Available on the CERN CDS information server **CMS PAS JME-13-007**

CMS Physics Analysis Summary

Contact: cms-pog-conveners-jetmet@cern.ch

2014/02/11

Boosted Top Jet Tagging at CMS

The CMS Collaboration

Abstract

Multiple techniques for reconstructing highly Lorentz-boosted, hadronically-decaying top quarks are studied. These techniques, including the CMS Top Tagger, the HEP Top Tagger, and N-subjettiness, use jet substructure and jet mass observables to identify “top-jets”. The efficiency and misidentification rate of these algorithms are compared, both with and without identification of b-quark subjets. The efficiency is measured in data and in simulation, and the ratio of these measurements is presented as a data-simulation scale factor.

1 Introduction

Numerous scenarios for physics beyond the Standard Model (BSM) predict massive particles with large couplings to third generation quarks. Frequently these scenarios result in the production of highly-energetic top quarks. Top quarks produced with an energy greater than twice their mass will be relativistic, resulting in a unique event topology in which the decay products are highly-collimated. If the Lorentz boost of the top quark is $\gamma = E/m$, the angle between the decaying W boson and b quark will be approximately $\Delta R = 2/\gamma$. If the W decays hadronically, the resulting jets (two from the decay of the W boson, and one from the hadronization of the b quark) will often be contained in an angular region less than $2/\gamma$. A jet clustering algorithm with a distance parameter (R) satisfying $R > 2/\gamma$ will typically cluster all of the decay products into a single jet. This resulting object is referred to as a "top jet".

Three algorithms used to identify top jets are presented here. The first is the CMS Top Tagger [1], based on the JHU Top Tagger [2]. The second is the HEP Top Tagger [3]. These algorithms rely on finding subjets and making kinematic selections consistent with a top quark decay. The third algorithm is based on a jet shape observable, N-subjettiness, which measures how consistent a jet is with having N or fewer subjets [4, 5]. In general, these algorithms reconstruct jets using a large distance parameter in order to increase the probability that all of the top decay products are reconstructed within one jet, and utilize the kinematics of the sequential decay $\mathrm{t} \rightarrow \mathrm{W} + \mathrm{b} \rightarrow \mathrm{qq'b}$. Using these identification features it is possible to discriminate top jets from QCD jets with a small mistag rate. Top tagging algorithms have been used at previous measurements at the LHC in 7 TeV collisions in Refs. [6–8], and in 8 TeV collisions in Refs. [9–11].

The CMS, HEP, and N-subjettiness based top-tagging algorithms are described in this document, along with performance measurements in control samples. Data-simulation scale factors, defined as the ratios of tagging efficiencies in data and in simulation, are presented. These scale factors should be used only with data samples that contain events with similar topologies to those in this paper.

2 CMS detector and reconstruction

The CMS detector [12] is a general-purpose detector that uses a silicon tracker, as well as finely segmented lead-tungstate crystal electromagnetic (ECAL) and brass/scintillator hadronic (HCAL) calorimeters. These subdetectors have full azimuthal coverage and are contained within the bore of a superconducting solenoid that provides a 3.8 T axial magnetic field. The CMS detector uses a polar coordinate system with the polar angle θ defined relative to the direction (z) of the counterclockwise proton beam. The pseudorapidity η is defined as $\eta = -\ln\tan(\theta/2)$, which agrees with the rapidity $y = \frac{1}{2}\ln\frac{E+p_zc}{E-p_zc}$ for objects of negligible mass, where E is the energy and p_z is the longitudinal momentum of the particle. Charged particles are reconstructed in the tracker for $|\eta| < 2.5$. The surrounding ECAL and HCAL provide coverage for photon, electron, and jet reconstruction for $|\eta| < 3$. The CMS detector also has extensive forward calorimetry that is not used in this analysis. Muons are measured in gas-ionization detectors embedded in the steel return yoke outside the solenoid.

We use the particle flow reconstruction algorithm [13]. The particle flow algorithm aims to reconstruct and identify each single particle with an optimized combination of all sub-detector information. Each particle is categorized into five types known as particle-flow candidates: muons, electrons, photons, charged hadrons, and neutral hadrons. The energy of photons is directly obtained from the ECAL measurement, corrected for zero-suppression effects. The energy of electrons 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 attached to the track. The energy of muons is obtained from the corresponding track momentum. The energy of charged hadrons is determined from a combination of the track momentum and the corresponding ECAL and HCAL energy, corrected for zero-suppression effects, and calibrated for the nonlinear response of the calorimeters. Finally, the energy of neutral hadrons is obtained from the corresponding calibrated ECAL and HCAL energy.

Muons are reconstructed using the information collected in the muon chambers and the tracking detectors [14]. Tracks from muon candidates must be consistent with a muon originating from the primary vertex and passing quality of fit requirements. The muons have selection criteria based on quality of the muon reconstruction, as well as a kinematic selection of $p_T > 45$ GeV/c and $|\eta| < 2.1$. They are required to be contained in the highest-p_{T} primary vertex.

2.1 Sequential jet clustering algorithms

Jets are defined through sequential, iterative jet clustering algorithms that combine four-vectors of input pairs of particles until certain criteria are satisfied and jets are formed. For the jet algorithms considered in this paper, for each pair of particles i and j, a "distance" metric between the two particles (d_{ij}), and the so-called "beam distance" for each particle (d_{iB}), are computed:

$$d_{ij} = \min({p_{\mathrm{T}}}_i^{2n}, {p_{\mathrm{T}}}_j^{2n}) \Delta R_{ij}^2 / R^2 \tag{1}$$

$$d_{iB} = {p_{\mathrm{T}}}_i^{2n}, \tag{2}$$

where ${p_{\mathrm{T}}}_i$ and ${p_{\mathrm{T}}}_j$ are the transverse momenta of particles i and j, respectively, "min" refers to the lesser of the two p_{T} values, the integer n depends on the specific jet algorithm, $\Delta R_{ij} = \sqrt{(\Delta y_{ij})^2 + (\Delta \phi_{ij})^2}$ is the distance between i and j in rapidity ($y = \frac{1}{2} \ln(E + p_z)/(E - p_z)$) and azimuth ($\phi$), and R is the "size" parameter of order unity [15], with all angles expressed in radians. The particle pair (i, j) with smallest d_{ij} is combined into a single object. All distances are recalculated using the new object, and the procedure is repeated until, for a given object i, all the d_{ij} are greater than d_{iB}. Object i is then classified as a jet and not considered further in the algorithm. The process is repeated until all input particles are clustered into jets.

The value for n in Eqs. (1) and (2) governs the topological properties of the jets. For $n = 1$ the procedure is referred to as the k_{T} algorithm. Jets reconstructed with the k_{T} algorithm tend to have irregular shapes and are especially useful for reconstructing jets of lower momentum [15]. For this reason, they are also sensitive to the presence of low-p_{T} pileup (PU) contributions, and are used to compute the mean p_{T} per unit area (in (y, ϕ)) of an event [16, 17]. For $n = -1$, the procedure is called the anti-k_{T} algorithm, with features close to an idealized cone algorithm. The anti-k_{T} algorithm is used extensively in LHC experiments and by the theoretical community for finding well-separated jets. For $n = 0$, the procedure is called the Cambridge–Aachen (CA) algorithm. This relies only on angular information, and, like the k_{T} algorithm, provides irregularly-shaped jets in (y, ϕ) [18, 19].

The jets used in this analysis are clustered using the Cambridge-Aachen (CA) algorithm [18, 19]. The clustering is done using FastJet version 3 [20]. The CMS Top Tagger algorithm uses jet distance parameter $R = 0.8$, and the HEP Top Tagger algorithm uses $R = 1.5$. The N-subjettiness algorithm is calculated using both $R = 0.8$ and $R = 1.5$ jets. Jet corrections are derived for the anti-k_T jet clustering algorithm [15] with distance parameter $R = 0.7$. Studies in simulation and in semileptonic $\mathrm{t\bar{t}}$ events validate these corrections.

The combined secondary vertex (CSV) b-tagging algorithm [21] is applied to each subjet [22] found by the top tagging algorithms in order to identify subjets originating from b-hadrons. The subjet with the maximum b-tag discriminant is used to tag the top jet. Two subjet b-tagging working points (WP), CSV-loose and CSV-medium, are defined by requiring CSV b-discriminant greater than 0.244 and 0.679 respectively [21]. Subjet b-tagging is studied using the subjets found by both the CMS Top Tagger and the HEP Top Tagger.

3 Reconstruction of top jets

3.1 CMS top-tagging algorithm

The CMS top tagging algorithm is based on the algorithm developed by Kaplan et al. [2] and modified according to Ref. [23]. It has been shown to have comparable performance to other top tagging algorithms[24]. Cambridge-Aachen $R = 0.8$ jets and their particle flow constituents are used as inputs to the top tagging algorithm. The input CA jets are hereby referred to as the "hard jets." The algorithm has two steps: the primary decomposition, in which the algorithm attempts to split the hard jet into two subclusters, and the secondary decomposition, in which the algorithm attempts to split the subclusters found by the primary decomposition.

The decomposition procedure is as follows:

1. The pairwise clustering sequence which was used to form the jet is examined in reverse order to find two subclusters.

2. Continue to the next step if the two subclusters satisfy $\sqrt{(\Delta\eta)^2 + (\Delta\phi)^2} > 0.4 - A \times p_{\mathrm{T}}^{\mathrm{C}}$, where $p_{\mathrm{T}}^{\mathrm{C}}$ is the transverse momentum of the original jet (subcluster) for the primary (secondary) decomposition, and the slope parameter $A = 0.0004$ was optimized using simulated events. If this selection is not satisfied the subclusters are too close and the decomposition fails. The primary decomposition uses the jet p_{T}, while the secondary decomposition uses the p_{T} of the subclusters found by the primary decomposition.

3. If the two subclusters satisfy the p_{T} fraction criterion $p_{\mathrm{T}}^{\mathrm{cluster}} > \delta_{\mathrm{p}} \times p_{\mathrm{T}}^{\mathrm{hardjet}}$, then decomposition succeeds. The default p_{T} fraction parameter is $\delta_{\mathrm{p}} = 0.05$. This value was found to be optimal by Kaplan et al. [2] for moderate to highly boosted tops and was verified to minimize mistag rate for a given signal efficiency in CMS simulation studies.

4. If only one of the subclusters satisfies the criterion $p_{\mathrm{T}}^{\mathrm{cluster}} > \delta_{\mathrm{p}} \times p_{\mathrm{T}}^{\mathrm{hardjet}}$, then the decomposition process is repeated on the passed subclusters, ignoring the constituents from the failed subcluster. This decomposition is repeated until both subclusters pass, both subclusters fail, or the subcluster consists of a single constituent.

5. If, after this iterative process, there is no subcluster with $p_{\mathrm{T}}^{\mathrm{cluster}} > \delta_{\mathrm{p}} \times p_{\mathrm{T}}^{\mathrm{hardjet}}$, or the subcluster is a single constituent, the decomposition fails.

The primary decomposition recursively declusters the hard jet until it finds two subclusters (define them as A and B) which are well separated and contain a significant fraction of the hard jet momentum. If the primary decomposition fails then this jet has one subjet (the original jet). If the primary decomposition succeeds, then the secondary decomposition is applied to subcluster A and B. If subcluster A and subcluster B can not be further decomposed then they become the only two subjets comprising the jet. If both subcluster A and subcluster B are successfully decomposed then the jet has four subjets. If the secondary decomposition succeeds on one subcluster and fails on the other, than this jet has three subjets.

The following variables, defined for each jet passing the algorithm, can be used to tag top jets:

- **Jet Mass** m_{jet} - The mass of the four-vector sum of the constituents of the hard jet.
- **Number of Subjets** N_{subjets} - The number of subjets found by the algorithm.
- **Minimum Pairwise Mass** m_{min} - The three highest p_{T} subjets are taken pairwise, and the invariant mass of each pair is calculated via $m_{ij} = \sqrt{(E_i + E_j)^2 - (\vec{p}_i + \vec{p}_j)^2}$. m_{min} is the mass of the pair with the lowest invariant mass ($m_{\text{min}} = \min[m_{12}, m_{13}, m_{23}]$). This variable is not defined for jets with less than three subjets.

Jets that have mass close to the top mass, at least three subjets, and minimum pairwise mass close to the W mass are tagged as top jets. Only jets with a transverse momentum greater than 350 GeV/c are considered, as at lower momenta the decay products of the hadronically decaying top are not expected to be merged in one single jet with a distance parameter of $R = 0.8$.

3.2 N-subjettiness

N-subjettiness is a jet shape variable designed to measure how consistent a jet is with a hypothesis of having N subjets [4][5]. The N-subjettiness jet shape variable is defined by:

$$\tau_N = \frac{\sum_{i=1}^{n_{\text{constituents}}} p_{\text{T},i} \min\{\Delta R_{1,i}, \Delta R_{2,i}, ..., \Delta R_{N,i}\}}{\sum_{i=1}^{n_{\text{constituents}}} p_{\text{T},i} R} \tag{3}$$

Here N represents the number of subjets in the hypothesis being tested. The summation runs over all particle flow jet constituents ("i"). $p_{T,i}$ is the transverse momentum of constituent i. The quantity $\min\{\Delta R_{1,i}, ..., \Delta R_{N,i}\}$ is the minimum of the ΔR distances between the ith constituent and each subjet axis in the hypothesis. R is the jet distance parameter. The denominator is a normalization factor to ensure $0 < \tau_N < 1$.

The τ_N variable is therefore the p_{T} weighted sum of the angular separation between each jet constituent and the closest subjet axis. Small values of τ_N represent jets which are consistent with having N or fewer subjets. In this case the jet constituents are closely aligned with the subjet axes. Subjet axes are determined by a one-pass optimization procedure which minimizes τ_N[5].

N-subjettiness becomes a more effective discriminator by taking the ratio of jet shapes: τ_N/τ_{N-1}. A top jet is expected to have 3 subjets and thus τ_3/τ_2 provides powerful top jet discrimination.

Selecting jets based on their N-subjettiness value (τ_N) is infrared (IR) safe [25], however selecting jets based on the ratio τ_N/τ_{N-1} is not IR safe [25] but is calculable[26]. The τ_3/τ_2 selection can be made IR safe by also making a cut on τ_2/τ_1 [25]. We find after tagging a top jet with the requirement $\tau_3/\tau_2 < 0.55$, additionally requiring $\tau_2/\tau_1 > 0.1$ is close to 100% efficient for both signal and background jets and provides IR safety.

3.3 HEP top-tagging algorithm

The HEP Top Tagger uses a collection of Cambridge/Aachen jets with a distance parameter $R = 1.5$ ('fat jets'). To identify top jets with the HEP Top Tagger algorithm [3], the following procedure is performed. For each fat jet j, the substructure of the jet is identified by stepping backward through the clustering history of the jet and applying a mass drop decomposition criteria. The jet is decomposed according to the last clustering step into two subjets j_1 and j_2, with $m_{j_1} > m_{j_2}$. If the mass drop condition $m_{j_1} < 0.8 m_j$ is satisfied then the algorithm attempts to further decompose j_1 and j_2. If the mass drop condition fails then j_2 is discarded and the algorithm attempts to further decompose j_1. The procedure is iterated for each subjet found, as long as its mass is greater than 30 GeV/c^2. If the subjet has mass less than 30 GeV/c^2 or if the subjet has only one constituent then the subjet is saved for the next step of the algorithm. Any number of subjets can be found by the mass drop procedure. If less than three mass drop subjets are found the jet is discarded and the HEP top tagging algorithm fails. A filtering algorithm is then applied to each combination of three subjets found by the mass drop procedure. The filtering algorithm reclusters the constituents with variable distance parameter $R_{\text{filt}} = \min(0.3, \Delta R_{ij}/2)$, where i and j are the closest subjets in ΔR. The five reclustered subjets with the largest p_{T} are kept. These kept subjets are referred to as the filtered subjets. The filtered mass is defined by the mass of these five subjets. The combination of mass drop subjets with filtered mass closest to the top quark mass is kept and all other combinations are discarded. The constituents of the filtered subjets are then reclustered again using the exclusive Cambridge/Aachen algorithm which forces the jet to have exactly three final subjets.

The HEP Top Tagger uses these three final subjets to tag top jets by making selections on the quantities:

- m_{123} – The invariant mass of the sum of the 4-vectors of the three final subjets. This is referred to as the HEP top jet mass.
- m_{12} – The invariant mass of the sum of the 4-vectors of the two highest p_T subjets.
- m_{13} – The invariant mass of the sum of the 4-vectors of the highest p_T and the lowest p_T subjets.
- m_{23} – The invariant mass of the sum of the 4-vectors of the two lowest p_T subjets.

The HEP Top Tagger identifies top jets by making top mass selections on the mass of the final three subjets (m_{123}) and W mass selections on the subjet pairwise masses (m_{12},m_{13},m_{23}). The top mass selection requires m_{123} to be in a top mass window and the W mass selection requires the jet pass at least one of the following three conditions:

$$0.2 < \arctan\frac{m_{13}}{m_{12}} < 1.3 \quad \text{and} \quad R_{\min} < \frac{m_{23}}{m_{123}} < R_{\max}$$

$$R^2_{\min}(1 + (\frac{m_{13}}{m_{12}})^2) < 1 - (\frac{m_{23}}{m_{123}})^2 < R^2_{\max}(1 + (\frac{m_{13}}{m_{12}})^2) \quad \text{and} \quad \frac{m_{23}}{m_{123}} > 0.35$$

$$R^2_{\min}(1 + (\frac{m_{12}}{m_{13}})^2) < 1 - (\frac{m_{23}}{m_{123}})^2 < R^2_{\max}(1 + (\frac{m_{12}}{m_{13}})^2) \quad \text{and} \quad \frac{m_{23}}{m_{123}} > 0.35$$

with $R_{\min} = (1 - f_W) \times m_W/m_t$ and $R_{\max} = (1 + f_W) \times m_W/m_t$. Here f_W is the W mass window width. By default f_W=0.15 but alternative values are considered in Section 5.3. These selections are motivated by three body kinematics of the top decay.

Only fat-jets with a transverse momentum greater than 200 GeV/c are considered, as at lower momenta the decay products of the hadronically decaying top are not expected to be merged in one single jet with a distance parameter of $R = 1.5$.

3.4 Combined top-tagging algorithms

It is shown in Section 5 that additional top-jet discrimination can be obtained by combining multiple top-tagging algorithms. The CMS Combined Tagger is defined as the join application of the CMS Top Tagger, N-subjettiness, and subjet b-tagging on the same Cambridge Aachen R=0.8 jet. The HEP Combined Tagger is defined as the join application of the HEP Top Tagger, N-subjettiness, and subjet b-tagging on the same Cambridge Aachen R=1.5 jet.

3.5 Tagging Partially Merged Top Jets

In addition to using top-jet tagging algorithms to identify fully merged top quark jets ('type 1' tags), CMS uses a procedure to identify top quark decays that are only partially merged, consisting of one jet corresponding to the W boson and a second jet corresponding to the b-quark, known as a 'type 2' tag. This top tagging algorithm also uses the large $R = 0.8$ jets, clustered with the Cambridge-Aachen algorithm. To identify the W jet, a jet pruning algorithm [27, 28] is used to remove soft and wide-angle constituents during the clustering process. After the pruning algorithm is applied, the following selections are used to identify jets consistent with W bosons:

- **Number of Subjets N_{subjets}** – Two subjets are required in the pruning algorithm, and are identified by reversing the final step in the clustering process.
- **Pruned Jet Mass m_{jet}** – The jet mass after the pruning algorithm is applied is required to be consistent with the W boson mass, $60 < m_{\text{jet}} < 100$ GeV$/c^2$
- **Mass Drop μ** – The mass drop requirement is defined as $\mu = m_1/m_{\text{jet}}$, and quantifies the energy spread between the two subjets. Here, m_1 is the harder subjet of the two. We expect the mass of W jets to be roughly evenly spread between the two subjets, while QCD jets generally have a massive central core containing a higher fraction of the entire jet mass. We require $\mu < 0.4$ to identify single jets consistent with W decays.

The W jet can be combined with the nearest jet to reconstruct the full top mass. To pass this 'type 2' tagger, we require the combined two-jet mass to be in the top mass window $140 < m_{\text{jet}} < 250$ GeV$/c^2$. Plots of the quantities for this algorithm can be seen in Section 6.

4 Data and Simulated Samples

Top quark events, produced via the strong and the electroweak interaction, are generated with three tools for comparison: MADGRAPH 5.1.1 interfaced to PYTHIA 6.4.26 [29][30], using the MLM matching scheme [31], the next-to-leading-order generator POWHEG v1 interfaced to PYTHIA 6.4.26 for the showering [32][30], and MC@NLO 3.4.1 interfaced to HERWIG 6.520 for showering[33][34]. QCD dijet samples are generated with PYTHIA 6.4.26. MADGRAPH 5.1.1 interfaced to PYTHIA 6.4.26, using the MLM matching scheme [31], is used for W and Z boson production in association with jets. Exclusive samples with 1, 2, 3 or 4 additional partons generated in the matrix element are used. Diboson processes (WW, WZ and ZZ) are generated with PYTHIA 6.4.26 for matrix element and showering.

Events were generated at the center of mass energy of 8 TeV and use the CTEQ6L parton distribution function (PDF) [35], except for the top samples simulated with POWHEG, which use the CT10 PDF set [36]. Resolutions and efficiencies, including those for lepton identification, b-quark tagging, jet energy and angular resolution, are corrected to match the ones measured in data [14, 37–40]. All generated events are propagated through the simulation of the CMS detector based on GEANT 4 [41]. All simulated samples include in-time pileup as well as out-of-time pileup. Simulated samples have been re-weighted to reflect the actual pileup conditions determined after data taking. The jet energy resolution is observed to be 10% lower in data than in simulation[38].

5 Algorithm Performance Results

The observables used by top tagging algorithms to identify merged boosted top jets are studied in simulation. The top-jet tagging efficiency and QCD background mistag rate are defined. In order to compare algorithm performance, the algorithmic selections on the top tagging observables are varied and the minimum mistag rate for a given signal efficiency is determined. Algorithms with lower mistag rate for a given efficiency perform better. Only jets from the $\mathrm{t\bar{t}}$ simulation which are matched to a hadronically decaying top are used.

5.1 Discriminating Variables for the CMS Top Tagger and N-subjettiness

The CA R=0.8 jet mass in QCD and $\mathrm{t\bar{t}}$ samples is shown for jets with $p_{\mathrm{T}} > 500\,\mathrm{GeV}/c$ in Figure 1a. The majority of jets passing this selection contain fully merged tops, and thus have mass close to the top mass. In some cases the W boson decay products are reconstructed within one jet while the b quark is reconstructed within a second jet. In this case the merged W decay products produce a W jet and a shoulder in the jet mass distribution at approximately the W mass. The W mass shoulder in the $\mathrm{t\bar{t}}$ jet mass distribution can be removed by requiring that the jets pass the substructure tagging selections from the CMS Top Tagger and N-subjettiness algorithms, as shown in Figure 1b.

The minimum subjet pairwise mass m_{min} is also an effective discriminator between QCD jets and merged top jets (Fig. 1c). Jets with mass in the top mass window and small τ_3/τ_2 maintain some discrimination in the m_{min} variable, and thus these discriminators can be used in tandem (Fig. 1d).

Similarly, the N-subjettiness jet shape ratio τ_3/τ_2 maintains discrimination both before (Fig. 1e) and after (Fig. 1f) the CMS Top Tagger selection has been applied.

5.2 Discriminating Variables for the HEP Top Tagger and N-subjettiness

The HEP top jet mass (m_{123}) in QCD and $\mathrm{t\bar{t}}$ samples is shown for jets with $p_{\mathrm{T}} > 200\,\mathrm{GeV}/c$ in Figure 2a. The $\mathrm{t\bar{t}}$ sample peaks at the top jet mass and has a low mass tail due to un-merged tops. The QCD distribution exhibits a peak at low mass and a secondary peak at the top mass. The top mass peak in the QCD sample results from the HEP top tagging algorithm procedure in which the combination of mass drop subjets with filtered mass closest to the top mass is chosen and all other combinations are discarded. The HEP top jet mass distribution (m_{123}) after the HEP W mass selection is shown in Figure 2b.

We also explore the possibility of using N-subjettiness in combination with the HEP Top Tagger as an additional discriminating variable. N-subjettiness is calculated using the constituents of the CA R=1.5 jet. The N-subjettiness jet shape ratio τ_3/τ_2 is shown for QCD and $\mathrm{t\bar{t}}$ samples in Figure 2c for $R = 1.5$ jets with no top tagging selection and in Figure 2d after the HEP Top Tagger top and W mass selections are required.

5.3 Algorithm Performance Comparison in Simulation

The algorithm tagging selections are varied iteratively and the efficiency and mistag rate are measured for each iteration. The selection with the smallest mistag rate for a given signal efficiency is kept. The algorithms are compared by plotting the minimum mistag rate versus the signal efficiency. Working points (WP) are defined as example selections which provide the minimum mistag rate for a given signal efficiency.

The denominator in the efficiency calculation is defined as the number of jets which are matched to a simulated top or anti-top quark that decays hadronically and passes the p_{T} selection. Sim-

Working point	m_{jet} selection	m_{min} selection	subjet b-tag WP	τ_3/τ_2 selection
CMS Tagger WP0	140-250 (GeV/c^2)	$>$ 50 (GeV/c^2)	none	none
CMS Combined WP1	140-250 (GeV/c^2)	$>$ 50 (GeV/c^2)	CSV-loose	$<$ 0.7
CMS Combined WP2	140-250 (GeV/c^2)	$>$ 50 (GeV/c^2)	CSV-loose	$<$ 0.6
CMS Combined WP3	140-250 (GeV/c^2)	$>$ 50 (GeV/c^2)	CSV-medium	$<$ 0.55
CMS Combined WP4	140-250 (GeV/c^2)	$>$ 65 (GeV/c^2)	CSV-medium	$<$ 0.4

Table 1: Working points for the CMS Top Tagger and CMS Combined Tagger (CMS + N-subjettiness + subjet b-tag)

ilarly, the mistag rate denominator is defined as the number of jets matched to a simulated quark or gluon from the hard scatter which passes the p_{T} selection. The numerator for both the efficiency and mistag rate is defined by the number of jets from the denominator which pass the top tagging selection. The tagging rate denominators are defined with respect to the matched generator level particle in order to compare algorithms using different jet collections. The mistag rate is dependent on the jet flavor and the results presented here are applicable only to event topologies with a quark/gluon mixture similar to the selection defined in Section 5.

Using these definitions, the minimum mistag rate for a given signal efficiency is shown in Figures 3 and 4. The algorithm tagging selections are varied as follows. The CMS Top Tagger curve is determined by fixing the m_{jet} and N_{subjets} selections ($140\,\text{GeV}/c^2 < m_{\text{jet}} < 250\,\text{GeV}/c^2$, $N_{\text{subjets}} \geq 3$), and varying the m_{min} selection. The N-subjettiness algorithm curve is determined by varying the τ_3/τ_2 selection with no jet mass selection. The subjet b-tagging curve is determined by varying the selection on the maximum subjet CSV discriminant with no jet mass selection. The HEP Top Tagger curve is determined by fixing the m_{123} selection ($140\,\text{GeV}/c^2 < m_{123} < 250\,\text{GeV}/c^2$) and varying the width of the W mass selection (f_W). The remainder of the curves demonstrate the combined application of two or more of these algorithms. In these cases all of the selections mentioned above are simultaneously varied and the minimum mistag rate for a given signal efficiency is determined.

The optimal top tagging algorithm depends on the p_{T} of the matched parton and on the chosen efficiency. In the lowest p_{T} range considered (jets matched to partons with $p_{\mathrm{T}} > 200$ GeV/c^2), only the HEP Top Tagger and HEP Combined Tagger are studied (Fig. 3). Here the best performance is provided by the HEP Combined Tagger, which combines the HEP Top Tagger, subjet b-tagging and N-subjettiness. The CMS Top Tagger and CMS Combined Tagger are not considered in this p_{T} range because the top decay products are rarely merged within R=0.8 jets. For jets matched to partons with $p_{\mathrm{T}} > 400$ GeV/c^2, the majority of boosted top quarks are fully merged into a R=1.5 jet, while a smaller fraction are reconstructed within R=0.8 jets. Therefore the HEP Top Tagger performs very well for high efficiency selections in this p_{T} range, especially when combined with N-subjettiness and subjet b-tagging (HEP Combined Tagger) . The combination of the CMS Top Tagger, N-subjettiness, and subjet b-tagging (CMS Combined Tagger) results in the best performance for low efficiency selections in this range (Fig. 4a). For jets matched to partons with $p_{\mathrm{T}} > 600$ GeV/c^2 or $p_{\mathrm{T}} > 800$ GeV/c^2, the CMS Combined Tagger performs best (Figures 4b and 4c).

Example selections with close to optimal performance for a given tagging efficiency are defined as working points (WP). These WP are defined in Table 1 for the CMS Top Tagger and CMS Combined Tagger and Table 2 for the HEP Top Tagger and HEP Combined Tagger. WP performance is shown by points in Figures 3 and 4. The CMS Top Tagger working point (CMS WP0) has been used for numerous analyses [9, 11, 42].

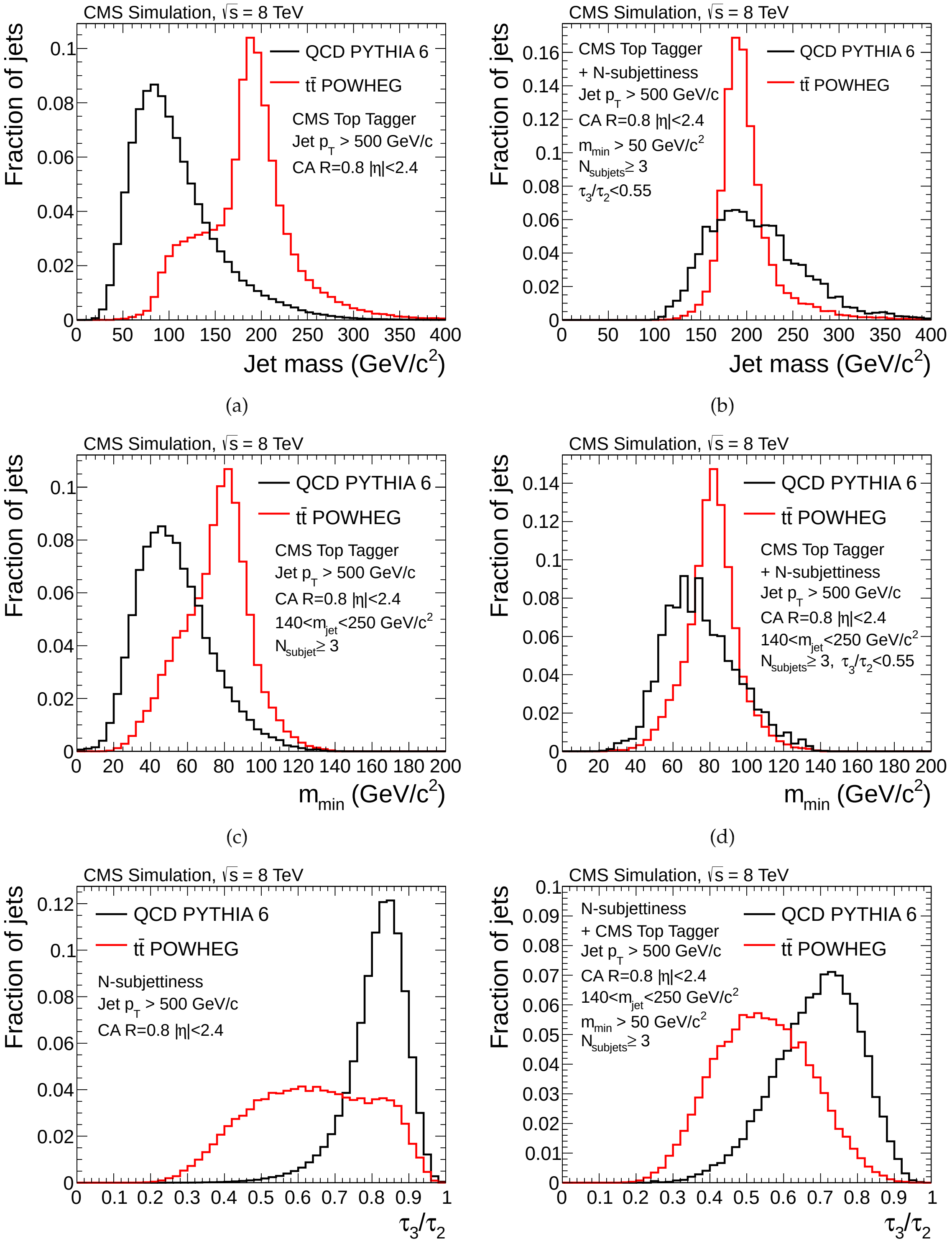
Working point	m_{123} **selection**	f_W **selection**	**subjet b-tag WP**	τ_3/τ_2 **selection**
HEP WP0	140-250 (GeV/c^2)	0.495	none	none
HEP Combined WP1	140-250 (GeV/c^2)	0.495	CSV-loose	none
HEP Combined WP2	140-250 (GeV/c^2)	0.15	CSV-medium	none
HEP Combined WP3	140-250 (GeV/c^2)	0.15	CSV-medium	< 0.63

Table 2: Working points for the HEP Top Tagger and HEP Combined Tagger (HEP Top Tagger + N-subjettiness + subjet b-tag)

Tagging selection	**Efficiency (%) vs.** N_{vtx} **slope**	**Mistag rate (%) vs.** N_{vtx} **slope**
CMS Tagger WP0	-0.031 ± 0.034	0.095 ± 0.006
$\tau_3/\tau_2 < 0.55$ (R=0.8)	-0.429 ± 0.031	-0.031 ± 0.001
Subjet b-tag CSV-medium	-0.049 ± 0.033	0.006 ± 0.002
CMS Combined Tagger WP3	-0.213 ± 0.024	-0.002 ± 0.0002
HEP Tagger WP2	-0.180 ± 0.028	-0.010 ± 0.006
HEP Combined Tagger WP3	-0.463 ± 0.0236	-0.001 ± 0.002

Table 3: Slope of tagging efficiency and mistag rate (in percent) vs N_{vtx} for different tagging selections and $p_{\text{T}}^{\text{jet}} > 500\,\text{GeV}/c$.

A moderate dependence on pileup is observed for these top tagging algorithms. The change in tagging rate as a function of the number of primary vertices is approximately linear. The slope of a linear fit to this distribution can be used to quantify the algorithm performance with pileup (Table 3). N-subjettiness has the largest dependence on pileup. New tools to decrease this dependence are under study [43].



(a)

(b)

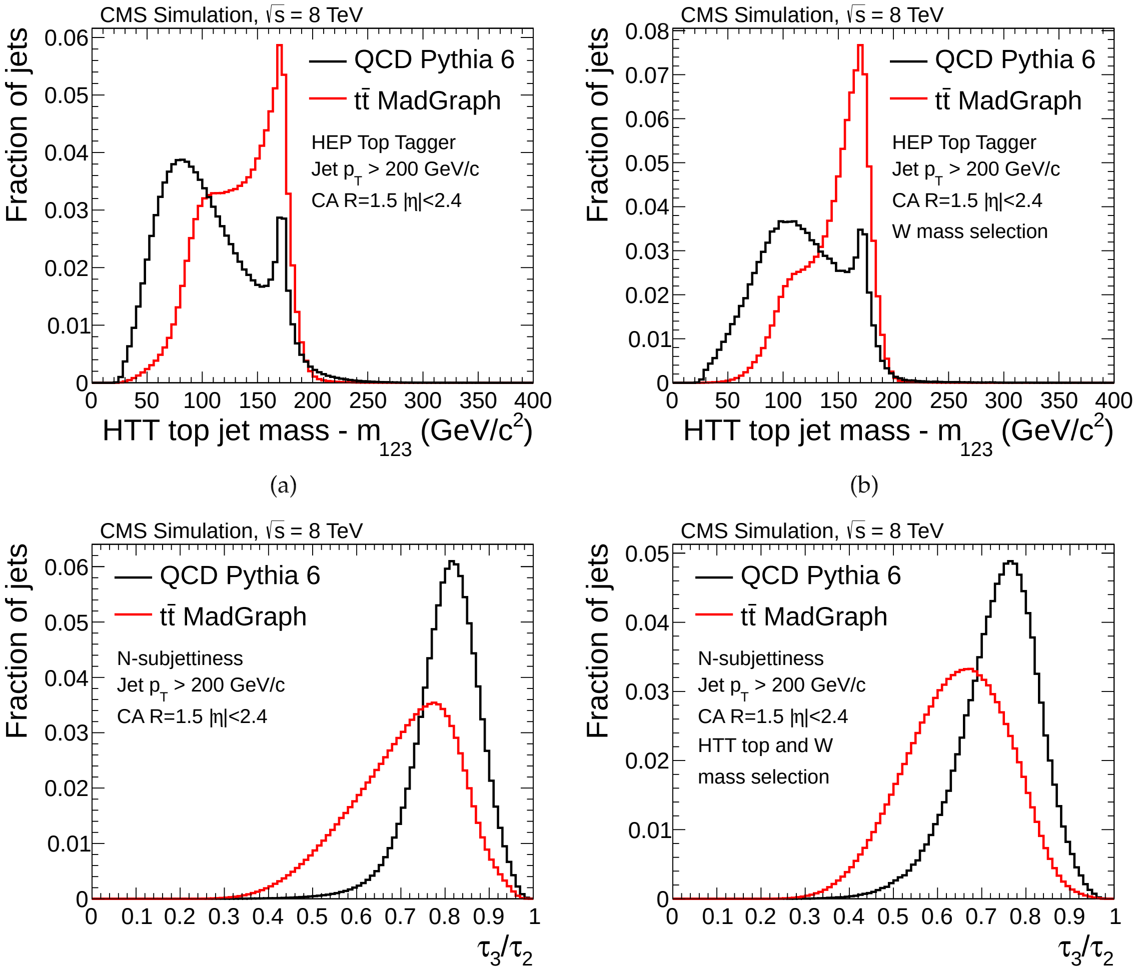
(c)

(d)

(e)

(f)

Figure 1: Top tagging variables for jets with $p_{\mathrm{T}} > 500$ GeV/c from a simulated $t\bar{t}$ POWHEG sample and from a simulated QCD PYTHIA 6 sample: (a) jet mass, (b) jet mass after tagging selections ($N_{\mathrm{subjets}} \geq 3$, $m_{\mathrm{min}} > 50$, $\tau_3/\tau_2 < 0.55$), (c) m_{min}, (d) m_{min} after N-subjettiness selection ($\tau_3/\tau_2 < 0.55$, $140 < m_{\mathrm{jet}} < 250\,\mathrm{GeV}/c^2$), (e) τ_3/τ_2, (f) τ_3/τ_2 after the CMS Top Tagger selection ($N_{\mathrm{subjets}} \geq 3$, $m_{\mathrm{min}} > 50$, $140 < m_{\mathrm{jet}} < 250\,\mathrm{GeV}/c^2$)



(a)

(b)

(c)

(d)

Figure 2: Top tagging variables for CA R=1.5 jets with $p_{\mathrm{T}} > 200$ GeV/c from a simulated $\mathrm{t\bar{t}}$ MADGRAPH sample and from a simulated QCD PYTHIA 6 sample: (a) HEP top jet mass (m_{123}), (b) HEP top jet mass (m_{123}) after W mass tagging selections, (c) τ_3/τ_2, (d) τ_3/τ_2 after the HEP Top Tagger top and W mass selections

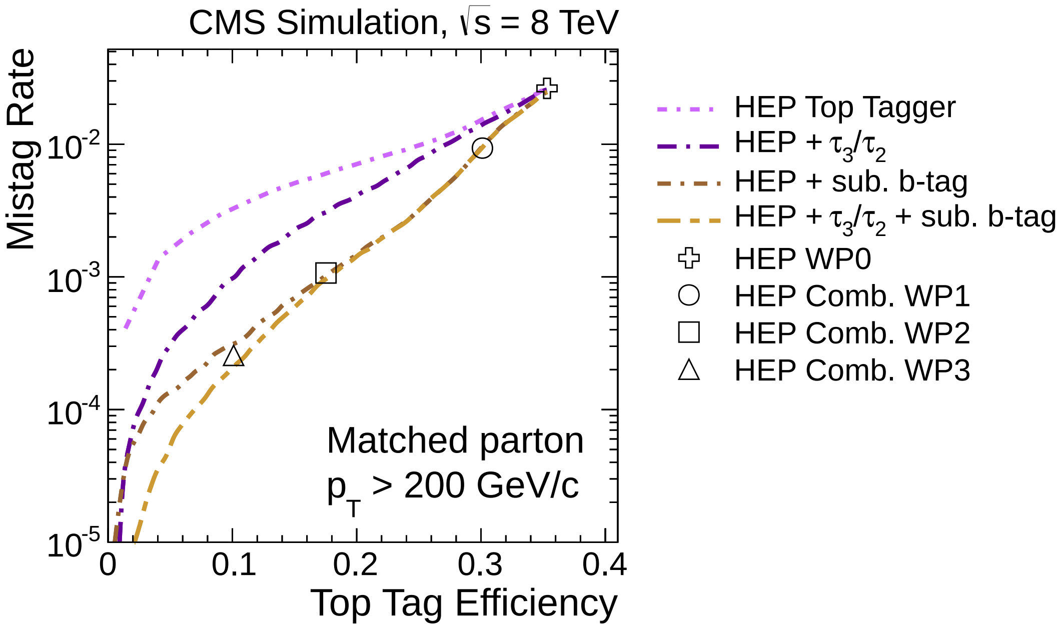


Figure 3: Mistag rate vs. top-jet tagging efficiency for the HEP Top Tagger and HEP Combined Tagger for jets matched to generated partons with $p_{\mathrm{T}} > 200\,\mathrm{GeV}/c^2$. The mistag rate is measured from QCD PYTHIA 6 Monte Carlo while the efficiency is measured with POWHEG $\mathrm{t\bar{t}}$ Monte Carlo. $R = 1.5$ jets are used for all algorithms. A mass cut of $140 < m_{123} < 250\,\mathrm{GeV}/c^2$ is required. Signal jets are matched to simulated all-hadronic generated top quarks, while background jets are matched to simulated partons from the hard scatter. The CMS Top Tagger and CMS Combined Tagger are not considered for this p_{T} range because the decay products of tops with p_{T} close to 200 GeV/c are rarely merged within R=0.8 jets.

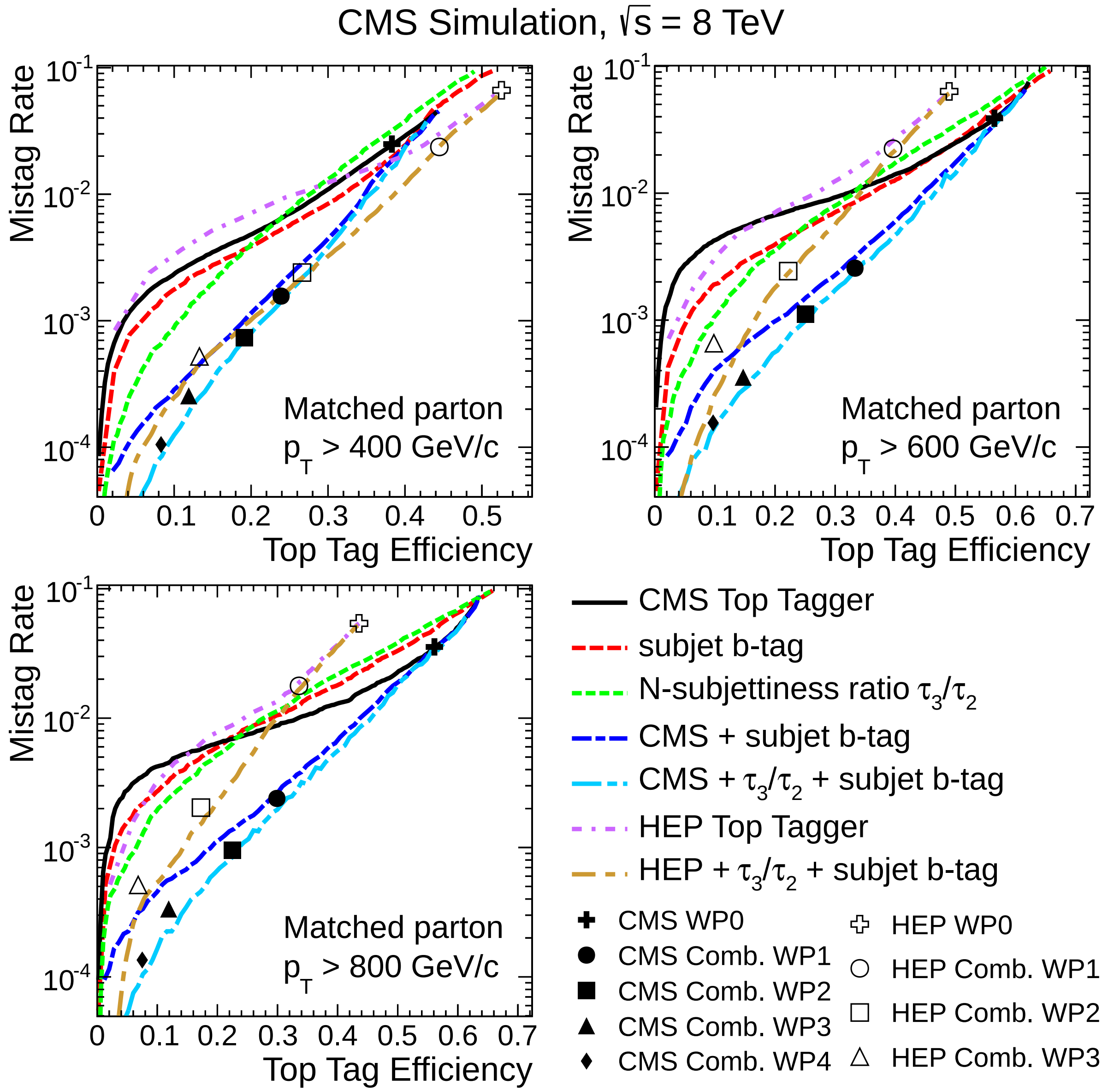


Figure 4: Mistag rate vs. top-jet tagging efficiency as measured from QCD PYTHIA 6 Monte Carlo and POWHEG $t\bar{t}$ Monte Carlo, respectively. In the cases where a jet mass cut is applied, the cut is not varied and is fixed at $140 < m_{\text{jet}} < 250\,\text{GeV}/c^2$. N-subjettiness is calculated using $R = 0.8$ jets except when used in combination with the HEP Top Tagger in which case $R = 1.5$ jets are used. Signal jets are matched to simulated all-hadronic generated top quarks, while background jets are matched to simulated partons from the hard scatter. Distributions are shown for three p_T selections, where the p_T cut is applied to the matched generated parton.

6 Algorithm Performance Comparison in Collider Data

A muon+jets semileptonic $\mathrm{t\bar{t}}$ selection is used to study a pure sample of hadronic boosted top jets in data. This sample is then used to measure the top tagging efficiency data-simulation scale factor for the CMS Top Tagger, CMS Combined Tagger, HEP Top Tagger, and HEP Combined Tagger. This procedure was introduced in Ref. [42].

The muon+jets semileptonic $\mathrm{t\bar{t}}$ sample is selected in data by requiring exactly one high-p_{T} muon ($p_{\mathrm{T}} > 45\,\mathrm{GeV}/c$), and at least one jet tagged with the CSV b tagging algorithm at the medium operating point. This b-tagged jet must satisfy $p_{\mathrm{T}} > 30\,\mathrm{GeV}/c$ and is constrained to be within $\Delta R < \pi/2$ of the muon. The region $\Delta R \geq \pi/2$ of the muon is used to define top candidate jets. A top candidate jet is defined by the highest p_{T} jet in this hemisphere. The top candidate jet for the CMS Top Tagger and CMS Combined Tagger is a Cambridge Aachen R=0.8 jet which must satisfy $p_{\mathrm{T}}^{\mathrm{jet}} > 400\,\mathrm{GeV}/c$ and $|\eta| < 2.4$. The top candidate jet for the HEP Top Tagger and HEP Combined Tagger is a Cambridge Aachen R=1.5 jet which must satisfy $p_{\mathrm{T}}^{\mathrm{jet}} > 200\,\mathrm{GeV}/c$ and $|\eta| < 2.4$. The HEP Top Tagger and HEP Combined Tagger additionally require at least one b-tagged subjet in order to increase the fraction of fully merged top quarks at low p_{T}.

By investigating these top candidate jets, we can extract the top tagging efficiency in data and simulation. The denominator of this top tagging efficiency is then the number of top candidate jets, while the numerator is the number of top candidate jets that pass the top tagging selection. We can then define the top tagging scale factor as the ratio of the top tagging efficiency as measured in data to that measured in simulated events. The semileptonic selection is highly pure and therefore no background is subtracted for the efficiency measurement.

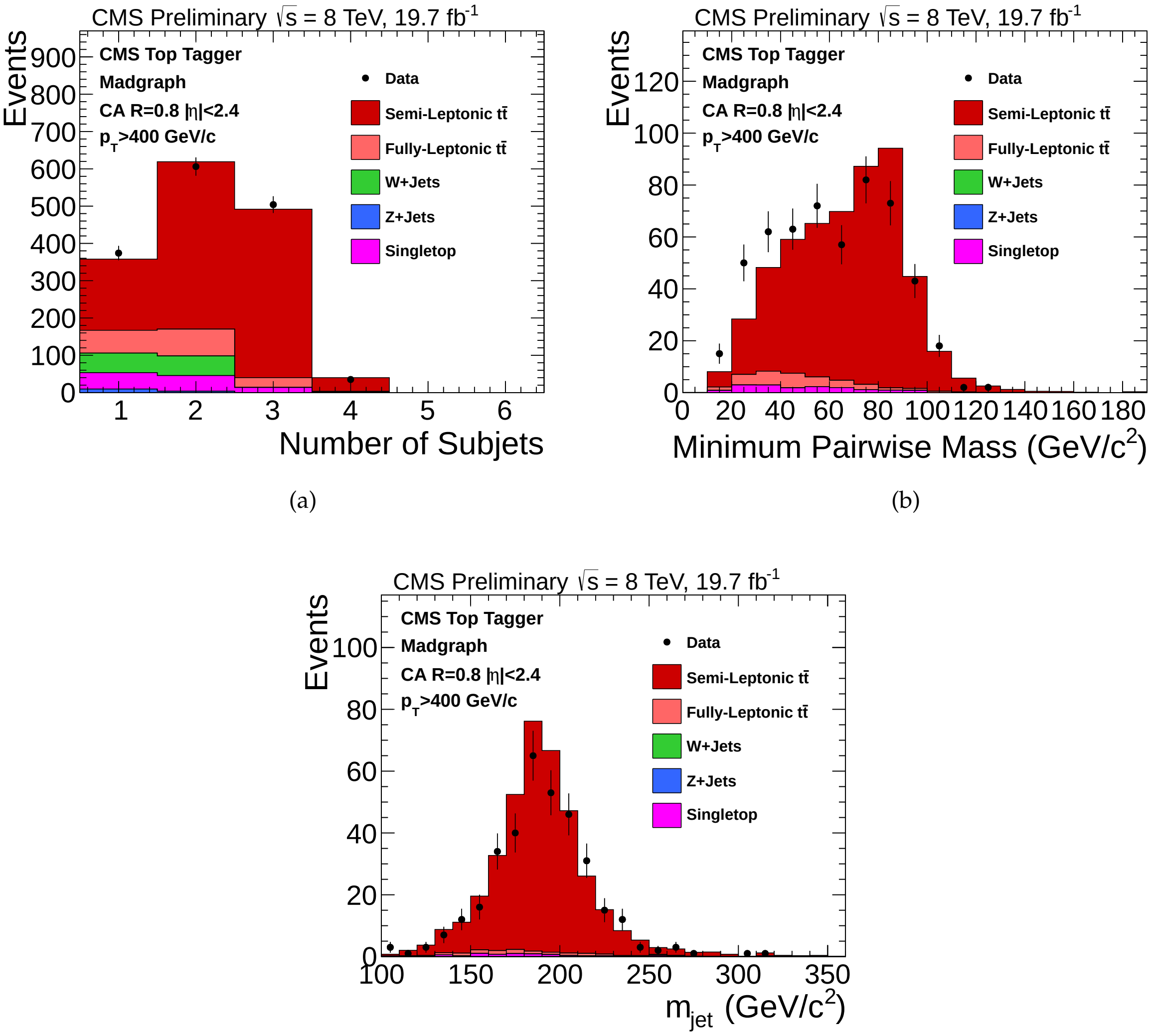
To correct for known differences in p_{T} between $\mathrm{t\bar{t}}$ data and simulation, events are weighted such that the leptonic top p_{T} distribution is identical between data and simulation in the full selection before top variables are investigated. The leptonic top p_{T} is defined as the sum of the muon p_{T}, b-tagged jet p_{T}, and missing transverse momentum.

6.0.1 CMS Top Tagger scale factor measurement

Figure 5 shows the CMS Top Tagger observables (number of subjets, minimum pairwise subjet mass m_{min}, and jet mass) using the semileptonic selection described in Section 6. The m_{min} distribution is not well modeled by the simulation and is most discrepant at large jet pseudorapidity. This effect may be due to mis-modeling of radiation within the top jet or merged subjets at very high jet momenta. The discrepancy is more evident in the pairwise mass of the two lowest p_{T} subjets m_{23} at large jet pseudorapidity. In order to account for this effect, we choose to measure a pseudorapidity-dependent scale factor, with one scale factor measured in the central region ($|\eta| < 1.0$) and a second for the forward region ($1.0 < |\eta| < 2.4$).

As demonstrated in Section 5.1, N-subjettiness can be used to further separate top and QCD jets after the CMS Top Tagger selection has been applied. Additional discrimination can be obtained by applying the CMS CSV b-tagging algorithm to the subjets found by the CMS Top Tagger algorithm. The N-subjettiness variable τ_3/τ_2 and the b-tagging discriminator of top candidate jets from the semileptonic selection are shown in Figure 6. The mass of the top candidate after the successive cuts on τ_3/τ_2 and subjet b-tagging discriminator is shown in Figure 7.

The top tagging scale factor is measured using the semileptonic selection as described in Section 6. Cumulative top tag efficiencies, measured after applying all CMS Tagger WP0 selections and after all CMS Combined Tagger WP3 selections, are shown in Table 4. The corresponding data-simulation scale factors are given in Table 5. The efficiency of each sequential selection of the



(a)

(b)

(c)

Figure 5: Distributions of top tagging variables for full merged top jet candidate in the hadronic hemisphere after the semileptonic selection: (a) number of subjets, (b) minimum pairwise subjet mass, and (c) and jet mass. $t\bar{t}$ is simulated with the MADGRAPH event generator. These distributions are used to evaluate the top-tagging efficiency SF. The m_{min} distribution is not well modeled by the simulation. This effect may be due to mis-modeling of radiation within the top jet or merged subjets at very high jet momenta. The discrepancy is most evident at large jet pseudorapidity and therefore we choose to measure a pseudorapidity-dependent scale factor.

Cumulative top-tagging efficiency - CMS Tagger, CMS Combined Tagger

$\mathbf{\lvert\eta\rvert < 1.0}$				
Selection	**Data**	**MadGraph**	**powheg**	**mc@nlo**
CMS Tagger WP0	0.252±0.013	0.256±0.013	0.215±0.012	0.244±0.014
Combined tagger WP3	0.092±0.009	0.103±0.009	0.086±0.009	0.098±0.010

$\mathbf{1.0 < \lvert\eta\rvert < 2.4}$				
Selection	**Data**	**MadGraph**	**powheg**	**mc@nlo**
CMS Tagger WP0	0.129±0.016	0.200±0.019	0.183±0.017	0.167±0.016
Combined Tagger WP3	0.042±0.009	0.061±0.011	0.047±0.010	0.053±0.009

Table 4: Cumulative efficiencies after requiring all selections of the CMS Top Tagger WP0 and all selections of the CMS Combined Tagger WP3 in data and simulation, as measured using the semileptonic selection. The efficiency is measured for three $\mathrm{t\bar{t}}$ Monte Carlo generators.

Cumulative data-simulation scale factor - CMS Tagger, CMS Combined Tagger

$\mathbf{\lvert\eta\rvert < 1.0}$			
Selection	**MadGraph**	**powheg**	**mc@nlo**
CMS Tagger WP0	0.985 ± 0.073	1.173 ± 0.092	1.033 ± 0.081
CMS Combined Tagger WP3	0.891 ± 0.118	1.063 ± 0.146	0.933 ± 0.129

$\mathbf{1.0 < \lvert\eta\rvert < 2.4}$			
Selection	**MadGraph**	**powheg**	**mc@nlo**
CMS Tagger WP0	0.644 ± 0.100	0.704 ± 0.110	0.768 ± 0.118
CMS Combined Tagger WP3	0.685 ± 0.199	0.906 ± 0.277	0.802 ± 0.230

Table 5: Data-simulation scale factors after requiring all selections of the CMS Top Tagger WP0 and after requiring all selections of the CMS combined tagger WP3 as measured using the semileptonic selection. The scale factor is measured for three $\mathrm{t\bar{t}}$ Monte Carlo generators.

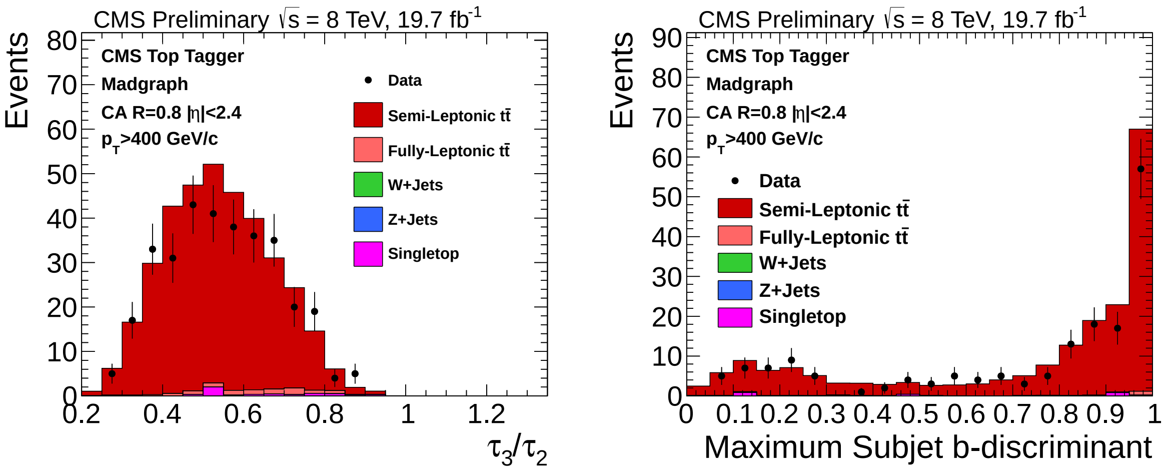
CMS Combined Tagger WP3 is shown in Table 6 and the corresponding data-simulation scale factors are given in Table 7.

Sequential selection top-tagging efficiency - CMS Tagger, CMS Combined Tagger

$\|\eta\| < 1.0$				
Selection	**Data**	**MADGRAPH**	**POWHEG**	**MC@NLO**
$N_{\mathrm{subjets}} \geq 3$	0.367±0.015	0.365±0.015	0.318±0.014	0.362±0.016
$m_{\mathrm{min}} > 50$ (GeV/c^2)	0.719±0.023	0.754±0.022	0.735±0.024	0.725±0.024
$140 < m_{\mathrm{jet}} < 250$ (GeV/c^2)	0.954±0.012	0.928±0.015	0.917±0.017	0.928±0.016
$\tau_3/\tau_2 < 0.55$	0.554±0.030	0.573±0.030	0.559±0.032	0.587±0.033
subjet b-tag CSV-medium	0.658±0.039	0.704±0.037	0.718±0.039	0.687±0.040

$1.0 < \|\eta\| < 2.4$				
Selection	**Data**	**MADGRAPH**	**POWHEG**	**MC@NLO**
$N_{\mathrm{subjets}} \geq 3$	0.326±0.022	0.323±0.022	0.314±0.021	0.291±0.019
$m_{\mathrm{min}} > 50$ (GeV/c^2)	0.456±0.041	0.661±0.040	0.619±0.040	0.615±0.037
$140 < m_{\mathrm{jet}} < 250$ (GeV/c^2)	0.866±0.042	0.936±0.025	0.939±0.025	0.936±0.024
$\tau_3/\tau_2 < 0.55$	0.362±0.063	0.453±0.053	0.428±0.053	0.447±0.051
subjet b-tag CSV-medium	0.905±0.064	0.680±0.074	0.595±0.080	0.702±0.070

Table 6: Efficiencies in data and simulation for each successive cut in the CMS Top Tagger, and for the addition of N-subjettiness and subjet b-tagging (CMS Combined Tagger WP3), as measured using the semileptonic selection. The denominator of each quoted efficiency is the number of events passing the previous selection. The efficiency is measured for three $t\bar{t}$ Monte Carlo generators (MADGRAPH, POWHEG, and MC@NLO).



(a)

(b)

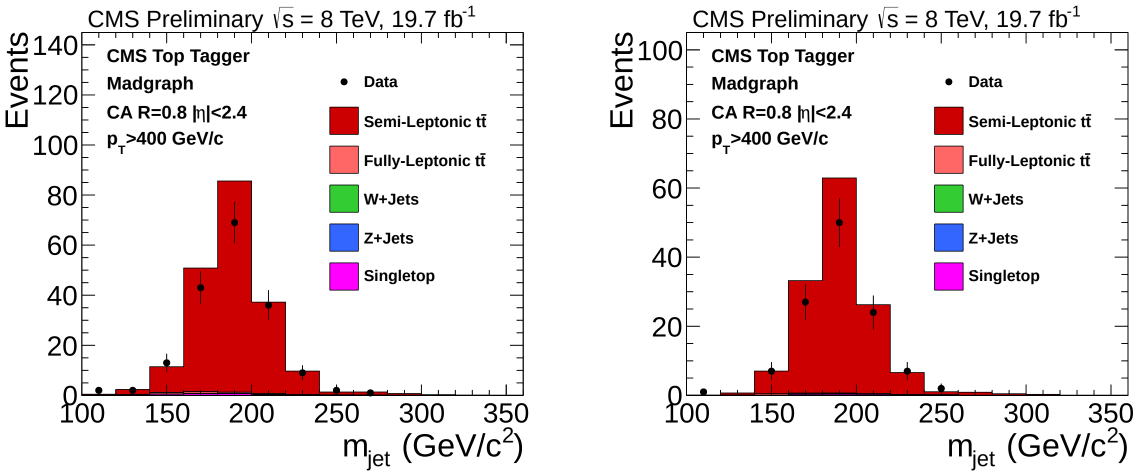
Figure 6: Tagging variables for fully-merged top candidates from the semileptonic sample: (a) τ_3/τ_2 (b) maximum subjet CSV b-discriminant. $t\bar{t}$ is simulated with the MADGRAPH event generator.

Data-simulation scale factor for sequential selections - CMS Tagger, CMS Combined Tagger

$\|\eta\| < 1.0$			
Selection	**MADGRAPH**	**POWHEG**	**MC@NLO**
$N_{\mathbf{subjets}} \geq 3$	1.006 ± 0.057	1.153 ± 0.069	1.014 ± 0.060
$m_{\mathbf{min}} > 50\ \mathrm{GeV}/c^2$	0.954 ± 0.041	0.978 ± 0.044	0.992 ± 0.046
$140\ \mathrm{GeV}/c^2 < m_{\mathbf{jet}} < 250\ \mathrm{GeV}/c^2$	1.028 ± 0.021	1.040 ± 0.024	1.028 ± 0.023
$\tau_3/\tau_2 < 0.55$	0.967 ± 0.073	0.990 ± 0.079	0.943 ± 0.073
subjet b-tag CSV-medium	0.935 ± 0.074	0.915 ± 0.074	0.957 ± 0.079

$1.0 < \|\eta\| < 2.4$			
Selection	**MADGRAPH**	**POWHEG**	**MC@NLO**
$N_{\mathbf{subjets}} \geq 3$	1.010 ± 0.097	1.037 ± 0.099	1.122 ± 0.106
$m_{\mathbf{min}} > 50\ \mathrm{GeV}/c^2$	0.689 ± 0.075	0.737 ± 0.081	0.741 ± 0.081
$140\ \mathrm{GeV}/c^2 < m_{\mathbf{jet}} < 250\ \mathrm{GeV}/c^2$	0.925 ± 0.051	0.922 ± 0.051	0.925 ± 0.051
$\tau_3/\tau_2 < 0.55$	0.800 ± 0.168	0.845 ± 0.181	0.810 ± 0.169
subjet b-tag CSV-medium	1.331 ± 0.172	1.52 ± 0.232	1.289 ± 0.157

Table 7: Data-simulation scale factors for three Monte Carlo generators for each successive cut in the CMS Top Tagger WP0, and for the addition of N-subjettiness and subjet b-tagging (CMS combined tagger WP3), as measured with the semileptonic selection. The scale factor measured in each row is with respect to the events passing the selection in the previous row. The scale factor is measured for three $t\bar{t}$ Monte Carlo generators (MADGRAPH, POWHEG, and MC@NLO). Subjet b-tagging is the last sequential selection and therefore this measurement has low statistics in data and in simulation. If measured before the m_{min} selection, and therefore with a larger sample of top jets, the scale factor for POWHEG is 0.97 ± 0.13.



(a) (b)

Figure 7: Jet mass for fully-merged top candidates from the semileptonic sample after successive selections: (a) $\tau_3/\tau_2 < 0.55$ (b) subjet b-tag CSV-medium and $\tau_3/\tau_2 < 0.55$. $t\bar{t}$ is simulated with the MADGRAPH event generator.

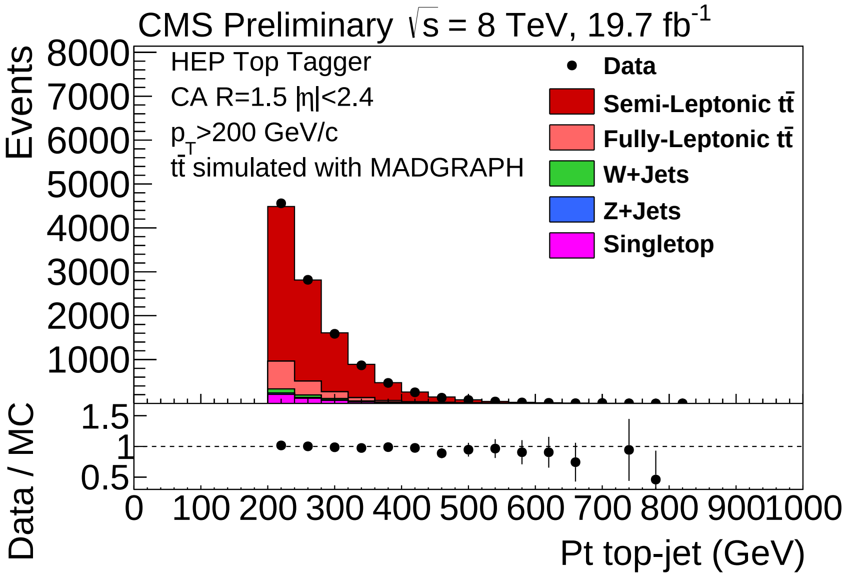


Figure 8: Distribution of the p_T of the Cambridge Aachen R=1.5 top candidate jet after all the selection cuts, but before applying the HEP Top Tagger. t$\bar{t}$ is simulated with the MADGRAPH event generator.

6.0.2 HEP Top Tagger Scale factor measurement

The HEP Top Tagger algorithm, as described in Section 3.3, utilizes jets with a very large distance parameter in order to increase the probability of catching the top decay products. The semileptonic t$\bar{t}$ selection defined in Section 6 is applied in order obtain a sample of merged top jets in data. The same selection is applied to simulated events, which are used to study the variables used by the algorithm to tag top jets. The selection is then used to calculate data–simulation scale factors.

The p_T distribution of top-jet candidates is shown in Figure 8. The HEP top jet mass (m_{123}) distribution of the top candidates from the semileptonic selection with subjet b-tagging but before requiring HEP Top Tagger tagging selections is shown in Figure 9 for different p_T bins. The distribution peaks at the top mass and has a low mass tail due to small fraction of unmerged top jets. The HEP top jet mass (m_{123}) distribution after requiring the HEP Top Tagger W mass selection defined in Section 3.3 but before requiring the top mass selection is shown in Figure 10. The low mass tail is reduced after requiring the W mass selection. The m_{123} top mass peak is not well-modeled by simulation.

The HEP Top Tagger algorithm finds three subjets within the jet and selects top jets based on the pairwise and three-way subjet masses. These cuts are applied to the two-dimensional plane defined by the ratio m_{23}/m_{123} and the inverse tangent of m_{13}/m_{12}. The distribution in this plane for simulated t$\bar{t}$ signal events is shown in Figure 11a for jets with $p_T > 200\,\text{GeV}/c$. The dotted regions represent the area selected by the tagger. In this low p_T region a large fraction of the jets are unmerged and thus fall outside the selection region. The distribution for simulated background events (including all processes) is shown in Figure 11b. Unlike jets from t$\bar{t}$ events, only the tail of the QCD jet distribution extends into the dotted selection region. The distribution is shown for data in Figure 11c. The distribution for simulated t$\bar{t}$ signal is again shown in Figure 12 but with larger p_T requirements. The majority of the high p_T jets are fully merged and thus fall inside the dotted selection region.

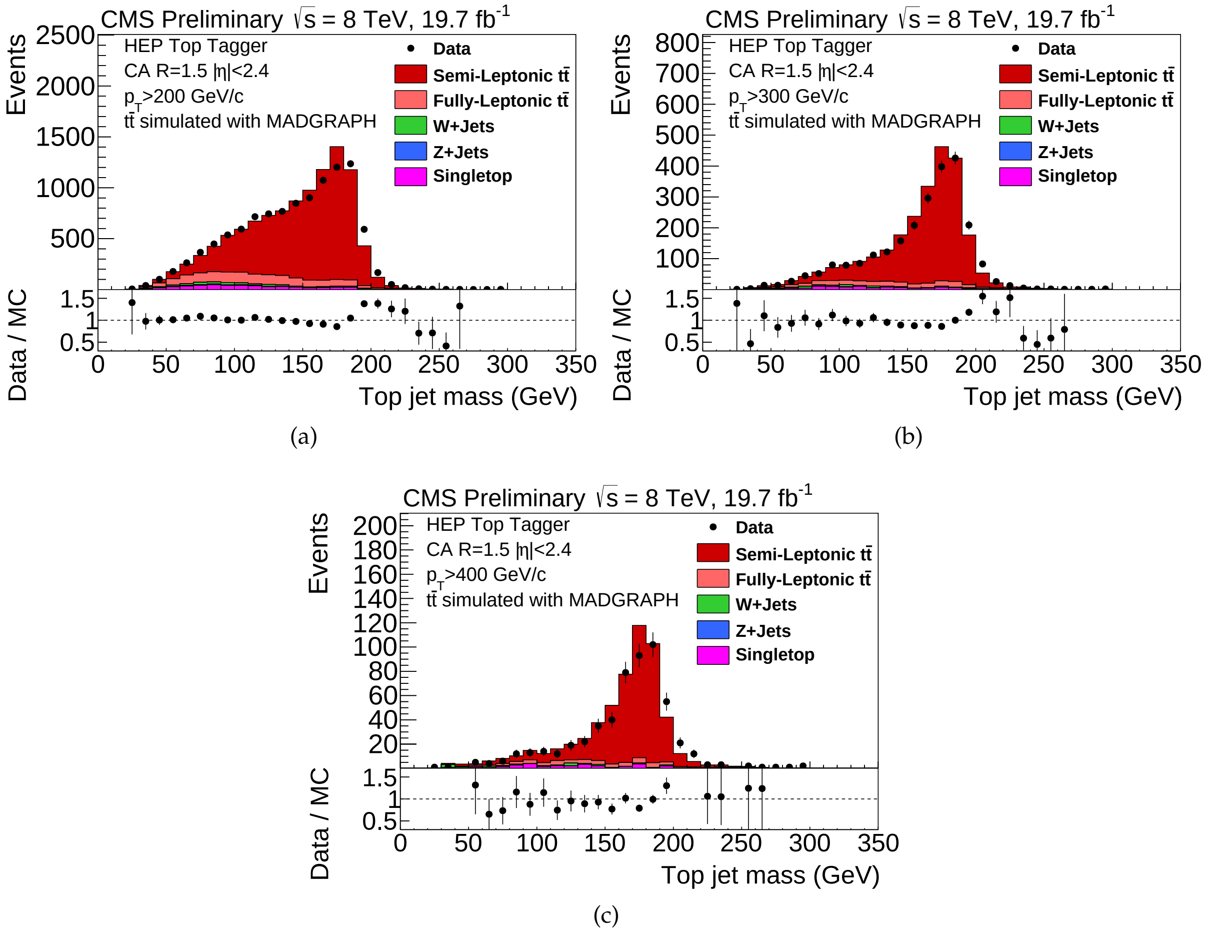
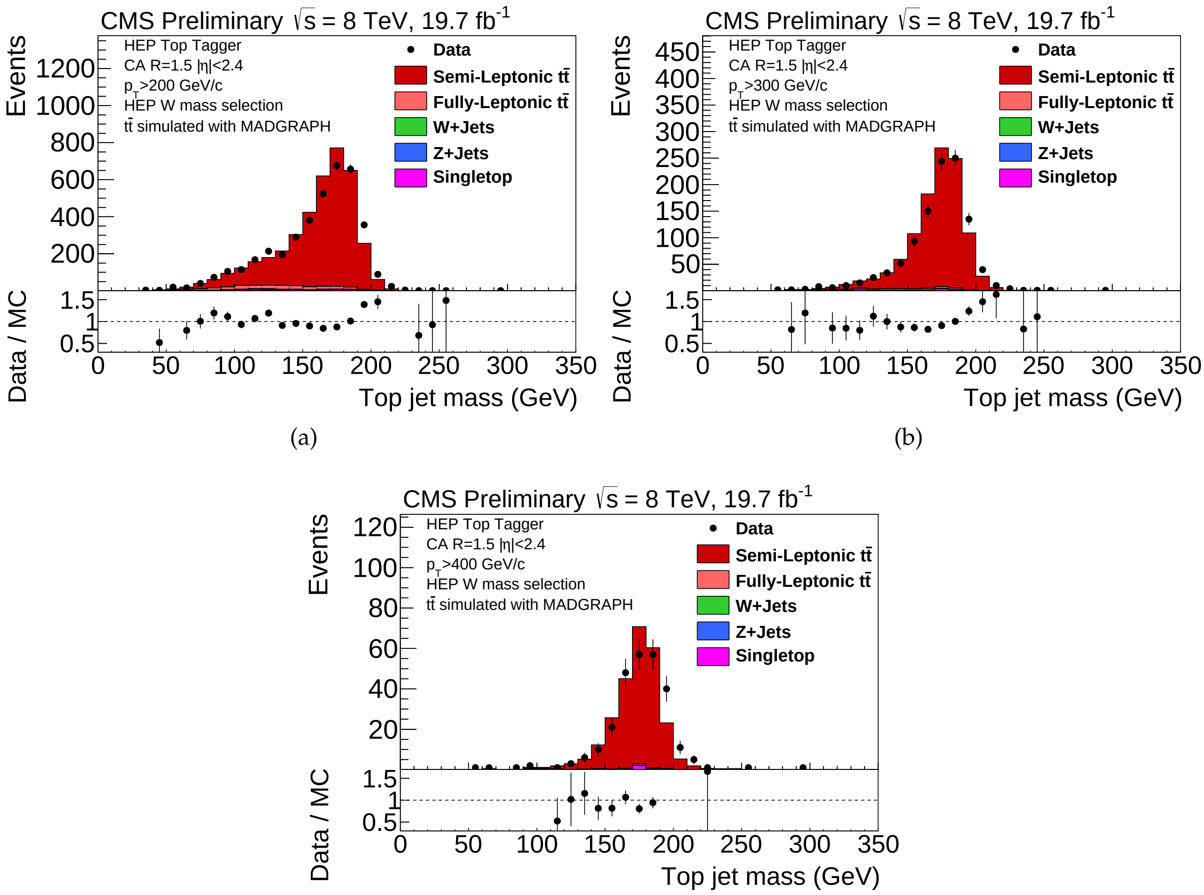


Figure 9: HEP Top Tagger top mass distribution (m_{123}) with different jet p_{T} cuts: (a) $p_{\mathrm{T}} > 200$ GeV/c, (b) $p_{\mathrm{T}} > 300$ GeV/c, (c) $p_{\mathrm{T}} > 400$ GeV/c. $\mathrm{t\bar{t}}$ is simulated with the MADGRAPH event generator. The m_{123} top mass peak is not well-modeled by simulation. This disagreement may be related to the jet energy resolution discrepancy observed between data and simulation[38].

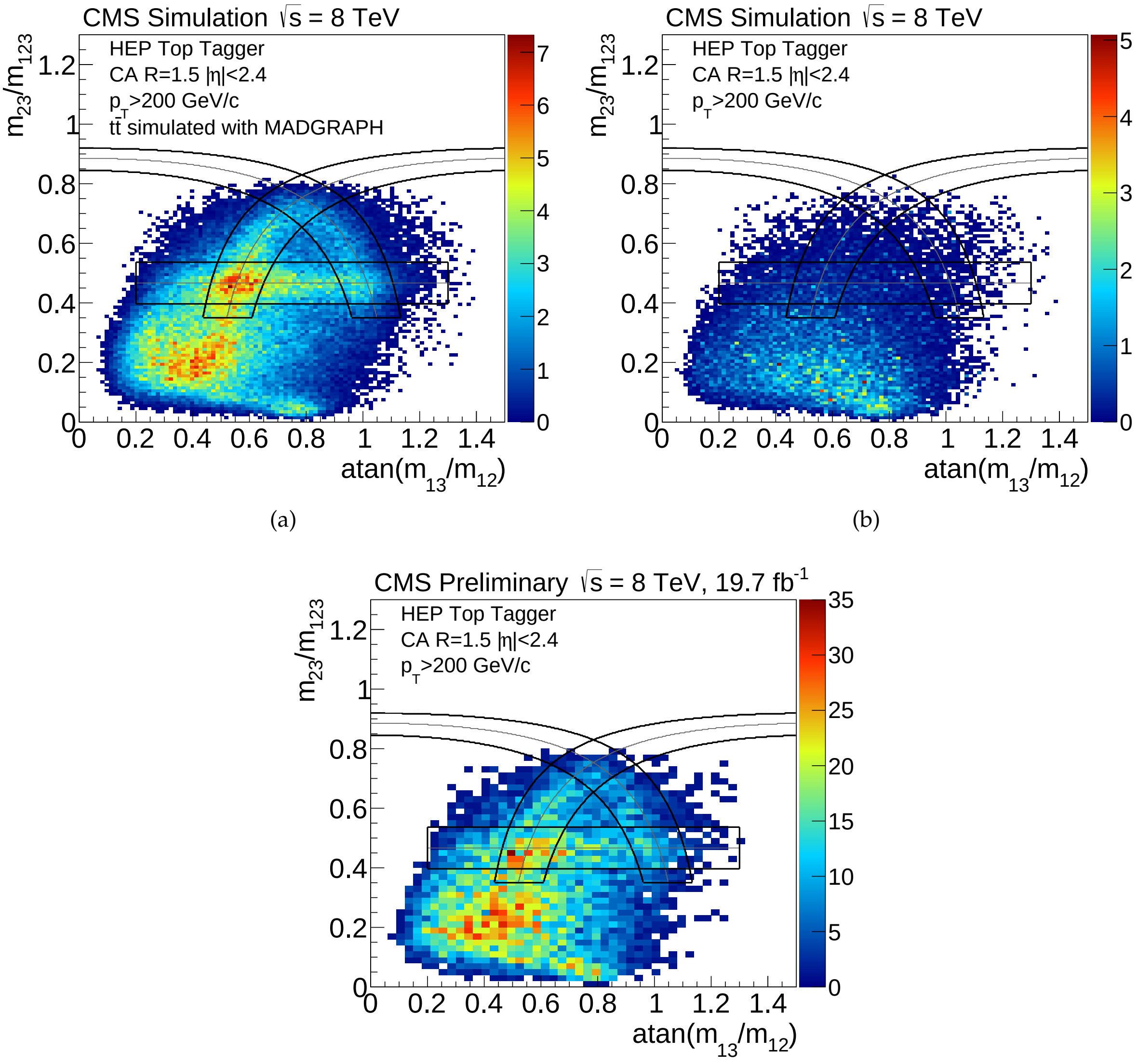


(a)

(b)

(c)

Figure 10: HEP Top Tagger top mass distribution (m_{123}) after the W mass selection but before the top mass selection defined in Section 3.3. The distribution is shown with three jet p_{T} selections: (a) $p_{\mathrm{T}} > 200$ GeV/c, (b) $p_{\mathrm{T}} > 300$ GeV/c, (c) $p_{\mathrm{T}} > 400$ GeV/c. $\mathrm{t\bar{t}}$ is simulated with the MADGRAPH event generator. The m_{123} top mass peak is not well-modeled by simulation. This disagreement may be related to the jet energy resolution discrepancy observed between data and simulation[38].

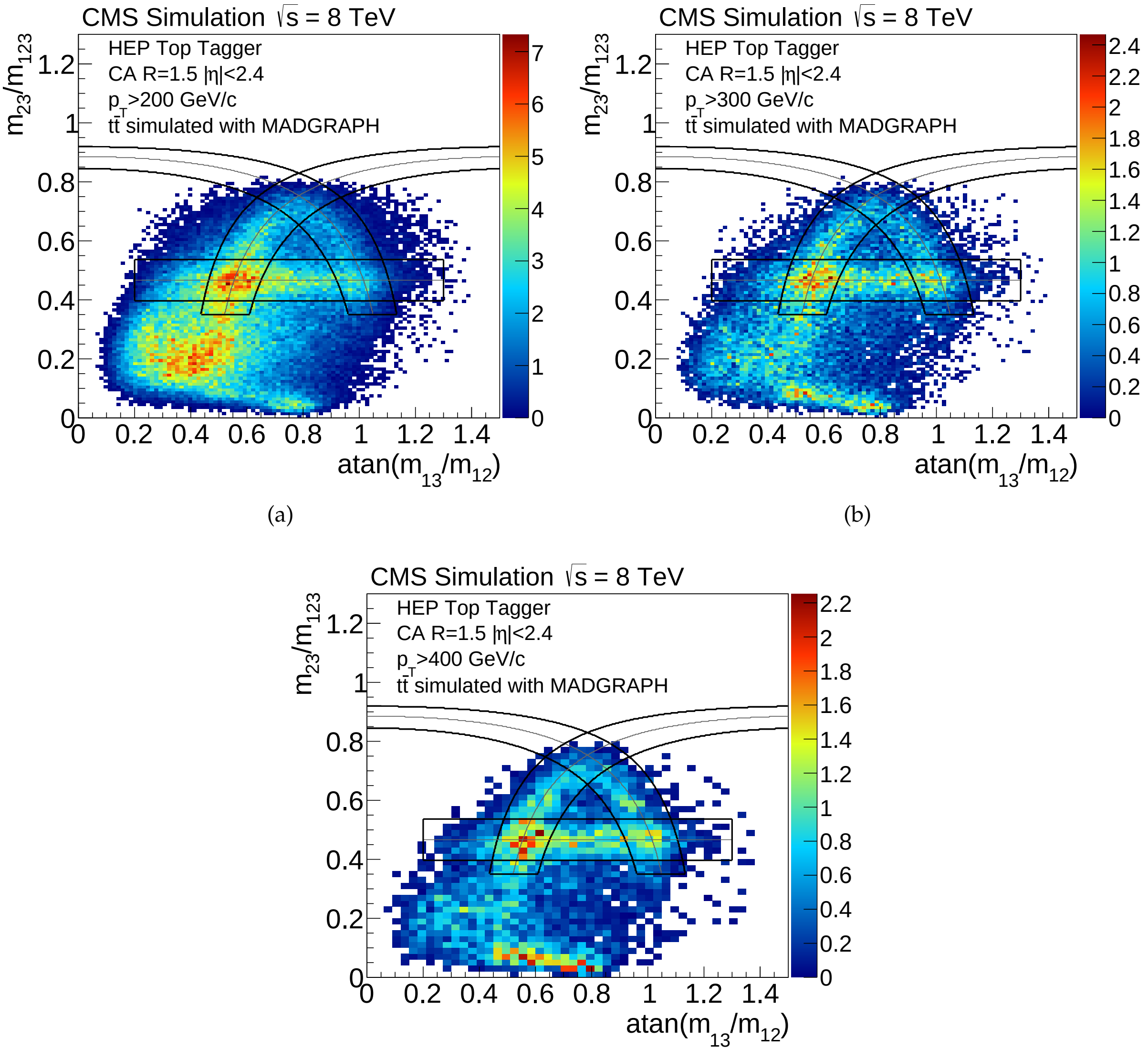


(a)

(b)

(c)

Figure 11: Bi-dimensional distributions of m_{23}/m_{123} vs. $\mathrm{atan}(m_{13}/m_{12})$ for HEP Top Tagger jets. The samples used are: (a) simulated $t\bar{t}$ (MADGRAPH), (b) background (cross section weighted boson+jets, diboson, single-top, $t\bar{t}$ all-hadronic, and $t\bar{t}$ leptonic production), and (c) data (semi-leptonic selection). The area enclosed by the black lines denotes the region selected by the HEP Top Tagger W mass selection.



(a)

(b)

(c)

Figure 12: Bi-dimensional distributions of m_{23}/m_{123} vs. $\mathrm{atan}(m_{13}/m_{12})$ for HEP Top Tagger jets in $\mathrm{t\bar{t}}$ simulated with MadGraph in three p_{T}(top-jet) regions: (a) $p_{\mathrm{T}} > 200$ GeV/c, (b) $p_{\mathrm{T}} > 300$ GeV/c, (c) $p_{\mathrm{T}} > 400$ GeV/c. The area enclosed by the black lines denotes the region selected by the HEP Top Tagger W mass selection.

Cumulative data-simulation scale factor - HEP Top Tagger, HEP Combined Tagger

$\|\eta\| < 1.0$				
Tagger	**p_T bin (GeV/c)**	**MADGRAPH**	**POWHEG**	**MC@NLO**
HEP Combined WP2	$200 < p_T < 250$	0.91 ± 0.04	0.92 ± 0.04	0.88 ± 0.04
	$250 < p_T < 400$	0.93 ± 0.03	0.95 ± 0.03	0.93 ± 0.03
	$p_T > 400$	1.15 ± 0.07	1.36 ± 0.07	1.19 ± 0.07
HEP Combined WP3	$200 < p_T < 250$	0.86 ± 0.05	0.86 ± 0.05	0.86 ± 0.05
	$250 < p_T < 400$	0.91 ± 0.04	0.93 ± 0.04	0.93 ± 0.04
	$p_T > 400$	0.98 ± 0.09	1.10 ± 0.12	1.10 ± 0.12

$1.0 < \|\eta\| < 2.4$				
Tagger	**p_T bin (GeV/c)**	**MADGRAPH**	**POWHEG**	**MC@NLO**
HEP Combined WP2	$200 < p_T < 250$	0.95 ± 0.05	0.93 ± 0.06	0.93 ± 0.05
	$250 < p_T < 400$	0.91 ± 0.04	0.95 ± 0.05	0.95 ± 0.04
	$p_T > 400$	0.85 ± 0.11	0.95 ± 0.15	0.99 ± 0.13
HEP Combined WP3	$200 < p_T < 250$	1.02 ± 0.07	1.00 ± 0.08	0.96 ± 0.07
	$250 < p_T < 400$	0.90 ± 0.05	0.97 ± 0.06	0.93 ± 0.05
	$p_T > 400$	0.85 ± 0.16	1.00 ± 0.22	0.99 ± 0.19

Table 8: Data-simulation scale factors after requiring all selections of the HEP Combined Tagger WP2 and after requiring all selections of the HEP Combined Tagger WP3. The scale factor is measured for three $t\bar{t}$ Monte Carlo generators in three p_T regions and in two η regions. The scale factor is measured with the semileptonic selection with subjet b-tagging.

The top tagging scale factor is defined as the efficiency for tagging top jets in data divided by the efficiency for simulated events. The efficiency denominator is defined by the number of top candidate jets in the semileptonic selection with subjet b-tagging. The cumulative HEP Top Tagger scale factor, measured after all selections of HEP Top Tagger WP2 and WP3, is shown in Table 8 for three different p_T bins and two η bins. The scale factor for each sequential selection of the HEP Top Tagger WP3 is shown in Table **??**. The dependence of the scale factor on jet η observed with the CMS Top Tagger is not observed with the HEP Top Tagger, although some mis-modeling is also observed in the pairwise mass of the two lowest p_T subjets m_{23}. The variable used by the CMS Top Tagger to make W mass selections, the minimum pairwise mass of subjets $m_{\min}$, is more susceptible to mis-modeling than the HEP Top Tagger W mass selection.

6.0.3 Type 2 top tag validation

The semileptonic event selection is also used to validate the ‘type 2’ tagging prescription described in Section 3.5. The leading p_T jet in the hadronic hemisphere is the W jet candidate. Figure 13a shows the jet mass distribution of the W jet candidate. The distribution is dominated by merged W jets and therefore the jet mass peaks at the W boson mass. The mass drop distribution of the W jet candidate is shown in Figure 13b. Figure 13c shows the pairwise mass of the W jet candidate and the closest jet in ΔR. This pairwise mass reconstructs the top quark candidate, and the distribution is seen to peak near the top quark mass.

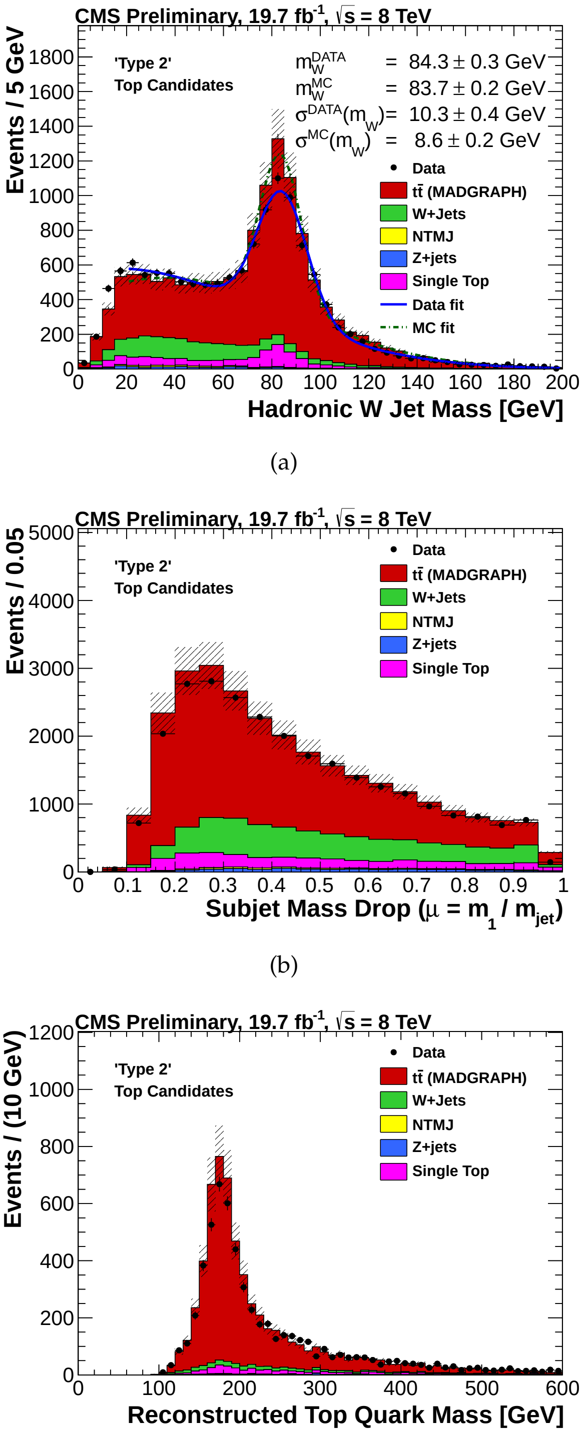
Additionally, this sample can be used to derive the subjet-energy scale and its uncertainty. This is done by comparing the position of the W mass peak in data and simulation after a fit to each mass distribution. We find good agreement in the W mass peak position, shown in Figure 13a.

Sequential data-simulation scale factor - HEP Combined Tagger WP3

$\lvert\eta\rvert < 1.0$				
Tagger	**p_T bin (GeV/c)**	MadGraph	powheg	mc@nlo
HEP top mass selection	$200 < p_T < 250$	0.92 ± 0.02	0.94 ± 0.03	0.92 ± 0.02
	$250 < p_T < 400$	0.93 ± 0.02	0.96 ± 0.02	0.94 ± 0.02
	$p_T > 400$	1.03 ± 0.03	1.07 ± 0.04	1.04 ± 0.04
HEP W mass selection	$200 < p_T < 250$	0.98 ± 0.03	0.98 ± 0.04	0.96 ± 0.03
	$250 < p_T < 400$	0.99 ± 0.02	0.99 ± 0.03	0.99 ± 0.02
	$p_T > 400$	1.11 ± 0.05	1.27 ± 0.07	1.14 ± 0.06
N-subjettiness selection	$200 < p_T < 250$	0.95 ± 0.04	0.93 ± 0.04	0.90 ± 0.03
	$250 < p_T < 400$	0.98 ± 0.03	0.98 ± 0.03	0.94 ± 0.03
	$p_T > 400$	0.85 ± 0.06	0.81 ± 0.07	0.84 ± 0.06

$1.0 < \lvert\eta\rvert < 2.4$				
Tagger	**p_T bin (GeV/c)**	MadGraph	powheg	mc@nlo
HEP top mass selection	$200 < p_T < 250$	0.89 ± 0.03	0.89 ± 0.04	0.90 ± 0.03
	$250 < p_T < 400$	0.92 ± 0.02	0.96 ± 0.03	0.97 ± 0.02
	$p_T > 400$	0.92 ± 0.07	0.98 ± 0.08	1.07 ± 0.08
HEP W mass selection	$200 < p_T < 250$	1.07 ± 0.04	1.04 ± 0.05	1.03 ± 0.04
	$250 < p_T < 400$	0.99 ± 0.03	0.99 ± 0.04	0.98 ± 0.03
	$p_T > 400$	0.92 ± 0.10	0.97 ± 0.13	0.92 ± 0.11
N-subjettiness selection	$200 < p_T < 250$	1.07 ± 0.05	1.07 ± 0.05	1.03 ± 0.05
	$250 < p_T < 400$	0.99 ± 0.04	1.03 ± 0.05	0.98 ± 0.04
	$p_T > 400$	1.01 ± 0.13	1.06 ± 0.17	1.00 ± 0.14

Table 9: Data-simulation scale factors for each sequential selection of the HEP Combined Tagger WP3. The scale factor measured for each selection is with respect to the number of events passing the previous selection. The scale factor is measured for three $t\bar{t}$ Monte Carlo generators in three p_T regions and in two η regions. The scale factor is measured with the semileptonic selection with subjet b-tagging.



(a)

(b)

(c)

Figure 13: Distributions of top tagging variables for partially merged 'type 2' boosted top topologies after the semileptonic selection. tt̄ is simulated with the MADGRAPH event generator. "NTMJ" represents non-top multijet backgrounds. These are measured in data by reversing the mass drop selection and normalizing through a fit to the H_T distribution[9]. The shaded regions represent the total uncertainty on the background model. (a) Pruned jet mass of the leading jet in the hadronic hemisphere. This is the hadronic W boson candidate. The W mass is measured in data and simulation in order to measure the subjet-energy scale. (b) Subjet mass drop μ for the W boson candidate in the hadronic hemisphere. (c) Pairwise mass of the W boson candidate and the closest jet in ΔR. This pairing is the "type 2" top quark candidate.

7 Conclusion

Several top tagging algorithms and jet observables are studied in both data and simulation, including the CMS Top Tagger, HEP Top Tagger, and N-subjettiness. A semileptonic $t\bar{t}$ selection is used to acquire a large, pure sample of merged boosted top jets and measure data–simulation scale factors for each top tagging algorithm. Studies of simulated events illustrate the performance of top tagging discriminators in separating top and QCD jets. Further sensitivity can be gained by utilizing a combination of these variables to improve top tagging performance.

The optimal top tagging algorithm is the one that provides the smallest mistag rate for the desired signal efficiency. The HEP Combined Tagger, combining the discriminating power of the HEP Top Tagger, subjet b-tagging, and N-subjettiness performs best for low p_T selections (jets matched to partons with $p_T < 400\,\text{GeV}/c$. The CMS Combined Tagger, combing the discriminating power of the CMS Top Tagger, N-subjettiness, and subjet b-tagging performs best for high p_T selections (jets matched to partons with $p_T > 400\,\text{GeV}/c$.

The top tagging algorithms considered here show a moderate decrease in performance at high pileup. In the near future it will be necessary to study the application of jet grooming and shape based pileup subtraction in order to decrease the dependence of these algorithms on pileup[27, 28, 43–45].

In the upcoming years, as the LHC is upgraded to run at a higher center-of-mass energy the mass reach for new particles will increase and top tagging algorithms will become essential elements of searches for new physics beyond the Standard Model.

References

[1] CMS Collaboration, "*A Cambridge-Aachen (C-A) based Jet Algorithm for boosted top-jet tagging*", *CMS Physics Analysis Summary* **CMS-PAS-JME-009-01** (2009).

[2] D. E. Kaplan, K. Rehermann, M. D. Schwartz, and B. Tweedie, "Top Tagging: A Method for Identifying Boosted Hadronically Decaying Top Quarks", *Phys. Rev. Lett.* **101** (2008) 142001, doi:10.1103/PhysRevLett.101.142001, arXiv:0806.0848.

[3] 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.

[4] J. Thaler and K. Van Tilburg, "Identifying Boosted Objects with N-subjettiness", *JHEP* **1103** (2011) 015, doi:10.1007/JHEP03(2011)015, arXiv:1011.2268.

[5] 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.

[6] CMS Collaboration Collaboration, "Search for anomalous t t-bar production in the highly-boosted all-hadronic final state", *JHEP* **1209** (2012) 029, doi:10.1007/JHEP09(2012)029, arXiv:1204.2488.

[7] ATLAS Collaboration Collaboration, "Search for resonances decaying into top-quark pairs using fully hadronic decays in pp collisions with ATLAS at $\sqrt{s} = 7$ TeV", *JHEP* **1301** (2013) 116, doi:10.1007/JHEP01(2013)116, arXiv:1211.2202.

[8] ATLAS Collaboration Collaboration, "A search for $t\bar{t}$ resonances in lepton+jets events with highly boosted top quarks collected in pp collisions at $\sqrt{s} = 7$ TeV with the ATLAS detector", *JHEP* **1209** (2012) 041, doi:10.1007/JHEP09(2012)041, arXiv:1207.2409.

[9] CMS Collaboration, "Search for Anomalous Top Quark Pair Production in the Boosted All-Hadronic Final State using pp Collisions at $\sqrt{s} = 8$ TeV", CMS Physics Analysis Summary CMS-PAS-B2G-12-005, (2013).

[10] ATLAS Collaboration, "A search for $t\bar{t}$ resonances in lepton plus jets events with ATLAS using 14 fb^{-1} of proton-proton collisions at $\sqrt{s} = 8$ TeV", ATLAS Conference Note ATLAS-CONF-2013-052, (2013).

[11] CMS Collaboration, "Search for top partners with charge 5e/3 in the same-sign dilepton final state", CMS Physics Analysis Summary CMS-PAS-B2G-12-012, (2013).

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

[13] CMS Collaboration, "Particle-Flow Event Reconstruction in CMS and Performance for Jets, Taus, and MET", *CMS PAS PFT-09-001* (2009).

[14] CMS Collaboration, "Performance of CMS muon reconstruction in pp collision events at $\sqrt{s} = 7$ TeV", (2012). arXiv:1206.4071. Accepted by JINST.

[15] 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.

[16] 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.

[17] 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.

[18] Y. L. Dokshitzer, G. D. Leder, S. Moretti, and B. R. Webber, "Better Jet Clustering Algorithms", *JHEP* **08** (1997) 001, arXiv:hep-ph/9707323.

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

[20] M. Cacciari, G. P. Salam, and G. Soyez, "FastJet user manual", *Eur.Phys.J.* **C72** (2012) 1896, arXiv:1111.6097.

[21] CMS Collaboration, "Identification of b-quark jets with the CMS experiment",.

[22] CMS Collaboration Collaboration, "Performance of b tagging at sqrt(s)=8 TeV in multijet, ttbar and boosted topology events",.

[23] CMS Collaboration, "*A Cambridge-Aachen (C-A) based Jet Algorithm for boosted top-jet tagging*", *CMS Physics Analysis Summary* **CMS-PAS-JME-009-01** (2009).

[24] A. Abdesselam et al., "Boosted objects: A Probe of beyond the Standard Model physics", *Eur.Phys.J.* **C71** (2011) 1661, arXiv:1012.5412.

[25] G. Salam, "Jet substructure - back to basics", 2013. CMS Jet Substructure Workshop, CERN, Switzerland.

[26] J. Thaler, “Theoretical progress in disecting jets: calculations past, present, and future”, 2013. BOOST 2013, Flagstaff, Arizona.

[27] 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.

[28] 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.

[29] J. Alwall et al., “MadGraph/MadEvent v4: The New Web Generation”, *JHEP* **09** (2007) 028, doi:10.1088/1126-6708/2007/09/028, arXiv:0706.2334.

[30] T. Sjöstrand et al., “High-energy physics event generation with PYTHIA 6.1”, *Comput. Phys. Commun.* **135** (2001) 238, doi:10.1016/S0010-4655(00)00236-8, arXiv:hep-ph/0010017.

[31] S. Hoeche et al., “Matching parton showers and matrix elements”, arXiv:hep-ph/0602031.

[32] S. Frixione, P. Nason, and C. Oleari, “Matching NLO QCD computations with Parton Shower simulations: the POWHEG method”, *JHEP* **0711** (2007) 070, doi:10.1088/1126-6708/2007/11/070, arXiv:0709.2092.

[33] S. Frixione and B. R. Webber, “The MC@NLO 3.4 Event Generator”, arXiv:0812.0770.

[34] G. Corcella et al., “HERWIG 6.5: an event generator for Hadron Emission Reactions With Interfering Gluons (including supersymmetric processes)”, *JHEP* **01** (2001) 010, arXiv:hep-ph/0011363.

[35] J. Pumplin et al., “New generation of parton distributions with uncertainties from global QCD analysis”, *JHEP* **0207** (2002) 012, arXiv:hep-ph/0201195.

[36] H.-L. Lai et al., “New parton distributions for collider physics”, *Phys.Rev.* **D82** (2010) 074024, doi:10.1103/PhysRevD.82.074024, arXiv:1007.2241.

[37] CMS Collaboration, “Electron reconstruction and identification at $\sqrt{s} = 7\,\mathrm{TeV}$”, CMS Physics Analysis Summary CMS-PAS-EGM-10-004, (2010).

[38] 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.

[39] CMS Collaboration, “b-Jet Identification in the CMS Experiment”, CMS Physics Analysis Summary CMS-PAS-BTV-11-004, (2012).

[40] CMS Collaboration, “Missing transverse energy performance of the CMS detector”, *JINST* **6** (2011) P09001, doi:10.1088/1748-0221/6/09/P09001, arXiv:1106.5048.

[41] S. Agostinelli et al., “GEANT4–a simulation toolkit”, *Nuclear Instruments and Methods in Physics Research Section A: Accelerators, Spectrometers, Detectors and Associated Equipment* **506** (2003), no. 3, 250 – 303, doi:10.1016/S0168-9002(03)01368-8.

[42] CMS Collaboration, "Search for anomalous $t\bar{t}$ production in the highly-boosted all-hadronic final state", *JHEP* **1209** (2012) 029, doi:10.1007/JHEP09(2012)029, arXiv:1204.2488.

[43] G. Soyez et al., "Pileup subtraction for jet shapes", *Phys.Rev.Lett.* **110** (2013) 162001, doi:10.1103/PhysRevLett.110.162001, arXiv:1211.2811.

[44] D. Krohn, J. Thaler, and L.-T. Wang, "Jet Trimming", *JHEP* **1002** (2010) 084, doi:10.1007/JHEP02(2010)084, arXiv:0912.1342.

[45] 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.