolved. Comparisons between data and simulation at 8 TeV are made for several observables, and their combined performance and correlations are evaluated. Next, we turn to simulations of Run II conditions to study the performance of V tagging variables at high pileup and of substructure reconstruction for high-p_T jets. We conclude with an outlook for a next-generation V tagger that combines multiple jet property/substructure observables.

2 CMS detector

The CMS detector, described in detail in ref. [2], is a multi-purpose apparatus designed to study high-p_T physics processes in proton-proton and heavy-ion collisions. CMS uses a right-handed coordinate system, with the origin at the nominal interaction point, the x axis pointing to the center of the LHC, the y axis pointing upwards, perpendicular to the plane of the LHC ring, and the z axis along the counterclockwise beam direction. The polar angle θ is measured from the positive z axis and the azimuthal angle, ϕ, is measured in the $x - y$ plane. The pseudorapidity is defined as $\eta = -\ln[\tan(\theta/2)]$. A superconducting solenoid occupies its central region, providing a magnetic field of 3.8 T parallel to the beam direction. Charged-particle trajectories are measured by the silicon pixel and strip trackers, which cover a pseudorapidity region of $|\eta| < 2.5$. A lead tungstate ($PbWO_4$) crystal electromagnetic calorimeter (ECAL) and a brass/scintillator hadron calorimeter (HCAL) surround the tracking volume and cover $|\eta| < 3$. The steel/quartz-fiber Cherenkov hadron forward (HF) calorimeter extends the coverage to $|\eta| < 5$. The muon system consists of gas-ionization detectors embedded in the steel flux return yoke outside the solenoid, and covers $|\eta| < 2.4$. The first level of the CMS trigger system, composed of custom hardware processors, is designed to select the most interesting events in less than 4 μs, using information from the calorimeters and muon detectors. The high-level trigger processor farm then further reduces the event rate to a few hundred Hz.

3 Event samples and reconstruction

The studies to be presented are performed using samples of hadronically decaying W bosons for the signal and samples of quark- and gluon-initiated jets for the background. For studies at $\sqrt{s} = 8$ TeV, we use the full 2012 CMS dataset corresponding to an integrated luminosity of $19.7 \pm 0.5\,\text{fb}^{-1}$ [3]. Top-pair events with a single muon in the final state are selected to obtain a sample of W bosons, while a background-enriched sample is obtained by selecting $Z + \text{jets}$ events where the Z boson decays to dimuons. For physics simulation, $t\bar{t}$ and $Z + \text{jets}$ processes are generated with MADGRAPH 5 [4] and interfaced to PYTHIA 6 [5] for simulation of parton showering and hadronization. The generated events are passed through a GEANT4 [6] simulation of the CMS detector. In preparation for Run II conditions, simulated events are generated using PYTHIA 8 [7] at $\sqrt{s} = 13$ TeV with an average of 40 pileup interactions per event. Production of Randall-Sundrum (RS) gravitons with varying masses decaying to W boson pairs [8, 9] are considered for the signal and QCD multijet events are considered for the background. The masses are chosen for the p_T range being probed. We use a graviton mass of 1 TeV to study jets with p_T around 0.5 TeV, and graviton masses of 1-4 TeV to study jets above 1 TeV.

Observable particles in an event are reconstructed using the CMS particle-flow algorithm [10]. The algorithm links topologically connected clusters of energy deposited in the calorimeters, charged-particle tracks identified in the central tracking system, and information from muon detectors to reconstruct individual particles and to define quality criteria. Of the vertices reconstructed in the event, that with the largest $\sum p_T^2$, where the sum runs over all tracks associated with the vertex, is chosen as the primary vertex. The remainder of the section describes the reconstructed physics objects used for event selection.

3.1 Jets

Jets are reconstructed by clustering the set of objects (i.e. particle-flow candidates) reconstructed by the particle-flow algorithm [11]. Particle-flow candidates provide a more convenient and coherent representation of jet constituents than can be achieved with more basic objects such as calorimeter clusters and tracks. For accurate descriptions of jet properties, the reconstruction of charged hadron candidates, photon candidates and neutrals hadron candidates in the particle-flow algorithm is the most important. Typically a jet contains 60% of charged hadrons (mainly $\pi^{\pm}$), 20% of photons (mainly from π^0 decays), and 10% of neutral hadrons [12]. The anti-k_T algorithm [13] is employed and two different distance parameters are considered: R = 0.5 (AK5) and R = 0.8 (AK8).

Combinatorial background arises from low-p_T jets from pileup interactions that are inadvertently clustered into high-p_T jets. A multivariate (MVA) selection is applied to separate jets from the hard scatter from jets reconstructed from energy deposits associated with pileup interactions [14]. The discrimination stems from differences in the jet shapes, on the relative multiplicity of charged and neutral components, and on the different fraction of transverse momentum that is carried by the hardest components.

Within the tracker detector acceptance, tracks belonging to each jet are required to be compatible with originating from the primary vertex. To reduce the effects from multiple interactions occuring in the same bunch crossing, we remove charged PF candidates not associated to the primary vertex[1] before performing jet clustering and computating all jet observables (known as Charged Hadron Subtraction or CHS). The CHS procedure is applied for large cone jets (AK8). Jet energy corrections are applied as a function of jet p_T and η [15]. Jet energy scale and resolution gives rise to an uncertainty in the yields of 2% (5%) for the low (high) jet multiplicity events.

To effectively explore the substructure of a jet, it is necessary to identify subjets corresponding to the quarks from W boson decay. Subjet finding is performed by reprocessing the constituents of the jet with either the Cambridge-Aachen (CA) [16] or k_T clustering algorithm. The choice of clustering algorithm depends on the algorithm to compute a particular substructure observable.

3.1.1 b jets

Jets that originate from the hadronization of b quarks are referred to as "b jets". The CSV b-tagging algorithm [17] is used to identify such jets. The algorithm combines track impact parameters and information from secondary vertices within jets into a likelihood discriminant that can separate b jets from jets that originate from light quarks, gluons, and charm quarks. The output of this CSV discriminant has values between zero and one; a jet with a CSV value above a particular threshold (here, the CMS "medium" operating point) is referred to as being "b-tagged".

3.2 Missing transverse energy

The missing transverse energy vector ($\vec{E}_T^{miss}$) is defined as the negative vector sum of the transverse momenta of all reconstructed particles (charged or neutral) in the event, with $E_T^{miss} = |\vec{E}_T^{miss}|$. The magnitude of the vector, E_T^{miss}, can be used to discriminate the signal from backgrounds.

3.3 Muons

Muon candidates are identified from charged-particle tracks in the muon system that match to tracks reconstructed in the central tracking system [18]. Sufficient precision for momentum measurement from the curvature of the track in the magnetic field is ensured by imposing a minimum requirement on the number of sensor measurements and on the quality of the full track fit. Muons are required to be isolated to distinguish between prompt leptons from W/Z-boson decays and those from semileptonic hadron decays or from hadrons in jets that are misidentified as leptons. The variable $\Delta R = \sqrt{(\Delta\eta)^2 + (\Delta\phi)^2}$ is used to measure the separation between reconstructed objects in the detector. Isolation criteria are set based on the distribution of low-momentum particles in the (η, ϕ) region around the leptons. To remove the contribution from overlapping pileup interactions in the isolation region, charged particles included in the computation of the isolation variable are required to originate from the primary vertex. A correction is applied to the neutral component in the isolation ΔR cone that is based on the average energy density deposited by the neutral particles from additional interactions [19].

4 Resolved jets

Resolved jets from hadronic decays of W and Z bosons occur when the boson p_T is sufficiently low so that each quark is reconstructed separately as a jet. It should be noted that two jets can still be reconstructed if the jets from the boson decay only partially overlap. For distance parameter R = 0.5, the effect of overlapping jets becomes relevant for W bosons with $p_T \approx$ 160 GeV. As will be discussed, this has implications on the behavior of some jet observables. In this section, we present studies of several variables that show discriminating power for W boson decays to dijets over the combinatorial background.

4.1 Event selection

A sample of dijets from W boson decays in $t\bar{t}$ production is obtained from events triggered by a single muon of $p_T > 24$ GeV and $|\eta| < 2.1$. The offline requirements are that the events have exactly one isolated muon and four or more AK5 jets, at least two of which are b-tagged. The event is vetoed if isolated electrons or additional muons are present. The offline muon is required to have $p_T > 30$ GeV and $|\eta| < 2.1$, and must be matched to the muon object from the online trigger reconstruction. Jets are required to have $p_T > 30$ GeV and $|\eta| < 4.7$. Pairs of non-b-tagged jets are considered to be W candidates if the dijet mass falls between 40 and 130 GeV.

4.2 Resolved jet observables

The observables studied are outlined below, with explicit definitions and discussion of results provided in subsequent sections.

- **Quark/Gluon Likelihood**, (QGL), first introduced by the CMS collaboration in [20], is a likelihood-based discriminator designed to distinguish between jets originating from quarks and gluons.
- **Jet Charge**, (q), is defined using the prescription in [1]. The quantity of interest is the sum of jet charges of the dijet pair.
- **Jet Pull Angle**, (θ_P), as defined in [21], characterizes the color flow between two jets. Weak vector bosons are color singlets, and thus their daughter quarks are color-connected. This contrasts with the combinatorial background, which in most cases is not color-connected.

4.2.1 Quark/Gluon likelihood

The QGL can be used to reduce combinatorial background involving jets originating from gluons. Figure 1 shows the QGL distributions for jets from W candidates; both jets contribute to the histogram. A correction for the simulation has been applied to account for the mismodeling of gluons in PYTHIA 6. This correction is derived from a QCD dijet control sample [20], and the residual discrepancy in the plot is attributed to the difference in quark/gluon composition between dijet events and $t\bar{t}$ events. The plots show the high purity of $t\bar{t}$ events in the selected sample, and that a significant contribution to the combinatorial background involves gluon jets. Figure 1 also provides a performance curve derived from a scan of QGL cut thresholds applied to both jets in simulated $t\bar{t}$ events. The signal is dijet combinations that are matched to both quarks from W decay at generator-level, while background is dijet combinations that are not matched.

4.2.2 Jet charge

The jet charge is defined as,

$$q = \sum_i q_i \cdot \left(\frac{(p_T)_i}{(p_T)_{\text{jet}}} \right)^{\kappa}, \tag{1}$$

where the sum is taken over the jet constituents and κ is a parameter that is freely chosen. Figure 2 shows distributions for the sum of jet charges of the dijet W candidates. Of the values explored for the κ parameter in the jet charge definition, the best separation ($\sim 1.8\sigma$ in 10-event hypothesis tests) between W^+ and W^- candidates in simulation is found with $\kappa = 1$. The charge sum distributions of data and simulation agree, and the separation power is found to have a mild dependence on W p_T. Figure 2 also includes a performance curve of the charge

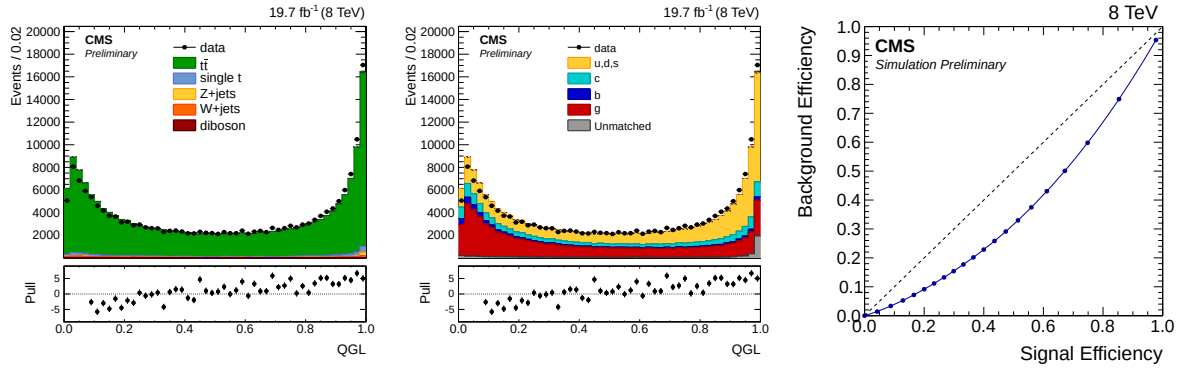

Figure 1: The QGL discriminant applied to resolved jets in data and simulation samples enriched in $t\bar{t}$. Left: the simulation is separated into contributions from different physics processes. Center: the simulation is separated into contributions from the underlying parton of the jets. Right: the performance in simulation of applying a cut on the QGL on both jets in the pair.

sum discriminant used for tagging W^+ and W^- bosons in simulated $t\bar{t}$ events. The curve is computed from scanning cut thresholds on the sum charge, accounting for the expected charge of the W candidate based on the charge of the identified muon. The signal is dijets matched to quarks from W decays at generator level and the background are unmatched dijets. Note that the discriminant performs better for W^- because there is more positively charged background in a proton-proton environment.

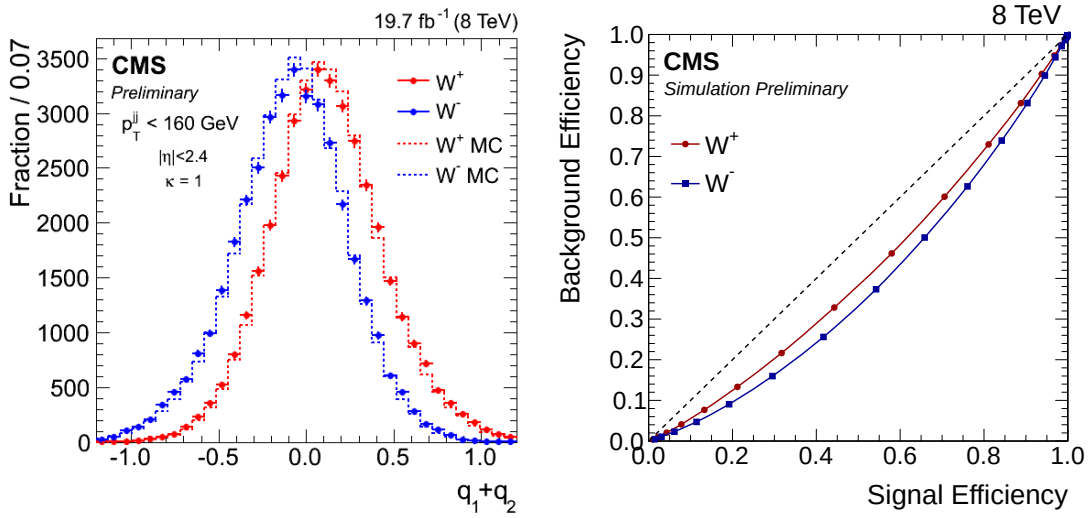

Figure 2: Left: Dijet charge distributions for W-enriched data and simulation. Right: The performance of applying a cut on the dijet charge sum in simulation.

4.2.3 Jet pull angle

Before defining the jet pull angle, it is necessary to define the jet pull vector [21], which is the weighted sum of jet constituent displacements from the jet axis:

$$\vec{t} = \sum_i \frac{(p_T)_i \, |r_i|}{(p_T)_{\text{jet}}} \vec{r}_i, \tag{2}$$

The sum is taken over the jet constituents and $\vec{r}_i$ is the displacement of a constituent from the jet axis in y, ϕ-space, with y denoting rapidity. The pull vectors of jets from color connected partons should tend to point towards each other. One way to quantify this behavior is to consider the angle between a jet's pull vector and the displacement vector between the two jet axes. This is the jet pull angle, θ_P. The pull angle distribution of color connected jet pairs peaks at zero, while for unconnected pairs the θ_P distribution is flat, as there is no correlation with the displacement vector. Figure 3 shows distributions of θ_P calculated from generator information in simulated $t\bar{t}$ events. Here the pull vector ($\vec{t}_1$) is calculated from the collection of final-state daughter particles in the lineage of the higher p_T quark from hadronic W decay. $\vec{J}_2$ and $\vec{B}$ correspond to the directions of the partner quark and the b quark from the same top decay, respectively. These distributions are consistent with expectation: pull angle is nearly uniformly distributed for non-color-connected particles, while particles that are color-connected exhibit a smaller pull angle on average. The plots also indicate that the θ_P distribution for quark pairs from W decays peaks more strongly at low angle with increasing p_T^W.

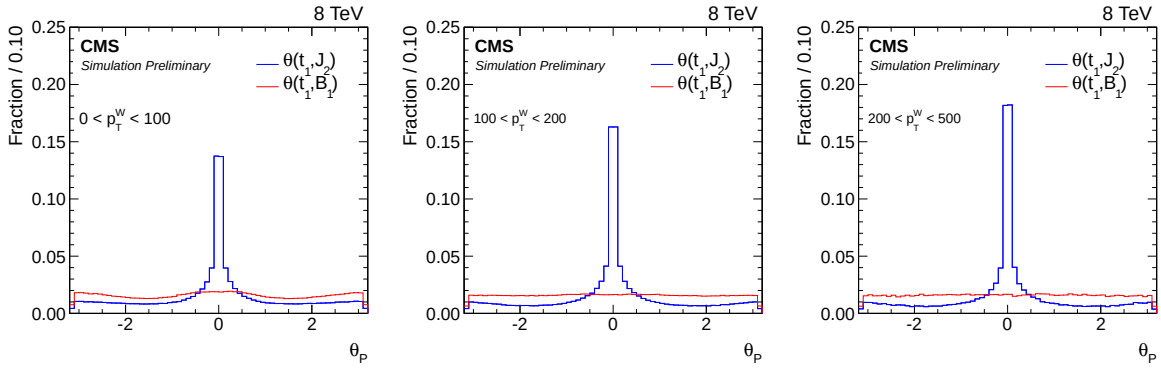

Figure 3: Pull angle (θ_P) calculated with generator-level particles in $t\bar{t}$ simulation. The pull vector is computed with respect to the leading quark from W decay. In each plot, two distributions are shown: the pull angle of the two quarks from W decay, and the pull angle of the leading quark from the W paired with the b quark from the same top decay.

Figure 4 shows distributions of θ_P from reconstructed W candidates in $t\bar{t}$ events in data and simulation. Also shown in Figure 4 is the pull angle between a jet from the W candidate with the leading b-tagged jet, which indicates the shape of a color unconnected background. The θ_P distribution of the pull vector from the leading reconstructed jet is markedly different than that from the subleading jet at high dijet p_T. In this regime the jets partially overlap, and the clustering algorithm can incorrectly assign particles to the jets. On average, the leading p_T jet will tend to have absorbed more constituents from the subleading jet, which enhances the small-angle peak for the former. This consequently removes constituents from the subleading jet, thus its pull vector will tend to point away from the leading jet.

The performance of $|\theta_P|$ as a discriminant is shown in Figure 5 for several ranges of dijet p_T. The pull angle here is computed using the pull vector from the leading jet. The signal is dijets matched to quarks from W decays at generator level and the background are unmatched dijets.

4.3 Summary

A set of jet observables have been studied in the context of identifying W boson decays into dijets. The three observables, QGL, dijet charge sum, and jet pull angle, do not show strong separation power individually; to combine the information in an optimal way, we introduce a W tagger using the aforementioned observables as input to a multivariate discriminator based

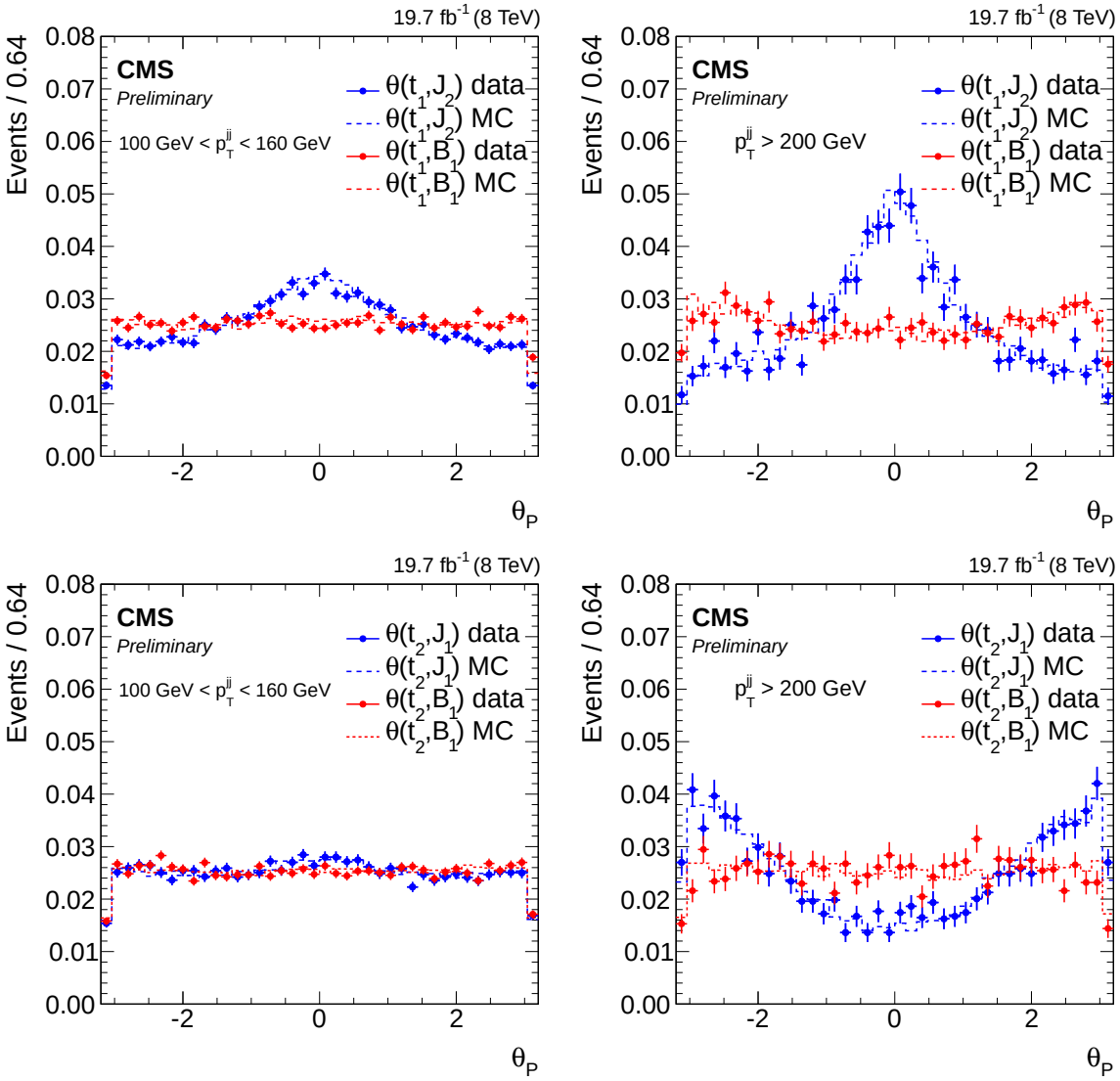

Figure 4: Top row: Distribution of pull angle computed from the leading jet of the W candidate. The slight rise in the background shape at $\pm\pi$ for high p_{T} corresponds to the cases where the b-tagged jet has absorbed some of the constituents that should have been assigned to the non-b-tagged jet. Bottom row: Distribution of pull angle computed from the subleading jet in the W candidate. The structure at high p_{T} can again be explained by overlapping jets with the leading jet absorbing extra constituents from the subleading jet.

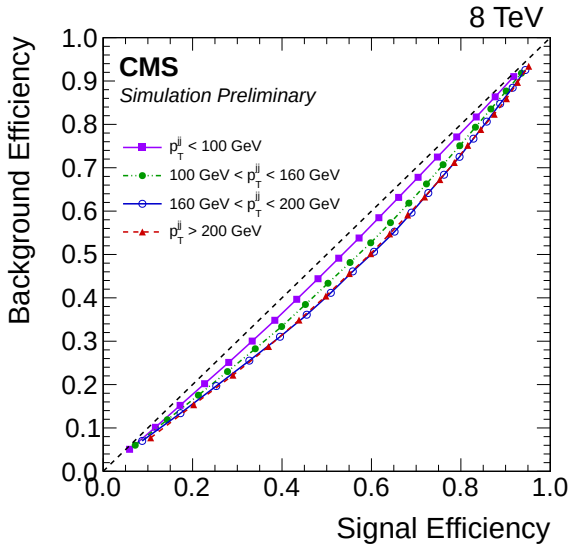

Figure 5: Performance curve in simulation for applying a cut on $|\theta_P|$ of the dijet system. The pull vector is computed from the leading jet.

on Boosted Decision Trees (BDT) using the TMVA framework [22]. We restrict to jets with $|\eta| < 2.4$ so that jet charge is relevant, and construct separate discriminators for dijet transverse momenta below and above 160 GeV to distinguish between where reconstruction effects are prominent to jet pull angle. These variables are found to be almost uncorrelated with each other, therefore each brings an independent handle to identifying W boson decays to dijets. The performance curve, shown in Figure 6, indicates roughly 25% background efficiency for a signal efficiency of 50%.

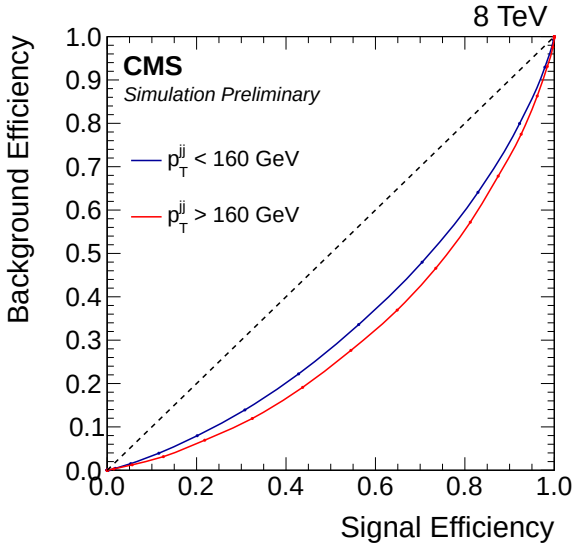

Figure 6: Performance curve for the multivariate W tagger using QGL, dijet charge sum, and jet pull angle as input.

5 Unresolved jets

For highly boosted weak vector bosons, the hadronic decay products can be merged into a single jet. For distance parameter R = 0.8, this occurs for boson p_T above 200 GeV. The radiation

profile of the individual, hard partons within a merged jet must be explicitly resolved for an accurate calculation of the boson mass. This contrasts with the resolved scenario, for which the boson mass can be determined simply from the properties of the individually reconstructed jets. A new class of observables has been developed for disentangling the radiation profiles of proximate partons. We explore these observables in this section.

5.1 Algorithms and observables

In this section the algorithms and related observables used to perform V tagging are discussed. These extend and complement previous studies [1]. The observables under study can be divided into two classes: the jet mass associated with a particular jet grooming algorithm, and variables generally related to the jet shower substructure.

Jet mass is the most natural discriminator between jets originating from V decays and those originating from single partons. Jet grooming techniques improve mass resolution by reducing the effects from pileup and underlying event. The following grooming algorithms are studied:

- **Filtering**, defined in [23], where we consider the three hardest CA subjets with R = 0.2.
- **Trimming**, defined in [24]. We consider subjets obtained with k_T clustering, R = 0.1, with a p_T fraction of the mother jet p_T exceeding 3%.
- **Pruning**, defined in [25, 26], with the the CA algorithm chosen as a measure for pruning. The minimal momentum fraction and maximal distance cuts are set to 0.1 and 0.5, respectively.
- **Soft-Drop**, defined in [27], where the soft threshold is fixed to 0.1 and the following values of angular exponent, β, are considered: $\{-1, 0, 1, 2\}$.

In addition to the various groomed mass algorithms, these other variables are studied:

- **Quark/Gluon Likelihood**, (QGL), is the same likelihood discriminator used for resolved jets in Section 4, and is applied on the pruned jet.
- **Subjet Quark/Gluon Likelihood** (Subjet QGL), is the QGL discriminant applied on the two leading pruned subjets. We also consider a combined discriminator (QGL Combo), defined as a linear combination of the leading subjet QGL with twice the second leading subjet QGL.
- **Energy Correlation Functions** (ECF), defined and studied in [28], can be combined into dimensionless double ratios (C_2^β) that can discriminate two-pronged jets against single-pronged jets. These C_2^β depend on the choice of the jet constituents angular weight β; for this study the considered values for β are $\{0, 0.2, 0.5, 1, 2\}$.
- **N-subjettiness** (τ_N), defined in [29, 30], is shown to be a powerful discriminator for hadronic V tagging when combined into ratios τ_N/τ_{N-1}. For the two-pronged topology the interesting ratio is τ_2/τ_1. We consider k_T sub-clustering with axis defined by the "winner-take-all" scheme [31].
- **Qjet volatility** ($\Gamma_{Q_{\text{jet}}}$), defined and studied in [32]. In this study, we adopt the same configuration used in [1], with the exception of the pre-clustering cut, which is removed. In order to maintain high processing speed, a new implementation of the algorithm is used following the Q_{jet} authors prescriptions. In particular, the truncation factor is set to 0.01.
- **Jet Pull Angle**, as defined for resolved jets in Section 4, is applied on the two leading pruned subjets.

- **Jet Pull Magnitude**, is the magnitude of the jet pull vector for pruned subjets.

5.2 Event selection

As done for the study of resolved jets, a $t\bar{t}$ selection is employed. In addition, a selection for $Z+$jets events is performed to obtain a background-enriched sample. In selecting both samples, an AK8 jet is required to have p_T larger than 250 GeV and $|\eta| < 2.5$, where the jet and the selected lepton(s) must be separated by $\Delta R > 0.3$.

5.2.1 Selection of $t\bar{t}$ events

As in the study of resolved jets, a single muon requirement is applied in the selection of $t\bar{t}$ events to obtain a sample of boosted W bosons. Two b-tagged AK5 jets with $p_T > 30$ GeV must be present in the event, and both jets must not overlap with the boosted AK8 jet. To reduce combinatorial background, the system constructed from the boosted jet and the b-tagged jet furthest from the lepton must have a mass below 300 GeV. The boosted jet is taken as the W candidate.

5.2.2 Selection of Z+jets events

This sample is selected using a set of high-p_T muon triggers and by applying the offline muon identification criteria described in Section 3. Two oppositely-charged muons with $p_T > 30$ GeV and $p_T > 10$ GeV for leading and trailing leptons, respectively, are required and the dimuon mass must lie within 15 GeV of the nominal Z boson mass [33]. In addition, the dilepton p_T is required to be larger than 100 GeV. With these requirements the background level is below 3%. This sample is relatively pure in quark jets: 70% of leading jets are expected to originate from light-quark hadronization.

5.2.3 Data vs simulation comparisons

The substructure observables are shown for the background dominated $Z+$jets sample in Figure 7. The simulation is broken down according to whether the jet is matched to a generator-level quark or gluon. The matching is performed by considering the closest generator-level parton in a cone of radius R=0.7 around the jet axis. The dashed band on top of the simulated spectra indicates the statistical uncertainty of the simulation. The agreement between data and simulation is good for most of the analyzed distributions. It is noticeable that, with respect to other observables, there is a worse modeling of the C_2^β functions.

Using the $t\bar{t}$ selection, signal jets in simulation are identified by matching to the quarks from W decays at generator level. Figure 8 shows the jet mass for different parameters of the soft drop grooming algorithms in this region. As expected, signal jets masses are concentrated around M_W, while background jets are characterized by a soft spectra. Due to limitations in the matching algorithm, a small number of W-jets are mis-identified as background and produce a structure around M_W. In general, good agreement is found for all the mass spectra obtained with different grooming techniques. The distributions are also proving that the grooming techniques perform well under the Run I pileup conditions. Note that in the figure with the $t\bar{t}$ selection, for the soft-drop mass ($\beta = -1$), the range is restricted to values greater than 4 GeV to enhance the comparison in the signal dominated region. For comparison, the soft-drop mass distribution with $\beta = -1$ is also shown for the background dominated $Z+$jets region.

The remaining comparisons for the $t\bar{t}$ region are presented in Figure 9; all distributions show good agreement between data and simulation apart from the ECF double ratios C_2^β. The distribution of the QGL for signal jets are particularly interesting: as the shower from W hadronic

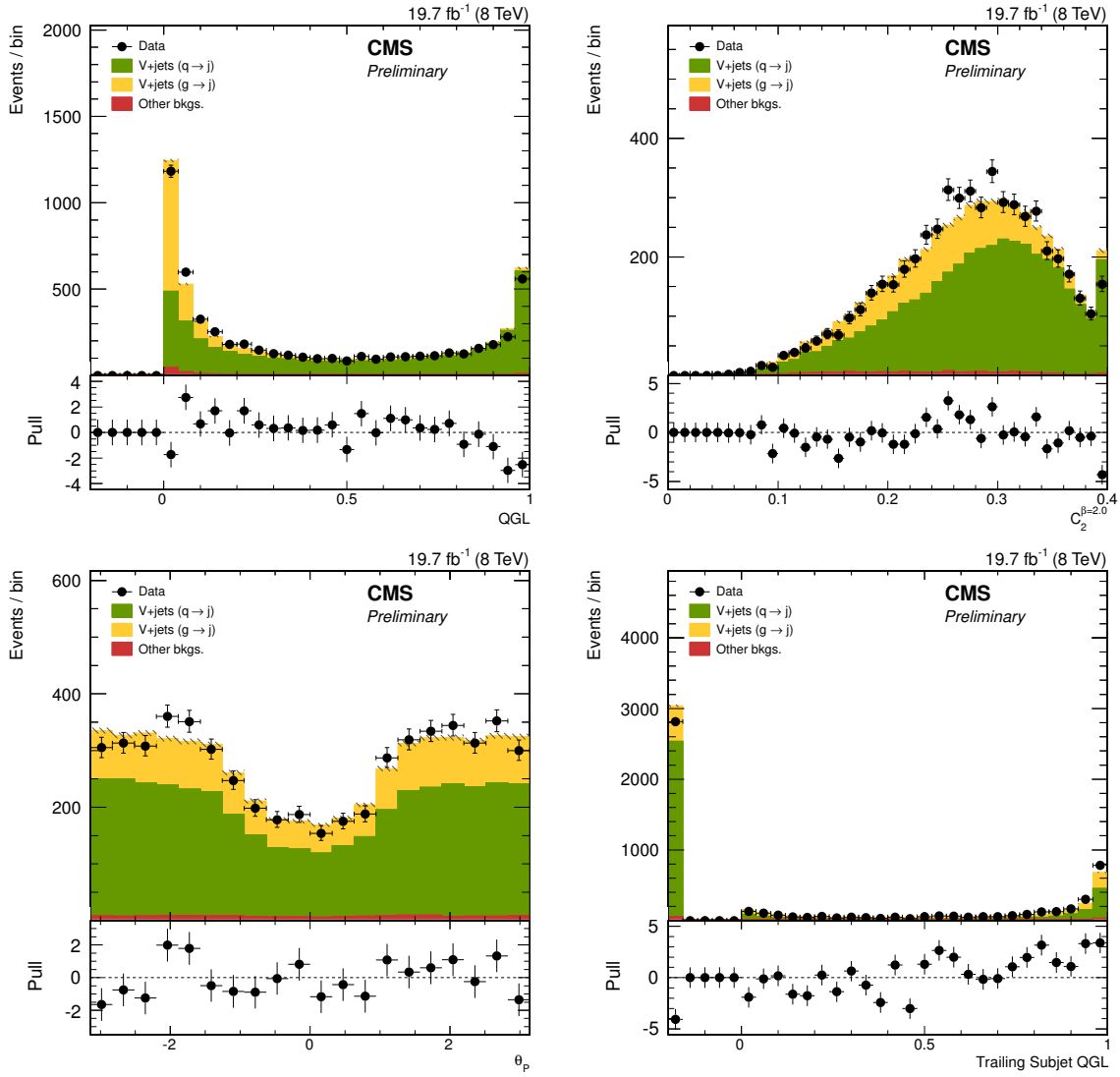

Figure 7: Jet substructure observables for a large cone jet in a region dominated by $Z+$jets events. From left to right, top row to bottom row, the plots show the CMS QGL, the ECF double ratio C_2 ($\beta = 2$), the jet pull angle (θ_P) and the Subjet QGL for the second highest p_T subjet. The simulation is normalized to the data. The contribution from $t\bar{t}$ and di-boson events has been considered in the simulation and is reported as "Other bkgs."

decays is fragmented and has high off-axis energy dispersion due to the presence of two hard legs, the QGL tags signal jets as more gluon-like than quark-like. This is expected as typical gluon showers are wider and more fragmented than quark showers.

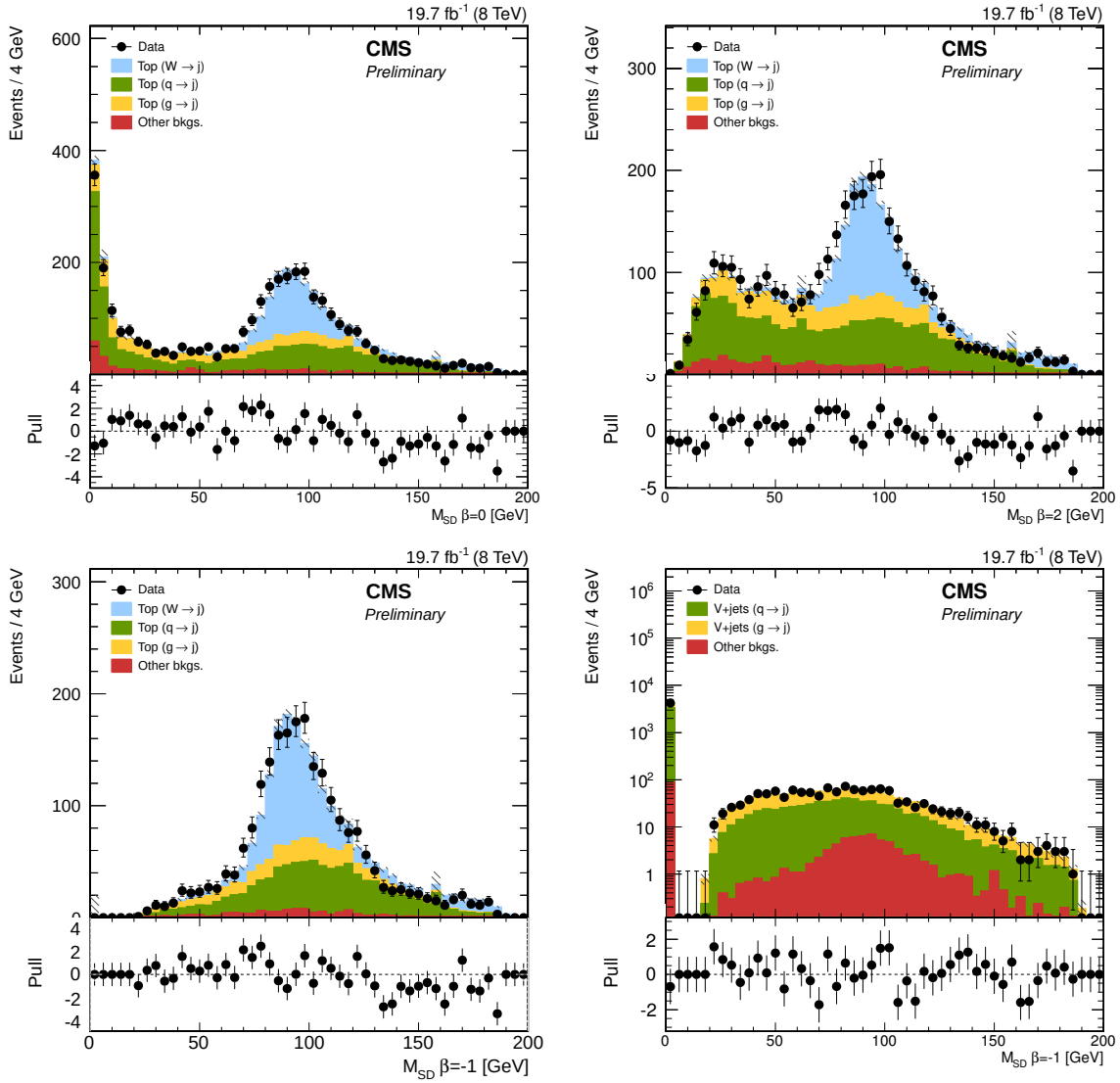

Figure 8: Jet mass for a large cone jet groomed with soft-drop method in a region dominated by $t\bar{t}$ events. The grooming algorithms used have a value of $\beta = 0$ (top-left), $\beta = 2$ (top-right), $\beta = -1$ (bottom-left). For the signal region where $\beta = -1$, the first mass bin is removed. For comparison, the Z + jets background region is shown for $\beta = -1$ (bottom-right). The contribution from single top, W/Z + jets and di-boson events has been considered in the simulation and is reported as "Other bkgs."

5.3 Performance studies

First, to understand which variables are the most interesting for the construction of a V tagger, the performance curve for each variable are shown in Figure 10. To compute the curves, signal jets are taken from simulation and background jets from the Z + jets sample in data. From the comparison of the mass obtained with different grooming algorithms, it appears that the most discriminating variables of this class are m_{Trim} and m_{Filt}. For what concerns jet substructure variables, N-subjettiness and Qjet volatility appear to perform best.

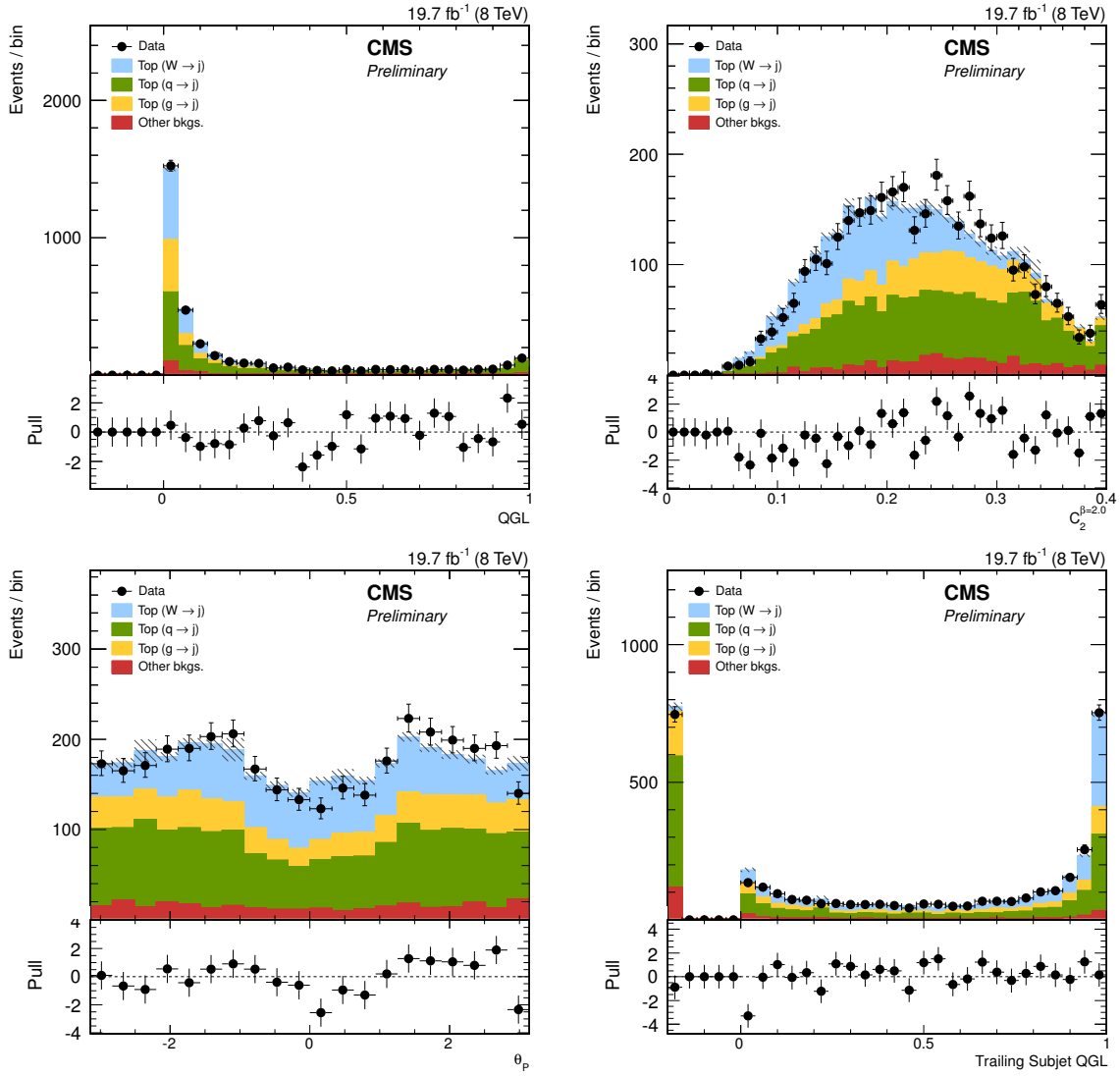

Figure 9: Jet substructure observables for a large cone jet in a region dominated by $t\bar{t}$ events. From left to right, top row to bottom row, the plots show the CMS QGL, the ECF double ratio C_2 ($\beta = 2$), the jet pull angle (θ_P) and the Subjet QGL for the second highest p_T subjet. The contribution from single top, $W/Z +$ jets and di-boson events has been considered in the simulation and is reported as "Other bkgs."

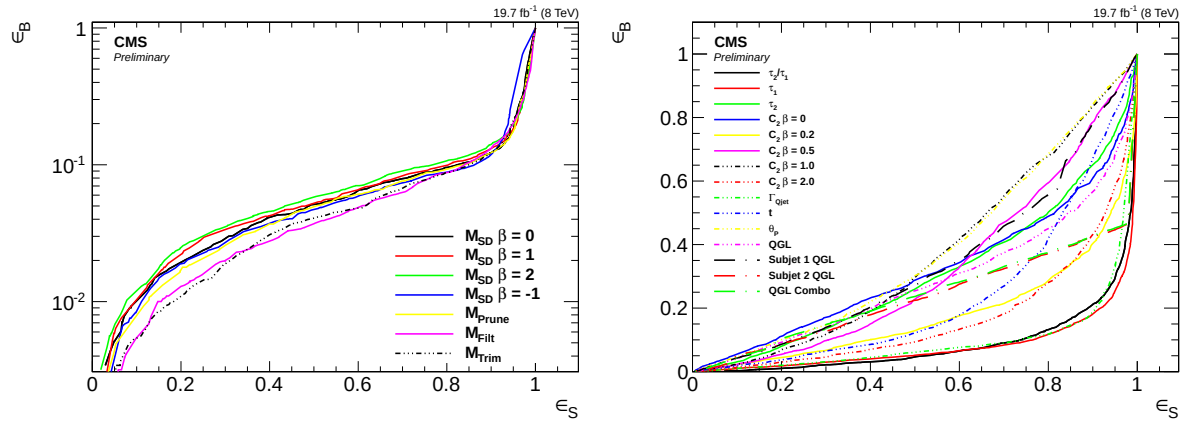

Figure 10: Performance curve for all single variables in mass (left) and other properties (right) used in merged W boson tagger, where signal is from simulated W-jets and background is from Z+jets events in data.

To study the performance of a tagger which employs a combination of all the variables, an MVA based on the BDT is employed. The working point that maximizes background rejection for a signal efficiency of 50% is taken as a benchmark for performance comparisons; the score **Z**, defined as the inverse of the background efficiency [1], is computed. To determine the relevant minimal set of observables that performs best, the procedure is repeated for all the pairs and triplets of observables.

To calculate the uncertainty on the metric **Z**, two sources of uncertainty must be accounted for. The first is the uncertainty induced by statistical effects, the second originates from the uncertainty in the training of the BDT. Since the same parameters are used for each BDT, we take this uncertainty to be of the order of the statistical uncertainty due to finite training sample size. The two uncertainties are added in quadrature. With this well defined metric, the performance of the variables are scanned. We find an uncertainty of roughly 10% for the **Z** comparisons against data and 2% for the **Z** comparisons against simulation.

In Figure 11 the score **Z** is shown for all the BDTs trained on pairs of different observables. The score is shown by considering two different regions for background jets, obtained by applying the $t\bar{t}$ and $Z+$jets event selections. From the figures and Table 1 we conclude that the best performing pairs are a mixture of a groomed mass and N-subjettiness, as was already noted in previous studies [1]. However, a mixture of a groomed mass and combination of QGL applied to subjets gives comparable performance within the uncertainties.

The comparison of peformance for BDTs trained on all triplets of the 22 studied variables shows that a mixture of a single groomed mass variable, N-subjettiness, and subjet QGL perform better than all the other combinations. Often it is found that pairs of mass variables along with either N-subjettiness or subjet QGL perform best. Figure 12 shows the **Z** distribution for background jets taken from the $Z+$jets selection in simulation. The best combination is given by m_{Filt}, m_{Prune}, and subjet QGL; however, many other combinations are statistically close in performance, as shown in Table 2 – which lists the top 20 highest scoring triplets. The resulting ranking of best triplets is quite interesting, as it shows that even though some variables, such as subjet QGL, seem not to have a good discrimination power when considered in isolation, they are in fact giving a boost to the tagging performances when combined to other variables, like mass and/or N-subjettiness.

[1] A score value of 50 corresponds to a 50% signal efficiency for a 2% fake rate.

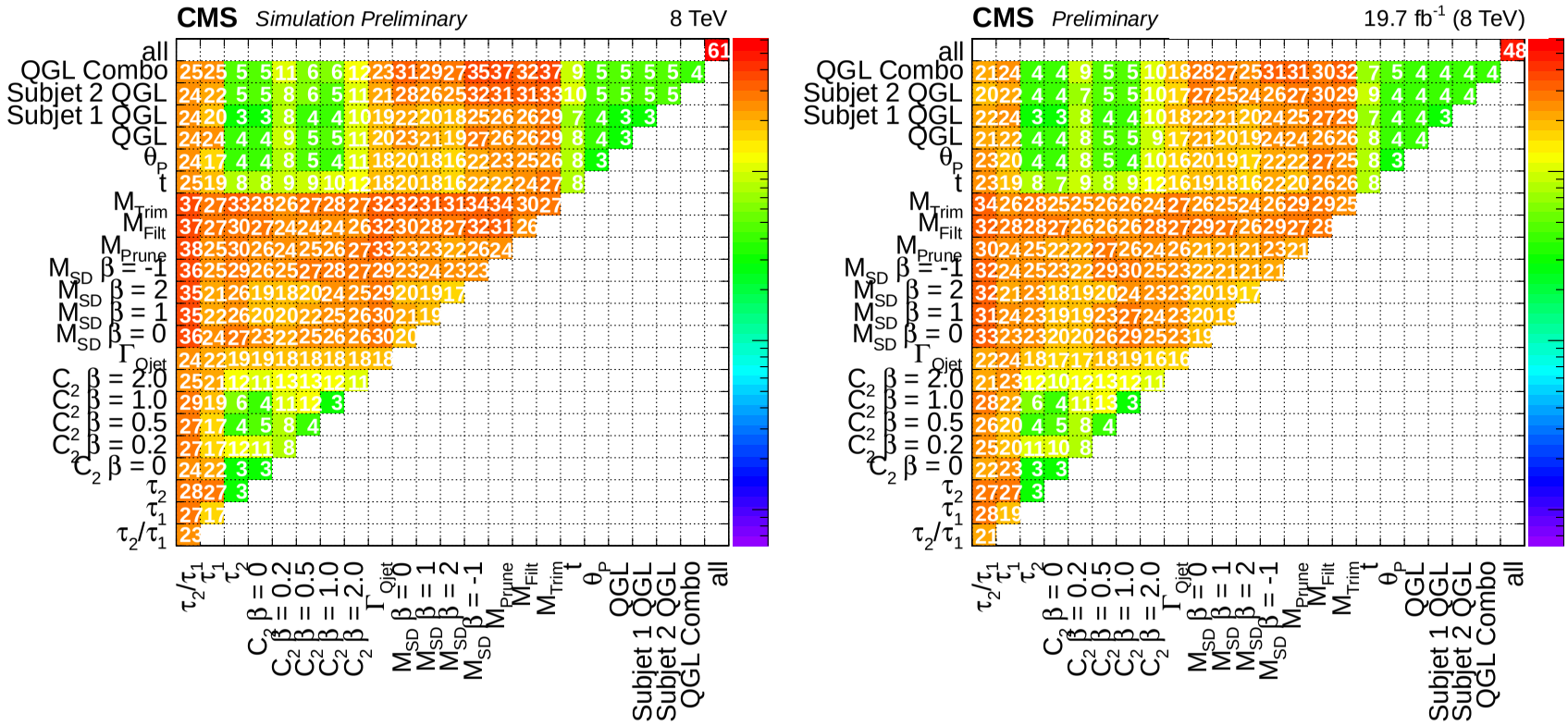

Figure 11: Performance score **Z** for V tagging using BDTs based on different pairs. For a signal selection of simulated W jets and jets from $Z+$jets events in simulation (left) and data (right). The score **Z** is defined as the $1/\epsilon_B$ where ϵ_B is the background efficiency for a signal efficiency of 50%.

Table 1: Performance of the top 10 pairs of variables in simulation.

Rank	**Z**	Pairs of variables	
1.	38.5 ± 1.6	M_{Prune}	τ_2/τ_1
2.	37.9 ± 1.6	M_{Filt}	τ_2/τ_1
3.	37.8 ± 1.6	M_{Trim}	τ_2/τ_1
4.	37.7 ± 1.6	M_{Trim}	QGL Combo
5.	37.2 ± 1.6	M_{Prune}	QGL Combo
6.	36.7 ± 1.5	M_{SD} $\beta = -1$	τ_2/τ_1
7.	36.3 ± 1.5	M_{SD} $\beta = 0$	τ_2/τ_1
8.	35.8 ± 1.5	M_{SD} $\beta = 2$	τ_2/τ_1
9.	35.3 ± 1.4	M_{SD} $\beta = 1$	τ_2/τ_1
10.	35.0 ± 1.4	M_{SD} $\beta = -1$	QGL Combo

Table 2: Performance of the top 20 triplets of variables in simulation.

Rank	**Z**	Triplets of variables		
1.	43.3 ± 2.0	M_{SD} $\beta = -1$	M_{Trim}	QGL Combo
2.	43.2 ± 2.0	C_2 $\beta = 2$	M_{SD} $\beta = 0$	τ_2/τ_1
3.	42.8 ± 1.9	M_{Prune}	QGL Combo	Γ_{Qjet}
4.	42.5 ± 1.9	M_{Prune}	M_{Trim}	QGL Combo
5.	42.5 ± 1.9	C_2 $\beta = 2$	M_{SD} $\beta = 1$	τ_2/τ_1
6.	42.5 ± 1.9	M_{Prune}	QGL Subjet 2	τ_2/τ_1
7.	42.4 ± 1.9	C_2 $\beta = 2$	M_{SD} $\beta = -1$	τ_2/τ_1
8.	42.1 ± 1.9	M_{Prune}	QGL Combo	τ_2/τ_1
9.	42.1 ± 1.9	M_{Filt}	M_{Prune}	QGL Combo
10.	42.0 ± 1.9	C_2 $\beta = 2$	M_{Prune}	τ_2/τ_1
11.	42.0 ± 1.9	C_2 $\beta = 2$	M_{SD} $\beta = -1$	QGL Combo
12.	41.8 ± 1.9	C_2 $\beta = 2$	M_{SD} $\beta = 2$	τ_2/τ_1
13.	41.8 ± 1.9	M_{Trim}	QGL Combo	τ_2/τ_1
14.	41.7 ± 1.9	M_{Filt}	M_{SD} $\beta = 0$	τ_2/τ_1
15.	41.5 ± 1.8	M_{Prune}	M_{Trim}	τ_2/τ_1
16.	41.5 ± 1.8	M_{Filt}	M_{SD} $\beta = -1$	τ_2/τ_1
17.	41.5 ± 1.8	M_{SD} $\beta = -1$	M_{Trim}	τ_2/τ_1
18.	41.5 ± 1.8	C_2 $\beta = 2$	M_{Trim}	τ_2/τ_1
19.	41.5 ± 1.8	M_{Filt}	M_{Prune}	τ_2/τ_1
20.	41.4 ± 1.8	C_2 $\beta = 1$	M_{SD} $\beta = -1$	QGL Combo

Figure 11 and Table 2 are also showing the difference in performance between the best pair/triplet and the full BDT, i.e. they also provide a quantitative estimation of the "minimality" of the selected sets of variables. As this difference is significant, we are interested in extending the

performance studies beyond 2 and 3 variable configurations. To accomplish that, we use the variable ranking, described in Section 5.4, to train a series of BDTs where each variable is sequentially appended in the order of their discriminating power. The performance gain versus appended variable is shown in Figure 12. This figure confirms again that the combination of mass and N-subjettiness provide the major contribution in W tagging. In addition, it suggests that saturation in performance can be attained with 9 variables, as we can group together all the soft-drop related observables given their very strong correlations. We can then conclude that in order to reach the highest performance we have to rely on a combination of a high number of observables. This situation, where single variables are characterized by low discrimination power but their combination is performing well, typically arises in datasets where the variables are loosely correlated. Hence, we performed a detailed study of variables correlation and present this in the next section.

The ability to simulate the variables to accurately reflect the data has a critical importance. The level of agreement between data and simulation for the BDT score of the best performing triplet and the all variables is shown in Figure 13. The signal-background separation deteriorates slightly in data, but the overall agreement is very good.

5.4 Study of correlations

Correlations are an important means by which the discrimination of the individual variables can be better understood. To minimize the effect of hidden correlations, we first transform the variables by running a BDT on each variable, and then use the resulting BDT discriminant as a transformed version of the original variable. These transforms yield functions that are monotonically increasing in signal yield versus background yield.

Following the transformation, we compute the correlation matrix of the transformed variables. This matrix is shown in Figure 14. The Z-axis of the figure shows the linear correlation coefficient of a BDT transformed variable labeled in the x-axis with the variable labeled in the y-axis. The correlation coeffieint is then multiplied by one hundred. From the correlation matrix, a few prominent features appear. First, the mass variables are all strongly correlated, with trimmed mass being the least correlated among them. Second, subleading subjet QGL, pull angle, and pull magnitude are not correlated with any of the other variables. Finally, the energy correlation functions C_2 are correlated with each other, but have limited correlations with other variables. As already noted in the Section 5.3, this indicates that building larger dimensional discriminators will largely benefit from the additional information associated with these uncorrelated variables.

Along with the correlation matrix, Figure 14 shows a final column that indicates the correlation of each variable with a BDT trained on all variables. The correlation of each variable with the BDT trained on all variables gives a measure for the importance of each variable in the final discrimination. We can use this correlation to form a ranking of the individual variables, taking the maximally correlated variables as the most discriminating. This ranking was used to choose the ordering in the iterative scan of discriminator performance (right plot in Figure 12). In the iterative scan, the benefits from the uncorrelated variables is observed in stepwise performance increases, which is most clear for the Subjet QGL combination.

5.5 Summary

We have studied 22 different observables to improve the V tagger performance, including several that are new to CMS. The new variables are the subjet quark/gluon likelihood, the pull angle and pull magnitude, and the soft-drop mass. They are found to be well modeled by the

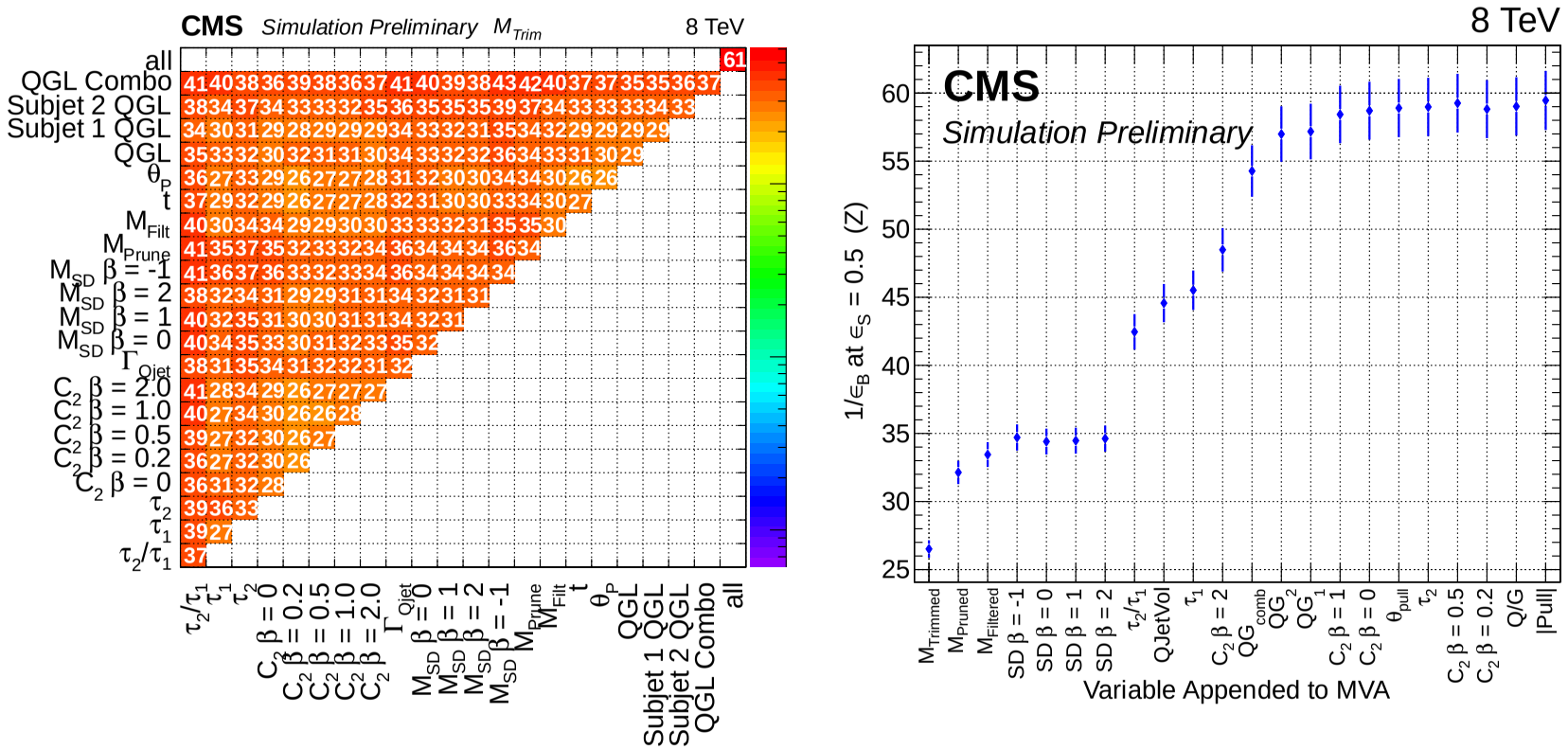

Figure 12: Performance score **Z** for V tagging using BDTs based on different triplets. The score is defined as the inverse background efficiency ($1/\epsilon_B$) for a signal efficiency of 50%. Each triplet contains a variable of type mass, which is limited to values around M_W when making these plots. The plot on the left is show for simulated W jets versus jets in Z+jets simulation. The right plot shows the performance of the same metric for a BDT performed by iteratively adding a variable following the order listed on the x-axis.

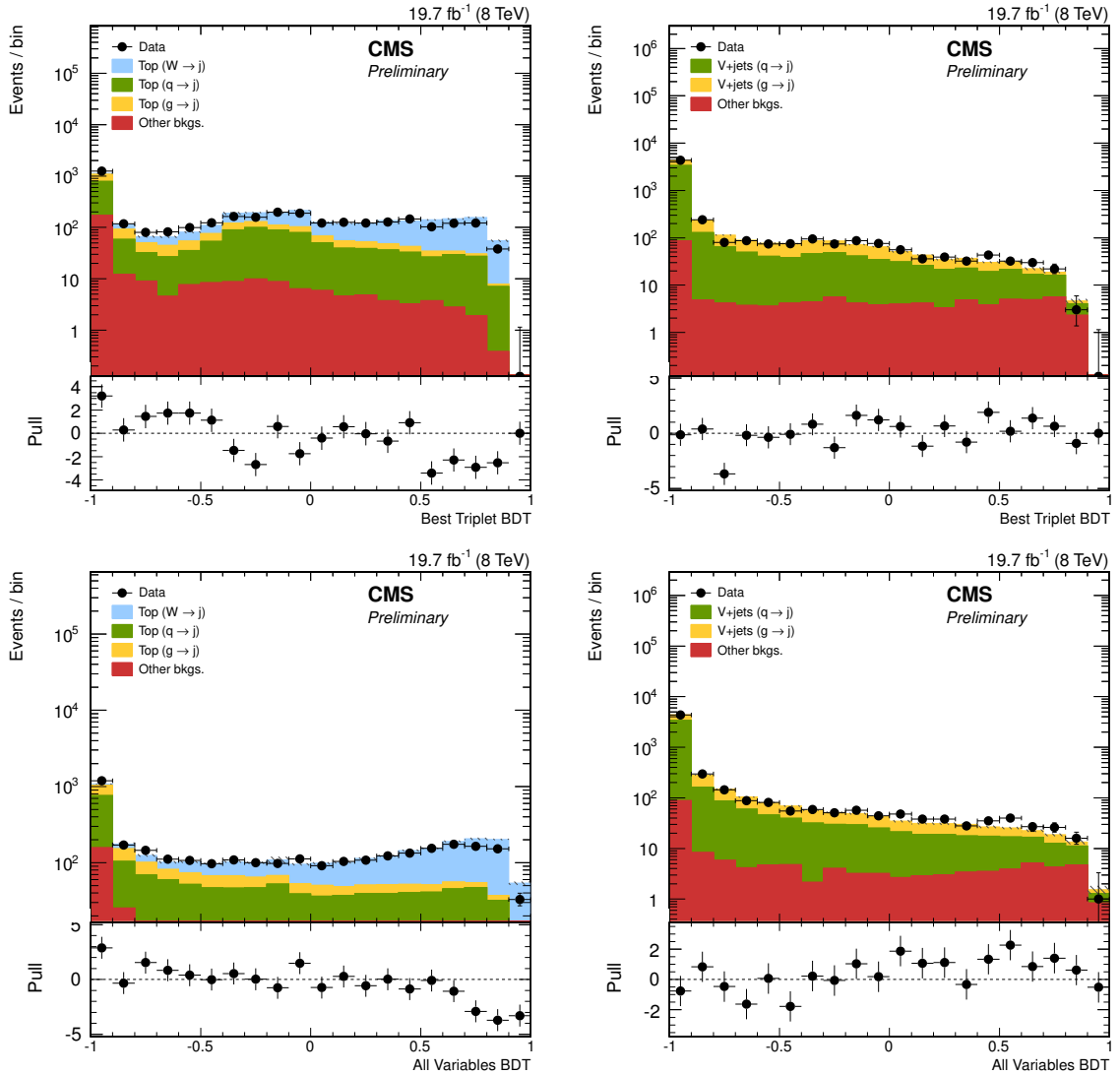

Figure 13: BDT score for $t\bar{t}$ (left column) and $Z+$jets (right column) events. Training for the best performing triplet $(M_{SD}^{\beta=-1}, M_{trim}, L_{q/g}^{comb})$ and all variables are shown on the top and bottom rows, respectively. For both the regions the subleading processes contributions have been considered in the simulation and reported as "Other bkgs." For $t\bar{t}$ events this component includes single top, $W/Z+$jets and di-boson processes. For $Z+$jets events this component includes $t\bar{t}$ and di-boson processes.

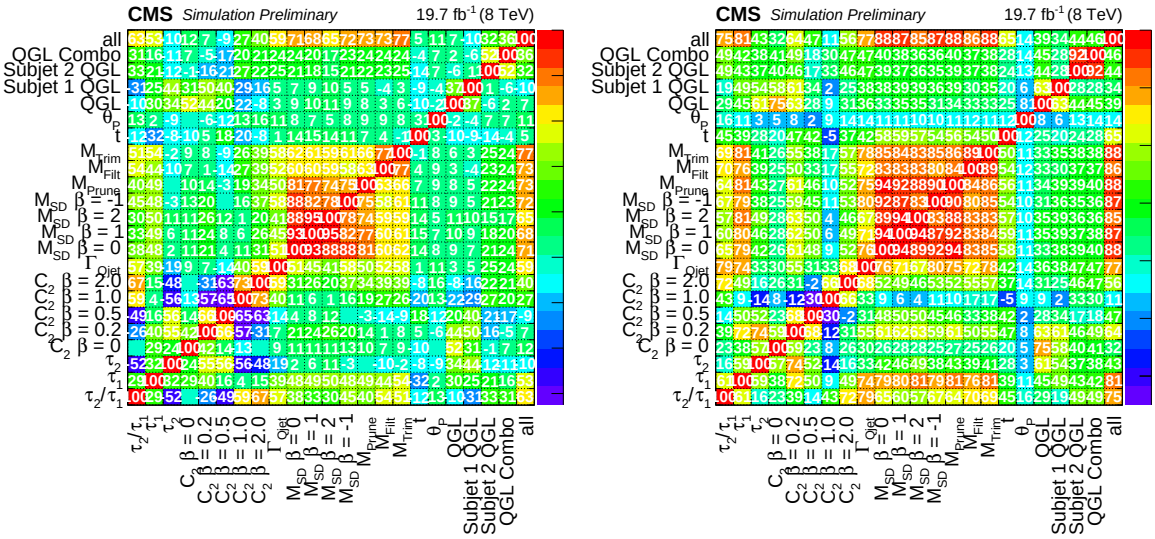

Figure 14: Correlation matrix of the BDTs of single variables with respect to each other for W jet signal (left), and background jets from dimuon data (right). The top row and rightmost column (labeled all) indicates the correlation of the BDT trained on all variables with respect to each set of variables.

current simulation.

We have scanned all combinations of pairs and triplets of the 22 variables to evaluate their discrimination of W jets against background. We find that the best pairwise combinations consist of a groomed mass variable and either τ_2/τ_1 or the newly introduced subjet QGL. The best triplet combinations always include a pair from a small set of variables combined with one of the groomed masses. This small set of varibales that form the pairs are τ_2/τ_1, subjet QGL, $C_2^{\beta=1}$, $C_2^{\beta=2}$, $\Gamma_{Q_t extjet}$ and another groomed mass.

We have also combined the variables into one MVA tagger and scanned the estimate of the most optimal combination over all the variable multiplicities: 1, 2, 3 , ..., 22. We find that combining just 9 of the 22 variables into a one multivariate tagger, at a signal selection efficiency of 50%, improves the background rejection by more than 50% with respect to the old tagger.

6 Performance at high pileup

The increase of the instantaneous luminosity (by roughly a factor two) will produce extreme pileup (PU) conditions, with possibly up to 50 interactions in a beam crossing. While the performance of legacy W tagging schemes has been shown to degrade as a function of jet p_T and pileup [1], substructure reconstruction techniques have presently only been applied on an average of 22 PU interactions. In this section, we demonstrate the performance of CMS jet substructure reconstruction techniques for Run II and describe the improvements with respect to legacy algorithms.

6.1 Simulation and event reconstruction

In this section, we study jet substructure performance in higher pileup conditions and over a range of moderate jet p_{T} ($475 - 600$ GeV), using simulated samples of QCD multijet as background jet source, while RS Graviton (M_G = 1 TeV) decaying in WW final state is adopted as source of W jet. The choice of slightly higher p_{T} than in Section 5 allows us to compare the change in variable performance at a higher p_{T} regime.

As described in Section 3, simulated samples of Run II conditions contain a Poisson distributed sample about 40 PU interactions occurring in the same bunch crossing as (i.e. "in-time" with) the primary interaction. This corresponds to the estimated average number of in-time PU interactions foreseen in Run II. To compare high pileup and low pileup, the samples are split into two distinct regions :

- **low pileup region** The sample requiring a pileup having $N_{PU} < 40$; this region has a mean pileup of *31*.
- **high pileup region** The sample requiring a pileup having $40 < N_{PU}$; which has a mean pileup of *47*.

This difference in pileup studied here is slightly smaller than the expected change in pileup going from Run I to Run II of 18 additional collisions.

The so-called "out-of-time" pileup occurs when the integration time of signals from some detectors (e.g. HCAL) is longer than the time difference between two different bunch crossings. Like in-time pileup, out-of-time can degrade the performance of W tagging. We expect the degradation to be less than that of the in-time. Out-of-time pileup is not considered here, but will be addressed in upcoming studies.

6.2 Jet Charge vs grooming and pileup

Jet grooming, particularly trimming, has been shown to minimize the effects of pileup. In light of a consistent use of grooming, we consider first the W jet observables, jet charge q, defined in Section 4.2, under different grooming and pileup scenarios. To compare the performance under grooming, W bosons are studied separately for each W boson charge, an aspect that will be of importance when separating same- and opposite sign diboson production in the semi-leptonic channel. Figure 15 shows the distribution and discrimination of W^+ against W^- for standard particle flow jets and for several trimming algorithm parameters. Overall charge discrimination is found to be independent of grooming algorithm, however the shape of the jet charge distribution does change as a result of jet grooming. When comparing the change against pileup 30 events with pileup 50 events the performance of the jet charge discriminant changed by only a few percent.

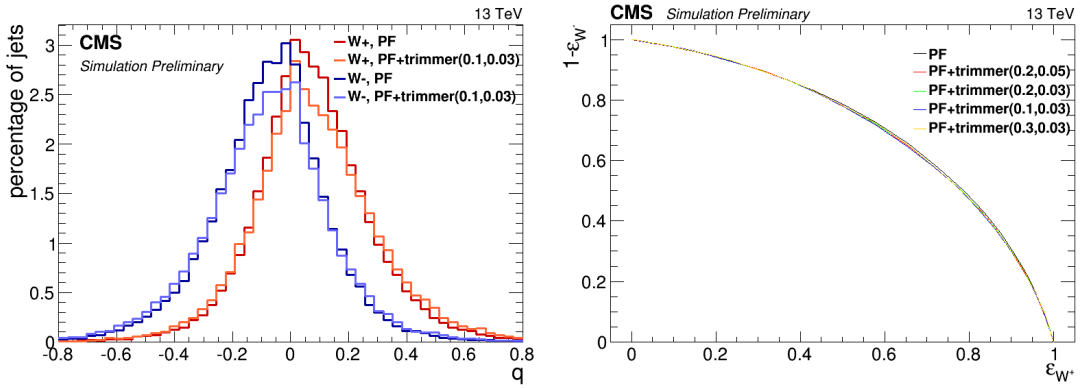

Figure 15: (Left) shape of the jet charge before and after trimming is applied. (Right) performance of charge discrimination for different trimmer parameters.

6.3 Vector boson tagging performance vs pileup

In the interest of preserving the performance under high pileup, we extend the previous section with new techniques to mitigate the influence of pileup. The groomed masses (namely: trimmed, pruned and soft drop mass) are corrected for pileup effects using the "safe" four-vector subtraction scheme [34]. This approach is an extension of the area subtraction technique and includes a correction for jet masses [35, 36]:

$$p^{\mu}_{sub} = p^{\mu} - \rho A^{\mu} - \rho_m A^{\mu}_m \tag{3}$$

where ρ and ρ_m are measures of the average pileup density and the ρ_m, multiplied by the z and E components of A^{μ} (A^{μ}_m), is required due to the non-zero hadron masses of the particle flow inputs. For subtraction on the grooming algorithms of soft drop and trimming, the four-vector subtraction corrects the jet p_T and mass at each step in the algorithm. For pruning, the correction is applied to the final product using the pruned jet area.

In addition to the observables described in Sec. 5.1, one additional pileup mitigation tool for jet mass has also been considered:

- **Constituent Subtraction**: as defined in [37]. This technique involves a local subtraction of pileup at the level of individual jet constituents, which contrasts with area or shape-based subtraction techniques.

Jet masses obtained from constituent subtraction are not further corrected using four-vector safe subtraction. In addition to constituent subtraction, jet cleansing [38] was also investigated, but found to perform poorly on charge hadron subtracted jets; it is not shown in this document.

In Figure 16, we show the performance of different groomed mass observables in the considered jet p_T range for low ($N_{PU} < 40$) and high ($N_{PU} \geq 40$) pileup regions. The performance of raw jet mass, trimming ($R_{filt} = 0.1$, $p_T^{frac} = 0.03$), pruning ($z_{cut} = 0.1$, $R_{cut} = 0.5$), constituent subtraction and soft drop (β = 0,1,2) are compared. From the figures no significant change in the performance is present.

In Figure 17 we show the performance of τ_2/τ_1, τ_2, τ_1, $C_2(\beta = 1, 1.5, 2)$, Γ_{Qjet} and QGL discriminator applied on the jet and on the trailing pruned subjet. No significant degradation in the performance is found as we extend to higher pileup.

Regarding the discrimination at a higher p_T than in Section 5, we conclude that with a signal

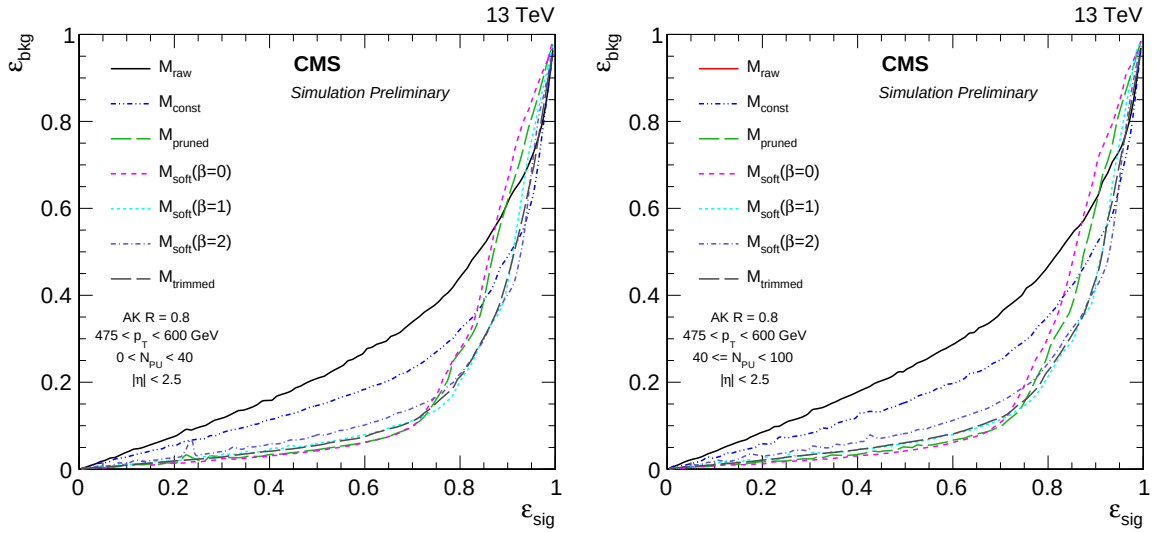

Figure 16: Comparison of various discriminant mass observable performance in the 475-600 GeV jet p_T bin, obtained considering simulated RS Graviton decaying in WW final state as source of W jet. Left: a comparison in the low pileup region $N_{PU} < 40$, and right: a comparison in the high pileup region $40 \leq N_{PU}$.

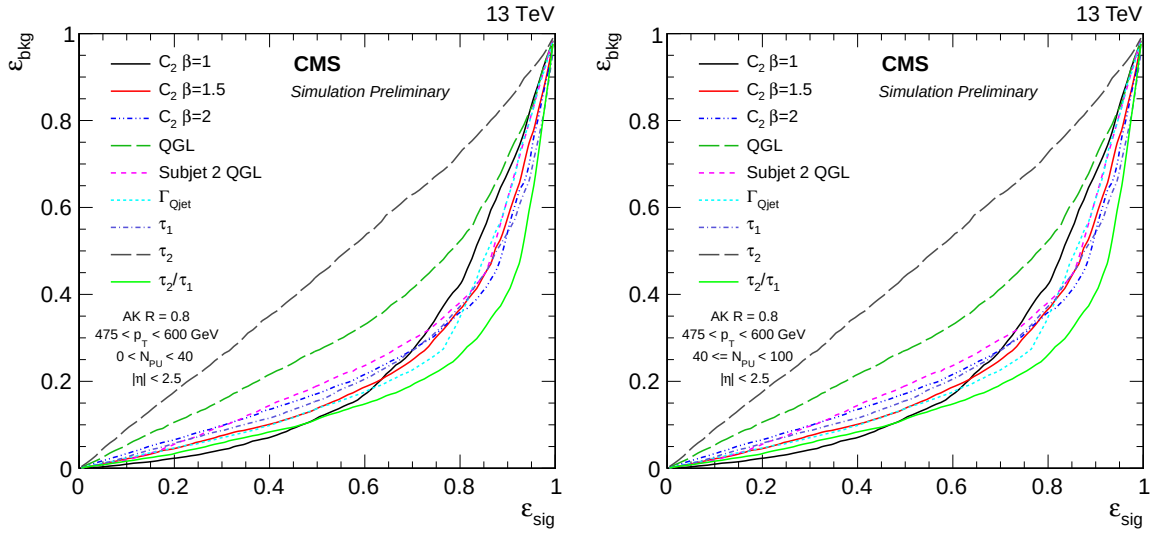

Figure 17: Comparison of various substructure observable performance in the jet p_T bin, 475-600 GeV, considering W jet in RS $G \rightarrow$ WW ($M_G = 1$ TeV) events as signal. (Left) Comparison in the low pileup region $0 < N_{PU} < 40$, (Right) high pileup one $40 \leq N_{PU} < 100$.

efficiency benchmark (ε_{sig}) at 70% efficiency the best performing mass variables are $m^{\beta=0,1}_{softdrop}$, and pruned mass. The trimmed mass is known to be less discriminating as one extends to higher p_T [1]. Of the various substructure observables, τ_2/τ_1 is clearly the best performing.

In light of exploiting the additional variables beyond the mass one, one observable case, we perform a multi-dimensional analysis based on the BDT, following a similar approach to Section 5.3. Again, a working point that maximizes the background rejection for a signal efficiency of 50% is determined and the score **Z**, defined as the inverse of the background efficiency, is computed. To determine the relevant minimal set of observables that provide good performance, the procedure is repeated for all pairs and triplets of observables in each pileup bin.

The left plot in Figure 18 shows the score **Z** for all BDTs trained on pairs of different observables for low pileup. The best discrimination is achieved by combining jet mass variables $M^{\beta=1}_{softdrop}$, M_{pruned} with τ_2/τ_1. This is consistent with results presented in [1]. The second most powerful combination of variables comes from the newly introduced combination of the subjet QGL. The right plot in Figure 18 shows BDT performance combining triplets of observables in events with leading jet p_T between 475 and 600 GeV. The best combination of triplet variables follows the trend observed in Section5.3 where a large class of variables lead to the roughly the same overall discrimination. These variables include M_{pruned} with τ_2/τ_1 and Γ_{Qjets}, or τ_2/τ_1 and a second mass variable, or Subjet QGL with $C_2^{\beta=1}$.

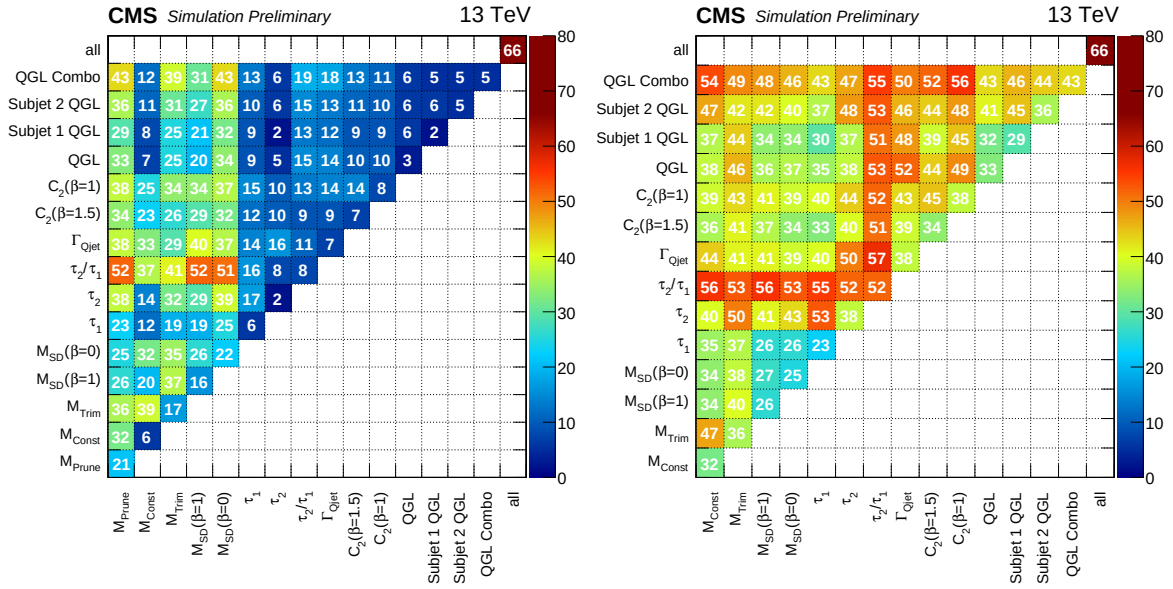

Figure 18: Performance score **Z** for V tagging using BDTs based on different pairs (left) and triplets (right) in events with leading AK8 jet p_T in the range 475 to 600 GeV: $0 < N_{PU} < 40$. The score **Z** is defined as the $1/\epsilon_B$, where ϵ_B is the background efficiency for a signal efficiency of 50%

When extending to higher pileup, we find a degradation in the performance of the overall discrimination. In addition, when scanning over the different variables, we find varying levels of pileup degradation. The performance of each variable is assessed by taking the ratio of **Z** for the high pileup control region ($N_{PU} \geq 40$) with the low pileup region ($N_{PU} < 40$). This ratio is shown in Figure 19 in the form of a two-dimensional map of the BDT pair and triplet trainings. We find that the W-tagging performance degradation with pileup is limited at the level of 20% over 16 pileup interactions when all variables are combined in a single BDT. A few variables,

however show a rather large degradation in the pileup performance. The two worst variables are the Q-jet volatility, Γ_{Qjet}, and the correlation function $C_2^{\beta=1}$. Both these variables lead to degradation on the 30-40% level over a change in average pileup of 16 interactions. Fortunately, the variables, Γ_{Qjet} and $C_2^{\beta=1}$, do not directly affect the performance of the BDT trained on all variables due to that fact that other variables contribute more to the overall discrimination. This study was extended for lower p_T jets, 300-450 GeV, where a smaller degradation with pileup was found.

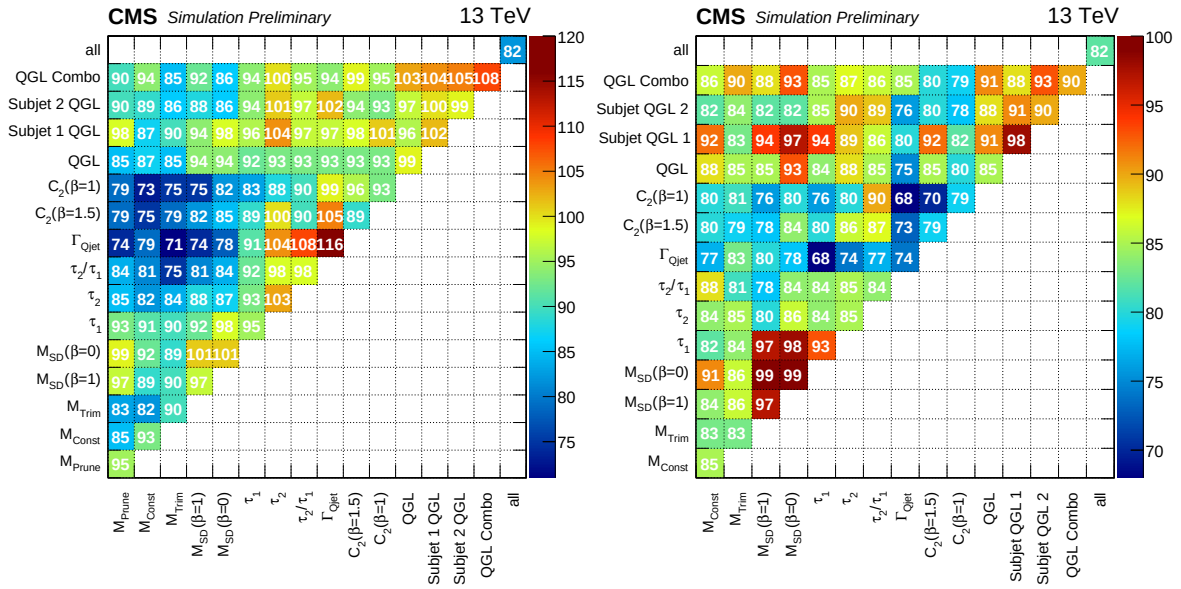

Figure 19: Percent ratio in **Z** score between high and low pileup scenario: (Left) result on BDT trained on pairs of variables, (Right) result on BDT trained on triplets. The triplet includes all combinations with the M_{pruned}.

6.4 Correlation and pileup

As stated in Section 5.4, a correlation matrix of the BDT trained on all variables can be computed with respect to the BDT of any single observable. This allows for a measure of the composition of the maximally discriminating BDT into the individual variable contributions, and provides a means of ranking the individual observable in terms of discriminating power. To determine the affect of pileup on the correlation matrix, we compute the correlation matrix in the lower pileup region and compare it with the high pileup region. The difference in the correlation is shown in Figure 20, shown here for leading jet p_{T} in the range 475 to 600 GeV. This plot allows us to understand were the degradation in performance is coming from. To see the degradation, consider the top line of the correlation matrix labeled all. This is the change in the total discrimination correlation with respect to each individual variable. For both signal and background we find that the variables with the largest reduced correlation are the $C_2^{\beta=1}$ function and the subjet quark gluon discriminators. This implies that the pileup dependence of these variables is causing the overall degradation in performance of the total discriminant. The positive correlations, on the other hand, partially compensate, mitigating the overall degradation with pileup. The largest positive correlation for the discrimination against all is coming from the pruned mass and soft drop ($\beta = 1$).

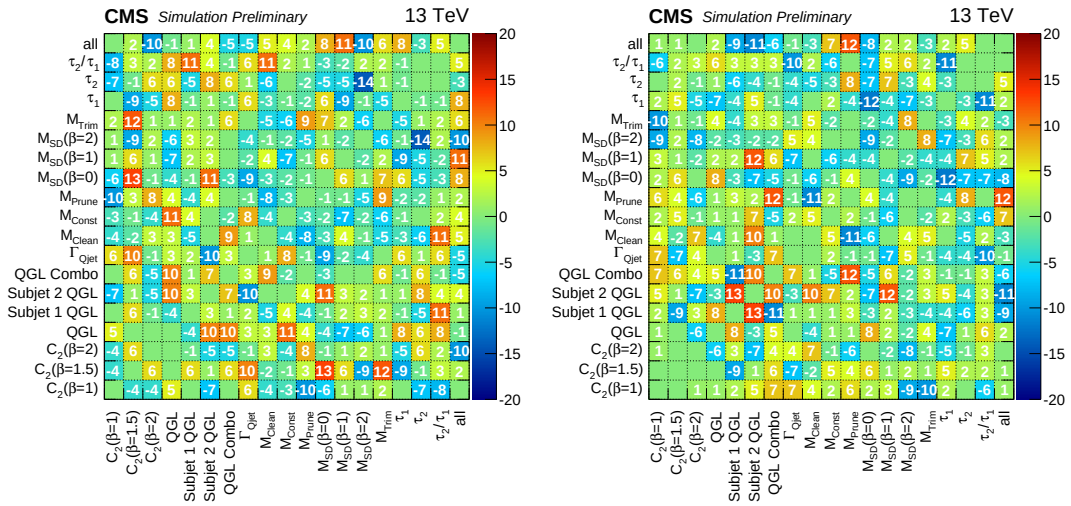

Figure 20: Difference between high pileup ($0 < N_{PU} < 40$) and low pileup ($40 \leq N_{PU} < 100$) in the correlation matrix of the BDTs of single variables with respect to each other for signal (left) and background (right).

6.5 Summary

We have done a systematic study of W jet observables with pileup. We have found the overall degradation in background rejection is expected to be less that 20% at a signal efficiency of 50% over a range of 15 pileup interactions. However, in the context of individual variables, a few variables stand out in their pileup dependence, namely Γ_{Qjet} and the correlation function $C_2^{\beta=1}$. The loss in performance with these variables and others in the overall discrimination is reduced due to compensating nature of the MVA.

7 Performance at high transverse momenta

As described in Section 6, the LHC is scheduled to restart in 2015 with a center of mass collision energy of 13 TeV, enabling searches and measurements of jets with transverse momenta above 4 TeV. Two representative variables, previously used for W boson identification, are used to quantify the performance of jet substructure reconstruction at high transverse momentum, namely the pruned jet mass [25, 26] and N-subjettiness ratio τ_2/τ_1 [29, 30]. It turns out the observed performance loss is actually due to breakdowns in particle reconstruction techniques, rather than the jet substructure algorithms. This section will describe how the CMS reconstruction has been optimized for jet substructure reconstruction for Run II. The performance at high transverse momenta is quantified in terms of W boson identification variables.

7.1 Particle-flow performance

This study uses anti-k_T jets with R = 0.8. We analyze the leading p_T jet in the event which is required to have $|\eta| < 2.4$. W-jets are required to be matched with a generated W within $\Delta R < 0.2$.

The particle-flow algorithm was described in Section 3, but some more technical details will be introduced here, as these have particular relevance to the current discussion. In the case of jets, it is typical that topologically linked clusters and tracks correspond to the traversal of more than one particle through that region of the detector. The particle-flow algorithm resolves the linked clusters and tracks as follows:

1. Charged hadrons are formed from tracks and calorimeter clusters using the best measurement of either the track momentum or the calorimeter energy calibrated for hadrons.
2. If the total calorimeter energy is consistent with the measured track momenta within resolution, no further candidate particles are formed.
3. If a significant excess of ECAL energy, but not HCAL energy, is found, additional photon candidates are formed.
4. If a significant excess of ECAL and HCAL energy is found, additional neutral hadron candidates are formed.

In very high p_T jets, particles are very collimated and have high momentum. For such particles, the track p_T resolution degrades and the angular resolution of the HCAL is not sufficient to measure the angular separation between particles. Not surprisingly, a reduction in the charged constituent multiplicity of the jet is observed with increasing p_T. Typically, the mean charged multiplicity response drops from +3% at 0.5 TeV to -10% at 3.5 TeV and the multiplicity resolution degrades from 10% to 25%. However, the calorimeters compensate and the combination of tracking and ECAL information maintains the measurement of particle multiplicity up to $p_T = 3.5$ TeV, though the mean changes by 11%.

In Figure 21, the distribution of the energy fraction of reconstructed charged hadron candidates with respect to the jet energy and the corresponding response with respect to the generated energy fraction of charged hadrons are shown. The mean response drops from +2% at 0.5 TeV to -68% at 3.5 TeV and the resolution is non-Gaussian already above $p_T = 1$ TeV. Due to the degraded track resolution at high p_T, one cannot maintain an accurate measurement of the charged hadron energy fraction at high p_T. Therefore, the measurement of the spatial energy distribution in a jet is dominated by the tracker for jet $p_T < 1.5$ TeV, while for higher p_T it is dominated by the ECAL.

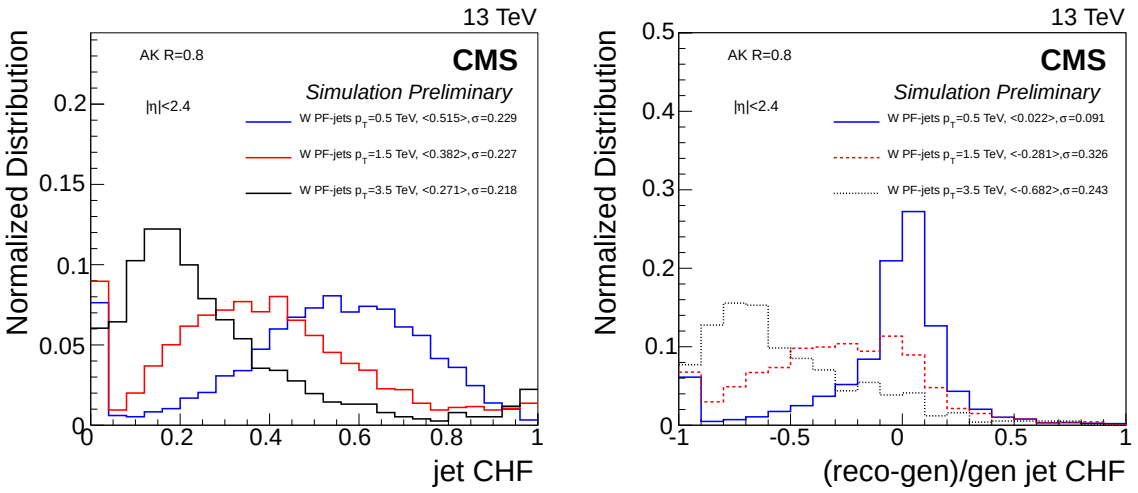

Figure 21: (left) Reconstructed charged hadronic energy fraction (CHF) in W-jets without simulation of PU interactions. (right) Relative difference between reconstructed and generated charged hadronic energy fraction in W-jets without simulation of PU interactions.

The CMS ECAL granularity is 5 times finer than HCAL in both η and ϕ. This finer granularity can be used to improve the spatial resolution needed to perform high-p_T jet substructure. In Figure 22 we compare the number of reconstructed particles in a jet for three scenarios with W-jets at $p_T = 2$ TeV. The three scenarios differ in the treatment of an excess in ECAL and HCAL

energy after reconstructing charged hadrons. In the first scenario, referred to as the *merged PF neutrals*, the excess is assigned to a single merged neutral particle, which is either a photon or neutral hadron. In the second scenario, referred to as the *split PF photons*, we make full use of the ECAL granularity by creating a photon for each ECAL cluster (in the case of an excess in ECAL only). In the third scenario, referred to as the *split PF neutrals*, we treat a simultaneous excess in ECAL and HCAL differently. In such cases, a neutral hadron is created in the direction of each ECAL cluster and the total energy measured in ECAL and HCAL calibrated for hadrons is distributed to the neutral hadrons according to the fraction of the ECAL cluster energy with respect to the sum of all ECAL clusters considered. While the second scenario is basically just using all available detector information, the third scenario makes the assumption that the direction of neutral hadrons can be well approximated by their energy deposits in the ECAL. This increase in the number of particles has a large impact on the performance of W tagging, to be discussed in the next section.

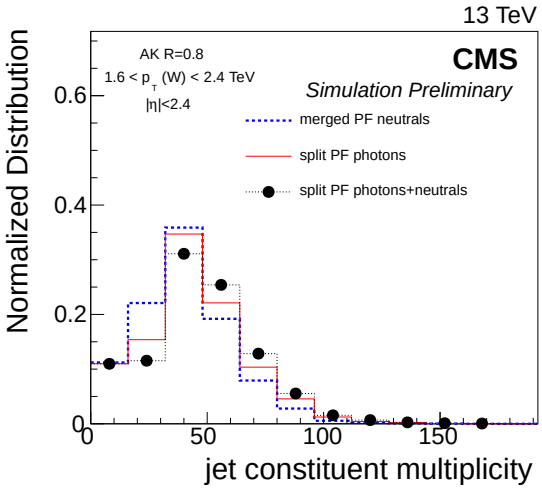

Figure 22: Reconstructed particle multiplicity in W-jets of $p_{\mathrm{T}} = 2$ TeV without simulation of PU interactions. Three neutral particle reconstruction configurations are compared.

Track reconstruction inside a high-p_{T} jet, where particles are very collimated, with large momenta and small track curvatures, can be improved with dedicated algorithms. While such algorithms are particularly important for b-hadron identification in jets, they also improve the measurement of the charged hadron energy fraction in a jet and can extend the region where tracking dominates the jet substructure measurement to higher p_{T}. In Figure 23 we compare the charged hadronic energy fraction in a jet for three scenarios for W-jets of $p_{\mathrm{T}} = 1.5$ TeV. In all scenarios the tracking modifications are studied in addition to splitting neutral hadrons. The first scenario, referred to as the *default tracking*, is the iterative tracking algorithm as used in the reconstruction of the LHC Run I data. In the second scenario, referred to as the *jet core tracking*, an additional tracking iteration is run, where the reconstructed jets are taken as additional seeds for track reconstruction. In the third scenario, referred to as the *jet core tracking and cluster splitting*, clusters in the pixel detector are split up such that they can be used to reconstruct very collimated tracks. For the second scenario, the mean response increases from -28% to -17% and the resolution changes from 31% to 25%. For the third scenario, the mean response increases from -17% to -10% and the resolution changes from 25% to 20% with respect to the second scenario. The impact of this on the performance of W-tagging will be discussed in the next section.

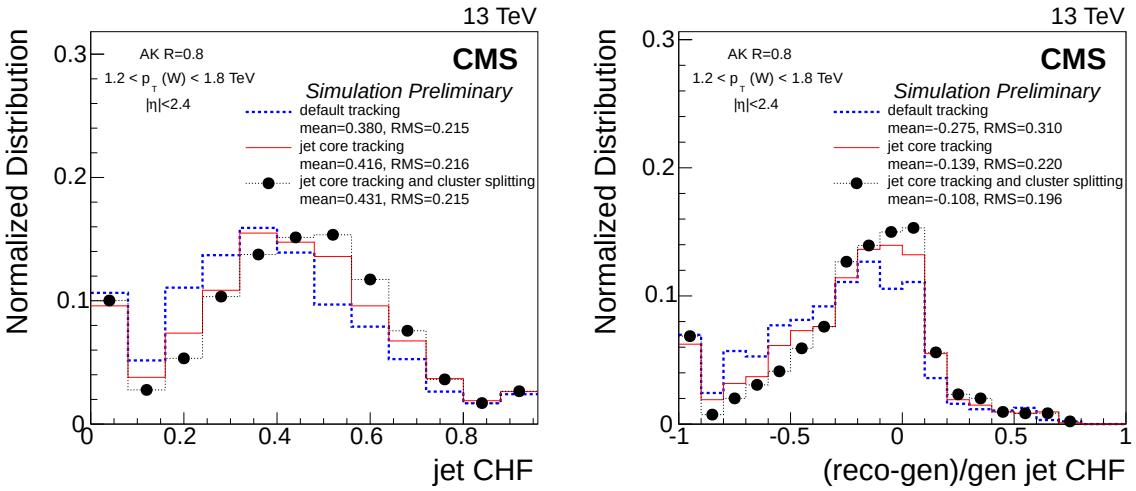

Figure 23: (left) Reconstructed charged hadronic energy fraction in W-jets of $p_T = 1.5$ TeV without simulation of PU interactions. Three track reconstruction configurations are compared. (right) Relative difference between reconstructed and generated charged hadronic energy fraction in W-jets without simulation of PU interactions.

7.2 W tagging performance at high p_T

The performance of jet substructure reconstruction is quantified by analyzing the reconstruction performance of pruned jet mass and τ_2/τ_1. Figure 24 shows how the three scenarios of neutral particle reconstruction perform in terms of pruned jet mass and τ_2/τ_1 reconstruction for W-jets and quark/gluon-jets of $p_T = 2\,TeV$. The pruned jet mass resolution for W-jets is significantly improved by splitting both photons and neutral hadrons. Also, the pruned jet mass for quark/gluon-jets is significantly reduced by splitting both photons and neutral hadrons, giving better discrimination from W-jets. The τ_2/τ_1 distribution is pushed to the left for W-jets and to the right for quark/gluon-jets by splitting both photons and neutral hadrons, giving better discrimination between quark/gluon-jets and W-jets. Adding the improvements from tracking, we find the effect of the tracking scenarios on pruned jet mass and τ_2/τ_1 to be small. This is mainly due to the fact that the ECAL-dominated jet substructure reconstruction, which is used if tracks are not reconstructed with good resolution, performs very similarly to the tracking-dominated reconstruction.

Figure 25 shows the pruned jet mass scale and resolution as a function of p_T. The scale and resolution are obtained by iteratively fitting to the response a Gaussian function in the range of $\pm 1.5\sigma$ around the Gaussian mean. The pruned jet mass scale is stable up to the highest jet p_T for both W-jets and quark/gluon-jets. The pruned jet mass resolution for W-jets increases from 5 GeV to 8 GeV while increasing the jet p_T from 0.5 TeV to 3.5 TeV. Given the three orders of magnitude difference between jet p_T and pruned jet mass, this is a very good precision. The pruned jet mass resolution for quark/gluon-jets is about 4 GeV higher than for W-jets, which can be attributed to the fact that quark/gluon-jets are narrower at the same p_T.

Figure 26 shows the W-tagging efficiency and fake rate using a pruned jet mass $60 < m_{pruned} < 100$ GeV and $\tau_2/\tau_1 < 0.6$ selection as a function of p_T. The pruned jet mass selection efficiency increases up to 1 TeV as a function of p_T because below this value some jets do not fully contain the showers of the two quarks from the W boson decay as discussed in Ref. [1]. Note that for the pruned jet mass selection, the fake rate is, within the uncertainties, consistent with being constant. Additionally, for the combined M_{pruned} and τ_2/τ_1 selection, the efficiency on signal decreases as a function of p_T from roughly 55% at jet $p_T = 1$ TeV to about 45% at 3.5 TeV, while

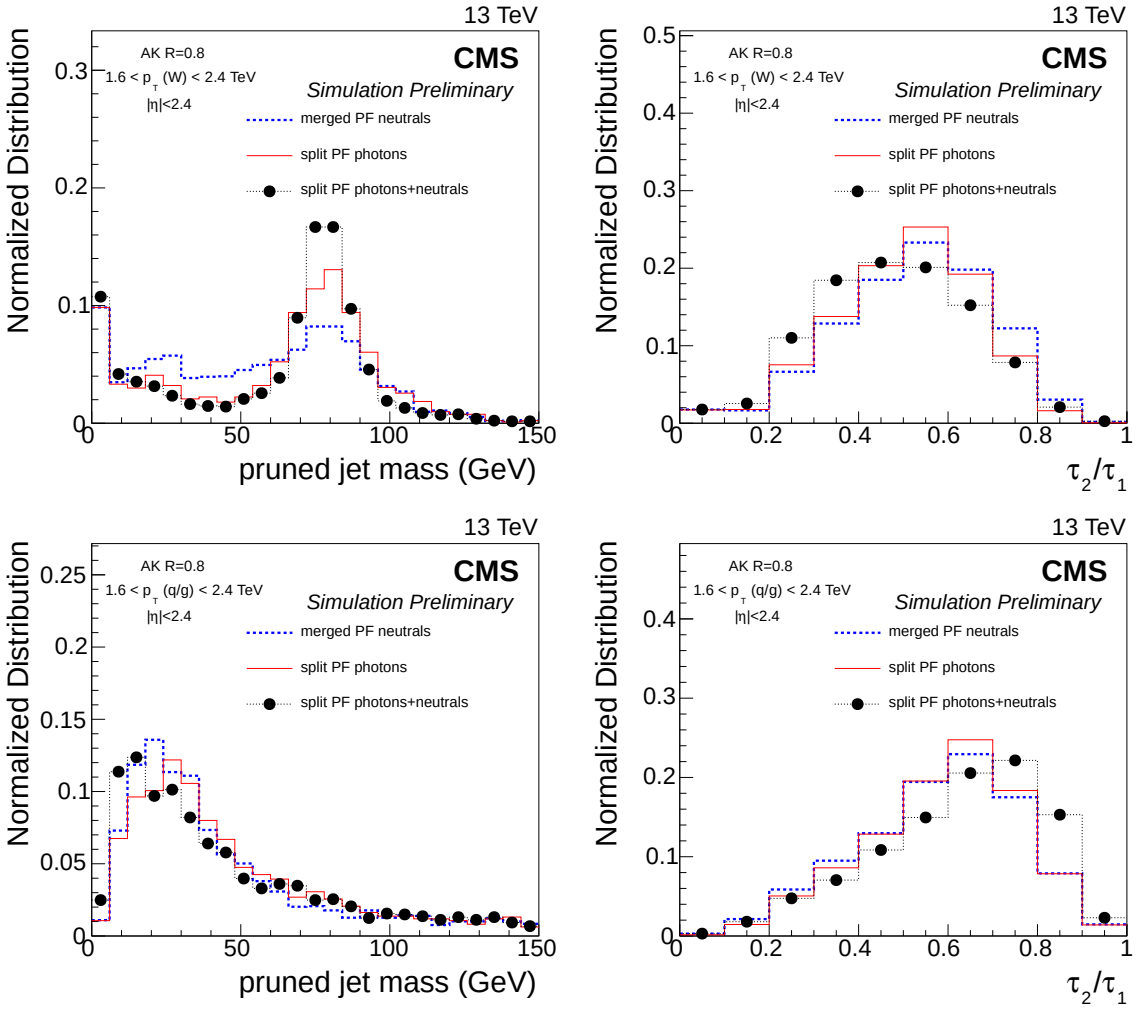

Figure 24: Reconstructed pruned jet mass (left) and N-subjettiness τ_2/τ_1 (right) of W-jets (top) and quark/gluon-jets (bottom) of $p_T = 2$ TeV with 40 simulated PU interactions. Three neutral particle reconstruction configurations are compared.

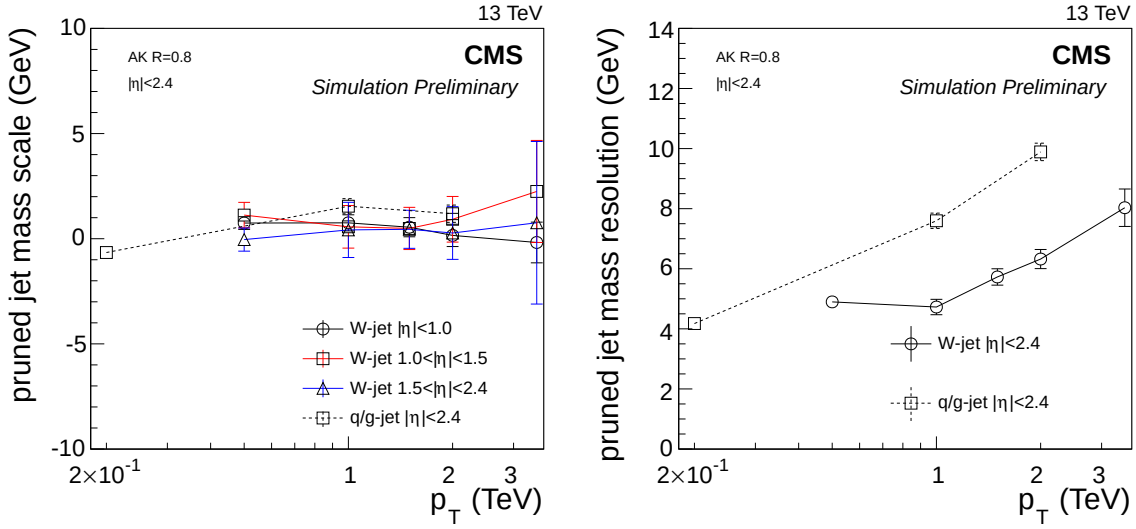

Figure 25: Reconstructed pruned jet mass scale (left) and resolution (right) of W-jets and quark/gluon-jets with 40 simulated PU interactions.

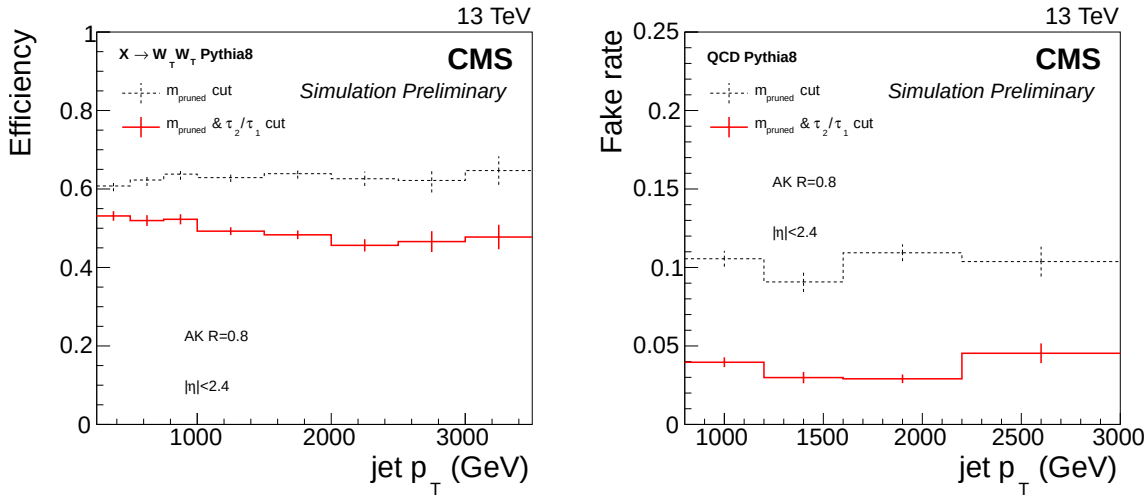

Figure 26: (left) Efficiency of pruned jet mass $60 < m_{pruned} < 100$ GeV and N-subjettiness $\tau_2/\tau_1 < 0.6$ selection on W-jets as a function of p_T with 40 simulated PU interactions. (right) Fake rate of pruned jet mass and τ_2/τ_1 selection on q/g-jets as a function of p_T.

the fake rate remains constant. This feature is again attributed to the jet properties and not to detector effects, since we observe that the generator-level τ_2/τ_1 distribution shifts as a function of p_T.

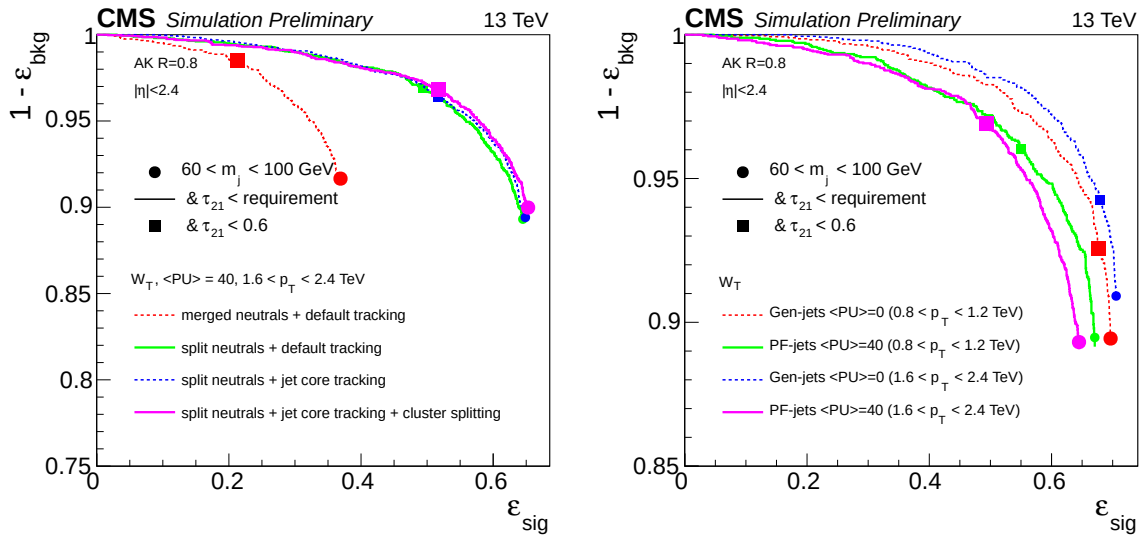

Figure 27: (left) Performance curve for pruned jet mass and N-subjettiness τ_2/τ_1 selection for different neutral particle and tracking scenarios. (right) Performance curve for pruned jet mass and N-subjettiness τ_2/τ_1 selection for two different jet p_T of 1 TeV and 2 TeV with detector simulation with 40 PU interactions and at generator level without PU interactions.

We summarize the performance of different neutral particle and tracking scenarios in terms of a performance curve in Figure 27 (left). The splitting of neutral particles gives a large improvement in terms of discrimination power over the whole range of efficiency, while the various tracking configurations perform very similarly.

Finally, we summarize the performance for two different jet p_T values, 1 TeV and 2 TeV, in terms of a performance curve in Figure 27 (right), comparing it also with the performance expected from generator-level jets. A degradation of performance from 1 TeV to 2 TeV of about 5% signal efficiency at the same fake rate is observed. Comparing the signal efficiency at the same fake rate, the reconstructed performance is about 5% worse than what one can achieve at generator-level without pileup interactions.

7.2.1 Summary

With simple modifications, the CMS particle-flow algorithm can maintain the current V tagging performance up to $p_T = 3.5$ TeV and an average of 40 pileup interactions. The CMS ECAL, thanks to its high precision and granularity, is the key detector component to the performance of high-p_T jet substructure reconstruction. It follows that CMS will effectively provide substructure observables across the full range of jet p_T expected in Run II.

8 Conclusions

As running extends to higher energies, V tagging will become increasingly important. A robust and well performing V tagger across all p_T ranges and pileup conditions is essential to searches for new physics. In preparing for the LHC Run II, we have demonstrated our expanded capability to perform V tagging on all fronts. At low p_T, we have introduced a resolved W tagger that extends identification of hadronic decays of W bosons beyond the traditional method of a

dijet mass cut. For intermediate p_T, where W tagging methods have been well established, we have advanced beyond the single mass variable, single observable technique of the previous CMS W tagger by systematically incorporating all discriminating observables into an MVA. At high p_T, we have overcome the difficulties of high boost, further extending V tagging robustness from 1.5 TeV to 3 TeV. In the interest of harmonizing approaches across all p_T ranges, we have introduced a suite of new observables primarily focused on color flow and subjet properties. Finally, we study these techniques in the high pileup environment, showing performance reductions of less than 20% under LHC Run II conditions.

These developments have improved the W tagging background rejection by more than 50% at a fixed efficiency of 0.5 across the full p_T range. Furthermore, by expanding the base of V tagging we have built a strong foundation for which future studies can build upon as the field continues to evolve.

References

[1] CMS Collaboration, "Identifying hadronically decaying vector bosons merged into a single jet", CMS Physics Analysis Summary CMS-PAS-JME-13-006, 2013.

[2] CMS Collaboration, "The CMS experiment at the CERN LHC", *JINST* **3** (2008) S08004, doi:10.1088/1748-0231/3/08/S08004.

[3] CMS Collaboration, "CMS Luminosity Based on Pixel Cluster Counting - Summer 2013 Update", CMS Physics Analysis Summary CMS-PAS-LUM-13-001, 2013.

[4] J. Alwall et al., "MadGraph 5: going beyond", *JHEP* **06** (2011) 128, doi:10.1007/JHEP06(2011)128, arXiv:1106.0522.

[5] T. Sjöstrand, S. Mrenna, and P. Skands, "PYTHIA 6.4 physics and manual", *JHEP* **05** (2006) 026, doi:10.1088/1126-6708/2006/05/026, arXiv:hep-ph/0603175.

[6] GEANT4 Collaboration, "GEANT4—a simulation toolkit", *Nucl. Instrum. Meth. A* **506** (2003) 250, doi:10.1016/S0168-9002(03)01368-8.

[7] S. M. T. Sjöstrand and P. Skands, "A Brief Introduction to PYTHIA 8.1", *Comput. Phys. Comm.* **178** (2008) 852, doi:10.1016/j.cpc.2008.01.036, arXiv:0710.3820.

[8] L. Randall and R. Sundrum, "A large mass hierarchy from a small extra dimension", *Phys. Rev. Lett.* **83** (1999) 3370, doi:10.1103/PhysRevLett.83.3370, arXiv:hep-ph/9905221.

[9] L. Randall and R. Sundrum, "An alternative to compactification", *Phys. Rev. Lett.* **83** (1999) 4690, doi:10.1103/PhysRevLett.83.4690, arXiv:hep-th/9906064.

[10] CMS Collaboration, "Particle–Flow Event Reconstruction in CMS and Performance for Jets, Taus, and E_T^{miss}", CMS Physics Analysis Summary CMS-PAS-PFT-09-001, 2009.

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

[12] CMS Collaboration, "Determination of jet energy calibration and transverse momentum resolution in CMS", *JINST* **6** (2011) P11002, doi:10.1088/1748-0221/6/11/P11002, arXiv:1107.4277.

[13] 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.

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

[15] CMS Collaboration, "Determination of jet energy calibration and transverse momentum resolution in CMS", *JINST* **6** (2011) P11002, doi:10.1088/1748-0221/6/11/P11002.

[16] 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.

[17] CMS Collaboration, "Identification of b-quark jets with the CMS experiment", *J. Instrum.* **8** (2012) P04013, doi:10.1088/1748-0221/8/04/P04013.

[18] CMS Collaboration, "Performance of CMS muon reconstruction in pp collision events at $\sqrt{s} = 7$ TeV", (2012). arXiv:1206.4071. Submitted to *J. Inst.*

[19] 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.

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

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

[22] A. Hoecker et al., "TMVA: Toolkit for Multivariate Data Analysis", *PoS* **ACAT** (2007) 040, arXiv:physics/0703039.

[23] 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.

[24] D. Krohn, J. Thaler, and L.-T. Wang, "Jet Trimming", *JHEP* **02** (2010) 084, doi:10.1007/JHEP02(2010)084, arXiv:0912.1342.

[25] S. D. Ellis, C. K. Vermilion, and J. R. Walsh, "Techniques for improved heavy particle searches with jet substructure", *Phys. Rev. D* **80** (2009) 051501, doi:10.1103/PhysRevD.80.051501, arXiv:0903.5081.

[26] 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. D* **81** (2010) 094023, doi:10.1103/PhysRevD.81.094023, arXiv:0912.0033.

[27] A. Larkoski, S. Marzani, G. Soyez, and J. Thaler, "Soft drop", *JHEP* **05** (2014) 146, doi:10.1007/JHEP05(2014)146.

[28] A. Larkoski, G. Salam, and J. Thaler, "Energy correlation functions for jet substructure", *JHEP* **06** (2013) 108, doi:10.1007/JHEP06(2013)108.

[29] J. Thaler and K. Van Tilburg, "Identifying boosted objects with N-subjettiness", *JHEP* **03** (2011) 015, doi:10.1007/JHEP03(2011)015, arXiv:1011.2268.

[30] 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.

[31] A. Larkoski, D. Neill, and J. Thaler, "Jet shapes with the broadening axis", *JHEP* **04** (2014) 017, doi:10.1007/JHEP04(2014)017.

[32] S. D. Ellis et al., "Qjets: A Non-Deterministic Approach to Tree-Based Jet Substructure", *Phys. Rev. Lett.* **108** (2012) 182003, doi:10.1103/PhysRevLett.108.182003, arXiv:1201.1914.

[33] Particle Data Group, K. Nakamura et al., "Review of particle physics", *J. Phys. G* **37** (2010) 075021, doi:10.1088/0954-3899/37/7A/075021.

[34] M. Cacciari, G. P. Salam, and G. Soyez, "On the use of charged-track information to subtract neutral pileup", arXiv:1404.7353.

[35] M. Cacciari and G. P. Salam, "Pileup subtraction using jet areas", *Phys. Lett. B* **659** (2008) 119–126, doi:10.1016/j.physletb.2007.09.077, arXiv:0707.1378.

[36] G. Soyez et al., "Pileup subtraction for jet shapes", *Phys. Rev. Lett.* **110** (2013), no. 16, 162001, doi:10.1103/PhysRevLett.110.162001, arXiv:1211.2811.

[37] P. Berta, M. Spousta, D. W. Miller, and R. Leitner, "Particle-level pileup subtraction for jets and jet shapes", *JHEP* **06** (2014) 092, arXiv:1403.3108.

[38] D. Krohn, M. Schwartz, M. Low, and L.-T. Wnag, "Jet Cleansing: Pileup Removal at High Luminosity", arXiv:1309.4777v1.